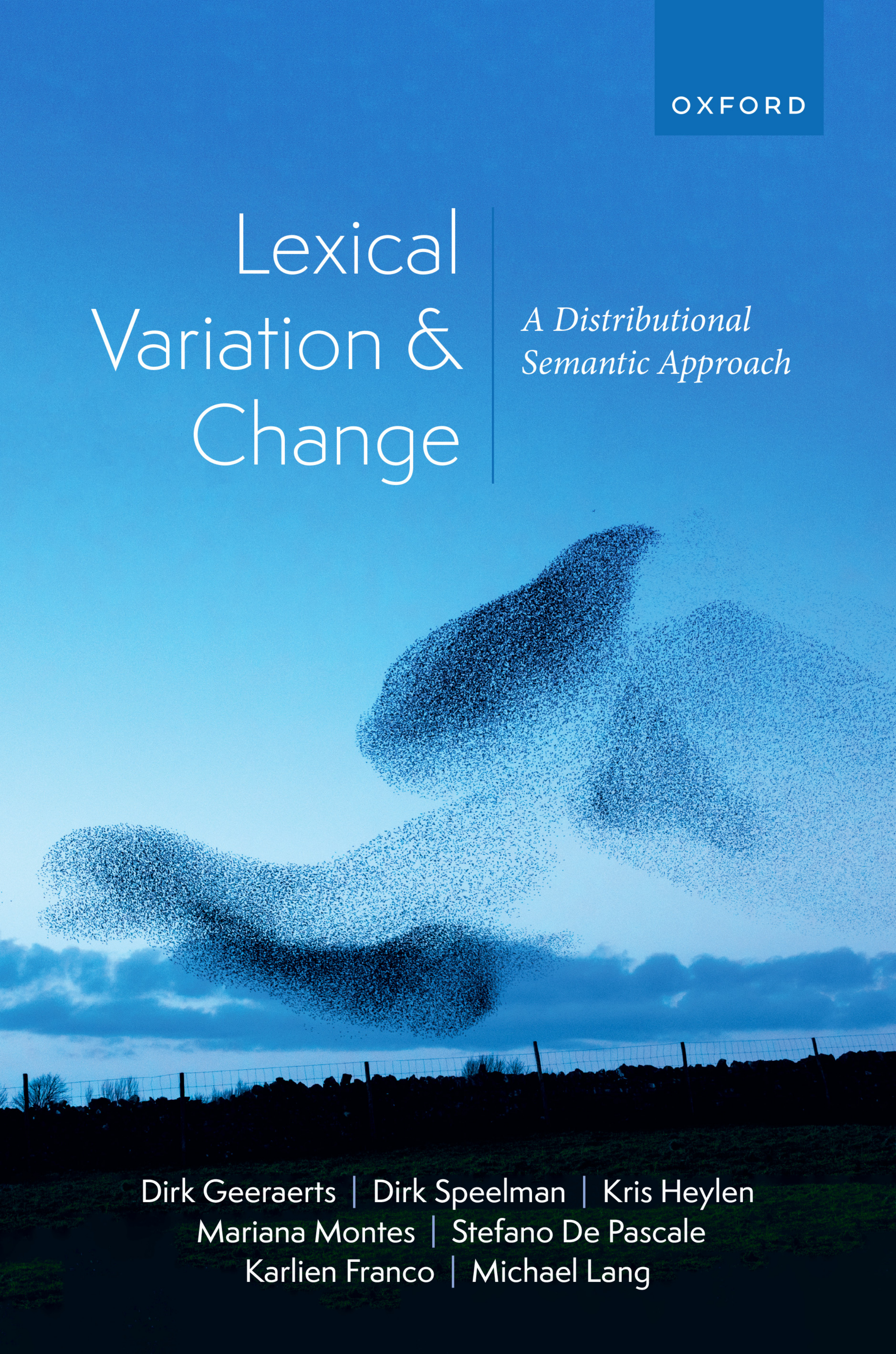
OXFORD

Lexical Variation & Change

A Distributional Semantic Approach



Dirk Geeraerts | Dirk Speelman | Kris Heylen
Mariana Montes | Stefano De Pascale
Karlien Franco | Michael Lang

Lexical Variation and Change

Lexical Variation and Change

A Distributional Semantic Approach

DIRK GEERAERTS
DIRK SPEELMAN
KRIS HEYLEN
MARIANA MONTES
STEFANO DE PASCALE
KARLIEN FRANCO
MICHAEL LANG

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2023937657

ISBN 9780198890676

DOI: 10.1093/oso/9780198890676.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

PART I. THEORETICAL PRELIMINARIES

PART II. DISTRIBUTIONAL METHODOLOGY

PART III. SEMASIOLOGICAL AND ONOMASIOLOGICAL EXPLORATIONS

PART IV. LECTOMETRIC METHODOLOGY

PART V. LECTOMETRIC EXPLORATIONS

List of figures

Colour versions of figures can be consulted via the free PDF download at https://global.oup.com/academic/product/lexical-variation-and-change-9780198890676 or via OUP's online platform at https://doi.org/10.1093/oso/9780198890676.001.0001.

List of tables

Introduction

In corpus linguistics, distributional semantics embodies the idea that the context in which a word occurs reveals the meaning of that word. By way of illustration, consider the words *underground* and *subway*, both referring to subterranean railway systems. The synonymy relationship that exists between the words may be recognized distributionally because they both co-occur frequently with words like *line*, *station*, *terminal*, *urban*, *crosstown*, *northbound*, *passenger*, *transit*, *train*, *run*, *operate*. That is to say, the similar distribution of the words *underground* and *subway* over contexts featuring items like *line*, *station*, *terminal*, and so on tells us something about the meaning of the two words. Importantly, there are computational techniques that allow us to identify the similarity in the distributional patterning of *underground* and *subway*. Those techniques can recognize that *underground* and *subway* are semantically closer than, say, *subway* and *sunshine*. But *underground* also has the meaning 'a secret organization fighting the established government or occupation forces', which co-occurs with words like *clandestine*, *resistance*, *insurrection*, *attack*, *army*, *hidden*, and which thus blurs the synonymy relationship with *subway*. A more fine-grained distributional approach then tries to model, not the overall similarity between *underground* and *subway*, but the similarity between the occurrences of *underground* in the sense 'subterranean railway' and those in the sense 'resistance movement'. Such a more detailed type of distributional semantics is called a *token-based* approach, where a token is any of the specific occurrences of the words, in contrast with a *type-based* approach that only looks at the level of the words as a whole. Computationally, token-based approaches group occurrences together based on their semantic (read: distributional) similarity, just like a type-based approach groups words as such together. So in the case of *underground*, you expect to come across a group of tokens for the 'subterranean railway' sense and another for the 'resistance movement' sense, and when you add the occurrences of *subway* to the model, you expect to find them intermingled with the group of *underground* tokens that represents the 'subterranean railway' sense. If we refer to such clusters of grouped-together tokens as clouds—token clouds—then the distributional approach consists of analysing configurations of token clouds to see what light they shed on the meanings of the expressions.

One major goal of the present monograph, then, is to explore the ins and outs of a distributional, token-cloud-based approach to word meaning. What does it involve, in what flavours does it come, how efficiently can it be implemented, and what exactly is its semantic import? The stakes for corpus semantics are high: if distributional modelling at the level of individual tokens of words works well, the automated or semi-automated analysis of meaning in large text corpora can be brought to a next level of detail and precision. There is also a very practical side to the methodological objectives of the book. The tools and algorithms that we will use are made available for public use, and so the book can also be seen as a portfolio of sample studies that might inspire other researchers. At the same time, we will point out the restrictions on the kind of distributional modelling that we have implemented and argue for some caution regarding its introduction in linguistic semantics. It turns out that the semantic information picked up by distributional models does not correspond in a stable and straightforward way with the information a linguist may be looking for and this recognition calls for specific measures as to how distributional models may be incorporated into a linguistic workflow.

But apart from this methodological purpose, the book has an equally important theoretical goal. Our exploration of distributional semantics continues a lexicological line of research that was developed over the past quarter century in the Quantitative Lexicology and Variational Linguistics (QLVL) research group at the University of Leuven. Situated within the broad context of cognitive linguistics, this research line translates the cognitive linguistic interest in categorization phenomena and semantic variability into a research programme that takes the interplay of semasiological, onomasiological, and lectal variation as its core question. To briefly and simplistically unpack this terminological triad (details follow in a separate chapter): semasiological variation looks from a word to its meanings; it studies polysemy, like the various senses of *underground*. Onomasiology reverses the perspective and describes how a given meaning can be expressed by various words, like the synonymy of *underground* and *subway* in the 'subterranean railway' sense. Lectal variation involves the way in which diversity along sociolinguistic, stylistic, geographical, and so on dimensions influences semasiological and onomasiological phenomena, like the observation that *underground* is typically British English and *subway* typically American English. This lectal perspective includes a so-called lectometric one: measuring the frequencies of *underground* and *subway* as expressions for 'subterranean railway' in British and American texts allows us to calculate how close lexical usage in the two varieties is with regard to each other, and to address the question whether they are growing together or apart. The present volume will detail this framework and examine how token-based distributional techniques might be used to scale up the research to the level of large-scale corpora. Although we will not exhaustively cover all the dimensions of the programme, the various studies showcasing the distributional method will treat crucial components of the theoretical frame of reference: the detection of polysemy, the interplay of semasiological and onomasiological variation, the treatment of lexical variation as a sociolinguistic variable, and the use of those variables to measure convergence or divergence between language varieties.

The book is structured in five parts of two chapters each. The first set of two chapters, *Theoretical preliminaries,* introduces the framework. Chapter 1 describes the various perspectives that may be taken in lexical variation research, and how these have so far been covered in existing research. Chapter 2 lays out the conceptual foundations of a token-based distributional method. The remaining eight chapters fall into two groups. A first set of two times two chapters deals with semasiology and onomasiology, that is, with the relationship between lexical expressions and their meanings, and how this may differ over chronological periods and language varieties. A second group of two times two chapters reverses the perspective. In Chapters 3 to 6, we are interested in how lectal variation may influence lexical variation. In Chapters 7 to 10, we are interested in what lexical variation has to say about lectal variation. In each set of two times two chapters, the first pair of chapters is devoted to methodological issues while the second pair illustrates the methodology with case studies. Accordingly, the *Distributional methodology* part introduces, in Chapter 3, the technical specifics of the distributional semantic workflow we will use, and in Chapter 4 the visualization tool that we have developed to explore its outcome. The chapters in the *Semasiological and onomasiological explorations* part put this exploration into practice. Using Dutch materials, Chapter 5 examines how far a distributional approach can take us on the path of semantic analysis, and Chapter 6 applies the distributional method to the interplay of semasiology and onomasiology in lexical semantic change. The final four chapters are similarly split up between two methodological and two descriptive chapters. The *Lectometric methodology* part introduces the various steps in a lectometric workflow. While Chapter 7 introduces the formulae that use lexical variation to quantify the relationship between language varieties, Chapter 8 specifies how a token-based distributional method identifies the sets of synonymous expressions that provide the basis for that quantification. The chapters in the final part, *Lectometric explorations*, illustrate the lectometric workflow. Chapter 9 looks diachronically at the evolution of Dutch. Chapter 10 presents a synchronic view of international varieties of Spanish. The book closes with a conclusion detailing in what ways the research programme can be further developed—and readers beware: there are plenty of them.

In light of this overview, we believe the book offers the following unique and innovative features. First, it presents a *comprehensive view of lexical variation*, based on the distinction between semasiology and onomasiology, and the addition of a lectal dimension. By describing how these distinctions define different perspectives for lexical research, and how the different phenomena interact, the book draws a more adequate picture of the richness and complexity of lexical phenomena than can be found in the existing literature. In particular, by treating lexical variation as a sociolinguistic variable in the sense of variationist sociolinguistics, the relationship between language varieties can be quantified at an aggregate level based on such variables. The monograph shows how such a lexical lectometry can be developed, and how it can profit from distributional methods.

Second, by comparing the semantic classifications produced by count-based distributional models with manually annotated disambiguated data, we offer a *critical insight into the machinery of distributional modelling.* Whereas a computational perspective on distributional methods is primarily concerned with their success in modelling linguistic phenomena, we aim for a deeper understanding of the mechanisms behind those results: how technical choices with regard to the distributional process influence which textual information is picked up by the models, and how that relates to a human interpretation of the data. Crucially, our analysis demonstrates, first, that there is no one-to-one relationship between the token clusters that fall out of a distributional modelling and what would traditionally be considered different senses, and second, that there is no single choice of model-building parameters that is optimal across the board, that is, that yields the best possible solution (the one closest to a human perspective) for any lexical item.

Third, the book is accompanied by a set of *digital tools* supporting the analytic workflows demonstrated in the case studies. On the one hand, some of these tools involve Python 3 and R packages used to extract information from corpora, create distributional models, and apply clustering and other statistical, viz. lectometric, analyses. On the other, visualization tools have been developed within the context of the semasiological workflow for the qualitative examination of token-level models. The availability of these tools greatly enhances the relevance of the book as a source of further research.

These assets suggest for which groups of readers the monograph may be of interest. Semanticists and lexicologists will be interested in the formulation of a comprehensive view of lexical variation, in the exploration of the possibilities and limits of token-based distributional semantics, and in the tools we offer for the incorporation of token-based distributional modelling in lexical and semantic research. Computational linguists will be interested in the distributional workflows we offer, with their accompanying tools, and our exploration of the possibilities and limits of a token-based distributional approach. Sociolinguists and historical linguists will be interested in our treatment of lexical variation as a sociolinguistic variable, and the synchronic and diachronic lexical lectometry based on it.

Because we intend to reach a diverse audience of linguists, the text is written with minimal assumptions regarding background knowledge. Specifically, the first two chapters are meant to bridge the gap between descriptively oriented linguists, who may need an introduction to the modus operandi of distributional semantics, and more technically minded researchers, who may be unfamiliar with the variety of perspectives in descriptive lexical and semantic research. In addition, because the trajectory we will describe is one with many optional turns and sideways, we will end each chapter with a summary that will help the reader to track the progress of the argument.

The project from which this monograph emanates was funded by the Research Council of the University of Leuven (project C16/15/023, with Dirk Geeraerts as principal investigator). Apart from the authors of the present volume, participants in the project included Benedikt Szmrecsanyi, Stefania Marzo, Weiwei Zhang, Tao Chen, Christian Andersen, and Kristina Geeraert. Although the present text is a collective product, resulting from several years of joint research efforts, the authors have contributed in different degrees to the various chapters. Dirk Geeraerts was lead author for Chapters 1, 2, and 7, Mariana Montes for Chapters 4 and 5, and for Chapter 3 together with Kris Heylen. Karlien Franco took the lead for Chapter 6, Stefano De Pascale for Chapter 9, and Michael Lang for Chapter 10. Stefano De Pascale and Karlien Franco were jointly responsible for Chapter 8.

PART I
THEORETICAL PRELIMINARIES

Two interwoven strands of research determine the organization of our monograph: a descriptive one, focusing on lexical variation, and a methodological one, focusing on distributional corpus semantics. In this first part of the book, two chapters present the basics and the background of both strands, with Chapter 1 introducing the descriptive framework, and Chapter 2 informally explaining the essentials of distributional vector semantics. Both chapters not only lay out the conceptual groundwork for these topics, but also situate them in a wider context of existing linguistic research.

1

Lexical variation and the lexeme-lection-lect triangle

As our investigation is situated at the crossroads of lexical variation research and distributional semantics, we have a double background to describe. In this chapter, we introduce the first of these two backdrops: what model of lexical variation do we start from, where do we situate our own research within that field, and how do we relate to previous research? The first section of the chapter charts various conceptual perspectives that may be taken in lexical variation studies; specifies the focus of our research in light of those alternatives; and indicates how our choice of perspective translates into the structure of the monograph. The second and third section then detail our choice of focus. The third section in particular introduces the lectometric perspective that plays a central role in later chapters, from Chapter 7 onward. The final two sections sketch the research background: on one hand, lexical studies in the broader context of linguistic variation research, on the other, our local research context. The present study continues a long-term research line within the Quantitative Lexicology and Variational Linguistics research group at the University of Leuven, and accordingly, we need to provide some detail about previous work and how the present approach builds on earlier achievements.

1.1 Choices of lexicological perspective

Imagine a pair of trousers ending just below the knee, tightened round the leg so that the bottom end is slightly baggy. How would they be called? Several terms exist: *knickerbockers*, *knickers*, and *breeches*. At the same time, they could simply be referred to as *trousers*, but then the item in question would be categorized differently. It would then not be identified as a member of the specific category BREECHES ‘pair of trousers ending just below the knee, tightened round the leg (etc.)’ that receives a unique, category-specific name with *knickerbockers* or *knickers* or *breeches*, but it would be identified as a member of the broader category TROUSERS ‘garment extending from the waist down to the knee or the ankle, covering each leg separately’. (Typographically, we will be using small caps for concepts or categories, specifically when they are represented by various synonymous expressions. Italics are used for lexical forms, and definitions, glosses, or

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0001

explanations will appear within quotes.) But how unique are terms like *knickerbockers* and *knickers*? At least for *knickers*, there is a polysemy to be considered, because it may also signify 'underpants', and the synonymy between *knickers* and *knickerbockers* does not extend to this second sense of *knickers*. A similar situation actually holds with regard to *trousers*: it is synonymous with *pants*, but in a polysemous sense, *pants* is synonymous with the 'underwear' reading of *knickers*. In addition, there is lectal variation in the distribution of the terms. Without being too detailed about it, we may note that *trousers* is typically British English whereas its synonym *pants* (like *knickerbockers* in comparison to *breeches*) is American English, and accordingly, the 'underwear' sense of *pants* is not common in American English (like that of *knickers*). Terms like *typically* are important here: the lexical choices are seldom of a black-and-white nature, but more often involve preferential patterns.

This brief example, to which we will come back in Section 1.2, is structured along two basic dimensions. The first one links linguistic forms to readings, whereas the second one brings in different language varieties and describes how the association between form and semantics differs according to the dialect (in the broadest possible sense of the term) under consideration. Crucially, both dimensions can be traversed in two directions. If you start from a lexical item and describe the semantics of how it is used, you take a *semasiological* perspective and your interest basically lies with polysemy. But if you focus on synonymy, you look from the semantic level to the level of forms, describing how a meaning can be expressed by various lexical items; this is an *onomasiological* perspective. The variational dimension can similarly be subjected to a perspectival switch. On the one hand (and this is the most common view), you can take the association of forms and meanings as a response variable and investigate how that association changes when you compare different language varieties. On the other hand, the relationship between those varieties can be your response variable: if you aggregate over a larger part of the vocabulary and its semasiological/onomasiological characteristics, what does that tell you about the language varieties in which that vocabulary appears? How close are they, and if you look over time, are they growing apart or growing together? The first of these perspectives, looking from varieties to variable word-meaning pairs, may be called *variationist*, because its outlook corresponds with that of variationist linguistics as the major branch of sociolinguistics initiated by Labov's work from the 1960s. The second perspective is a *lectometric* one, because it focuses on measuring distances among lects. *Lect* in this definition is a cover term for all kinds of language varieties. In the terminology of Coseriu (1981), this variety of varieties may be structured along four cross-classifying dimensions: a diatopic one, involving the dialects, regiolects, chronolects, national varieties, and so on, used in different parts and locations of a linguistic area; a diastratic one, involving sociolects belonging to different social groups; a diaphasic one, involving the differences of style and register that show up in different speech situations and communicative contexts; and a diachronic one, involving the chronological development and the historical stages of a language. Lectometry has so far primarily been an enterprise with a diatopic perspective, but in accordance with a generic conception of *lect*, we think of it as a generalization of that dialectometric tradition. (On dialectometry, see Goebl 2011, Wieling and Nerbonne 2015, and the discussion in Section 1.3.)

Given these two dimensions and the associated perspectival switches (semasiological-onomasiological, variationist-lectometric), the scope of our study can be described in terms of what we will call the *lexeme-lection-lect triangle*. Terminologically, *lexemes* are the lexical items under investigation, and a *lection* is the specific reading with which such a word appears in a text (like whether, to come back to the example, *knickers* is used in an 'underwear' reading or a 'breeches' reading). In the sense intended here, *lection* is a rather outdated philological term, and we are admittedly selecting it largely for its alliterating qualities. But the definition it receives in The New Shorter Oxford English Dictionary as 'a particular way of reading or interpreting a passage; a reading found in a particular copy or edition of a text', adequately captures what is of concern to us here, viz. the meaning-in-context of a word, the particular interpretation with which it is used in a given text passage. *Lect*, as indicated, is a general term for all kinds of language varieties.

Lexemes, lections, and lects interact, and talking about a *lexeme-lection-lect triangle* provides us with a handy image to schematically represent the various aspects of that interaction—or perhaps more precisely, the combinations of the two perspectival dimensions that we introduced above: see Figure 1.1. At the base of the triangle, the difference between a semasiological and an onomasiological perspective is expressed by the direction of the arrow linking lexeme and lection.

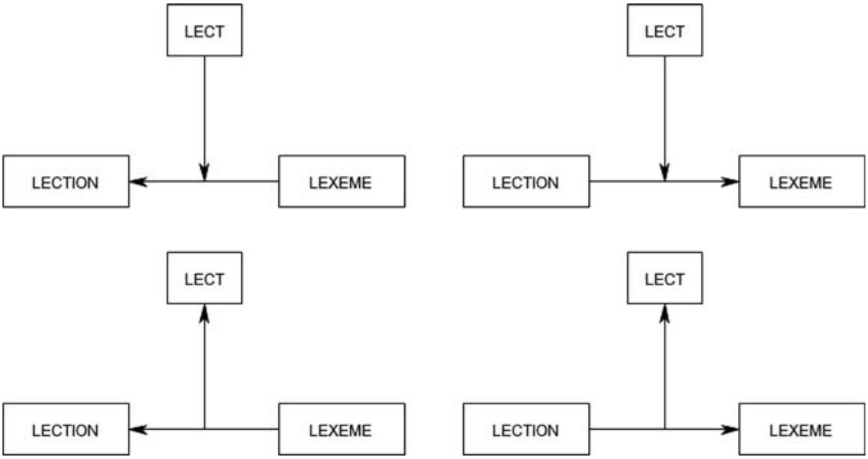


Figure 1.1 Research perspectives within the lexeme-lection-lect triangle

The panels on the left-hand side embody a semasiological perspective: looking from lexemes to their readings. The panels on the right embody the converse, onomasiological perspective: looking from readings to the forms through which they are expressed. Orthogonal to the semasiological/onomasiological dimension, the perpendicular line represents the other basic perspective. In the top panels, lectal variation is an explanatory variable: if you look at either semasiological or onomasiological variation, to what extent is it influenced by lectal diversity? In the bottom panels, the perspective is reversed, and lectal variation becomes a response variable: if you aggregate over either semasiological or onomasiological variation, which lectal structure emerges?

The various parts of the present monograph take their starting point in these perspectives. Part III, *Semasiological and onomasiological explorations*, focuses on the top-left and the top-right approaches. Part V, *Lectometric explorations*, deals with the bottom-right approach. The bottom-left perspective—semasiological lectometry—will not feature separately in the volume (but see Speelman and Heylen 2017 for an example). There are two reasons for the omission. First, if you study a sample of the vocabulary that is large enough, the lectal structure that emerges will be the same, regardless of whether you sum over semasiological differences or whether you sum over onomasiological differences: every semasiological difference between lect A and lect B will also show up if you start from the onomasiological side, and vice versa. Of course, this is only an argument in principle, because studying the entire vocabulary is not feasible. Second, however, there is a tradition in contemporary variationist linguistics to study lectal differences from a formal point of view, that is, to assume that linguistic differences between dialects, sociolects, and what have you are best seen in alternative lectal preferences for functionally equivalent forms of expression. This idea is captured by the notion of sociolinguistic variable. Put simply, a sociolinguistic variable in the sense of contemporary sociolinguistics (see Labov 1966) is a set of alternative ways of expressing the same linguistic function or realizing the same linguistic element, where each of the alternatives has social significance: 'Social and stylistic variation presuppose the option of saying "the same thing" in several different ways: that is, the variants are identical in reference or truth value, but opposed in their social and/or stylistic significance' (Labov 1972: 271). As such, a sociolinguistic variable is a linguistic element that is sensitive to a number of extralinguistic independent variables like social class, age, sex, geographical location, ethnic group, or contextual style and register. Classical cases of sociolinguistic variables involve pronunciation. Pronouncing the *t* in *butter* as a glottal stop is indicative of a Cockney accent, just like a full pronunciation of the *n* in *chemin* is typical of southern French in contrast with standard French. Examples like these had been studied for a long time in traditional dialectology, but modern sociolinguistics as it emerged in the 1960s enlarged the scope of investigation beyond the traditional diatopic dialects to other lects. If you apply the concept of a sociolinguistic variable to the lexicon, you inevitably reach an onomasiological perspective, because onomasiology (and more specifically, formal onomasiology) precisely involves alternative lexical expressions for the same sense.

Two more things need to be said about the way we will cover the terrain outlined above. In the first place, the subsequent parts of the text build on each other. Part I, *Theoretical preliminaries*, lays the groundwork. Parts II and III then focus on the semasiological and onomasiological perspectives that belong to the upper layer of Figure 1.1, whereas Parts IV and V take a lectometric point of view as in the lower layer of the figure. In each of these two sets, the first part is devoted to methodological issues while the second illustrates the methodology with case studies. Thus Part II, *Distributional methodology*, introduces the particulars of the distributional semantic workflow, together with the visualization tool that we will use to explore its outcome. Part III, *Semasiological and onomasiological explorations*, puts this exploration into practice. It examines how far a distributional approach can take us on the path of semantic analysis (as we shall see, there are a number of restrictions on distributional information that will make us adopt a certain amount of caution for the further steps) and applies the distributional method to the interplay of semasiology and onomasiology in lexical semantic change. Part IV, *Lectometric methodology*, introduces the various steps in a lectometric workflow: how to determine the relevant sets of alternating expressions and the contexts in which they alternate as equivalents (what sociolinguistics refers to as the *envelope of variation*), and how to feed the distribution of the competing expressions within the envelopes into a calculation of lectometric distances. Part V, *Lectometric explorations*, illustrates this workflow. Overall then, the structure of the text embodies a gradual build-up. It is not just that the chapters in Part II smooth the way for those in Part III, and those in Part IV for Part V, but (to the extent that identifying lexical sociolinguistic variables requires a semantic analysis) Parts II and III together also prepare the ground for Parts IV and V.

In the second place, the degree to which we will cover the perspectivally defined domains schematically represented in Figure 1.1 will by no means be complete, even apart from the absence of a semasiological lectometric approach. Our purpose is to define, illustrate, and explore a research programme, not to treat it exhaustively—if that would be possible at all. Throughout the chapters, we will explicitly point to open issues and possibilities for further investigation.

In the following two sections of the present chapter, we will look more deeply into the two dimensions and the associated questions that shape the structure of the book and that are graphically summarized in Figure 1.1. Along the semasiology/onomasiology dimension, Section 1.2 will consider the status of a vector space approach from the point of view of semantic and conceptual analysis. Along the variationist/lectometric dimension, Section 1.3 details what it implies to treat lexical variation as a sociolinguistic variable in the Labovian sense and to use that variation as the basis for lexical lectometry.

1.2 Semasiology, conceptual onomasiology, formal onomasiology

To move beyond the simple view of semasiology and onomasiology that was introduced in Section 1.1, and to better appreciate the way in which distributional corpus semantics can be incorporated into lexicology, we have to highlight the role that frequency and salience play in contemporary lexical theory. The distinction between onomasiology and semasiology is a fundamental one in the European tradition of lexicological research, invoking the Saussurean conception of the sign as consisting of a formal *signifiant* and a semantic *signifié*: semasiology starts out from the *signifiant* and considers the various *signifiés* associated with it, while onomasiology takes the reverse perspective. Kurt Baldinger, a prominent lexicologist from the structuralist era (see Geeraerts 2010a for an overview of the various theoretical stages in the history of lexical studies), described the distinction as follows: 'Semasiology . . . considers the isolated word and the way its meanings are manifested, while onomasiology looks at the designations of a particular concept, that is, at a multiplicity of expressions which form a whole' (1980: 278). The distinction between semasiology and onomasiology, in other words, equals the distinction between meaning and naming: semasiology takes its starting point in the word as a form, and charts the meanings that the word can occur with; onomasiology takes its starting point in a concept, and investigates by which different expressions the concept can be designated, or named. Between both, as we have emphasized, there is a difference of viewpoint: semasiology starts from the expression and looks at its meanings, onomasiology starts from the meaning and looks at the different expressions it occurs with. Characteristically, a traditional, structuralistically inspired view of semasiology and onomasiology considers only two levels: that of linguistic forms and that of the concepts expressed by those forms. But if we go back to the example discussed in Section 1.1, we may note that at least implicitly there is yet another level to consider: the denotational one, where we situate the things that are being talked about (and *thing* should evidently be taken broadly here, as anything that can be talked about to begin with). In Figure 1.2, the overall situation is represented by including, at the bottom level, a picture of a real-world pair of breeches. That denotational level lies outside of language, and in a simple view of the lexicon, the knowledge associated with it hardly matters. The definition of *knickerbockers* belongs to the language, but what else we know about breeches is encyclopaedic knowledge that does not belong to linguistic structure. The angle of view of such a structuralist conception is so to speak restricted to the upper levels in the figure. Since the emergence of prototype theory in the 1970s, however, linguistic semantics has shifted to a position in which the relevance of the lower level is explicitly envisaged. In the following pages, we will first present an overview of the new perspectives triggered by this shift. Next, we will discuss how this point of view can be translated to a distributional corpus approach and indicate which of the perspectives under consideration will play a role in the rest of the book.

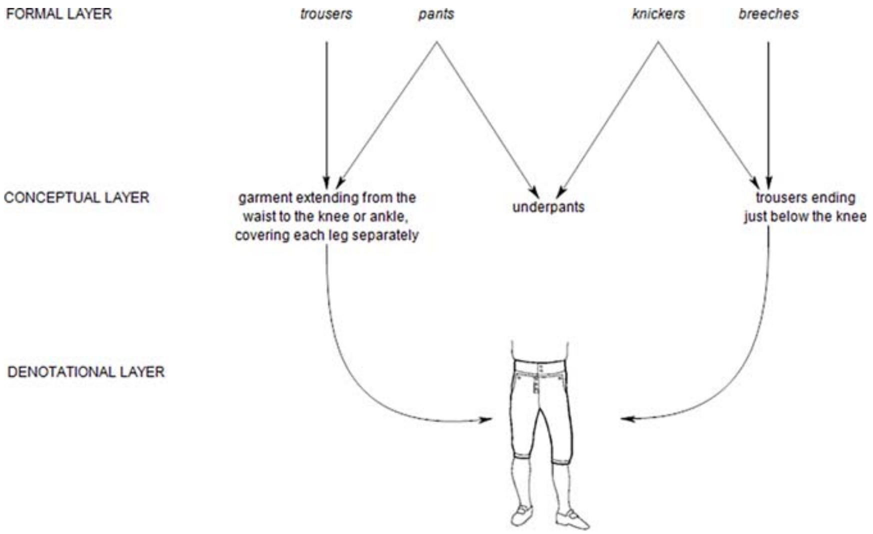

FORMAL LAYER
trousers
pants
knickers
breeches
CONCEPTUAL LAYER
garment extending from the waist to the knee or ankle, covering each leg separately
underpants
trousers ending just below the knee
DENOTATIONAL LAYER

Figure 1.2 An example of denotationally expanded lexicology

Regarding semasiology, the incorporation of the denotational level drew the attention to a number of interrelated prototypicality effects. The prototype-based conception of categorization originated in the mid-1970s with Rosch's psycholinguistic research into the internal structure of categories (see among others Rosch 1975) and was later elaborated in linguistic lexical semantics (see Geeraerts 1989; Taylor 1989; Hanks 2013). Four prototype-theoretical characteristics are frequently mentioned in the linguistic literature. First, prototypical categories cannot be defined by means of a single set of criterial (necessary and sufficient) attributes. Second, prototypical categories exhibit a family resemblance structure, that is, a structure like the similarities that exist between relatives (some have the same typical hair colour, some have the same typically shaped nose, some have the same typical eyes, but none have all and only the typical family traits); the different denotational uses of a word have several features in common with one or more other referents, but no features are common to all. Third, prototypical categories exhibit degrees of category membership; not every member is equally representative for a category. And fourth, prototypical categories may be blurred at the edges.

By way of example, consider *fruit* as referring to a type of food. If you ask people to list kinds of fruit, some types come to mind more easily than others. For American and European subjects (there is clear cultural variation on this point), oranges, apples, and bananas are the most typical fruits, while pineapples, watermelons, and pomegranates receive low typicality ratings. This illustrates the third characteristic mentioned above. But now, consider coconuts and olives. Is a coconut or an olive a fruit in the ordinary everyday sense of that word? For many people, the answer is not immediately obvious, which illustrates the fourth characteristic: if we zoom in on the least typical exemplars of a category, membership in the category may become fuzzy. A category like *fruit* should be considered not only with regard to the exemplars that belong to it, but also with regard to the features that these category members share and that together define the category. Types of fruit do not, however, share a single set of definitional features that sufficiently distinguishes fruit from, say, vegetables and other natural foodstuffs. All are edible seed-bearing parts of plants, but most other features that we think of as typical for fruit are not general: while most are sweet, some are not, like lemons; while most are juicy, some are not, like bananas; while most grow on trees and tree-like plants, some grow on bushes, like strawberries; and so on. This absence of a neat definition illustrates the first characteristic. Instead of such a single definition, what seems to hold together the category are overlapping clusters of representative features. Whereas the most typical kinds of fruit are the sweet and juicy ones that grow on trees, other kinds may lack one or even more of these features. This then illustrates the second characteristic mentioned above.

The prototype-theoretical expansion of the scope of lexical semantics to the denotational level was extrapolated in two different directions: towards the conceptual level, and towards the onomasiological perspective. The first extrapolation involves identifying prototype effects between senses rather than within a given sense. In its original form, and in the way we have so far described it, prototypicality involves the relationship between the exemplars of a single sense of an item, that is to say, the entities that are situated at the denotational level of Figure 1.2 and that belong together under the umbrella of one of the senses situated at the conceptual level. After all, *fruit* can also be used with other meanings than the one referring to food, like when you would talk metaphorically about the *bitter fruit* (the results, the consequences) of bad behaviour or sorry mistakes—but all the prototypicality effects mentioned previously were situated *within* the food sense. However, prototype theory as it developed in linguistics was applied not just to the internal structure of a single word meaning, but also to the structure of polysemous words, that is, to the relationship between the various senses that a lexical item exhibits at the conceptual level. In particular, it was pointed out that the structure of polysemy may take the form of a set of clustered and overlapping meanings, which may be related by similarity or by other associative links, such as metaphor or metonymy. Because this clustered set is mostly built up round a central meaning, the term *radial network* is often used for this kind of polysemic structure. Radial networks are a popular representational format in lexical semantics; see Brugman (1988) for an early and influential example, and see Geeraerts (1995) for a comparison with alternative forms of representation in cognitive semantics. This extrapolation from a within-senses level to a between-senses level implies that an indiscriminate use of *prototypicality* may sometimes be confusing, when the word is used both for the dominant sense in a polysemous network and the central case in a denotational set (Kleiber 1990). At the same time, the potential issues go further than the terminological scope of *prototypical*: as we shall discuss in Chapter 2, the mutual demarcation of semantic entities—senses, and within-senses versus between-senses levels—is not unproblematic in its own right.

The second extrapolation incorporates the denotational level into the onomasiological perspective. This implies that one should not just look from the conceptual level to the formal level, identifying relations of synonymy and near-synonymy among expressions, but that one should also look from the denotational level to the conceptual level, identifying alternative conceptualizations of the same chunk of reality. Terminologically speaking, lexical semantics has not yet settled on a conventional name for this phenomenon, but *categorization* or *conceptual construal* should be good candidates. In the example of Figure 1.2, for instance, the item of clothing depicted on the denotational level may be categorized as breeches, but it may also be identified as a pair of trousers, and then it is construed as a member of a different, broader category than when using *breeches*. The resulting picture of the interrelated terms is presented in Table 1.1; the corresponding subfields of lexicology in Table 1.2. The extension of onomasiology to the denotational level implies that we may now also distinguish between *conceptual onomasiology*, focusing on the relationship between the denotational level and the conceptual level, and *formal onomasiology*, focusing on the relationship between the conceptual and the formal level. Whereas conceptual onomasiological variation involves the choice of different conceptual categories for a referent, formal onomasiological variation merely involves the use of different synonymous names for the same conceptual category. The names *jeans* and *trousers* for denim leisure-wear trousers constitute an instance of conceptual variation, for they represent categories at different taxonomical levels. *Jeans* and *denims*, however, are no more than different (but synonymous) names for the same denotational entity. Onomasiological variation as a sociolinguistic variable in the Labovian sense, then, belongs to formal onomasiology.

The extrapolation of prototype theory to onomasiology goes one step further, though. A major consequence of prototype theory is to give frequency and salience a place in the description of semasiological structure. Next to the qualitative

Table 1.1 Terminological distinctions in denotationally expanded lexicology: phenomena

	SEMASIOLOGY	ONOMASIOLOGY
words w.r.t. concepts	polysemy	synonymy
concepts w.r.t. referents	prototypicality	categorization

Table 1.2 Terminological distinctions in denotationally expanded lexicology: subfields

	SEMASIOLOGY	ONOMASIOLOGY
words w.r.t. concepts	between-sense semasiology	formal onomasiology
concepts w.r.t. referents	within-sense semasiology	conceptual onomasiology

relations among the elements in a semasiological structure (like metaphor and metonymy), a quantifiable centre-periphery relationship is introduced as part of the architecture. Central exemplars, or recurrent features, carry more weight than others; all fruits are equal as fruits, but some fruits are more equal than others. This quantitative way of looking at semantic phenomena can then also be applied to onomasiological relations. The initial step in the introduction of onomasiological salience is the basic-level hypothesis (Berlin 1978, 1992). This hypothesis is based on the ethnolinguistic observation that folk classifications of biological domains usually conform to a general organizational principle, in the sense that they consist of five or six taxonomical levels. The highest rank in the taxonomy is that of the 'unique beginner', which names a major domain like *plant* and *animal*. The domain of the unique beginner is subdivided by just a few general 'life forms' like *tree* or *fish*, which are in turn specified by 'folk genera' like *pine*, *oak*, *beech*, *ash*, *elm*, *chestnut*. A folk genus may be further specified by 'folk specifics' (*white pine*) and 'varietal taxa' (*western white pine*). To the extent that the generic level is the core of any folk biological category, it is the basic level: 'Generic taxa are highly salient and are the first terms encountered in ethnobiological enquiry, presumably because they refer to the most commonly used, everyday categories of folk biological knowledge' (Berlin 1978: 17). The generic level, in other words, is onomasiologically salient: within the lexical set defined by the taxonomy, the generic level embodies a naming preference; given a particular referent, the names situated at the basic level are more likely to be selected for that referent from among the alternatives provided by the taxonomy. Apart from embodying a concept of onomasiological salience, basic-level categories are claimed to exhibit a number of other characteristics. From a psychological point of view, they are conceptualized as perceptual and functional gestalts. From a developmental point of view, they are early in acquisition, that is, they are the first terms of the taxonomy learned by the child. From a linguistic point of view, they are named by short, morphologically simple items. And from a conceptual point of view, Rosch (1978) argues that the basic level constitutes the level where prototype effects are most outspoken, in the sense that they maximize the number of attributes shared by members of the category, and minimize the number of attributes shared with members of other categories.

Although the basic-level hypothesis was only formulated for natural categories, it can be extrapolated to artefacts. If a particular referent—say, a particular piece of clothing—can be alternatively categorized as a garment, a skirt, or a wrap-around skirt, the choice will be preferentially made for *skirt*, which may then be considered a basic level term. But the extrapolation can go further: differences of onomasiological preference may also occur among categories on the same level in a taxonomical hierarchy, and not just between different levels in the taxonomy. If a particular referent can be alternatively categorized as a wrap-around skirt or a miniskirt, there could just as well be a preferential choice: when you encounter something that is both a wrap-around skirt and a miniskirt, the most natural way of naming that referent in a neutral context would probably be *miniskirt*. To illustrate this notion of generalized onomasiological salience, we may have a look at some of the results obtained in Geeraerts, Grondelaers, and Bakema (1994). Let us note first that calculating conceptual onomasiological salience assumes that the referents of the expressions can be identified. In the 1994 study of clothing terms, this is achieved by using a 'referentially enriched' corpus: rather than using just a text corpus, the study uses illustrated magazines, so that the pictures accompanying the text provide independent access to the entities being named. This allows us, for instance, to spot cases where *trousers* refers to a pair of breeches, even if they are not named as such—an indispensable piece of information for applying the definition of conceptual onomasiological salience. Once such referential identification is available, the conceptual onomasiological salience of a competing category may be simply defined as the frequency with which that category is used relative to the overall frequency of the members of the category: of all the breeches appearing in the dataset, how many are actually named by the term *breeches* and its synonyms? Table 1.3, then, shows how the onomasiological salience of categories on the same taxonomical level may differ considerably. In the upper part of the table, *short*, *bermuda*, *legging*, and *jeans* are co-hyponyms, as they all fall under the hyperonymous category *broek* 'trousers'. However, the onomasiological salience of the different concepts differs considerably: that of the concept JEANS, represented by the synonyms *jeans*, *jeansbroek*, *spijkerbroek*, doubles that of LEGGING, represented by the synonyms *legging, leggings, caleçon*. This means that a potential member of the category JEANS is twice as likely to be designated by an expression that names the category JEANS than a member of the category LEGGING would be likely to be designated by an expression that names the category LEGGING.

In Geeraerts, Grondelaers, and Bakema (1994) it is further shown how the choice for one lexical item rather than the other as the name for a given referent is determined by the semasiological salience of the referent (i.e. the degree of prototypicality of the referent with regard to the semasiological structure of the category), by the overall onomasiological salience of the category represented by the expression, and by contextual features of a classical sociolinguistic and

Table 1.3 Differences in conceptual onomasiological salience among co-hyponyms

CATEGORY	TERMS	SALIENCE
TROUSERS	*broek*	46.47
SHORTS	*short, shorts*	45.61
BERMUDA	*bermuda*	50.88
LEGGINGS	*legging, leggings, caleçon*	45.50
JEANS	*jeans, jeansbroek, spijkerbroek*	81.66

Figures reproduced from Geeraerts, Grondelaers, and Bakema (1994).

geographical nature, involving the competition between different language varieties. Zooming in on the last type of factor, we are back to onomasiological variation as a sociolinguistic variable, not least because quantifying alternative preferences is part and parcel of the tradition in variationist sociolinguistics. But we will come back to that in Section 1.3. Once we recognize the relevance of frequency and salience for an onomasiological perspective, there are two more remarks to be made (see also Geeraerts 2016a).

First, with regard to onomasiology, we can distinguish an indirect, oblique form of conceptual onomasiology alongside the direct categorial choices of the type illustrated above. Such indirect indications for conceptual onomasiological salience come in two kinds. On the one hand, the textual context in which a topic appears may reveal aspects of how the topic is thought of. It would make a difference, for instance, whether *breeches* is predominantly accompanied by *uncomfortable* rather than *leisurely*. On the other hand, the categorial labels with which a phenomenon are named may themselves embody a specific way of looking: while *skirt* does not express a specific perspective, *miniskirt* highlights the length of the garment, and *wrap-around skirt* profiles the method of fastening. Specifically when the designations have a figurative value, as with metaphors, looking for salient patterns in the semantic motifs expressed by the words used may show us something of how the phenomenon in question is conceptualized.

Second, the importance of frequency—for establishing centrality effects in the semasiological domain, for identifying naming and categorization preferences in the onomasiological domain—implies that the field moves away from a structuralist to a pragmatic, usage-based conception of lexical research. It is not feasible to determine semasiological or onomasiological weights unless you take into account the actual linguistic behaviour of language users. In terms of the history of linguistics, this is an important shift. The focus shifts from an investigation of language structure to an investigation of language in use, or in Saussurean terms, from an investigation of *langue* to an investigation of *parole*. The structural perspective deals with sets of related expressions and asks the question: what elements should we distinguish in the linguistic system and what are the relations among these entities? The usage-based conception deals with the actual choices made between the available entities and asks the question: which expressive options are preferred and what factors determine the choice for one or the other alternative? So, we have what we might call a 'qualitative' and a 'quantitative' perspective respectively, and as Table 1.4 shows, these can be applied both semasiologically and onomasiologically.

Now that we have a better view of semasiology and onomasiology and the shift from a structuralist to a usage-based framework, the next question to consider is how the latter relates to the overview of Section 1.1, and how a distributional corpus approach fits into it. We will first consider the position of vector space semantics in a usage-oriented perspective, and then consider the position of conceptual onomasiology. Using corpus data has an evident appeal for a usage-oriented lexicology: text corpora are repositories of actual acts of *parole*. Accordingly, we can think of the instances of a given word in a text or a collection of texts as instances of the concept(s) expressed by the lexical item in question, similar to actual examples of *knickerbockers* or other clothing terms. As such, distributional corpus semantics looks at concepts in an *extensional* rather than intensional way: a concept is represented by its instantiations rather than by its definition. The overall approach is analogous to the method used in Geeraerts, Grondelaers, and Bakema (1994): in the 1994 study, the extension (the set of available instances) of a lexical item is described by means of descriptive features characterizing the actual referents that we find in the sources, such as the length, width, material, elasticity, type of fastening, and so on, of trousers. Comparably, in the corpus approach, each element of the extension is described in terms of the words that a target item co-occurs with in the utterances (the elements of the extensional set) that it occurs in. Vector space semantics works by grouping together similar instantiations. If *chair* is the target item, *this store sells kitchen chairs* and *we need to buy new chairs for the dining room* will end in each other's vicinity, because the neighbourhoods in which *chair* appears are similar: *kitchen* and *dining room* are semantically related, and so are *buy* and *sell*. But *why*

Table 1.4 Structural and usage-oriented perspectives in lexical research

	SEMASIOLOGY	ONOMASIOLOGY
qualitatively investigating structure: elements and relations	senses and semantic links (metaphor, metonymy, etc.)	relations among lexemes (fields, taxonomies, etc.)
quantitatively investigating usage: typicality and salience	typicality effects within and between senses	salience effects between categories and levels

don't you take a chair? will wind up at a certain distance from both sentences, because the neighbourhood is different. This reveals a second feature that aligns a corpus-based distributional approach with the usage-based view of semasiology: the extension that instantiates the concept is not a homogeneous mass, but is a structured set, and the description should take that structure into consideration.

The analogy will be further detailed in Section 2.1, but at this point, we may conclude that we have a double alignment between a vector space approach and the model schematized in Table 1.1: we analyse the conceptual level by looking through the lens of its instantiations, and we assume that there is relevant structure in that level. But that correspondence also raises questions, and those are the focus of the first of the two main research questions of this study. Textual occurrences and vector representations of words are not the same as the direct referential data used in Geeraerts, Grondelaers, and Bakema (1994): utterances like the ones above may indirectly tell us something about chairs, but they are not a direct representation of chairs (like a picture of knickerbockers in a fashion magazine would be). So, how far can the analogy go? How exactly should we think of the extensional information provided by corpus-based vector spaces, and what do they tell us about semasiology? This question will be central in Chapter 5, but already in Chapter 2, major steps will be taken in the direction of a reply.

Let us now shift the focus to onomasiology. In principle, the distinction between formal and conceptual onomasiology doubles the lexeme-lection-lect triangle, and so the question arises what the role will be of a conceptual onomasiological perspective in the architecture of the book as introduced in Section 1.1. For several reasons, the following chapters will concentrate on semasiological and formal onomasiological variation, that is, we will not include conceptual onomasiology as a topic in its own right. This is to some extent a purely practical choice. Putting it simply, there will be enough to say about semasiology and formal onomasiology as such. More importantly, issues of a conceptual onomasiological kind will be difficult to separate from the semasiological and formal onomasiological case studies that we will consider. In other words, the conceptual onomasiological perspective will be inevitably intertwined with the others, for descriptive and methodological reasons. The descriptive interweaving is illustrated in Chapter 6, which includes an example demonstrating how semasiological and onomasiological changes go hand in hand.

The methodological interlacing of conceptual onomasiology with the other types turns round the question whether it is always possible to neatly distinguish between both onomasiological levels. In the tradition of variationist sociolinguistics, the problem of semantics was identified early on, in an important article by Beatriz Lavandera. She argued that 'it is inadequate at the current state of sociolinguistic research to extend to other levels of analysis of variation the notion of sociolinguistic variable originally developed on the basis of phonological data. The quantitative studies of variation which deal with morphological, syntactic, and lexical alternation suffer from the lack of an articulated theory of meanings' (1978: 171). In practice, variational sociolinguistics has not been fast in developing such an approach. Most of variational sociolinguistics is focused on pronunciation, and sociophonetics is its dominant research field. Lavandera's remark, in other words, seems to have worked less as an incentive to get one's semantic hands dirty than to stay in the relative safety of more manageable variation. But with our lexical interests, we cannot escape the issue. In the context of the present book, we may forego the question whether the problem is as outspoken in morphology and syntax as it is in lexicology, but the lexical problem is definitely real: how exactly do we establish whether a number of potential synonyms actually express 'the same thing', or whether, by contrast, they are merely referential near-synonyms? Differentiating between formal and conceptual onomasiology is a central concern if we want to treat lexical variation as a sociolinguistic variable, but it raises a fundamental methodological question: how consistently can the distinction be made? A major purpose of the present book is to investigate how distributional corpus semantics contributes to the development of a methodology for variational lexicology. Specifically for the identification of lexical variation as a sociolinguistic variable, Chapter 8 will present a procedure to that effect. But at the same time, we want to issue a warning against an excess of methodological confidence. In Chapter 2 we will show that the usage-based conception of semantics comes with a certain degree of methodological underdetermination. This underdetermination is primarily of a semasiological kind, but it has inevitable consequences for the demarcation of formal onomasiological variation. If the question 'What is a different meaning?' may sometimes be difficult to answer on the level of usage events (the level where we situate lections), it follows that determining the contextual synonymy of two or more items may also face a degree of uncertainty.

1.3 Onomasiological profiles and lectometry

Given that we are interested in studying the relationship between language varieties based on patterns of onomasiological variation, how do we quantify similarity and difference of word choice? In this section, we introduce the basics of our lexical lectometry, and give it more body with a case study on the evolution of contemporary Dutch and its main varieties. The measure of lexical overlap that we illustrate in this section, first introduced in Geeraerts, Grondelaers, and Speelman (1999), is based on the notions *onomasiological profile* and *uniformity*. The onomasiological profile of a concept in a particular source (like a collection of textual materials representing a language variety) is the set of synonymous names for that concept in that source, differentiated by relative frequency. Uniformity is then defined as a measure for the correspondence between two onomasiological profiles. In its most extreme form, lexical uniformity in the naming of a concept obtains when two language varieties have an identical name for that concept, or several names with identical frequencies in the two varieties. Table 1.5 presents a toy example, taking inspiration from the sound poems of Dada artists like Tristan Tzara and Hugo Ball. Let us say that the onomasiological profiles for the concept NONSENSE in the Tzara and Ball subcorpus of the International Corpus of the Dada Language (all fictitious, needless to say) are as indicated in Table 1.5. (The actual word forms are not fictitious, though. They are taken from Hugo Ball's poem *Seepferdchen und Flugfische*.)

In our corpus sample for the Tzaran lect (you can think of Tzara as a place or a region, or a sociological group, or a specific style—any type of language variation will do) NONSENSE is named by *tressli* in 35 observations, by *bessli* in 21, and by *nebogen* in 14. In the Ballish lect, the frequencies are 20, 12, and 8 respectively. In absolute terms there is no identity between the language use in both varieties: Tzara uses *tressli* 35 times, Ball only 20. Still, the proportion with which the terms are used is the same in both: both lects use *tressli* in 50% of all observations, *bessli* in 30%, and *nebogen* in 20%. By looking at the relative frequencies of the competing items, we act as if we have 100 observations for each lect, rather than 70 for Tzara and 40 for Ball, as is the case in the raw data. The identity of the relative frequencies of the three terms means that there is a complete overlap between the choices made in both lects: out of 100 instances in which speakers of the Tzaran lect and speakers of the Ballish lect need to make a choice on how to refer to NONSENSE, all 100 cases are decided in the same way, that is, with the same probability for selecting *tressli*, *bessli*, or *nebogen*. This way of describing the relationship between the varieties in terms of overlap is a handle for describing situations in which there is no complete identity of the choices, as in Table 1.6. Focusing once again on the relative figures, the question then becomes: in how many cases of the entire set of 200 observations do the speakers of the Arpian and those of the Picabian lect make the same choice? This can be answered on an item-by-item basis. For *tressli*, there are ten events out of the total set of 200 in which the Arpians do not behave like the Picabians, for *bessli* 5, and likewise for *nebogen*. On a total of 200, then, 20 fall outside the area of overlap between the Arp and the

Table 1.5 Onomasiological profiles for NONSENSE in the fictitious Tzara and Ball dialects

	TZARA		BALL	
tressli	n=35	50%	n=20	50%
bessli	n=21	30%	n=12	30%
nebogen	n=14	20%	n=8	20%

Table 1.6 Onomasiological profiles for NONSENSE in the fictitious Arp and Picabia dialects

	ARP		PICABIA	
tressli	n=28	40%	n=20	50%
bessli	n=28	40%	n=12	35%
nebogen	n=14	20%	n=8	15%

Picabia lect, which means that there is a uniformity of 90% in the lexical choices made by the two groups of speakers.

We can turn this rationale into a formula when we note that the resulting percentage corresponds to the sum of the minimal relative frequencies of each of the alternative terms. For *tressli* the lowest relative frequency, comparing both lects, is 40%. For *bessli* it is 35%, and for *nebogen* it is 15%, with 90 equal to the sum of 40, 35, and 15. The formula takes the form as in (1.1).

(1.1) *Uniformity for a single concept*

$$U_Z(Y_1, Y_2) = \sum_{i=1}^{n} \min\left(F_{Z,Y_1}(x_i), F_{Z,Y_2}(x_i)\right)$$

In this formula, Y_1 and Y_2 refer to the lects we intend to compare, or more precisely, to the datasets that represent the varieties in question. Z is a concept that may be expressed by n competing expressions, from x_1 to x_n. The frequency of an expression x_i in naming Z in the dataset Y_1 is represented by $F_{Z,Y_1}(x_i)$. The $\min(F_{Z,Y_1}(x_i), F_{Z,Y2}(x_i))$ part of the formula refers to the minimum value of the relative frequencies of x_i for Z in Y_1 and Y_2, as illustrated in the toy example above. That minimum value needs to be established for all n items, and then all those minima are summed, as indicated by the sigma sign. If more than one concept is investigated (as would be the obvious thing to do if you want to get a balanced view of the relationship among lects), the uniformity index U is defined as the average of the uniformity indexes of the separate concepts, as in Formula (1.2). In this formula, Z_1 to Z_n are the various concepts, n in number, that are included in the calculation: for each concept Z_i a U-value is determined on the basis of Formula (1.1), and these uniformity values are then averaged.

(1.2) *Average uniformity for a set of concepts*

$$U(Y_1, Y_2) = \frac{1}{n} \sum_{i=1}^{n} U_{Z_i}(Y_1, Y_2)$$

By calculating the uniformity between two datasets as the straightforward mean of the uniformities that hold for individual concept, we imply that all concepts are equally important for the overall relationship between Y_1 and Y_2, in the sense that each concept has an equal share in the calculation of the uniformity. This makes perfect sense in a system-oriented conception of language, in which language structure and language use are well separated, and in which the focus of linguistic inquiry falls on the system. The lexicon of a language is then a collection of mutually linked elements, and all elements have so to speak equal rights as fully fledged components of that collection. From a usage-based perspective, conversely, it could be argued that the frequency with which concepts appear in actual communication should be considered. The degree of lexical uniformity for a high frequency concept has more impact on the commonality between language varieties than that of infrequent concepts. For instance, a uniformity of 50% for concept Z implies that there is a chance of miscommunication in half of the speech events in which Z is mentioned between speakers of the lects at stake, but it makes a real practical difference whether that half is taken from a set of 200 or from a set of 20 utterances. To accommodate such a usage-based perspective on the aggregation of U-values, we introduce Formula (1.3), which defines uniformity index U' as a weighted average. The relative frequency of each concept in the combined datasets, expressed as $G_{Zi}(Y_1 \cup Y_2)$ in the formula, is used as a weighting factor for the uniformity index of each concept separately.

(1.3) *Weighted average uniformity for a set of concepts*

$$U'(Y_1, Y_2) = \sum_{i=1}^{n} U_{Z_i}(Y_1, Y_2) \cdot G_{Z_i}(Y_1 \cup Y_2)$$

Formulae like these are but a first step towards the type of lexical lectometry that we would like to develop. In Chapter 7, the perspective is presented in more detail, while Chapter 8 discusses how it can incorporate a distributional identification of onomasiological profiles. As mentioned earlier, a lectometric approach has so far primarily been an enterprise with a diatopic perspective, but in accordance with a generic conception of *lect*, lectometry (or 'sociolectometry') can be thought of as a generalization of that dialectometric tradition. Dialectometry is not primarily lexical, but variation of vocabulary is regularly included. While the 'Salzburg school' of dialectometry (Goebl 2011) does not usually include quantifiable lexical variation within a single dialect in the analyses, the 'Groningen school' (Wieling and Nerbonne 2015) does. For instance, Wieling, Montemagni, Nerbonne, and Baayen (2014) use survey data with a geographical and sociodemographic stratification to map out, literally, the relationship between dialect areas in Italy. Recent examples of corpus-based lexical lectometry may be found in Grieve, Asnaghi, and Ruette (2013), Grieve (2016), Ruette, Ehret, and Szmrecsanyi (2016), Grieve, Nini, and Guo (2018), and Grieve, Montgomery, Nini, Murakami, and Guo (2019). Like the approach we will demonstrate in later chapters, the latter paper incorporates vector space semantics in the lectometric workflow.

1.4 The lexicon in language variation research

As noted by Durkin (2012), the relative neglect of lexical variation in variationist sociolinguistics contrasts both with the massive popular interest in differences of word use, and with the long standing lexicographic tradition of detailed vocabulary description. This peripheral status of sociolinguistic lexical variation research may well be caused by the specific challenge of studying lexical variation. This challenge derives from the simple fact that words have meaning even before they acquire social meaning. Sociolinguistics broadly speaking describes the socially meaningful association between linguistic and social variables, but words—more conspicuously than any other element of linguistic structure—have a semantic value regardless of whether their distribution and use carries a social significance. *Trousers* is primarily meaningful because it lexicalizes a concept referring to a two-legged outer garment covering the lower part of the body, regardless of the fact that it may secondarily signal Britishness in contrast with *pants*, which is the more common alternative in American English. The presence of this primary level of meaning complicates the sociolinguistic perspective: it asks for a conceptual clarification of the different semantic phenomena involved and the viewpoints from which they can be studied. The previous sections have tried to provide such a clarification, and we can now use that framework to give a short overview of existing areas of lexical variation research. The bibliographical references that follow are meant to be exemplary and illustrative, but in no way exhaustive. Specifically, lexical and semantic change are only mentioned if the diachronic perspective goes hand in hand with a sociolinguistic one. (For an overview of diachronic lexicology per se, see Geeraerts 2015.) Also, while our perspective will be restricted to academic research laid down in papers and monographs, it should throughout be kept in mind that next to these, lexicography provides a major source of information on lexical variation, in the form of labelled senses in alphabetic dictionaries, in the form of usage notes in thesauri, in the form of lect-specific dictionaries like dialect, slang, technical dictionaries, and so on.

Taking our starting point in Table 1.1 (and moving counter-clockwise starting from the top left), four domains of study can be distinguished: lectal polysemy, referential prototypicality effects, conceptual and categorial variation, and lexical variation as a sociolinguistic variable.

LECTAL POLYSEMY—A lectal perspective on semasiological variation is mostly found in sociohistorical and dialectological contexts. In diachronic semantics studies, a lectal perspective takes a sociohistorical form, broadly defined: to what extent is a given semasiological change mediated by lectal factors? To be sure, this is an old topic in historical semantics: see already Meillet (1906) on the role of social factors in the emergence of new senses. Meillet describes how the meaning 'to reach one's destination, to arrive' of French *arriver* (which etymologically means 'to reach the shore, to disembark') arises when the word moves from its original social circle to the general language. Within the social group of sailors, disembarking implies reaching one's destination, but when the word is taken over by the larger community of language users, only the latter reading is retained. More recent examples of a similar perspective are Galván Torres (2021) on geographic factors in the semantic extension of the originally neutral Spanish term *macho*, Wright and Langmuir (2019) on the modifier *neat* in different communities of practice in early-19th-century newspapers, or Kiesling (2004) on the social dynamics behind the development of *dude* as a form of address. Robinson (2010, 2012) illustrates the apparent-time counterpart of such real-time studies, showing how the incidence of the newer senses of *gay* and *awesome* obeys an age-related pattern. Given that age was but one of the demographic factors included in the survey, next to the education, occupation, gender, and place of residence of the respondents, Robinson's work is a scarce example of applying an all-out synchronic sociolinguistic framework to semasiological variation.

In dialectological and geolinguistic research, studies with a semasiological focus are less common than in diachronic research, in the sense that lexical variation is usually studied on the level of the vocabulary rather than that of the individual word. The description and analysis of lexical variation then automatically combines a semasiological and an onomasiological focus, as for instance in Lötscher's study on the dialect vocabulary of Swiss German (2017) or McColl Millar, Barras, and Bonnici's description of lexical variation and attrition in Scottish fishing communities (2014). Dollinger (2017) and Gillmann (2018) are some recent examples of studies concentrating on the polysemy of a single expression.

REFERENTIAL VARIATION—Describing within-sense extensional structure from a variationist point of view is not common. Even Labov's early exploration of prototype effects in the items *cup* and *mug* (1973)—an excursion into lexical variation that remained isolated in his work—does not feature an outspoken variationist dimension. Methodologically speaking, the studies that do exist fall into two categories.

First, in line with work like that of Rosch and Labov that put prototype effects on the lexicological map, surveys, interviews, and experimental paradigms collect data about naming practices, often using pictures of artefacts to map out boundary and centrality effects in categories. The relevance of social variables for category structure was pointed out in Kempton's study of pottery terms in rural Mexico (1981). He observed that gender, professional expertise, modernity of the village, and age all systematically affected the referential structure of ceramic vessel categories. Further examples mainly derive from psycholinguistic studies, providing further evidence for the dimensions mentioned by Kempton: age (Verheyen, Ameel, and Storms 2011; White, Storms, Malt, and Verheyen 2018), gender (Stukken, Verheyen, and Storms 2013; Biria and Bahadoran-Baghbaderani 2016), expertise and familiarity (Malt and Smith 1982; Johnson 2001). Studies like these may also point to demographic variables that are less common from a sociolinguistic viewpoint. Ameel, Malt, Storms, and Van Assche (2009) for instance point to categorization differences between monolingual and bilingual speakers of the same language.

Second, while studies like the above rely on various kinds of elicitation, observational studies make use of spontaneous language use in the form of existing texts. The extensional perspective obviously works best if the texts are 'referentially enriched', that is, when there is some form of access to the denotata of the words. In the study of Dutch clothing terms presented in Geeraerts, Grondelaers, and Bakema (1994), this is achieved by illustrated magazines, so that the pictures accompanying the text provide independent access to the entities being named. The real-world characteristics of the garments (like the length, width, fabric, type of fastening, etc. of trousers) are then transformed into a feature database that allows for the identification of lectal and chronological differences in the category structure of clothing terms. Other examples are Anishchanka, Speelman, and Geeraerts (2015a, 2015b), in which digitized colour information from webpages is used to explore the range and mutual relationship of colour terms. Among other things, these studies reveal that in actual usage, the extension of colour terms may differ across webpage types, like clothing catalogues versus car adverts.

CONCEPTUAL VARIATION—A major consequence of prototype theory is to give frequency and salience a place in the description of semasiological structure. We saw earlier how that idea can be extrapolated to onomasiology. Onomasiological salience may then be roughly defined as the likelihood that a particular categorization will be chosen to talk about a given piece or reality rather than another potentially applicable one. So, what does conceptual onomasiological salience mean from a sociolinguistic point of view? To begin with, we may note that research into the lexicon of a particular group very often presents a mixture of typical (or unique) synonyms and typical (or unique) concepts. The description of the lexicon of, say, contemporary urban youth gangs or 19th-century farriers will usually focus indiscriminately on expressions that are characteristic for the group either because they name generally familiar concepts in unfamiliar ways, or because they name concepts that are less familiar to the general outgroup. Because the former involves the secondary, social meaning of words, and the latter their primary meaning, a methodologically rigorous approach would be well served by separating both. Concepts are represented by a set of synonyms, and accordingly, the presence or weight of the concept needs to be expressed in terms of that formal onomasiological range. The initial publications in culturomics (Michel and Lieberman Aiden 2010), for instance, identifying cultural trends based on the mere relative frequency of words, illustrate an approach that could profit from a stricter distinction between lexical and conceptual variation.

Further, the indirect conceptual perspective that we talked about in Section 1.2 shows up in the many studies that focus on the discursive representation of reality, as in Gabrielatos and Baker (2008) on the image of refugees and asylum seekers in the British press, Nerlich and Koteyko (2009) on climate science, or Peirsman, Heylen, and Geeraerts (2010) on religion names before and after 9/11—the examples could be multiplied ad libitum. This line of research received a methodological stimulus from corpus linguistics, facilitating the identification of collocations as significant elements in the context of a target item, and a theoretical one from the increased interest in metaphor and figurative language triggered by the rise of cognitive linguistics. In disciplinary terms, this way of looking at conceptual variation in discourse aligns with all kinds of 'framing' research in the social sciences.

LECTAL VARIATION AS A SOCIOLINGUISTIC VARIABLE—The lectal distribution of synonyms is the proper focus of lexical variation as a sociolinguistic variable, and any of the many studies describing the lexicon of a particular lect (be it a dialect, natiolect, sociolect, register, specialized language, etc.) inevitably includes the notion of cross-lectal synonymy. However, to arrive at a full parallel with the way sociophonetics studies pronunciation variables, two more things are necessary: the traditional focus on what is typical for a given lect needs to be abandoned in favour of an approach that considers all the alternating equivalent forms that occur in language use, and above all, the relationship between those competing forms needs to be quantified. Such a quantitative approach is still quite rare, though. The following four examples of studies performed in the past decade illustrate the various methodological bases that can be used: surveys, existing resources, either with a lexical focus or not, and text corpora. None of these is ideal, if 'ideal' involves a dataset that consists of spontaneous, non-elicited language use, that allows for the study of a large number of lexical variables, and that is rich in speaker information.

Escoriza Morera (2015) interviewed 72 participants from Cadiz, stratified according to gender, age, and education. Fifteen concepts with three or four lexical variants each (like *empezar*, *comenzar*, and *iniciar* for the concept 'to start, to begin') were presented in texts with different degrees of formality (like a letter to an official institution in contrast with a personal letter). The respondents had to indicate a contextual preference among the available alternatives. The advantage of a dedicated design of this type is the level of control the researcher can exert over the demographic factors and the lexical variables, but an influence of the test situation on the results cannot be excluded. Also, because a forced choice task of this kind tests passive knowledge only, it would have to be supplemented with a production task; avoiding the observer's paradox in such a production task would probably come at the expense of the number of variables that can be tested. Interestingly, contemporary crowdsourcing technology allows to scale up survey-based research of this kind. Leemann, Kolly, and Britain (2018) describe the use of a mobile app inviting users to indicate which variants of 26 words they use; the application then guesses their local dialect. While the number of lexical variants in the survey is still very limited, the number of participants reaches no less than 47 000. Also, the survey includes metadata on the ethnicity, age, educational level, and gender of the participants.

Beal and Burbano-Elizondo (2012) extracted the traditional dialect terms *lad* and *lass*, together with their counterparts *boy/son* and *girl/daughter*, from an existing dataset focused on detecting phonological variation in the modern urban dialects of Newcastle upon Tyne and Sunderland. Because the data were collected in unstructured conversation recorded by an unobtrusive researcher, spontaneity was ensured, and because the conversational pairs tended to talk about friends and family, the concepts were represented by a sufficiently large number of tokens. Likewise, the demographic characteristics of the 35 participants were well known. On the downside, the concepts that can be studied in this way are necessarily small in number and beyond the control of the researcher. A similar approach is found in Franco and Tagliamonte (2021).

Franco, Speelman, Geeraerts, and Van Hout (2019a) used the digitized databases of two large-scale dialect dictionaries of Dutch, viz. the Dictionary of the Brabantic Dialects and the Dictionary of the Limburgish Dialects, to investigate the effect of concept characteristics (vagueness, salience, affect, semantic field) on lexical diversity. These dictionaries, based on systematic surveys conducted between 1960 and 1980, are onomasiologically organized, which allows for the inclusion of large numbers of lexical variables in the analysis. By contrast, the external variables that can be included are restricted to geographic ones; no demographic data are available. Also, the context of the dialect questionnaire probably introduced a ‘typicality bias’ in the responses. (See Pickl 2013 for a similar approach.)

Zenner, Speelman, and Geeraerts (2012) based an analysis of the factors contributing to the success of English loanwords in Dutch on a newspaper corpus of 1.6 billion words, stratified by country (Netherlandic Dutch versus Belgian Dutch) and journalistic type (quality newspapers versus popular ones). The success of borrowed forms was measured in terms of the proportion of the anglicism in the set of onomasiological alternatives representing a concept. As in the previous case type, the size of the corpus allows for the study of many concepts, with the additional advantage that it consists of entirely non-elicited language use. Speaker-related information is still largely absent, though, as would mostly be the case with text corpora of this kind. Again, as in the previous example, additional explanatory variables are included: concept-related ones like the semantic field and the frequency of the concept, and lexeme-related ones like the length of the loanword and its source language frequency.

Overall, these examples show that the methods available for sociovariationist lexical research range from ones that stay close to standard sociolinguistic designs to ones that link up with the tradition of corpus linguistics. The latter are attractive for the further development of sociolexicology, because they hold the promise of exploiting the wealth of digitized usage data that are currently becoming available. Two additional remarks are due at this point.

First, if one uses corpus data, the frequency of a concept (the concept whose synonymous expressions one wants to count) on a given stretch of text will be much lower than that of phonetic variables: you need only a short text to come across most phonemes of a language, but for a fair portion of even the high frequency concepts of a language, you need a much longer text. In other words, corpora will need to be big enough to avoid data sparseness. With the amount of digitized data currently around that is not necessarily a problem, but then another hurdle appears: the bigger the corpus, the more important it becomes to minimize the amount of manual processing that the data may have to be subjected to. The main methodological prerequisite for a corpus-based formal onomasiological approach, in other words, is not just having a big enough dataset, but also a method for establishing semantic equivalence. And that, of course, is where distributional semantics can play a role.

Second, some of the studies in the overview above illustrate well how the various forms of meaning relevant for lexical variation bite each other's tail. Both Franco, Speelman, Geeraerts, and Van Hout (2019a) and Zenner, Speelman, and Geeraerts (2012) introduce variables into the analysis of the formal variation that do not belong to the regular sociovariationist repertoire of demographic and situational dimensions but that relate to the concepts demarcating the set of alternating expressions. To treat lexical variation as a sociolinguistic variable—in other words, to look for social meaning in the lectal distribution of synonyms—means controlling for the meaning of the items, but that control needs to go further than merely identifying synonym sets as items sharing their primary meaning: the characteristics of those meanings (like semantic field, familiarity, recency, abstractness, affect, vagueness, etc.) may have an impact on the kind of formal onomasiological variation one encounters. Two more examples of papers exploring this interaction are Swanenberg (2001), who looks at underlying prototype effects in the variation of bird names in the Dutch dialects, and Franco and Geeraerts (2019), who look at the amount of lexical dialect variation found in names for naturally occurring plants in ecologically consistent geographical regions in the northern part of Belgium and explore the relationship between that variation and the experiential salience of the concepts (i.e. the degree to which the plants in question occur frequently in the everyday environment of the speakers). Current research has only scratched the surface of questions like these, but they need to be emphasized, because they highlight the specificity of lexical variation research in contrast with, say, sociophonetics. Meaning permeates lexical variation research well beyond establishing primary semantic equivalence in synonyms: conceptual onomasiological variation shows that the distribution of primary meanings itself may be socially meaningful, and formal onomasiological variation may depend on the characteristics of the concept underlying the sociolinguistic variable.

1.5 From cognitive linguistics to cognitive sociolinguistics

The framework for the study of lexical variation presented here continues a long-term research programme that was developed since the early 1990s in the Quantitative Lexicology and Variational Linguistics research group of the University of Leuven. To complete this first introductory chapter, the present section provides a synopsis of how the programme emerged and unfolded. Defining publications are the monographs *Diachronic Prototype Semantics* (Geeraerts 1997), *The Structure of Lexical Variation* (Geeraerts, Grondelaers, and Bakema 1994), and *Convergentie en divergentie in de Nederlandse woordenschat* (Geeraerts, Grondelaers, and Speelman 1999). These three are not mentioned in a strict chronological order here, because the 1997 book predominantly draws on research executed before the research project that led to the 1994 book. More importantly, the 1997 publication is also conceptually prior to the others. It takes a semasiological point of view, identifying different prototypicality effects and analysing how they play a role in diachronic language variation. The 1994 volume then effectuates the crucial shift from a semasiological to an onomasiological perspective. More precisely, it mirrors the usage-based analysis of prototype effects with an analysis of conceptual and formal onomasiological salience, and systematically pairs these with lectal variation. The 1999 book further adds the lectometric perspective, with a study of the changing relationship between the vocabulary of Netherlandic Dutch and Belgian Dutch (as referenced in Section 1.3 above).

In addition to laying out the conceptual framework for the research programme, the trilogy exhibits a variety of methods that is relevant for the present study. Most of the case studies in the 1997 monograph are based on historical texts, while the analysis takes the form of a definitional interpretation in the spirit of traditional philological and lexicographical research. The 1994 book goes beyond this, both with regard to the basic data and with regard to the method of analysis. As described above, it is based on manually collected and 'referentially enriched' data, in which the pictures illustrating clothing terms in fashion magazines and the like are translated into featural descriptions of the garments in question. Although statistically unsophisticated, the analysis of this database rests on a quantitative inspection of those feature configurations and their lexicalizations; we will come back to this with more detail in Section 2.1. With the development of a dedicated text corpus (see Grondelaers, Deygers, Van Aken, Van den Heede, and Speelman 2000), the 1999 study marks a move towards digital corpus data. The semantic analysis of the data is done manually, but the introduction of lectometric measures as in Section 1.3 increases the quantitative refinement of the analysis. In the light of this methodological trajectory, the position of the present volume will be clear. Like the main parts of the 1997 and 1999 monographs it is text-based, without referential enrichment, but at the same time, as explained earlier, it takes an extensional outlook that is similar to the 1994 one: lexical items and senses are represented by the set of their instantiations, which in our case are text occurrences rather than actual referents. Also, by using vector space modelling, the approach intends to minimize the role of manual intervention in the semantic description of those occurrences and maximize the use of quantitative analysis.

Following the initial formulation of the research programme in the 1990s trilogy, the dimensions were elaborated predominantly through various PhD projects. Along the *semasiological* dimension, methodological advances took the form of a digitized form of 'referential enrichment' in which colour information from webpages was used in a study of colour terms (Anishchanka, Speelman, and Geeraerts 2014, 2015a, 2015b), while Glynn (2014, 2016) explored clustering techniques on manually annotated data (more on this in Section 2.2). Along the *onomasiological* dimension, a descriptive focus lay on the role of loanwords in onomasiological variation (Zenner, Speelman, and Geeraerts 2012, 2013, 2014, 2015) and on the effect of concept characteristics—like vagueness, familiarity, affect—on the degree of lexical variation and its lectal distribution (Speelman and Geeraerts 2009b; Geeraerts and Speelman 2010; Franco and Geeraerts 2019; Franco, Geeraerts, Speelman, and Van Hout 2019a, 2019b). Methodologically important for the current project was the introduction of a type-based vector analysis in the identification of synonymy. In the present book, we pursue a token-based approach, meaning that the entities for which we build vectors are individual occurrences of words. In a type-based approach, the vector representation involves words as a whole rather than the individual instances thereof. (For more detail, see Section 2.1 and Chapter 3.) In the evolution of the research programme, such a type-based approach constitutes an intermediate step between the pre-2000 groundwork and the current study: see Heylen, Peirsman, and Geeraerts (2008); Heylen, Peirsman, Geeraerts, and Speelman (2008); Peirsman, De Deyne, Heylen, and Geeraerts (2008); Peirsman, Heylen, and Geeraerts (2008); Peirsman and Geeraerts (2009); Peirsman, Geeraerts, and Speelman (2010, 2015). Along the *lectometric* dimension, methodological attention went to the statistical refinement of the descriptive techniques (Speelman, Grondelaers, and Geeraerts 2003, 2006) and to the incorporation of a type-based vector approach (Ruette, Speelman, and Geeraerts 2011, 2014; Ruette, Geeraerts, Peirsman, and Speelman 2014; Ruette, Ehret, and Szmrecsanyi 2016). Descriptively, the observations about the developing relationship between the main varieties of Dutch were elaborated in Daems, Heylen, and Geeraerts (2015), Daems, Zenner, and Geeraerts (2016), and Daems (2022). Following the method and the model of the 1999 book, Soares da Silva (2010, 2014) describes the evolution of Portuguese as a pluricentric language.

While the current monograph continues the central line of lexical semasiological, (formal) onomasiological, and lectometric research, it should be mentioned that in the course of the past two decades, the framework was also developed in a number of tangential directions that will not, or not conspicuously, be represented in the following chapters. Four types of related research lines may be mentioned and illustrated with a few representative publications.

To begin with, several publications illustrate *conceptual onomasiological* investigations as defined earlier, specifically, indirect conceptual onomasiology. In the present context, Peirsman, Heylen, and Geeraerts (2010) is a particularly relevant example because it uses type-based word space models. Using such models, it identifies the words that are saliently associated in a Dutch newspaper corpus with the concepts ISLAM and CHRISTIANITY before and after the attacks of 11 September 2001. Comparing their degrees of association before and after 9/11 reveals how the event triggered changes in the conceptualization of religions and the use of religious terms. Other studies in this group cover a bigger diachronic range. Zhang, Geeraerts, and Speelman (2015, see also Zhang 2016) trace diachronic variation in the metonymic patterns with which target concepts like WOMAN, BEAUTIFUL WOMAN, or WOMAN BELONGING TO THE IMPERIAL HOUSEHOLD have been expressed throughout the history of Chinese. The relative frequency of different metonymical patterns (like LOCATION FOR LOCATED or ACTION FOR AGENT) in the total set of metonymic expressions that occur for a given target concept in a specific historical period points to changes of conceptualization. For instance, changes in the metonymic patterns used for the expression of the target concept BEAUTIFUL WOMAN suggest a historical and cultural shift of the beauty ideal from intrinsic attributes to external decorative attributes. A similar approach is illustrated by a series of papers on the history of English *anger* and the relative frequency of metaphorical, metonymical, literal construals of the concept (Geeraerts and Gevaert 2008; Geeraerts, Gevaert, and Speelman 2012).

A second branch consists of *onomasiological studies applied to non-lexical variables*: which factors explain the variation between functionally equivalent expressions in the morphological, syntactic, constructional realm, and to which extent does that alternation evince an interaction with lectal variation? Beginning with Grondelaers' PhD on the Dutch presentative article *er* (see Grondelaers, Geeraerts, and Speelman 2002, 2008), several doctoral theses were completed within the Quantitative Lexicology and Variational Linguistics group pursuing this line. Tummers (Tummers, Speelman, and Geeraerts 2004, 2005; Tummers, Speelman, Heylen, and Geeraerts 2015) studied inflectional variation in adjectives modifying definite neuter nouns in Dutch. De Sutter (De Sutter, Speelman, and Geeraerts 2005, 2008) looked at word order variation in Dutch clause-final verb phrases. Heylen (2005) performed a quantitative corpus analysis of word order variation in the middle field of the German clause. Levshina (Levshina, Geeraerts, and

Speelman 2013a, 2013b) compared causative constructions with the Dutch auxiliaries *doen* and *laten*. Applying a quantitative usage-based methodology, these studies show how alternating constructions appear to be sensitive to a variety of factors: shades of meaning, sentence structure, and lexical environment, discourse, lectal differences. Again, some of these publications incorporate a type-based distributional approach, see for instance Levshina and Heylen (2014).

Third, the programme was expanded to *lectometry beyond the lexicon*. Because the colloquial variety of Belgian Dutch is specifically characterized by an extensive catalogue of morphological phenomena, Plevoets (Plevoets, Speelman, and Geeraerts 2007, and see Geeraerts 2010b) applied a lectometric perspective to non-lexical markers of informality in colloquial Belgian Dutch. A more far-reaching addition to the framework, in a sense, is to complement production-based research with the study of perceptions and attitudes. A complete picture of a sociolinguistic situation requires not only knowledge of what people actually do with their language, but also of how they perceive and evaluate language diversity. Thus, Impe, Geeraerts, and Speelman (2009), and Speelman, Impe, and Geeraerts (2014) map out the mutual intelligibility of regional Dutch accents in Belgium and the Netherlands. Speelman, Spruyt, Impe, and Geeraerts (2013), and Rosseel, Speelman, and Geeraerts (2018, 2019a, 2019b) explore evaluative attitudes regarding colloquial Dutch, with a methodological focus on newer experimental paradigms from the field of social psychology.

Finally, distributional corpus analysis was put to work in an applied linguistic setting, in the context of *terminology research*: see Bertels and Speelman (2014); Heylen and De Hertog (2015); Heylen and Bertels (2016); Grön and Bertels (2018).

Situating it in a wider disciplinary context, the research programme with all its ramifications embodies the framework of *cognitive sociolinguistics* as represented by among others Kristiansen and Dirven (2008); Geeraerts, Kristiansen, and Peirsman (2010); Geeraerts (2005, 2016b, 2018b); and Kristiansen, Franco, De Pascale, Rosseel, and Zhang (2021). Emerging from the field of cognitive linguistics (as covered by handbooks like Geeraerts and Cuyckens 2007; Littlemore and Taylor 2014; Dąbrowska and Divjak 2015), cognitive sociolinguistics combines the interest in the social aspects of language that characterizes the tradition of sociolinguistics with the conception on meaning that lies at the heart of cognitive linguistics. That conception crucially incorporates an extensionally expanded and usage-based view as described in Section 1.2, with flexibility, salience, and prototypicality effects as structural features of meaning. (The methodological relevance of these characteristics will be further analysed in Chapter 2.) Such a combination of a social and a cognitive semantic view is intrinsically motivated from both disciplinary sides. As we noted, sociolinguistics is still in need of an articulated theory of meaning and a matching methodology, not just because meaning (and the lexicon) have been understudied from the sociolinguistic point of view, but also because semantic and functional equivalence is constitutive of the sociolinguistic variable as one of its central concepts. On the other side, embracing a usage-based viewpoint like cognitive linguistics does inevitably entail a description of variation in language use—and that variation will most often be shaped by social factors.

Against this background, a major motivation for the present book is the deep-seated conviction that lexical variation research has an exemplary position for cognitive sociolinguistics. For defining concepts and perspectives and for trying out methods, the lexicon, with its long and strong tradition of word meaning research, constitutes a testing ground par excellence for any approach that tries to take meaning seriously in the description of language.

The bottom line

- Lexical variation research is organized on the basis of two dimensions, one linking linguistic forms to meaning, and an orthogonal one considering that association from the point of view of lectal variation. Crucially, both dimensions can be traversed in two directions. This produces a distinction between semasiology and onomasiology for the first dimension, and a distinction between lectal variation as an explanatory or as a response variable for the second dimension.
- The addition of an extensional, usage-based layer to the traditional distinction between words and senses adds new phenomena to the scope of lexical research: on the semasiological side, the internal structure of senses (like prototypicality effects among the exemplars of a category), on the onomasiological side, the distinction between conceptual and formal onomasiology.
- Treating lexical variation as a sociolinguistic variable in the sense of variational sociolinguistics is situated at the level of formal onomasiology. On an aggregate level, this can be the basis of a lexical lectometry.
- In usage-based corpus semantics, the individual occurrences of words in specific utterances act as the exemplars instantiating the category. For a vector space approach, this implies modelling meaning at token level rather than at type level.
- The approach to lexical variation research presented in this book is part of a broader framework, cognitive sociolinguistics, that combines a cognitive linguistic conception of semantics with a sociolinguistic interest in linguistic variation.

2

Distributional semantics and the fog of meaning

The line of investigation developed in this book lies at the crossroads of lexical variation research and distributional semantics. Just as Chapter 1 sketched the lexicological framework of what is about to follow, the present chapter provides some background regarding distributional semantics and its methodological position in word meaning research. Two main points will be discussed. To begin with, in Sections 2.1 and 2.2, we will introduce the specifics of our distributional approach, situate it in the context of alternative methods in lexical semantics, point out alternative forms of distributional corpus semantics, and motivate the specific choice we are making. In particular, we will introduce the distinction between so-called count-based and prediction-based models and argue that our selection of a count-based approach (in spite of it being the least popular in computational linguistics) may be justified by the need for more transparency in the use of distributional modelling. To support such a search for more transparency, later chapters of the book will use a flexible visualization tool that will allow us to explore the effect of different parameter choices in distributional models. The tool itself will be described in Chapter 4, and the parameters that we will include in our models, together with a technical specification of the workflow we will follow, are the subject matter for Chapter 3. In the present chapter, we anticipate on that discussion by giving a rough and largely informal outline of the architecture and the methodological background of the count-based distributional models that we will be exploring.

The need for looking more closely and more analytically at distributional modelling does not however arise exclusively from a desire to see more clearly what is happening behind the screens of the distributional algorithms. A more fundamental reason lies in the object of modelling itself: meaning. The positive vibe currently surrounding vector space models suggests that they tap directly into meaning, but meaning is notoriously untraceable. Of all the levels and dimensions of linguistic structure, it is methodologically the most difficult to pin down, and accordingly, it is adamant to scrutinize the epistemological status of the information unearthed by distributional models. This exploration will be found in Chapter 5 (and obviously, the exploratory tool introduced in Chapter 4 will have

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0002

a crucial role in the analysis), but in Sections 2.3 and 2.4 we will try to say a bit more about the conceptual background of that exploration. We will distinguish between different methodological inroads into meaning and provide evidence that these perspectives need not coincide. On a fundamental level, this raises a question about the indeterminacy of meaning: if none of the perspectives is by nature dominant or preferential, perhaps we may have to accept that there is an amount of indeterminacy at the core of semantics.

2.1 From contexts to clusters

For a better understanding of the vector space approach, we can develop an analogy with the referent-based methodology of *The Structure of Lexical Variation* (Geeraerts, Grondelaers, and Bakema 1994). As we discussed in Chapter 1, there is a basic similarity between the two in the sense that both take an extensional perspective. In both cases, the semantic description takes its starting point in the instantiations of a lexical item: denotata (actual items of clothing represented by pictures in magazines) in one case, textual occurrences in the other. Also, those extensional entities are described in terms of characteristic features: in one case, properties of real-world referents like the garments in *The Structure of Lexical Variation*, co-occurring words in the other. Grouping the extensional instances together on the basis of their similarity can then be achieved by comparing features: instantiations are similar to the extent that they share features. Here is how it worked in the 1994 book for the lexical item *vest* 'cardigan', such as it could be found in a number of Netherlandic fashion and lifestyle magazines. The relevant descriptive features are as follows:

LENGTH

[1] The garment is not longer than the waist.
[2] The garment is roughly as long as the waist.
[3] The garment is longer than the waist.

FASTENING

[1] The garment does not have a fastening; the panels cannot be attached to each other.
[2] The garment has a zipper fastening.
[3] The garment has a full, single-breasted button fastening.
[4] The garment has a full, double-breasted button fastening.

MATERIAL

[1] The garment is made of a relatively thick and smooth fabric.
[2] The garment is made of coarsely knitted material.
[3] The garment is made of finely knitted material.
[4] The garment is made of a towel-like material.

SLEEVES

[1] The garment does not have sleeves.
[2] The garment has long sleeves.

Different types of *vest* found in the data can then be summarily described with a notation like C2332. This describes a garment of type global C, that is, of the cardigan type (as contrasted for instance with trousers, which require a different set of descriptive features). The subsequent positions in the notation identify a specific value on the four dimensions listed above. So, the referential configuration C2332 is waist-long, has a single-breasted button fastening, is finely knitted, and has long sleeves.

Charting the similarities between the garments as pictured in the magazines can be done graphically with a plot as in Figure 2.1. The boxes indicate the various features that seem relevant in the structure of the item. Each box represents a specific feature, more particularly, the dominant value(s) of each of the four descriptive dimensions. Thus, the KNITTED box comprises exemplars with value 2 or 3 on the dimension MATERIAL. Each box contains the referential configurations that exhibit the feature represented by the box, together with the absolute frequency with which that configuration occurs in the data for *vest* 'cardigan'. Underlying the figure is a matrix of the kind illustrated in Table 2.1, in which each garment in the dataset is described individually. The figure establishes that there is a correlation between intensional and extensional salience: the kind of *vest* that combines most of the dominant values for each dimension (it can be found in the centre of the figure) is also the one that is most frequently compared to the other configurations of descriptive characteristics. The category as a whole appears to be structured in terms of a maximally overlapping high frequency core region surrounded by a peripheral area with low frequency and decreasing overlapping of attributes. This relationship between the frequency of descriptive features and the salience of members of a category is well known in the literature on prototypicality: 'The more an item has attributes in common with other members of the category, the more it will be considered a good and representative member of the category' (Rosch and Mervis 1975: 582).

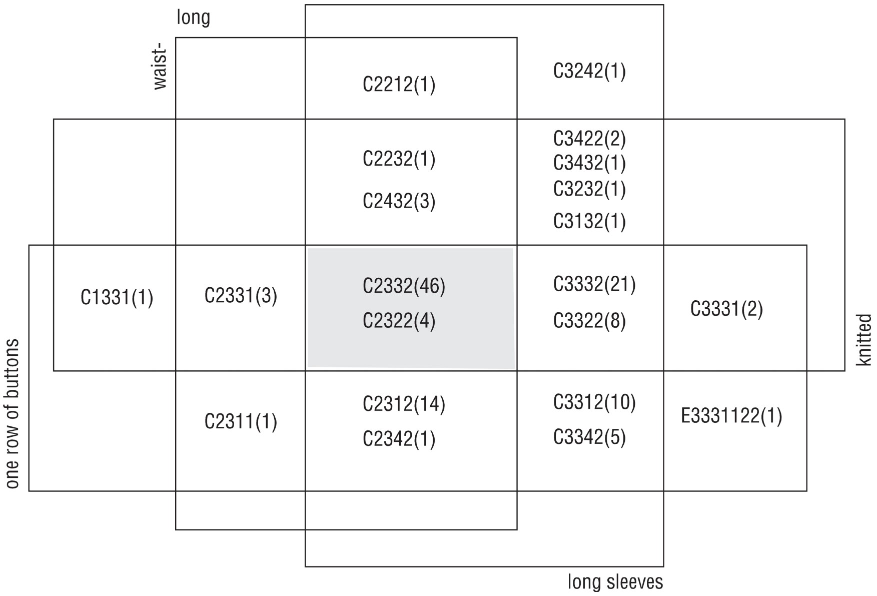


Figure 2.1 Semasiological structure of *vest*. Reproduced from Geeraerts, Grondelaers, and Bakema (1994)

Table 2.1 Partial matrix underlying the analysis of *vest* in Figure 2.1

	LENGTH	FASTENING	MATERIAL	SLEEVES
Exemplar 1	2	2	1	2
Exemplar 2	2	2	3	2
Exemplar 3	2	4	3	2
Exemplar 4	2	4	3	2
Exemplar 5	2	4	3	2
Exemplar 6	2	3	3	2
Exemplar 7	2	3	3	2

If we extrapolate this model to a corpus-based approach, three questions may bring out the similarities and differences with the denotational analysis. How do you decide on the relevance of features, how do you fill a gap in the features, and how do you determine the similarity between items? The first question involves the observation that the attributes in the analysis of *vest* are preselected. They were manually and preliminarily collected on the basis of real-world familiarity with the garment types and a summary inspection of the pictures, but as such, they already presuppose an initial categorization of the items as a cardigan type of clothing. Features that are relevant for a different major category (like for instance the length of the legs for trousers) are not included, and neither are attributes like the colour of the cardigan, because it is not considered distinctive for this type of garment. In the corpus-based set-up, by contrast, there is no such initial restriction. We want to analyse our targets in function of the words that they co-occur with, but it would be useful if we could concentrate on the most pertinent context words. To make the procedure more efficient, a mechanism therefore needs to be introduced to identify the most relevant context items. This is done by finding words that occur together with the target on a more than average basis. With a target word like *car*, the words *speed*, *traffic light*, and *driver* occur more often than, say, *otherworldliness*, and it therefore makes sense to consider *speed* and so on as relevant features, and *otherworldliness* not. Identifying relevant words can be done in various ways. For instance, one can only consider words with higher frequencies, or one can exclude function words and only take into account open word classes, or one can concentrate on words that have a specific syntactic relation with the target, like subjects and objects of verbs or the modifying adjectives of nouns. But while these are a priori selections, one can also use the corpus itself to determine items that are particularly relevant for the target: what does the corpus itself say about relevant words? This information can be extracted by statistical means. There are actually various ways of achieving this (see Evert 2009; Wiechmann 2008 for overviews), but we will here present an approach that we will often refer to in the more technical chapters of the book, and that has also played a seminal role in the evolution of corpus semantics.

In Church and Hanks (1989), the Pointwise Mutual Information index or *pmi* is defined in terms of the probability of occurrence of the combination x, y, compared to the probabilities of x and y separately. The probability P(x) of x and P(y) of y in a corpus is given by their relative frequency in the corpus. Given these probabilities, the theoretical probability of x and y occurring together is, by a general law of probability theory, the product of P(x) and P(y). But we can also measure the actual probability of x, y, by determining its relative frequency in the corpus. Then, we compare P(x, y) with P(x)*P(y): if the probability P(x, y) of the combination is bigger than what we might expect on the basis of the probabilities P(x) and P(y) of the constituent parts, we have an indication that the observed combination is not just due to chance. If there is an actual combinatory association between x and y, then the joint probability P(x, y) will be much larger than chance P(x)*P(y). Passing over a number of technical refinements (like the fact that the calculation of P(x, y) takes into account the span within which combining elements are sought to the left and the right of the node), the bottom line is this: a statistical analysis of co-occurrences can help researchers to pinpoint relevant context words: they are the ones with a high pmi value. Measures of associative strength like pmi can be used to select context words as features, or they can be used to weigh context words, in particular by multiplying the frequency contribution of a context element with its pmi value.

Now it could be asked whether such an identification of significantly occurring context words might not be sufficient. We are looking for appropriate context words with which to characterize the occurrences of words in specific, individual utterances, but why would we want to describe individual instances of lexical usage (and look for similarities among those) if the regularly occurring context words can be retrieved anyway on a less granular level? Why look beyond the overall patterns of co-occurrence retrieved by measures like pmi? Don't we have enough information already if we establish that our target occurs together in a significant way with such-and-such words? The answer lies in the potentially misleading character of such an overall analysis: there may be differences of meaning or usage hiding within the total set of pertinent context words. *Vest* may again serve as an example. Mostly in Belgian Dutch, it does not only refer to cardigan-like garments, but also to jackets as worn as the upper part of a suit, and other blazer-like items. The characteristics of this class of garments are different from the cardigan kind, though: jackets typically have a lapel while cardigans do not, and conversely, cardigans are mostly knitted, which is unusual for jackets. A global tally of the features of exemplars of *vest* might then show that the frequency of lapels is as high as the frequency of knitted fabrics. But this would miss the fact that these characteristics belong to different overall classes, that is, that there are subsets of characteristics that often occur together, as the 'jacket' class of *vest* and the 'cardigan' class of *vest*. This difference does show up if you map out the way features co-occur in individual exemplars. In the terminology of distributional semantics, this is the distinction between a type-based approach and a token-based approach. A *type-based* approach looks globally at the words that co-occur with the target, while a *token-based* approach looks at the words that co-occur with the target in each specific instance of use, and then tries to group those tokens on the basis of their similarity.

The next question to consider derives from differences rather than similarities between a referential and a corpus approach. If we describe garments with features like the ones listed earlier, we can be fairly sure that each of the descriptive dimensions can be assigned to each and every exemplar that we come across: all cardigans have a certain width, fabric, and so on. But in the corpus data, the descriptive dimensions are co-occurring words, and we can be sure that *car* does not always occur together with *driver*. On the contrary, it is precisely because it does not always occur together with *driver* that the presence of *driver* in the utterance has a distinctive value. Also, the value that we describe for each dimension (the potentially co-occurring words) indicates the presence or absence of the context word in question, and not a qualitative specification like 'zipper' for the dimension FASTENING in the case of *vest*. On top of that, the number of potential dimensions is big, even if we restrict the potential context words in one of the ways described a moment ago. In a more technical perspective this means that a matrix like that in Table 2.1, when applied to corpus data, will look highly unpopulated. Imagine characterizing instances of *car* on the basis of context words like *driver*, *traffic*, *hood*, *windshield*, *passenger*, *bicycle*, *speed*, *drive*, *licence*, *road*, *highroad*, *accident*, *trip*, *repair*, *commute*, *jam*, and so on: in each individual utterance featuring *car*, at most a few of these context words will be present. For a single occurrence of *car* you will get a long string of dimensions, with values that are mostly zero, because the potential context word in question does not appear near to the target token. In such a sparse matrix, it is difficult to find groups of similar exemplars because the basis of the similarity (a few words out of the many) is so small.

There are several mutually non-exclusive ways to arrive at a more populated matrix. One is through statistical techniques such as singular value decomposition, which reduces the dimensionality of a matrix and turns sparse vectors into denser ones. Another is to enrich the matrix with secondary information about the context words. The rationale is as follows. When *car* occurs with a context word like *driver*, we would like to establish the similarity with utterances that contain near-synonyms like *chauffeur*, *motorist*, *automobilist*. An utterance that combines *car* and *driver* will be closer to one that combines *car* and *chauffeur* than to one that combines *car* and *accident*. But you will not notice that if you merely indicate the presence or absence of *driver*, *chauffeur*, *accident* in the context of *car*. Now if we look at type level at *driver*, *chauffeur*, *accident*, we are likely to find that based on their global co-occurrence patterns, *driver* and *chauffeur* are closer with regard to each other than with regard to *accident*. The trick then consists of including that type-level information in the description of the specific tokens in which *car* combines which *chauffeur* or combines with *driver*. How this is achieved will be described more technically in Chapter 3, but because it is such a fundamental aspect of the methodology, and also because it is a slightly complicated idea, we will now go through an informal presentation of the procedure.

In the previous chapter, we used a fancy example referring to the Dada art movement. To continue that artistic line, let us now return to the Parisian avant-garde of the early 20th century, and imagine that we witness Picasso at three different occasions, in the company of various people as recorded in Table 2.2 (in which plus and minus signs indicate presence or absence in the encounter). If you take into account the fact that Georges Braque and Juan Gris are painters belonging to the same Cubist movement as Picasso, while Eric Satie is a composer and Guillaume Apollinaire a poet, it may be concluded that the first and the second meeting are more similar to each other than to the third. But if you don’t know Braque, Gris, Satie, and Apollinaire, you first need to explore their background, for instance by keeping track of the company they repeatedly appear in: how often have Braque, Gris, Satie, and Apollinaire met with, say, Constantin Brancusi, Sonia Delaunay, Marc Chagall, Jean Cocteau, Gertrude Stein, Blaise Cendrars, Francis Poulenc, Nadia Boulanger, Arthur Honegger in the same year in which you saw Picasso? The result could be as in Table 2.3. Note that this is a type-level matrix, in which we describe Braque, Gris, Satie, and Apollinaire in general, not specific occasions as in the original matrix with the Picasso encounters. Also note that Brancusi, Delaunay, and Chagall are visual artists; Cocteau, Stein, and Cendrars are literary figures, and Poulenc, Boulanger, and Honegger are composers. You don't need to know that, though, to observe that Braque and Gris dwell in the same circles more than the other two. We can then use that knowledge in our classification of the Picasso encounters. This is done by plotting the Picasso events, not against the four dimensions represented by Braque, Gris, Satie, and Apollinaire directly (as in Table 2.2), but against the same dimensions that we use to characterize Braque, Gris, Satie, and Apollinaire in Table 2.3. In that table, Braque, Gris, Satie, and Apollinaire are each represented by their own row. So, an event featuring Braque and Gris can be described by a combination of the row representing Braque and the row representing Gris. There are several ways of achieving that combination, but for this example, we average over the values on the dimensions. In other words, for the first Picasso encounter, we build up a new row (a vector) in which the value for the dimensions (Brancusi, Delaunay, Chagall, Cocteau, Stein, Cendrars, Poulenc, Boulanger, and Honegger) is the mean of the values of Braque and Apollinaire for those dimensions (because Braque and Apollinaire but not Gris and Satie were

Table 2.2 Toy example of Picasso social encounters

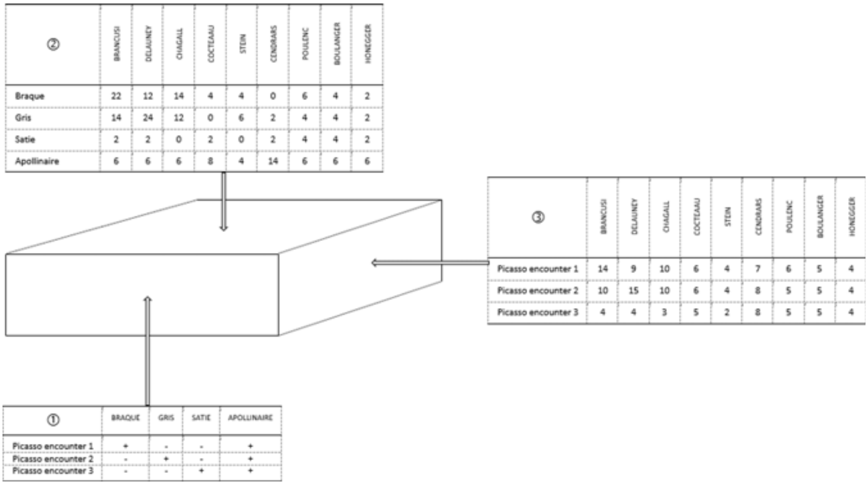
	BRAQUE	GRIS	SATIE	APOLLINAIRE
Picasso encounter 1	+	−	−	+
Picasso encounter 2	−	+	−	+
Picasso encounter 3	−	−	+	+

Table 2.3 Second-order vectors for Picasso's companions

	BRANCUSI	DELAUNEY	CHAGALL	COCTEAAU	STEIN	CENDRARS	POULENC	BOULANGER	HONEGGER
Braque	22	12	14	4	4	0	6	4	2
Gris	14	24	12	0	6	2	4	4	2
Satie	2	2	0	2	0	2	4	4	2
Apollinaire	6	6	6	8	4	14	6	6	6

Table 2.4 Enriched matrix for Picasso's social encounters

	BRANCUSI	DELAUNEY	CHAGALL	COCTEAAU	STEIN	CENDRARS	POULENC	BOULANGER	HONEGGER
Picasso encounter 1	14	9	10	6	4	7	6	5	4
Picasso encounter 2	10	15	10	6	4	8	5	5	4
Picasso encounter 3	4	4	3	5	2	8	5	5	4



②	BRANCUSI	DELAUNEY	CHAGALL	COCTEAAU	STEIN	CENDRARS	POULENC	BOULANGER	HONEGGER
Braque	22	12	14	4	4	0	6	4	2
Gris	14	24	12	0	6	2	4	4	2
Satie	2	2	0	2	0	2	4	4	2
Apollinaire	6	6	6	8	4	14	6	6	6

③	BRANCUSI	DELAUNEY	CHAGALL	COCTEAAU	STEIN	CENDRARS	POULENC	BOULANGER	HONEGGER
Picasso encounter 1	14	9	10	6	4	7	6	5	4
Picasso encounter 2	10	15	10	6	4	8	5	5	4
Picasso encounter 3	4	4	3	5	2	8	5	5	4

①	BRAQUE	GRIS	SATIE	APOLLINAIRE
Picasso encounter 1	+	-	-	+
Picasso encounter 2	-	+	-	+
Picasso encounter 3	-	-	+	+

Figure 2.2 Graphical representation of the steps in the distributional workflow

Picasso's companions in the first encounter). If we repeat that procedure for the other Picasso encounters, we get Table 2.4.

In Figure 2.2, the procedure is represented graphically as a process in which we gradually build up a cube (or more precisely, a rectangular prism). The dimensions of the matrix on the front side become rows of the matrix on the top, and then the third side is added by combining the rows of the front side with the dimensions of the matrix on the top.

Table 2.4 is more densely populated than the original matrix in Table 2.2, and this makes it easier to calculate the degree of similarity between the three Picasso encounters. This can be achieved by using distance measures similar to the ones that we saw in Section 1.3. Mathematically speaking, vectors (the rows in our tables) define the position of a point in a multidimensional space, more precisely, a space with as many dimensions as there are columns in the matrix. If the values on two vectors are similar, the distance between the points defined by those vectors is small. So, if we simply calculate the Euclidean distance between the three Picasso encounters, we get the following values:

encounter 1 with regard to encounter 2:	7.34
encounter 1 with regard to encounter 3:	13.45
encounter 2 with regard to encounter 3:	14.52

These figures show that enriching an initial matrix as in Table 2.2 with second-order information as in Table 2.3 reveals patterns that would otherwise remain hidden. If you only consider Table 2.2, the three encounters are similar, with just two companions in each case. But if you include the relations between those companions as registered by Table 2.3, more structure emerges. For our lexical semantic purposes, we will act similarly, then: the 'Picasso encounters' are the instances of word use that we would like to group according to their similarity; Picasso's companions are the context words we include in the analysis of that similarity, and the information of Table 2.3 consists of the type vectors that enrich the initial token matrix.

To round off this initial description of the distributional approach that we will pursue (more detail, of a technical kind, follows in Chapter 3), two remarks are due. First, the distance measure we will use in the following chapters will not be Euclidean. In a distributional framework it is customary to use cosine distance. Without going into mathematical detail, this can be understood as follows. A vector defining a point in a multidimensional space can be thought of as an arrow connecting that point with the origin of the space, which is simply the point where all the dimensions have value zero. This is easy to imagine for a three-dimensional space: if we have a vector with three positions, each of the values positions a point with regard to each of the dimensions of the 3D space, and we can draw an arrow from point zero to the point so defined. In a multidimensional space, it is no longer possible to imagine this visually, but the idea is the same: a vector is an arrow connecting a point to the origin of the space. Given two different points in the space, the arrow that connects them to the origin will have a certain angle with regard to each other, and cosine is a standard geometrical measure to describe that angle. The closer lines are with regard to each other, the smaller their cosine value, and so, cosine can be used as a distance measure.

And second, when we have more than three points to compare, grouping the points together based on their distances is an additional step. It involves building a new matrix—a distance matrix—comprising all the pairwise distances between the tokens under consideration, and then identifying the groups of closely related tokens. We will not illustrate the procedure here: see Chapter 3 with the technical description of the workflow we will follow.

2.2 The diversity of distributional semantics

Distributional corpus semantics does not always take the form described in the previous section, and lexical semantics does not always take the form of distributional corpus semantics. Next to a text-based research type, there are two other major methodological perspectives: the referential one as in the 1994 clothing terms study that we referenced above, and a psycho-experimental one as illustrated by some of Rosch's original prototypicality studies. In Sections 2.3 and 2.4, we will discuss the methodological relationship between these three fundamental sources of semantic information. In the present section, we focus on the various appearances of distributional corpus semantics, and our own position within that spectrum (which will be schematically represented by Table 2.5 near the end of the section). As a first step, we may have a look at the historical lineage of distributional corpus semantics. In current publications, it is customary to refer to Zellig Harris (1946, 1951, 1954) and John Rupert Firth (1957), if not as founding fathers then at least as crucial inspirations for the contemporary development of the distributional method. These standard references are, however, somewhat misleading, for the following reasons.

To begin with, for reasons that are detailed in Geeraerts (2017), the distributionalism of Harris does not fit the current approaches very well. The analysis of usage at the level of actual utterance tokens, the use of statistical methods, the rejection of a strict layering of linguistic structure, and the overall semantic rather than formal goals of current approaches are substantially different from the Harrisian original. The final point in particular is relevant: the role of semantic analysis in Harrisian and contemporary distributional approaches differs in important respects. In current approaches, establishing semantic equivalence is the goal par excellence of the analysis. For Harris, by contrast, the main target is establishing the formal structure of the language, consisting of entities and patterns for combining them at different structural levels. All those formal entities are assumed to have a meaning or function of their own which is reflected precisely in their distributional properties, but identifying the forms is the fundamental goal. As a correlate of this difference of focus, semantic equivalence is an input datum for Harris, rather than an output observation; it is a tool rather than a target—but how is it established? To determine whether two phonetic segments can be considered variant realizations of a single phoneme, Harris defines a 'repetition test'. The repetition test is similar to the identification of minimal pairs in phonology but focuses on the identification of identity rather than distinctiveness. If a phonetic variation between word form *a* and word form *a*' does not prevent us from saying that *a*' is a repetition of *a*, then the phonetic variants between *a* and *a*' can be seen as instantiations of a single phoneme. But clearly, whether *a* and *a*' are repetitions of each other turns on their semantic equivalence. Thus, the distinction between, say, *pet* and *bet* is phonemic, because the semantic differences between both allow us to say that they are different words, rather than repetitions of the same entity. Conversely, a slightly more or slightly less aspirated pronunciation of the p in *pet* does not correlate with semantic differences, and so the more or less aspirated forms can be considered repetitions. This procedure is then the only point where for Harris semantic considerations enter the analysis: 'In principle, meaning need be involved only to the extent of determining what is repetition. If we know that *life* and *rife* are not entirely repetitions of each other, we will then discover that they differ in distribution (and hence in "meaning")' (1951: 7). The distinction with contemporary approaches will be clearer now: current distributionalism will try to establish whether *life* and *rife* are synonyms by looking at their textual distribution, where Harrisian distributionalism will take their non-synonymy (as detected by an intuitive repetition test) as the initial stepping stone for an incremental process leading to a multi-layered description of the formal structure of the language.

By contrast, the reference to Firth as a foundational figure is uncontested. His aphorism 'You shall know a word by the company it keeps' (1957: 11) adequately synthesizes what contemporary distributional semantics tries to do, and there is in fact a historical line leading from Firth to contemporary corpus semantics. We will come back to that presently, but not without noting—as a second nuance regarding a simplistic reference to Harris and Firth as founding figures—that Firth's ideas did not come out of the blue. For one thing, they derive theoretically from the structuralist framework that dominated a lot of linguistic thinking in the middle of the 20th century. The distinction between paradigmatic and syntagmatic relations is one of the central concepts in De Saussure (1916), roughly corresponding with similarity and combinability. Firth's essential insight is to recognize the syntagmatic relations between words (the way in which they occur together) as diagnostic for their meaning. For another, that diagnostic value is part and parcel of the tradition of lexicography. The great historical dictionary projects that were started in the mid-19th century were all, in their own painstakingly manual way, corpus-based: a dictionary like the *Oxford English Dictionary*, the German *Deutsches Wörterbuch*, or the Dutch *Woordenboek der Nederlandsche Taal* rests on a huge collection of quotations extracted from historical texts. And the method used by the historical lexicographers for analysing and classifying those quotations was based on the principle of interpretation in context, that is, on an examination of the elements co-occurring with the target word in the attested usage cases. How else, after all, could texts be interpreted and words defined than by carefully looking at the contexts in which a word presents itself? The systematicity with which the data are currently collected may have improved compared to these older dictionaries, but the idea itself of using a large repository of real language data as the empirical basis for semantic descriptions is rather a continuation of the finest traditions of philological and lexicographical work than a radical break with the past.

It is perhaps no surprise then that the major impact of Firth's views came about through a dictionary project. For his work on the *Collins Cobuild English Language Dictionary* (Sinclair and Hanks 1987; Sinclair 1991), John Sinclair combined a Firthian perspective on language use with digital technology: a 20 million word corpus of contemporary English was compiled as the empirical basis for the dictionary, and statistical methods for exploring the corpus were developed: we have already mentioned the introduction of the pmi measure. This effectively laid the foundation for contemporary quantitative corpus linguistics. Zooming in on corpus lexicology, we may then distinguish three methodological approaches, one of which is vector space modelling of the kind we focus on this book. Brought down to its essentials, that kind of modelling consists of two components: identifying relevant context features and clustering tokens based on their similarity. Importantly, quantitative considerations play a dominant role with regard to both components. Building a token matrix and enriching it by importing the type vectors of context words relies on measures of co-occurrence like pmi, and building a token-by-token similarity matrix and finding clusters in it has its basis in distance measures like Euclidean or cosine distance. The other two corpus linguistic approaches predominantly use quantitative information and mathematical operations with regard to one of those components only.

To begin with, corpus linguistics as it developed in Britain in the 1990s restricts the quantitative perspective to the first component, focusing on the statistical identification of collocating features, and then interpreting the patterns (the second component) chiefly in a manual way. Firth (1957) remarked that part of the 'meaning' of *cows* can be indicated by such collocations as *They are milking the cows, Cows give milk*. This observation is taken as a methodological starting point: the words co-occurring with another one help to identify the properties of the word under scrutiny. An example (taken from Stubbs 2002: 15) may illustrate the basic idea. A classic example of homonymy in English is the item *bank*, which is either a financial institution, or an area of sloping ground, specifically the raised ground on the side of the river or underneath a shallow layer of water. The sets of words that these two exemplars of *bank* normally occur with hardly overlap. Looking at compounds on the one hand, and on the other hand at co-occurring items within a few words to the left or right of *bank*, Stubbs comes up with the following lists:

(2.1) bank account, bank balance, bank robbery, piggybank cashier,
deposit, financial, money, overdraft, pay, steal

(2.2) sand bank, canal bank, riverbank, the South Bank, the Left Bank,
Dogger bank, Rockall Bank, Icelandic Banks cave, cod, fish, float,
headland, sailing, sea, water

The entities in the environment of the two homonyms appear to differentiate efficiently and effectively between the two meanings, and in that sense, a systematic analysis of the co-occurring items would appear to be an excellent methodological ground for lexical-semantic analysis. In theoretical terms, the essential concept here is that of *collocation*, defined as 'a lexical relation between two or more words which have a tendency to co-occur within a few words of each other in running text' (Stubbs 2002: 24). Collocations in this broad sense may take different shapes, depending on the level at which the co-occurrence of words (and sets of words) is defined. Sinclair (1991, 1996) distinguishes four types: collocation, colligation, semantic preference, and semantic prosody. Because the classification will turn out to be relevant for Chapter 5, we can introduce with some more detail.

Collocation in the most immediate sense is the co-occurrence of words or wordforms in a line of text. Terminologically, the target word is often called the *node*, and the co-occurring word the *collocate*. A common way of examining collocations is to produce a concordance of a text or a set of texts, that is, an alphabetical list of the words in those texts, presented in their immediate context. The node of a collocation analysis may be a word form or a word, if lemmatization can be applied, that is, if all the inflectional forms of a word are treated as instances of a single lexical unit. Also, nodes may themselves be multiword expressions or phrases.

Following Firth, Sinclair defines *colligation* as 'the co-occurrence of grammatical choices' (1996: 85), that is, the syntactic pattern with which a word appears. Co-occurrences, in other words, are now defined between the node and a syntactic class.

Semantic preference involves the relation between the node and a set of semantically related words. Unlike collocation, semantic preference involves a class of lexical items, not a limited set of one or a few. And unlike colligation, semantic preference involves a class of items that is defined by their semantic and not their syntactic properties. Semantic and syntactic restrictions can go hand in hand, though. In Sinclair (1996) the phrase *naked eye* appears to co-occur, on the third position preceding the node, with expressions that come predominantly from two classes: the top collocates in that position are *see/seen* and *visible/invisible*, but more verbs include *detect*, *spot*, *appear*, *perceive*, *view*, *recognize*, *read*, *study*, *judge*, *tell*, and more adjectives include *apparent*, *evident*, *obvious*, *undetectable*. Combining the level of colligation and that of semantic preference, we may then say that the third position to the left of *naked eye* is dominantly filled by a verb or adjective referring to (in)visibility.

Semantic prosody looks at co-occurrences not from a purely lexical perspective (as in collocation), nor from a syntactic perspective (as in the case of colligation, looking at grammatical categories), nor from a semantic perspective (as in semantic preference, looking at semantically defined lexical sets), but from a connotational perspective, that is, from the point of view of the emotive or evaluative attitude expressed by the surrounding words. It refers to the fact that words may tend to line up with either positively or negatively evaluated words.

Collocational analysis for lexical description was developed not just in a descriptive direction (as in Hoey 1991, 2005; Partington 1998; Stubbs 2002) and in a statistical direction (as by Church and Hanks 1989), but also in practical terms, as a software tool for lexicographers (see for instance Kilgarriff and Rundell 2002). But round the turn of the millennium, a different type of combination of quantitative analysis and corpus data came to the fore. As presented above, semantic vector modelling consists of two components: identifying relevant context features and finding patterns among the tokens based on their similarity with regard to those features. On the one hand, collocational analysis approaches the first component from a quantitative point of view. But the on the other hand, the statistics may also be concentrated in the second component. Usage cases are then manually (or semi-automatically) annotated for a wide array of potentially relevant characteristics, and subsequently the patterns are explored by statistical means. This method can be applied to a broad range of phenomena, and it actually emerged in a grammatical rather than lexical context, when the dissertations of Stefan Grondelaers (see Grondelaers, Speelman, and Geeraerts 2002) and Stefan Gries (see Gries 2003) independently pioneered the use of regression techniques to model grammatical phenomena. In the lexical domain, the multifactorial approach (i.e. a method that applies a statistical analysis for finding patterns and similarities to a dataset that is annotated for a variety of features fitting the purpose) goes by two names: the Behavioural Profile Approach (as in Gries 2006; Divjak 2006, 2010; Gries 2010; Jansegers and Gries 2017) and the Multifactorial Usage-Feature Analysis (Glynn 2008, 2009; Krawczak 2014; Krawczak and Glynn 2015; Glynn 2016). More important than this terminological variation is the difference in the entities that form the basis for the annotation. Three classes can be distinguished.

First, the annotated entities may be instances of near-synonyms or alleged synonyms, as in Divjak (2006, 2010). The quantitative analysis of the contexts of use of the competing expressions may then reveal to what extent they are indeed co-extensive: are they used in the same contexts, and are they used indiscriminately in those contexts? Or conversely, what does the difference in their dominant contexts of usage suggest about the difference in meaning between them? For instance, Speelman and Geeraerts (2009a; see also Levshina and Heylen 2014) examine whether the Dutch auxiliaries *doen* and *laten* express different types of causation. Geeraerts, Gevaert, and Speelman (2012) investigate the hypothesis that the historical substitution of *wrath* by *anger* is driven by an increasing individualization of society.

Second, one can focus on a pre-established list of senses for a given lexical item and annotate the occurrences of those senses for whatever contextual properties seem important. Thus, both Gries (2006) and Glynn (2014) examine the contextual correlates of the polysemy of *to run* by initially identifying lexical senses in the sample, annotating them, and statistically looking for patterns in the annotated examples.

As they presuppose an initial sense classification of the word under consideration, studies like Gries (2006) and Glynn (2014) do not yet offer a fully bottom-up distributional classification of the usages of a word. The third type, then, as illustrated in detail by Glynn (2016), does precisely that. Without making any a priori assumptions about the meanings of *to annoy*, usage cases are annotated for the cause, the patient and the agent of the annoying event or situation. Next, multivariate statistics (multiple correspondence analysis, hierarchical clustering, k-medoid cluster analysis) are employed to identify patterns among the annotated examples, that is, they are grouped together on the basis of their mutual similarity, with the similarity being measured as the distance between the vectors—in this case, strings of annotated features—representing the usage cases. This third approach is clearly closest to the perspective we have sketched for token-based distributional semantics: we are also aiming for a bottom-up identification of similar tokens based on their vector representation. Only, the vectors will themselves be produced through a quantitative analysis of the corpus data, rather than through manual annotation.

We can bring the various types of corpus linguistics together in the overview chart of Table 2.5. The table distinguishes between the two components—assigning context features and grouping tokens based on those features—and cross-classifies this distinction with the distinction between a chiefly manual and a chiefly statistical methodology. While we started from these distinctions to situate the various contemporary forms of corpus linguistics with regard to each other, the classification appropriately also provides a place for the older, fully manual kind of corpus analysis that we identified with the tradition of lexicography.

This overview does not, however, exhaust the variety of distributional methods. Within the set of vector space models, a distinction exists between two essentially different architectures: so-called count models and prediction models (see Baroni, Dinu, and Kruszewski 2014; for a general overview of distributional modelling of word meaning, see Lenci 2018). The framework that we have described above and will be developing in the following chapters illustrates the count-based approach. Prediction-based models, as the name suggests, make use of statistical analyses that predict the occurrence of a word on the basis of a neural network, that is, a computing system that intends to mimic the organization of the human brain. In its simplest form, such a neural network consists of a single hidden layer of

Table 2.5 The diversity of distributional semantics

		IDENTIFYING CONTEXT FEATURES	
		manual	***statistical***
GROUPING TOKENS	*manual*	lexicography	collocation analysis
	statistical	behavioural profile approach	vector space approach

nodes (artificial neurons) that predicts whether the outcome of interest will occur in a given context. So, to get back to our previous example, the question could be whether, given a record of a large number of dinner parties in the Parisian artistic scene, we can predict with some accuracy whether Picasso will participate in a specific one. In such a model, the vector for Picasso (the representation for Picasso as a dinner party guest) derives from the hidden layer of the model: a range of nodes in a neural network with for each node, its weight in predicting the outcome. Applied to language, the basic model predicts the presence of a target word after the words preceding it, and the corresponding vector is known as a 'word embedding'.

Building up such a vector is referred to as 'training' the model: with each new input of actual data, the weights of the nodes are adjusted to increase the predictive accuracy. Also, because the number of nodes is much lower than the original set of features that constitute the input, building up the embedding is a process transforming a sparse vector to a dense one with just a limited number of nodes. Indeed, out of thousands of possible participants (the entire beehive of bohemian Paris), a single dinner party will only feature a few people, so a vector characterizing any of those parties will be sparse. The neural network that is trained on those sparse descriptions and that predicts Picasso's presence, by contrast, is a dense vector with a limited number of nodes. In other words, transforming scarcely populated input vectors into richer and more compact ones is a feature of prediction-based approaches just like it is in count-based approaches. Only, the method is different. In a workflow like ours, it is the result of successive steps (feature selection and weighting, matrix building, dimension reduction), while in neural networks, it is achieved in a single step by building the network.

A further similarity follows from the observation that both in a count-based and prediction-based method a distinction can be made between type-level and token-level approaches. The original word embedding approach, first described by Mikolov (Mikolov, Sutskever, Chen, Corrado, and Dean 2013) and best known as the *word2vec* approach, yields a single, static embedding for each word. In more recent prediction-based models such as the highly popular BERT (Devlin, Chang, Lee, and Toutanova 2019), embeddings are contextualized. By using multiple hidden layers in the neural network, the classical embeddings for the words constituting a sentence can be used as input for a sentence-specific vector for a target word appearing in that sentence. The likeness with a count-based approach will be obvious: type-level vectors (as the case may be, static word embeddings) are used as building blocks for arriving at a token-level vector (as the case may be, a contextualized word embedding).

Despite these similarities, the crucial distinction between count-based and prediction-based approaches resides in the fact that the latter are more difficult to interpret than the former. The nodes in a hidden layer of a neural network do not have an immediate linguistic interpretation; it is difficult to determine what they correspond to from a linguistic point of view. That is the reason why in the present study we opt for count models. If we are interested in seeing how far vector space models can take us descriptively in lexical variation research, it is an advantage to employ a distributional method that is more transparent and more manipulable, that is, where we can have more control of the settings and can more closely look under the hood to see what is going on—even though, admittedly, count models are not fully transparent either, and prediction models are tremendously successful in natural language processing. From an application-oriented point of view, prediction-based models (specifically those of the latest BERT generation) are undoubtedly superior to count-based approaches. From the descriptive point of view that is ours, however, the bigger transparency of count-based modelling provides a better starting point. Notwithstanding an increasing interest in looking under the hood of prediction-based modelling, a natural language processing approach is often not heavily interested in the vectors as such and how to interpret them: the vector representations primarily have to serve machine translation, sentiment analysis, document classification, and other natural language processing tasks, and the success of different distributional models can be measured by how well the vectors support these tasks. By contrast, for reasons that will be made clearer in Sections 2.3 and 2.4, our perspective is principally exploratory: so much methodological uncertainty still surrounds the notion of meaning itself that a cautious and circumspect scrutiny of the way distributional modelling taps into it is called for.

2.3 Sense determination and semantic indeterminacy

Now that we have had a closer look at distributional methods, we can broaden the perspective and compare the distributional approach to other common methodological techniques in lexical semantics. We will do so by focusing on a central question for semantic research: how to establish the different senses of a word? An anecdotal illustration may introduce the non-trivial character of the question. In the first week of March 2013, immediately after the Italian parliamentary elections, the cover of the magazine *The Economist* bore a composite picture featuring Silvio Berlusconi and Beppe Grillo, under the caption 'Send in the clowns'. Semantically speaking, lots of things are going on here. Both men are clowns in a derived sense only, if we take the literal meaning of *clown* to be 'fool, jester, as in a circus or a pantomime; performer who dresses in brightly coloured unusual clothes and whose performance is meant to make the audience laugh'. Grillo entered parliament as the leader of the anti-establishment Five Star Movement, but as he originally is an actor and a comedian, the relevant sense of *clown* could be paraphrased as 'comic entertainer', that is, as a slightly looser, more general reading of the central meaning. Berlusconi on the other hand is a clown in a figurative sense: his populist political antics characterize him as a man acting in a silly and foolish way—a metaphorical buffoon, in short. But while we readily recognize that *clown* applies in different ways to Grillo and Berlusconi, this creates a problem when we try to define the precise meaning of the word in *Send in the clowns*. The plural suggests that there is a single sense of *clown* that applies to both men, but then what would that meaning be, given the differences that we just discussed? Should we define a meaning at all that covers both 'comic entertainer' and 'metaphorical buffoon', or should we rather say that the simultaneous presence of two distinct senses underlies the punning effect of *Send in the clowns*?

From the point of view of semantic theory, this simple example illustrates a crucial methodological question: we perceive the differences between the three interpretations of *clown*—'jester in a circus or pantomime', 'comic entertainer in general', 'someone acting so silly as to make a fool of himself'—but what arguments exactly do we have to say that these are different meanings of the word, and how do we determine what the meaning is in the context of a specific utterance? How, in other words, do we establish the polysemy of a word, or any other linguistic expression? The question is of obvious concern for the descriptive framework we have sketched in Chapter 1; it determines how we look at the conceptual level that mediates between the formal and the referential layer. In the wake of prototype theory and the emergence of usage-based linguistics, a major change has taken place in the way linguists think about the problem of polysemy. Roughly speaking, semantic theory has moved from a static conception of polysemy, in which senses are well-defined linguistic units (just like, say, phonemes or morphemes are discrete elements within the structure of a language) to a much more flexible and dynamic view of meaning. In this section and the next, we will explore the methodological aspects of this shift, and its consequences for a distributional approach to meaning. The argumentation (which reproduces large parts of Geeraerts 2016c) unfolds in two steps. In the present section, we will first gradually zoom in on the central questions of polysemy research, and then present an overview of the arguments that have led semanticists to abandon a static conception of polysemy. In the following section, we will spell out the effect on our distributional method.

To get a grip on the issues involved in the study of polysemy, we first need to introduce two sets of distinctions: that between polysemy, vagueness, and ambiguity on the one hand, and that between utterance meaning and systemic meaning on the other. The first of these distinctions involves the question whether a particular semantic specification is part of the semantic structure of the item, or is the result of a contextual, pragmatic process. For instance, *neighbour* is not considered to be polysemous between the readings 'male person living next door' and 'female person living next door', in the sense that the utterance *our neighbour is leaving for a vacation* will not be recognized as requiring disambiguation in the way that *she is a plain girl* does. In the latter case, you may be inclined to ask whether *plain* is meant in the sense of 'ordinary looking, of no particular beauty' or 'simple, unsophisticated'. In the former case, you may perhaps wonder whether the neighbour in question is a man or a woman, but you would not be inclined to ask something like: 'In which sense do you mean *neighbour*—male neighbour or female neighbour?' The semantic information that is associated with the item *neighbour* in the lexicon does not, in other words, contain a specification regarding gender; *neighbour* is vague, or 'unspecified', as to the dimension of gender, and the gender differences between neighbours are differences in the real world, not semantic differences in the language. This notion of *conceptual underspecification* must be kept distinct from three other forms of semantic indeterminacy.

Since at least some of these alternative forms of indeterminacy may themselves be referred to as *vagueness*, we need to be aware that the discussion of vagueness (as contrasting with polysemy) is beset by terminological pitfalls. First, conceptual underspecification as just illustrated differs from the *referential indeterminacy* that may characterize the individual members of a category. Think of a word like *knee*: it is impossible to indicate precisely where the knee ends and the rest of the leg begins, and so, each individual member of the category *knee* is not discretely demarcated. Second, referential indeterminacy may relate to entire concepts rather than just their individual members. Such *categorical indeterminacy* involves the fuzzy boundaries of conceptual categories, as illustrated by any colour term. In the same way in which we can think of the category *knee* as the set of all real and possible knees, we can think of a colour like *red* as the set of all individual hues that could be called *red*. But then, it will be very difficult to draw a line within the spectrum between those hues that are a member of the category *red* and those that are not: where exactly does the boundary between *red* and *orange* or *red* and *purple* lie? (This is a phenomenon—unclarity at the border of a category—that we have already come across when we discussed prototypicality.) Third, the conceptual underspecification of individual meanings differs from the *interpretative indeterminacy* that occurs when a given utterance cannot be contextually disambiguated. For instance, when the intended interpretation underlying *she is a plain girl* cannot be determined based on the available information, the interpretation is indeterminate, and the utterance is said to exhibit ambiguity. Ambiguity, in other words, may result from contextually unresolved polysemy.

A second distinction that is necessary to get a clear view on the problem of polysemy is that between meaning at the level of the linguistic utterance, and meaning at the level of the linguistic system—between the meaning, in other words, that is supposedly a stable part of the system of the language, and the meaning that is realized in the context of a specific speech situation. In a simple model, the distinction between vagueness and polysemy runs parallel to that between utterance meaning and systemic meaning. As the case may be, in the actual situation in which the sentence is uttered, *our neighbour is leaving for a vacation* might call up the idea of a man or a woman, when all involved know who is being talked about. But although either the concept 'male person living next door' or 'female person living next door' would then indeed be activated in the context of the utterance, one would still not say that they add to the polysemy of *neighbour*. One could call 'male person living next door' or 'female person living next door' the utterance meaning of *neighbour*, but this contextual specification is different from the systemic meaning, which would just be 'person who lives next door'. In this view, the systemic meaning belongs to the level of semantics, the utterance meaning to the level of pragmatics.

Does this imply that we can forget about utterance meaning? In a usage-based approach to language, that would obviously not be the case, but we know that embracing the usage level is a relatively recent turn in linguistic theory. In the Saussurean, structuralist framework, the core of linguistic enquiry is the system of the language, and in the Chomskyan, generative framework, it is the mental representation of language, the way language is represented in the mind. But, while both traditions tend to theoretically privilege systemic meaning, that does not mean that utterance meaning can be methodologically ignored. From a methodological point of view, utterance meaning could be ignored if we have direct access to the mental lexicon, that is, if we can introspectively establish the meaning of linguistic expressions at the level of the linguistic system. Wierzbicka (1985) for instance argues that to state the meaning of a word, one must introspectively study the structure of the concept which underlies and explains how the word can be used, and to understand the structure of the concept means to discover and describe fully and accurately the internal logic of the concept, through methodical introspection and thinking. To the extent that they understand language, language users have direct, unmediated access to the meaning of linguistic expressions; the semanticist's primary move, then, consists of attentively tapping into that immediate knowledge. (Note that 'introspection' as meant here does not simply equal 'interpretation'. On one hand, there is the notion that any form of semantics involves understanding, that is, mentally accessing the interpretation of an expression. On the other, there is the idea that this process of interpretation can take place directly at the systemic level. When we talk about 'introspection' here, it is the latter position that is at stake.)

Such an idealist methodological position needs to be treated with caution, though. First, as a rather down to earth rebuttal, we may consider the way in which such an introspective exercise would work. In practice, one would likely imagine different contexts in which the target expression is used and determine the definition of the word on that basis: if you want to know what *clown* means, you imagine circumstances in which you would use the word and try to find a common denominator for those usages. But that, of course, is basically a roundabout way of grounding the analysis in contextualized language use: rather than a direct access to systemic meaning or mental representations, introspection then merely provides an indirect access to utterance meaning. Second, we could ask how an introspective method can be validated, that is, how can we establish that it is a valid method, without simply assuming that it is? One possibility could be to compare the results of an introspective strategy with actual usage data: is the meaning that is intuitively identified the same that is activated in actual usage? But then again we would obviously be back to square one: we would again be using utterance meaning as a point of comparison, and we would need to establish what those utterance meanings are. It follows that including utterance meaning in the investigation is a methodological prerequisite: even if you are primarily interested in systemic meaning, *utterance meaning is the primary observational basis of semantics.*

But if we take that observation as our point of departure, there are two major perspectives we can take to solve our polysemy issue. On the one hand, we can take the idea that we have about the link between systemic meaning and utterance meaning to filter out those aspects of utterance meaning that do not correspond with systemic meaning. For instance, if we assume that only conventionalized meanings can be systemic, we can take conventionalization as the criterion to go from the utterance level to the systemic level. On the other hand, we can focus more directly on utterance meanings, and try to group them on the basis of specific polysemy tests. For instance, if we assume that readings that can be captured under the umbrella of a single definition can never be separate meanings, then we can apply a definitional criterion to distinguish vagueness from polysemy. In the following pages, we will explore both major perspectives—but in both cases, the results will not bring us peace of mind.

The link between systemic meaning and utterance meaning can be specified in two ways: as a distinction between conventional meaning and occasional meaning, and as a distinction between stored meaning and derived meaning. If we look more closely into these two distinctions, it will become clear that they blur the equation of 'polysemy versus vagueness' and 'systemic meaning versus utterance meaning'.

The distinction between conventional and occasional meaning was first made explicit by Hermann Paul at the end of the 19th century: the conventional meaning (*usuelle Bedeutung*) is the established meaning as shared by the members of a language community; the occasional meaning (*okkasionelle Bedeutung*) involves the modulations that the usual meaning can undergo in actual speech (1920: 75). If the 'usuelle Bedeutung' is like the semantic description that would be recorded in a dictionary (fairly general, and in principle known to all the speakers of a language), then the 'okkasionelle Bedeutung' is the concretization of that meaning in the context of a specific utterance. To mention just one of the examples listed by Paul, the word *corn* used to be a cover term for all kinds of grain, but was differently specialized to 'wheat' in England, to 'oats' in Scotland, and to 'maize' in the United States, depending on the dominant variety of grain grown in each of these countries: the context of use triggers the specialized meaning. But crucially, there exists a dialectic relationship between language system and language use: occasional meanings that are used very often may themselves become usual, that is, they may acquire an independent status. So, on the one hand, usual meanings are the basis for deriving occasional ones, but on the other, the contextualized meanings may become conventional and decontextualized. The clearest criterion for a shift from the occasional to the usual level is the possibility of interpreting the new meaning independently. If *corn* evokes 'wheat' without specific clues in the linguistic or the extralinguistic environment, then we can be sure that the sense 'wheat' has become conventionalized.

This dialectic relationship precludes a simple equation of 'conventional meaning versus occasional meaning' with 'polysemy versus vagueness'. To the extent that occasional meanings are just easily traceable contextual specifications, they fall under the heading of 'vagueness', and their relevance for linguistics is minimal. However, to the extent that occasional meanings might be on their way of becoming conventionalized, 'conventional' becomes a graded notion: meanings may be more or less conventional (and hence, more or less interesting from the systemic point of view). More generally, if we want to get a good idea of language change, occasional utterance meanings cannot be discarded as in principle less interesting: all changes of conventions begin as occasional changes on the utterance level. (For a contemporary formulation of the interplay between system and usage in polysemy research, see Hanks 2013.)

The distinction between conventional meaning and occasional meaning takes a predominantly social perspective on language: it looks at what is common in a community of speakers, and how those common patterns change over time. By contrast, we may look at language as an individual phenomenon as represented in the head of the language user. Within such a psychological perspective (a perspective that has been dominant in contemporary linguistics ever since Chomsky's definition of language as a cognitive phenomenon), economy of representation is often mentioned as an important criterion: a mental representation of the language that is parsimonious is supposed to be superior, and more specifically, linguistic phenomena that can be derived by some kind of generative, rule-based mechanism need not be stored separately in the mental representation. Applied to semantics, this implies that meanings that can be contextually derived need not be mentally stored as such. For instance, *chocolate* has two meanings: 'food made from cocoa beans, with a brown colour and a hard but brittle substance' and 'hot drink made from milk and powder containing chocolate (as defined before)'. It could then be argued that in the context *a mug of chocolate*, the presence of *mug* automatically triggers the second interpretation. The pattern *a mug of*—assumes that a mass noun will fill the slot, and specifically, a mass noun referring to a liquid. The meaning of *chocolate* is then, so to speak, automatically liquified. In terms of representation, if we know what *chocolate* means in its basic reading and what *a mug of*—demands of its slot filler, it would seem that it is not necessary to separately list the second meaning of *chocolate* in the mental lexicon: instead of selecting the meaning from a list of stored readings, the meaning is computed by applying the expectations that are activated by *mug* to the stored basic meaning of *chocolate*.

This kind of model, aiming at a parsimonious distinction between stored meanings and contextually derived meanings, appears in various theoretical quarters, from Ruhl's largely descriptive approach (1989) to Evans' version of cognitive semantics (2009) to Pustejovsky's formalized Generative Lexicon model (1995). Two problems are relevant in the present context. First, how important is it really to keep listed meanings and derived meanings separate? A parsimonious approach makes a distinction between semantic information that is stored in the (mental) lexicon, and readings that are derived pragmatically, in context. But if we take into account language change, such a strict distinction between what is stored and what is derived cannot be preserved. Pragmatic, context-dependent meanings must be able to permeate to the level of semantics, in the way in which Paul's *okkasionelle Bedeutung* can over time be promoted to the status of *usuelle Bedeutung*. This is not just a social process of conventionalization, but it is also an individual psychological process: one of the cognitive phenomena to be accounted for is the fact that some uses of a word may become psychologically more salient than others. Such a process requires that a reading that is at one point pragmatically derived leaves a trace in the mental lexicon of the language user: language users remember hearing/reading or saying/writing it, and the more they use it, the more cognitively entrenched it becomes. Just like in the case of conventional and occasional meanings, a strict separation between stored and derived readings (what Langacker 1991 refers to as the 'rule/list fallacy') is difficult to maintain.

Second, even if we were able to strictly keep up the distinction, it would not help us with the problem of sense individuation. The distinction between stored meanings and derived meanings does not coincide with that between conventional meanings and occasional meanings, nor does it coincide with that between polysemy and vagueness. Even if the 'hot drink' meaning of *chocolate* can be derived contextually, it is still considered a different reading (and a conventional one at that). In fact, it is precisely *because* it is considered a different reading that it makes sense to explore how it can be most economically represented, by listing it or by computing it. As a consequence (and this is a point that cannot be sufficiently emphasized), assuming a dynamic model of meaning distinguishing between listed and computed meanings does not as such solve the question how to distinguish vagueness from polysemy.

Would switching to the other major perspective clear up the situation? There again, if we look more closely at existing polysemy tests, a critical scrutiny of those tests tends to blur the distinction between vagueness and polysemy. An examination of different basic criteria for distinguishing between polysemy and vagueness reveals, first, that those criteria are in mutual conflict: they need not lead to the same conclusion in the same circumstances. Second, each of them taken separately need not lead to a stable distinction between polysemy and vagueness: what is a distinct meaning according to one of the tests in one context, may be reduced to a case of vagueness according to the same test in another context. (Geeraerts 1993 offers a more extended treatment of this argumentation.) In general, three types of polysemy criterion can be distinguished.

First, from the *truth-theoretical* point of view taken by Quine (1960: 129), a lexical item is polysemous if it can simultaneously be clearly true and clearly false of the same referent. Considering the readings 'harbour' and 'fortified sweet wine from Portugal' of *port*, the polysemy of that item is established by sentences such as *Sandeman is a port* (in a bottle), *but not a port* (with ships). Up to a point, we could say that this criterion basically captures a semantic intuition: are two interpretations of a given expression intuitively sufficiently dissimilar so that one may be said to apply and the other not?

Second, *linguistic* tests involve syntactic rather than semantic intuitions. Specifically, they are based on acceptability judgements about sentences that contain two related occurrences of the item under consideration (one of which may be implicit). If the grammatical relationship between both occurrences requires their semantic identity, the resulting sentence may be an indication for the polysemy of the item. For instance, the identity test described by Zwicky and Sadock (1975) involves 'identity-of-sense anaphora'. Thus, *at midnight the ship passed the port, and so did the bartender* is awkward if the two lexical meanings of *port* are at stake. Disregarding puns, it can only mean that the ship and the bartender alike passed the harbour, or conversely that both moved a particular kind of wine from one place to another. A mixed reading in which the first occurrence of *port* refers to the harbour, and the second to wine, is normally excluded. By contrast, the fact that the notions 'vintage sweet wine from Portugal' and 'blended sweet wine from Portugal' can be combined in *Vintage Noval is a port, and so is blended Sandeman* indicates that *port* is vague rather than polysemous with regard to the distinction between blended and vintage wines.

Third, the *definitional* criterion (as already informally stated by Aristotle in the *Posterior Analytics* II.xiii) specifies that an item has more than one lexical meaning if there is no minimally specific definition covering the extension of the item as a whole, and that it has no more lexical meanings than there are maximally general definitions necessary to describe its extension. Definitions of lexical items should be maximally general in the sense that they should cover as large a subset of the extension of an item as possible. Thus, separate definitions for 'blended sweet fortified wine from Portugal' and 'vintage sweet fortified wine from Portugal' could not be considered definitions of lexical meanings, because they can be brought together under the definition 'sweet fortified wine from Portugal'. On the other hand, definitions should be minimally specific in the sense that they should be sufficient to distinguish the item from other non-synonymous items. A maximally general definition covering both *port* 'harbour' and *port* 'kind of wine' under the definition 'thing, entity' is excluded because it does not capture the specificity of *port* as distinct from other words.

The existence of various polysemy tests is non-trivial for two fundamental, interlocking reasons. First, the three types of criteria may be in mutual conflict, in the sense that they need not lead to the same conclusion in the same circumstances. In the case of autohyponymous words, for instance, the definitional approach does not reveal an ambiguity, whereas the Quinean criterion does. A word is autohyponymous if one of its senses is a proper subset of one of its other senses. Thus, *dog* is autohyponymous between the readings 'Canis familiaris', contrasting with *cat* or *wolf*, and 'male Canis familiaris', contrasting with *bitch*. A definition of *dog* as 'male Canis familiaris', however, does not conform to the definitional criterion of maximal coverage, because it defines a proper subset of the 'Canis familiaris' reading. On the other hand, the sentence *Lady is a dog, but not a dog*, which exemplifies the logical criterion, cannot be ruled out as ungrammatical.

Second, each of the criteria taken separately need not lead to a stable distinction between polysemy and vagueness, in the sense that what is a distinct meaning according to one of the tests in one context, may be reduced to a case of vagueness according to the same test in another context. Without trying to be exhaustive, let us cite a few examples involving the linguistic criterion. Contextual influences on the linguistic test have been (implicitly or explicitly) noted by several authors. In fact, the recognition occurs relatively early in the literature on the subject. When Lakoff (1970) introduced the *and so*-construction as a criterion for polysemy, he argued that *hit* is ambiguous between an intentional and an unintentional reading, because *John hit the wall and so did Fred* would constitute an anomalous utterance in situations in which John hit the wall intentionally but Fred only did so by accident, or the other way round. Catlin and Catlin (1972), however, noted that the sentence could easily be uttered in a context involving imitation. A situation in which John hits his head against the wall after stumbling over his vacuum cleaner and is then comically imitated by Fred, might very well be described by the sentence in question. Nunberg (1979) further drew the attention to sentences such as *The newspaper has decided to change its size*, which features intuitively distinct senses of newspaper ('management, board of directors' and 'material publication'). Similar cases can be found involving co-ordination rather than anaphora. For instance, Norrick (1981: 115) contrasted the decidedly odd sentence *Judy's dissertation is thought provoking and yellowed with age* with the perfectly natural construction *Judy's dissertation is still thought provoking though yellowed with age*. If the co-ordination generally requires that *dissertation* be used in the same sense with regard to both elements of the co-ordinated predicate, the sentences show that the distinction between the dissertation as a material product and its contents may or may not play a role. Cruse (1982) noted that none of the following series of sentences containing co-ordination produces feelings of oddity: *John likes blondes and racehorses—John likes racehorses and fast cars—John likes* *cars and elegant clothes—John likes elegant clothes and expensive aftershave—John likes expensive aftershave and vintage port—John likes vintage port and marshmallows*. Coordinating the first item in the series with the last, however, does produce an awkward sentence. So, while the awkwardness of *John likes blondes and marshmallows* would normally be taken as evidence for the polysemy of *like*, the pairings mentioned above suggest that there is a continuum of meaning rather than a dichotomy. Cruse concludes that readings which are close together can be co-ordinated without oddity, but if they are sufficiently far apart, they are incompatible. If this picture is correct, it does not make sense to ask how many senses of *like* there are: 'There is just a seamless fabric of meaning-potential' (1982: 79).

From these and similar publications (Taylor 1992; Geeraerts 1993; Tuggy 1993; Kilgarriff 1997; Allwood 2003) it appeared, in other words, that the contextual flexibility of meaning may take radical forms: it does not just involve a context-driven choice between existing meanings, or the on-the-spot creation of new ones, but it blurs and dynamizes the very distinction between polysemy and vagueness. To come back to our initial example, Grillo is a clown in one sense but not in the other, and the reverse holds for Berlusconi, but in the right context, both seemingly incompatible senses can be combined.

2.4 Semantics without meaning

It may be noted that the critical deconstruction of the traditional distinctions—polysemy and vagueness, systemic meaning and utterance meaning—is roughly situated in the final decade of the previous century. It was then part of the emergence of post-structuralist conceptions of meaning: prototype theory, and more broadly, cognitive semantics. In the following decades this theoretical shift was enriched by a methodological shift towards various empirical approaches. Indeed, when you reach the conclusion that a complete model of linguistic meaning cannot be achieved without systematic attention to differences in contextualized meanings as they appear in actual usage, you should also face the fact that utterance meaning is clearly no more immediately transparent than stored meanings. Recall our opening example: it will not be easy to come up spontaneously with a definition of the meaning realized in *Send in the clowns*. Or consider the example *We are out of fruit*. We know that various features are associated with *fruit*: fruit is generally sweet, juicy, it is commonly used as dessert, and technically it is the seed-bearing part of a plant. But it is unlikely that all those features are activated in the mind when someone utters the statement *We are out of fruit*. In the context of *A lemon is a fruit*, only a subset of features is activated and conversely, others are backgrounded: a lemon is not sweet, and it is not used as dessert. But how would that mechanism of foregrounding and backgrounding work in *We are out of fruit*? When you use that phrase when you are drawing up your grocery list, the idea of a certain type of food will probably be prominent in your mind, but apart from that, is the idea of fruit that you have in your head at that point so clear that you can ascertain whether the fact that fruits are dominantly sweet was on your mind or not? Or perhaps you weren't thinking of fruit in terms of an abstract concept with definitional features, but you were thinking of it in terms of a collection of things like apples, strawberries, and bananas? But then again, is what passed through your head so clear that you would be able to tell without a doubt whether, for instance, oranges were part of the set you were thinking of?

The difficulty of such direct, introspective analyses strengthens the need for indirect measures of meaning: instead of studying meaning directly, we can study the behavioural correlates of meaningful language use and base our analysis on those. Three major perspectives for doing this have come to the foreground in the past three decades (see also Stefanowitsch 2010): an experimental, a referential, and a distributional corpus-based approach. The latter is the focus of the present book, and the referential one should also be familiar: it is the approach illustrated by the Geeraerts, Grondelaers, and Bakema (1994) study of Dutch clothing terms that we referred to earlier. Other examples are Anishchanka, Speelman, and Geeraerts (2014, 2015a, 2015b), in which digitized colour information from webpages is used to explore the range and mutual relationship of colour terms. Among other things, these studies reveal that in actual usage, the term *navy* is not a strict hyponym of *blue*: the referential range of *navy*, defined as the set of hues to which it applies, overlaps with that of *blue*, but is not a strict subset of it. The third main approach, experimental research, covers methods like reaction time experiments, naming tasks, association tasks, similarity judgements, self-paced reading, lexical decision tasks, sentence completion, eye tracking, neuroimaging, and so on. It constitutes the standard methodological paradigm in psycholinguistic research. Through the work of Rosch (1975) on category structure, this type of work had an indirect but considerable influence on the adoption of prototype models in linguistics, and more generally on the methods used in cognitive linguistics. An update on prototype-oriented experimental categorization research can be found in Hampton (2016), but this is just a small part of the psycho-experimental interest in the mental lexicon; a full overview is beyond the scope of the present text. Of particular interest though are types of experimentation with an outcome similar to the distributional methods discussed here. This is the case for word association data: participants in the experiment are presented with a word that serves as a cue and must respond with the first word that comes to their mind. For each cue, the full set of responses, with an associative weight based on the frequency with which they are given, is not unlike a type vector derived from distributional corpus data. Accordingly, kindred research questions can be answered by means of kindred techniques, like using the distance between vectors to determine the semantic similarity between words. For a description of what is currently probably the major word association database, see De Deyne, Navarro, Perfors, Brysbaert, and Storms (2019).

Interestingly, the three methodological perspectives are crucially similar to the three traditional polysemy tests that we distinguished earlier:

- The referential approach resembles the definitional test, to the extent that it too primarily looks at the extralinguistic situation that is referred to by the words.
- Like the corpus-based distributional method, the zeugma test looks at syntagmatic patterns in which a word occurs (but with a much narrower scope than the contemporary corpus approach, to be sure).
- Experimental psycholinguistic methods, like the logical test, explore the subjective spontaneous understanding of the language user.

Given the similarity among the newer empirical methods and the older polysemy tests, we should beware of difficulties that are like the ones we discussed in the previous section. In accordance with what we saw there, the issues to consider are of two kinds. On the one hand, how contextually stable is a distributional analysis of polysemy? In Chapter 5, we will argue that the instability that characterized the traditional polysemy tests also applies to a distributional approach. The distributional models that seem to perform well for one set of items may yield less satisfactory results for another, and more generally, a straightforward identification of token clusters with senses as one would traditionally think of them is largely misguided.

On the other hand, the relationship between the major methodological perspectives needs to be examined. Are there any divergences between the methodological perspectives, and if so, is the information they provide complementary or conflicting? This second major question will not be of central concern for our work, but even without going very deeply into the matter, we can see that the various approaches seem to capture different, non-overlapping phenomena. The referential method, for instance, will work best for material objects, events, processes, but a lot of the information that will be revealed by taking such a referential perspective may be absent from the corpus. The information that is encoded in texts is probably not all the information that language users rely on, and specifically, the kind of features that are prominent in a referential approach (like the shape of objects) may not be explicitly expressed in textual data. How easy, for instance, would it be to retrieve information from the corpus about the average length of leggings, or the dominant shades of a colour term like *navy*? In a similar way, at least some of the psycholinguistic experimental methods can gather information about on-line processing and mental representation that is inaccessible to the off-line perspective of a referential or a distributional method. But those types of information need not converge. Schmid (2010), for instance, argues that corpus frequencies need not directly reflect the psychological entrenchment of linguistic expressions. Similarly, a direct comparison of word association data and distributional data reveals differences between both. In Vankrunkelsven, Verheyen, Storms, and De Deyne (2018; see also Vankrunkelsven, Vankelecom, Storms, De Deyne, and Voorspoels 2021), a distributional semantic model derived from textual co-occurrences and a model based on a free word association task are compared in their ability to predict properties that affect lexical processing, viz. age of acquisition, concreteness, and three affective variables (valence, arousal, and dominance). Comparing both models to determine which is better at predicting the properties in question, both a study on Dutch data and a study on English data found the word association-based model to perform better. From there to conclude that word association models are best overall is too rash a conclusion, though. The point should rather be that the different methodological inroads into meaning should be systematically compared.

Our exploration of distributional corpus semantics should thus be seen as part of a broader research programme that aims at triangulating referential, psycho-experimental, and corpus-based perspectives on meaning. That triangulation is not a systematic component of what we will be doing here, but the overall perspective does have fundamental consequences for our modus operandi. If we place our exploration of distributional corpus semantics in the context of such a wider investigation, it follows that we will accept a high degree of *methodological underdetermination* in semantics. If the phenomenon usually known as 'meaning' (we will come back to this hedge in a moment) is anisomorphically multidimensional, we should be careful with considering any of the dimensions as an ultimate standard by which to evaluate the others. The various dimensions need to be compared, but as a working assumption we assume that at this point into the research programme it is impossible to assign a single one that can function as an indubitable yardstick. If we were simply to ask which method is the best at identifying utterance meanings, we should be aware by now that formulating the question in that way may be deceptive. Validating the methods in a straightforward way is only possible if we have an independent way of identifying utterance meanings—but that was the difficulty to begin with. In addition, if we think of the various methods as tools for identifying precisely delineated utterance meanings, we may well be repeating the mistake that originally came with the traditional model of systemic polysemy. We have given up the idea of discrete systemic meanings, but aren't we still thinking of utterance meanings as clear and distinct entities? Methodologically speaking, we are trying to get a clear picture of utterance meaning, but what if the thing we try to picture is intrinsically unclear? Are we looking at something through a fog, or is the fog the thing we are looking at?

So, although we *will* be interpreting distributional models against the background of an intuitive understanding of the texts, we must be careful with considering that intuitive understanding a stable standard. As we have illustrated in the previous section (and as will be recognized by anyone who has ever done lexicographical work or engaged in a close reading of texts) semantic intuitions about utterance meaning are as shifty as grammaticality judgements. Such an epistemological position is different from the approach taken in an engineering context. A computational linguist will look at alternative distributional models as competing approximations of the standard and evaluate the success of the model against that standard. By contrast, we assume that for the time being (and perhaps more fundamentally), any such standard is itself only an approximation.

The bottom line

- The token-based distributional workflow illustrated in this monograph is a specific instantiation of a broader set of distributional approaches to lexical semantics, including collocational analysis, the behavioural profile approach, and traditional fully manual text analysis.
- The choice for a count-based instead of a prediction-based distributional method is motivated by a transparency requirement: the apparent attractivity of the vector space approach necessitates a linguistically informed analysis of its operation.
- The broad set of distributional approaches in lexical semantics does not exhaust the methodological options; it is merely one of the main methodological perspectives, alongside referential and psycho-experimental approaches.
- These three broad methodological perspectives do not necessarily converge, and the more traditional intuition-based polysemy tests that correspond to them are known to yield divergent results. Accordingly, we must reckon with a fair amount of (potentially irreducible) methodological underdetermination in semantics.

PART II

DISTRIBUTIONAL METHODOLOGY

The alternation between a descriptive and a methodological perspective that we encountered in the first part of the book also shapes the further structure of the monograph, with this difference that for each step, we will now work with sets of two chapters. Thus, Chapters 3 and 4, which constitute the second part of the book, take a methodological point of view, whereas Chapters 5 and 6 have a descriptive orientation. Taken together, Chapters 3, 4, 5, and 6 all concentrate on semantic and lexical variation as such, without putting a heavy focus on the way in which semantic and lexical phenomena may exhibit variation across language varieties. In particular, expanding on the informal introduction of Chapter 2, Chapters 3 and 4 offer a technically oriented look at a distributional semantics. Chapter 3 shows how a count-based, token-based vector semantics has to make a wide number of choices with regard to the steps and parameters included in the workflow, and how that variation defines a space of possible outcomes. Chapter 4 introduces the visualization tool that we have developed to support the analysis of such spaces.

3
Parameters and procedures for token-based distributional semantics

Chapter 2 introduced the basic concepts behind the distributional approach to modelling lexical semantics on the level of individual occurrences, that is, tokens. The procedure consists of several steps. First, the words that a target item co-occurs with in a given token are considered indicative of semantic features expressed by the token. We refer to these as *first-order context words*. Second, the relevance of a specific context word to the target word's semantics is estimated via their mutual association strength in a corpus, for instance with pmi. By combining first-order context words and their association strengths, we build a *first-order vector representation* for the target. Third, by incorporating the type-level vectors of first-order context words into the first-order vector representation, as illustrated in Figure 2.2, we create a *second-order vector representation*. Finally, semantic similarity between tokens can be measured mathematically as the similarity between their second-order vector representations, and tokens can then be grouped together on the basis of their similarity. Such clusters of tokens reveal what semantic structure can be found in the data, for instance, whether we can identify clearly distinct groups that correspond with what we would normally think of as the meanings of the word.

In this chapter, we discuss this procedure in more detail. We introduce the distributional models that are used in this volume and offer an overview of the different parameters that play a role in building them. Distributional modelling indeed implies making a number of choices, from the source of the data and the unit of analysis to the definition of what counts as context and the selection of metrics and algorithms. Making these decisions explicit is crucial: on the one hand, they are necessary to interpret the models themselves, but on the other, they are essential for reproducibility (see also Levy, Goldberg, and Dagan 2015: 211–12). Accordingly, the chapter focuses on the choices and options that we have included in our workflow, and that shape the models that will appear in following chapters. First, Section 3.1 describes how token-level vector space models are created as mathematical representations of the occurrences of a lexical item. As explained in Chapter 2, we will focus on context-counting models, but this is by no means the only viable path. Other techniques, such as BERT (Devlin, Chang, Lee, and Toutanova 2019), that can also generate vectors for individual instances of a word, could be used for the first stage of this workflow. By and large, Section 3.1 goes

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0003

through the same steps as Section 2.1, but now from a technical and mathematical perspective. The next two sections provide detail about the variability built into the procedure, that is, about the alternative choices that can be made regarding the successive steps. These choices fall into two broad groups. Section 3.2 explores parameter settings that have a linguistic background or that can be interpreted linguistically. For instance, context words may be selected by considering their syntactic status, or not. Section 3.3 by contrast delves into more mathematical or statistical parameters that are harder to interpret from a linguistic perspective. Section 3.4 discusses the procedures that can be applied to token-by-context matrices, from computing the distances between the tokens to clustering them for the purpose of sense identification. In our approach, this step is supported by a visualization tool, which will be introduced and discussed in depth in Chapter 4. Finally, Section 3.5 offers an overview of the parameter settings used in the different chapters of the volume. Our case studies do not uniformly use a single set of parameters, but the parameter settings for each study depend on the topic of investigation and the corpus materials at hand. So, Section 3.5 charts where specific alternatives show up in the case study chapters.

The present chapter has a more technical nature than the previous ones, and some of the mathematical detail may be less accessible to readers with a linguistic rather than a computational background. This should, however, be no problem for an understanding of how the following chapters unfold. The essential information to be retained from this chapter pertains to two points: the origins of the variability built into the distributional method, and the terminology we will use to refer to the various aspects of that diversity. This diversity is in fact a fundamental issue for the following steps of our argument. In Chapter 5, we will explore if semasiologically optimal choices can be made from among the extensive set of possible distributional models, and the outcome of that exploration will inform the way in which we conduct the onomasiological and lectometric studies in the second half of the monograph.

3.1 From text to vector space

Context-counting vectors are essentially lists of association strength values. Each word is represented by its association strength to a long array of words that it might co-occur with, as illustrated in Table 3.1. Unlike in collocation studies, low values—or even the absence of co-occurrence—are not excluded but used in the comparison with other words. For example, Table 3.1 shows small vectors representing the English nouns *linguistics*, *lexicography*, *research*, and *chocolate*, as well as the adjective *computational*, with co-occurrence information obtained from the GloWbE corpus (Corpus of Global Web-based English, Davies and Fuchs 2015). The values are their Pointwise Mutual Information (pmi) with each of the lemmas

Table 3.1 Example of type-level vectors

TARGET	LANGUAGE/N	WORD/N	FLEMISH/J	ENGLISH/J	EAT/V	SPEAK/V
linguistics/n	4.37	0.99	–	3.16	–	0.41
lexicography/n	3.51	2.18	–	2.19	–	2.09
computational/j	1.6	0.08	–	−1	–	−1.8
research/n	0.2	−0.84	0.04	−0.5	−0.68	−0.38
chocolate/n	−1.72	−0.53	1.28	−0.73	3.08	−1.13

PMI values based on symmetric window of 10; frequency data from GloWbE. The letter to the right of the slash indicates the part of speech: *n* for nouns, *j* for adjectives, *v* for verbs.

in the columns: the higher the values, the stronger the attraction between the word in the row and the word in the column. From a collocational perspective, *linguistics* is strongly attracted to both *language* and *English*, that is, they occur very often in a span of ten words from each other, considering their individual frequencies. By contrast, it is less attracted to *word* and *to speak*, and does not co-occur with either *to eat* or *Flemish* within that window, in this corpus.

Each row in Table 3.1 is a vector coding the distributional information of the lemma it represents. ‘Lemmas’ in this case are lemmatized and part-of-speech tagged items. Lemmatization implies that all the morphological forms of a word, like singular *chocolate* and plural *chocolates*, are considered together. Part-of-speech tagging means that the word class of the item is identified: *research/n* involves the noun *research* but not the verb *to research*. (Lemmatization and part-of-speech tagging are not strictly required for a distributional analysis. Section 3.2 will discuss alternative definitions of the unit of analysis.) The vectors in Table 3.1 are meant to code the distributional behaviour of the linguistic forms they represent in order to operationalize the notion of distributional similarity and, consequently, model their semantic similarity. For example, the first two rows of Table 3.1, representing *linguistics* and *lexicography*, are similar to each other: both words have a similar attraction to *language* and to *English*, although *lexicography* is more strongly attracted to *word* and *to speak* than *linguistics*. More importantly, they are more similar to each other than to other rows in the table, which have lower values for those four columns and might even co-occur with *Flemish* or *to eat*. The basic idea behind a distributional methodology—sometimes referred to as the Distributional Hypothesis—rests on the observation that words that are distributionally similar, like *linguistics* and *lexicography*, are semantically similar or related, whereas words that are distributionally different, like *linguistics* and *chocolate*, are semantically different or unrelated.

The rows in this table are type-level vectors: each of them aggregates over a number of attestations of a given lemma in a given corpus to build an overall profile. As a result, a type-level vector collapses the internal variation of the lemma, that is, it does not distinguish between its different senses or other aspects of its semasiological structure. Some researchers solve this issue by means of factorization methods (Van de Cruys and Apidianaki 2011). In contrast, the studies in this volume build vectors for the individual instances or tokens by relying on the same principle underlying the type-level vectors: items occurring in similar contexts will be semantically similar. For instance, we might want to model the three occurrences of *to study* in (3.1) through (3.3), where the target item is in boldface and some context words are in italics.

(3.1) Would you like to **study** *lexicography*?

(3.2) They **study** this in *computational linguistics* as well.

(3.3) I eat *chocolate* while I **study**.

Given that, at the aggregate level of the entire corpus under consideration, a word can co-occur with thousands of different words, type-level vectors can include thousands of values. In contrast, token-level vectors that capture which context words occur in the immediate surroundings of the target can maximally only have as many non-zero values as the individual window size comprises. In a corpus as a whole, a word occurs together with many, many other words. But in a specific utterance, it co-occurs with only a few. This reduces the chances of overlap between token-level vectors drastically, making most tokens maximally different from each other. In fact, the three examples don't share any item other than the target, *to study*. As a solution, inspired by Schütze (1998), the context words around the token are replaced with their respective type-level vectors (see Chapter 2, and Heylen, Speelman, Wielfaert, and Geeraerts 2015; De Pascale 2019; Montes 2021a). Concretely, example (3.1) would be represented by the vector for its context word *lexicography*, that is, the second row in Table 3.1; example (3.2) by a combination of the vectors for *linguistics* (row 1) and *computational* (row 3); and example (3.3) by the vector for *chocolate* (row 5). This not only addresses the sparsity issue, ensuring at least some overlap between the vectors, but also allows us to find similarity between (3.1) and (3.2) based on the similarity between the vectors for *lexicography* and *linguistics*. As we will see in the following sections, we can additionally use the association strength between the context words and the target type to give more weight to the context words that are more characteristic of the lemma we try to model. The result of this procedure is a co-occurrence matrix like the one shown in Table 3.2. Each row represents an instance of the target lemma, like *to study*, and each column one of many selected lemmas occurring in the corpus. The values are the (sum of the) association strength between the words that occur around the token, that is, their first-order context words, and each of the words in the columns, that is, the second-order context words. In addition, all negative and missing values have been set to zero, due to the unreliability of negative pmi values (see Section 3.3).

Table 3.2 Small example of token-level vectors of three instances of *to study*

TARGET	LANGUAGE/N	WORD/N	ENGLISH/J	SPEAK/V	FLEMISH/J	EAT/V
$\text{study}_{(3.1)}$	4.37	0.99	3.16	0.41	0.00	0.00
$\text{study}_{(3.2)}$	5.97	1.07	2.16	0.00	0.00	0.00
$\text{study}_{(3.3)}$	0.00	0.00	0.00	0.00	1.28	3.08

Table 3.3 Cosine distance matrix between the three instances of *to study*

TARGET	$\text{STUDY}_{(3.1)}$	$\text{STUDY}_{(3.2)}$	$\text{STUDY}_{(3.3)}$
$\text{study}_{(3.1)}$	0.00	0.04	1.00
$\text{study}_{(3.2)}$	0.04	0.00	1.00
$\text{study}_{(3.3)}$	1.00	1.00	0.00

By this point we have obtained a token-level model, but keep in mind that models can be built in various ways: Sections 3.2 and 3.3 describe the kinds of choices that will lead us from the corpus to multiple alternative models. We are not done once we have a model, though. The next step in the workflow is to compare the items to each other. We can achieve this by computing cosine distances between the vectors (see Section 3.4 for the technical description). The resulting distance matrix, shown in Table 3.3, tells us how different each token is to itself, which takes the minimum value of 0, and to each of the other tokens, with a maximum value of 1. We can see that rows $\textit{study}_{(3.1)}$ and $\textit{study}_{(3.2)}$, representing examples (3.1) and (3.2) respectively, are very similar to each other, because they co-occur with similar context words, that is, *linguistics* and *lexicography*, but drastically different from $\textit{study}_{(3.3)}$, which was modelled based on *chocolate*. The specific selection of context words is crucial: if we had selected *computational* but not *lexicography* to model $\textit{study}_{(3.2)}$ it would have resulted in a larger difference with $\textit{study}_{(3.1)}$. Those are the choices discussed in Sections 3.2 and 3.3.

Table 3.3 is small and simple, but what if we had hundreds of tokens? The more items we compare to one another, the larger and more complex the distance matrix becomes. In order to interpret it, we need more stages of processing. Dimensionality reduction techniques, such as multidimensional scaling, t-SNE and UMAP, which will be discussed in Chapter 4, offer us a way of visualizing the distances between all the tokens by projecting them to a 2D space. We can then represent each distance matrix as a scatterplot, like in the plot (shown twice) of Figure 3.1, where each point represents a token and their distances in 2D space approximate their distances in the multidimensional space of the co-occurrence matrix (this plot is also discussed as Figure 5.19 in Chapter 5). Visual analytics, such as the tool described in Chapter 4, can then help us explore the scatterplot to figure out how tokens are distributed in space, why they form the groups they form, and so on.

Now, why do we show the same plot twice in Figure 3.1? The one on the left is the result of dimensionality reduction as just discussed. If there are any clusters emerging from the underlying distance matrix, they have to be identified visually (and obviously, the relevance of a dimension reduction technique is precisely to allow for such a visual analysis). But clusters of tokens can also be identified by mathematical means, by programs—clustering algorithms—that organize the data in groups on the basis of their closeness in the distance matrix. While dimensionality reduction can help us visualize an approximation of the relative distances between the individual points, clustering algorithms add to the analysis by automatically extracting groupings. In computational linguistics, this is the typical workflow in so-called word sense disambiguation tasks, where the resulting clusters are then matched to senses (see for instance Navigli 2012; Nasiruddin 2013; Amrami and Goldberg 2019). In most of the studies collected in this volume, the preferred clustering algorithm is HDBSCAN (see Section 3.4), because it includes information on the relative membership of each token to the assigned cluster and does not try to cluster all the tokens. The image on the right in Figure 3.1 reproduces the same model as the plot on the left but adds a colour code corresponding to the clusters returned by HDBSCAN. The clusters are mapped to colours and, in addition, the ε value (a proxy for density used by the HDBSCAN algorithm, see Section 3.4)

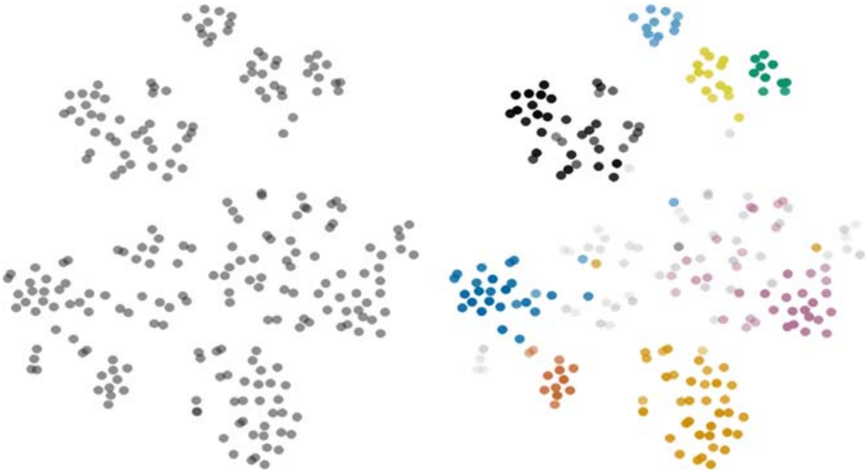


Figure 3.1 2D representation of Dutch *hachelijk* 'dangerous/critical'. Without colour coding on the left side, and with HDBSCAN clusters mapped to colours on the right side

A colour version of this figure and later figures can be consulted via the free PDF download at https://global.oup.com/academic/product/lexical-variation-and-change-9780198890676 or via OUP's online platform at https://doi.org/10.1093/oso/9780198890676.001.0001.

is mapped to the transparency of the dots. The tokens in grey in the right-hand plot are tokens that HDBSCAN discards as noise because they are not similar enough to the other clustered tokens; they are tokens that the algorithm finds hard to group with any of the other clusters. It will be noted that the plot on the right shows a fair amount of agreement between the HDBSCAN clustering solution and the spatial grouping returned by t-SNE: roughly speaking, the groups that may be recognized on the left have their own colour on the right. Thus, Montes (2021a) relies on the degree of agreement between HDBSCAN and t-SNE, among other features, to classify the shapes found in these kinds of plots. The technicalities of these procedures will be presented in more depth in Section 3.4.

3.2 Linguistically informed parameters

Some of the choices to make when designing vector space models will be informed by linguistic theory rather than statistics or mathematics. The first of these is the corpus to use, based on the language and/or lect we want to study and the availability of corpus resources. The corpora used in the case studies in this volume will be described in the context of the relevant chapters.

Another important choice involves the unit of analysis. The case studies illustrated in Chapter 5 and Chapter 10 make use of annotated corpora with information on lemma and part-of-speech so that we can define the unit of analysis as *lemma/pos*, that is, a lemmatized and syntactically labelled lexical item, as in the examples in Section 3.1. In such a case, all items at all levels of the procedure (the target, the first-order context features, and the second-order features) take a lemmatized and part-of-speech tagged form, and co-occurrence frequencies and association strength measures are always computed with the lemma/pos combination as unit. In contrast, the case studies illustrated in Chapters 6 and 9 rely on corpora that are not lemmatized or part-of-speech tagged, and the target units are defined accordingly. The target unit will then be defined by manually identifying the morphological variants of the lemma; the specific steps are described in the corresponding chapters. First- and second-order context words, on the other hand, will be different word forms or even different spelling variants of the word forms. As a result, the association strength values between the target type and its context words will be computed between the manually lemmatized target and context word forms, while the values in the vectors themselves will be computed between word forms or spelling variants.

Various considerations pertain to the choice between a fully lemmatized approach and one relying on word forms; see Turney and Pantel (2010: 155) and Sahlgren (2008: 47–8) for a discussion and Kiela and Clark (2014: 25) for performance comparisons. On the one hand, word forms of the same lemma/pos may tend to behave in different ways, which could be a descriptive argument for a word-form-based approach. From a lexicographic and lexicological perspective, on the other hand, it makes sense to use lemma/pos as a unit. It is the head of dictionary entries and a more typical unit of linguistic analysis. In addition to these descriptive considerations, there are practical ones to take into account. In our research, it proved easier to obtain meaningful distributional results on datasets with annotation. Furthermore, the (mis)match between word forms and lemmas strongly depends on the language under study: in languages like Spanish, French, Japanese, and Dutch, verbs can take many more different forms than in English; conversely, Mandarin lacks morphological variation or even spaces between what could count as words. The word form *hoop* in Dutch, for instance, can correspond to the noun meaning either 'hope' or 'heap', or the verb meaning 'to hope', which can also take other forms such as *hopen*, *hoopt*, *hoopte*, and *gehoopt* depending on person, number, and tense. If our interest, from a lexicological perspective, lies in studying the behaviour of the noun *hoop* and its meanings, conflating the noun with one of the verbal forms of the homographic verb needs to be avoided. But the automatic annotation of corpora is not always reliable, especially in the case of historical materials. So, even if a fully lemmatized approach might be preferred, a word form-based approach may be practically motivated: if the corpus resources are not annotated, text normalization and annotation might be too costly, and a choice can be made for relying on non-annotated word forms, as in Chapters 6 and 9.

Another choice with respect to the unit of analysis pertains to multi-word expressions. They can either be dealt with, like lemmatization, during the tokenization stage of the corpus, or, as in the studies in this volume, left to the post-hoc analysis of the token models, where they typically appear as specific clusters in the token spaces.

Once we have selected our corpus and defined our unit of analysis, we can move on to determine which context words to capture, both at the level of the token and for the type-level vectors representing its context features. First-order parameters are those that influence which elements in the immediate environment of the token will be included in modelling said token. In terms of example (3.1), this translates as deciding whether *lexicography*, *like*, *would*, *you*... should be included or excluded in the modelling of that particular occurrence of *to study*. Such decisions depend on the available information. The corpus used for the case study in Chapter 5 includes syntactic information, which expands the possibilities to syntactically informed models. Models that do not take syntactic information into account are called bag-of-words models. They may vary based on whether sentence boundaries are respected, which window size is chosen, and whether any part-of-speech filters are applied. In contrast, syntactically informed models can select context words based on the distance to the target in terms of syntactic relationships. The syntactic models considered in Chapter 5 are based on dependency-parsed corpora, and so we will refer to them as dependency-based models. Syntactically informed models could also find context words that match specific, predefined templates. Rather than using them as selection mechanisms for context words, models could also have syntactic patterns as context features, in line with the notion of 'colligation' mentioned in Section 2.2. This was not incorporated in any of the case studies discussed here, though.

The first decision in bag-of-words models distinguishes between those that include words outside the sentence of the target and those that do not; this typically does not make much of a difference in the final result. Another choice is the frequency threshold for context features: in this monograph, different case studies choose different thresholds based on their corpora and focus, but within each case study, all models use the same threshold. More relevant is the window size: the span of words to either side of the target from which context words are selected. Window sizes are typically larger for token-level models than for type-level models (Schütze 1998; De Pascale 2019). In our case studies they range between 3 and 15. Finally, some models refine their first-order selection with part-of-speech filters, for example only including nouns, adjectives, verbs, and adverbs. In Montes (2021a), bag-of-words models without part-of-speech filtering models tend to behave similarly to dependency-based models, while those with such a filter tend to be redundant with ppmi-based selection, described later.

The distinction between bag-of-words and dependency-based models does not only affect the number and type of context words included but also how tailored the selection is to each specific token. On the one hand, a closed-class element may be distinctive of particular usage patterns in which a term might occur. For example, a count-reading or a mass-reading of a noun could be distinguished by its article; different senses of a verb might co-occur with different prepositions, and so on. However, such a frequent and multifunctional context word could easily occur in the immediate context of the target without actually being related to it: a bag-of-words model would not distinguish between the relationship between *a* and *coffee* in *a cup of coffee*, *a coffee* or *some coffee at a bar*, whereas a syntactically informed model would. As these examples show, narrowing the window span does not solve the issue, and besides, it would also drastically reduce the number of context words available for the token and for any other token in the model. On the other hand, we might also be interested in syntactically related but graphically distant context words, such as *interested* and *words* in this very sentence. In the previous sentence, *words* is the head of the prepositional object of *interested*, but there are seven words in between. Widening the window to include such context words may add too much noise to the representation of this token and to that of any other token in the model.

A dependency-based model, instead, will only include context words in a certain syntactic relationship to the target, regardless of the number of words in between from a bag-of-words perspective. As illustration, Figures 3.2 and 3.3 represent the syntactic trees corresponding to examples (3.1) and (3.2), with the arrows going from the head to the dependent and their labels indicating the syntactic

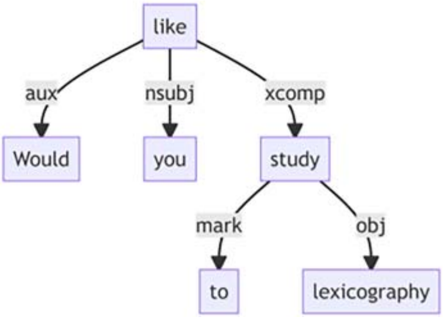


Figure 3.2 Syntactic tree of example (3.1)

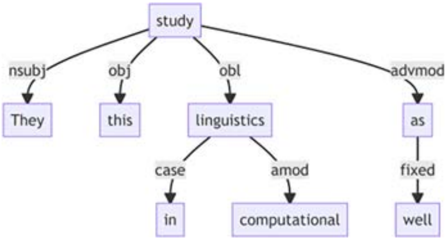


Figure 3.3 Syntactic tree of example (3.2)

relationship. A model could select the context words that are directly linked to the target element, that is, parents and children. For example (3.1) these would be *like*, *to*, and *lexicography*. We could also have a model that extends the threshold to two steps in the syntactic path, adding siblings such as *would* and *you* in example (3.1), or grandchildren such as *in* and *well* in example (3.2).

The dependency models used in Chapter 5 that use this approach have three possible settings: they accept up to two steps, they accept up to three steps, or they accept up to three steps but additionally weigh the contribution of the vectors based on the distance along the syntactic path. As an example, imagine that (3.1) is one of our occurrences but our target is actually *lexicography*. The only word linked by only one step is its parent, *study*. A model that only accepts two steps would additionally include its grandparent *like* and its sibling *to*. A model that also accepts three steps would also include the parent's siblings, *would* and *you*, as context words. With such a simple sentence, all the elements are taken by the model. The third kind of model, however, would not just add the vectors of these context words but give more weight to *study*, less weight to *like* and *to*, and even less weight to *would* and *you*. Other dependency models used in Chapter 5 use specific relationship patterns instead of distance along the syntactic path. For instance, given a target such as *study*, we might be more interested in collecting subjects and objects than functional complements. A model with such settings would then capture *lexicography* for example (3.1) but not *to*; it would also capture *they* and *this* in example (3.2). Because these models were originally designed for the research in Montes (2021a), the thresholds of the step-based models and the templates of the pattern-based models were informed by the manual annotation of the case studies analysed in that dissertation. Concretely, the annotators had the task of selecting the context words that helped them in the disambiguation. The distances for the thresholds and the patterns for the templates were based on the most frequent results.

In the studies collected here, all the first-order parameters produce filters to select the context words in the environment of each token. Alternatively, dependency information could have been included as a feature or dimension in its own right. For example, instead of selecting *lexicography* as context word of the token in (3.1) based on its bag-of-words distance, part-of-speech filter, or dependency relation to the target, we could use (*lexicography*, *object*) that is, 'has *lexicography* as direct object' as a first-order feature. Its type-level vector then would have information on all the other verbs that take *lexicography* as its direct object. For technical and practical reasons, this was not implemented in the studies discussed here, but we refer to Padó and Lapata (2007) as a general framework for dependency-based semantic modelling and more particularly to Erk and Padó (2008) for an implementation of dependency-based token vectors.

Regardless of whether we use a bag-of-words or dependency-based model, we can further implement filters based on association strength. As the productive field of collocation analysis suggests (see Evert 2009), this informativeness can be operationalized by association measures, such as ppmi; the specific choice of ppmi over alternative measures belongs to the more statistically informed choices described in Section 3.3 below. Such association strength measures could then be used to exclude words that are not sufficiently attracted to the target or, additionally, to give more influence to the words that are, since they are assumed to be more informative. As illustration, example (3.4) replicates (3.2) with ppmi values from the GloWbE corpus (based on a symmetric window size of 4) as subscripts.

(3.4) They$_{0.20}$ *study* this$_{0.0}$ in$_{0.39}$ computational$_{1.30}$ linguistics$_{3.66}$ as$_{0.0}$ well$_{0.15}$.

In a model that selects context words with a positive pmi, *this* and *as* are immediately excluded as context words. If we raise the threshold to 1, then *they*, *in*, and *well* are also excluded, and we are left with only *computational* and *linguistics*, as shown in (3.2) originally. In addition, a model that weighs the context words based on their ppmi to the target would multiply the vectors of these context words by said ppmi, reinforcing the contribution of those words most attracted to *study* and dampening that of those words less attracted. In such a situation, given other instances of *to study* co-occurring with *computational* or with *linguistics*, (3.4) would be more similar to those co-occurring with *linguistics* than to those co-occurring with *computational*, since its ppmi with *study* is much larger. In contrast, in a model without weighting the contribution of both *linguistics* and *computational* will be equal, and therefore the similarity to tokens with either *linguistics* or *computational* will be the same. Of course, in practice there would probably be other context words influencing that distance as well. It may be noted that ppmi is not the only measure that may be used for selecting or weighting context words. Heylen, Wielfaert, Speelman, and Geeraerts (2015) weigh the contribution of each context word by their ppmi with the target, while De Pascale (2019) adds ppmi and log-likelihood ratio (see Section 3.3) thresholds to the selection of context words.

Finally, the selection of second-order features determines the dimensions of the token vectors, that is, how the selected first-order features are represented. Next to the window size and association measure used to calculate the values of the vectors, which were typically fixed to a symmetric window of four and ppmi, there are two variable parameters. First, second-order context words can be filtered by frequency thresholds and, if part-of-speech tagging is involved, by part-of-speech filters. Second, we might reduce the length of the vector, that is, the number of second-order features. One way of doing this is by selecting, for example, the 5000 most frequent lemmas in the corpus, varying the number of dimensions we allow. Another option is to use the union of first-order context words captured for the modelled tokens as second-order dimensions. As a result, the second-order dimensions are tailored to the context of the sample, regardless of their frequency in the corpus; this also leads to a smaller number of dimensions, depending on the size of the sample and the strictness of the first-order filters.

The parameter settings described in this section are linguistically informed in the sense that linguistic theory can guide us in the selection of features, and we can expect the results from each parameter setting to inform, in turn, future linguistic theory. This involves thinking about what constitutes the context: how wide should a window span be, which words count, what is the role of syntactic relationships, what do we expect from strongly attracted context words. In the next section we will look at parameter settings with a less straightforward linguistic interpretation and a heavier statistical background.

3.3 Statistical parameters

In technical terms, that is, with respect to what algorithms work with, a vector space model is an item-by-feature matrix, with each row a vector representing some item and each column a context feature, storing in each cell a representation of the relationship between the row item and the column feature. The first distributional models counted the occurrences of words in documents and represented them in word-by-document matrices: each row represents a word at the type level, each column a document, and each cell the frequency with which the word occurs in the document or some other stretch of text. From such a matrix both word vectors and document vectors can be extracted: the former represent words based on their distribution across documents, whereas the latter represent documents based on the frequency of the words that constitute it. Turney and

Pantel (2010) offer an overview of different kinds of matrices, based on the items modelled and the features used to describe them. Besides matrices, vector space models can be tensors, which are generalizations of matrices for more dimensions that can allow for more complex interactions, like subject-verb-object triples in Van de Cruys, Poibeau, and Korhonen (2013); see also Lenci (2018).

Most of the models described in this volume are token-by-feature matrices: the rows are attestations of a lexical item, and the features are second-order co-occurrences, that is, context words of the context words of the token. Each model is defined by a configuration of parameter settings, that is, by the choices that guided the workflow from the corpus to the matrix. That said, the matrices further undergo changes that allow for clustering and visualization, which themselves imply technical choices. In this section we will go through these kinds of choices, mostly guided by an understanding of the mathematical properties of vectors and by parameter overviews such as Kiela and Clark (2014).

The first technical decision pertains to how the frequency information will be reported. The distribution of words in a corpus follows a power law: a few items are extremely frequent and most of the items are extremely infrequent. Association measures transform raw frequency information to measure the attraction between two items while taking into account the relative frequencies with which they occur. They typically manipulate, in different ways, the frequency of the node $f(n)$, the frequency of its collocate $f(c)$, their frequency of co-occurrence $f(n, c)$ and the size of the corpus N. Evert (2009) and Gablasova, Brezina, and McEnery (2017) offer an overview of how different measures are computed and used in corpus linguistics; Kiela and Clark (2014) compare measures used in distributional models. This technical decision affects two different steps in the workflow: the values of the second order vectors and filtering/weighting values when combining these vectors to represent a token. It is possible to use different measures in either step, combined with different thresholds for the filtering step. In all the studies in this volume, positive pointwise mutual information (ppmi) is used for the values of the second order vectors. When it comes to filtering or weighting first-order context words, instead, all the studies use ppmi and the one in Chapter 9 additionally uses log-likelihood ratio.

Pointwise Mutual Information (Church and Hanks 1989) is one of the most popular measures both in collocation studies and distributional semantics (Bullinaria and Levy 2007; Kiela and Clark 2014; Lapesa and Evert 2014; Jurafsky and Martin 2023). We have already explained the basic idea behind the pmi measure in Section 2.1, but we have to add a few words about the restriction to *positive* Pointwise Mutual Information. In ppmi, the negative pmi values are set to zeros; this is often preferred because negative pmi values tend to be unreliable (Bullinaria and Levy 2007; Kiela and Clark 2014; De Pascale 2019; Jurafsky and Martin 2023), but it also has the advantage of keeping the cosine distances (see below) between the vectors in the 0–1 range. One of the greatest disadvantages of pmi (and ppmi) is its bias towards infrequent events. Referring back to the explanation in Section 2.1, if either P(x) or P(y) is very low, pmi tends to be very high. If context word A is infrequent and always occurs with B, their pmi will be quite high regardless of the frequency of B, that is, even if the occurrence of A is not substantial from the perspective of B. In distributional semantics, the accuracy of models that rely on ppmi does not seem to be affected by the issue presented by this bias; moreover, in these studies the most infrequent words were excluded from any modelling to avoid too sparse, uninformative vectors. However, in collocation studies, pmi's infrequency bias is often counteracted by combining pmi filters with other measures that favour frequent co-occurrences, such as t-scores or log-likelihood ratio (McEnery, Xiao, and Tono 2010). In this volume, Chapter 9 will explore the use of log-likelihood ratio as an oppositely biased alternative to ppmi for filtering and weighting. More technically, whereas ppmi measures the *effect size* of the association between two items, log-likelihood ratio (Dunning 1993; Evert 2004; Lapesa and Evert 2014) measures the *strength of the evidence* that such an association exists. The more events, the more evidence there is, and hence the bias of log-likelihood towards highly frequent events. Log-likelihood also often takes much higher values than pmi, in the order of hundreds and thousands, which reinforces its tendency to overly magnify frequently occurring events.

As mentioned before, ppmi based on a symmetric window span of four words to either side is used for the values of the type-level vectors that represent the context words. However, some models in Chapter 9 also first reduce the dimensionality of the type-level matrix before retrieving the vectors of the context words. For this purpose, singular value decomposition (SVD) is used. Singular value decomposition has been a fundamental aspect of modelling semantic vector spaces since their very introduction (see De Pascale 2019: 227–30 for details), due to its beneficial impact on computing semantic similarity among type vectors. The goal is to reduce the redundancy caused by similar columns in the matrix, generating a new matrix with fewer dimensions that are linearly independent from each other. In addition, these new dimensions are weighted based on the amount of variation that they explain, which is captured by the so-called singular values. Moreover, this operation has the effect of 'smoothing' the original association between the rows and the columns of the matrix, that is, between the first-order context words and the second-order context words. Even though some associations might be unattested in the corpus, they might still be judged likely to be true in language at large, on the basis of knowledge about other target-context co-occurrences. This smoothing operation arguably allows for a better generalization of the semantic properties of the vectors. Following best practices, described in De Pascale (2019), the type-level matrix is reduced from a dimensionality of 20 000 to only the 200 highest (latent) dimensions. These dimensions are then multiplied by the square root of the singular values, so that the contributions of the first dimensions are dampened and those made by the lower dimensions are simultaneously increased

(De Pascale 2019: 230). The representation of the context words is then taken from this reduced, denser matrix.

When creating a token-level vector, the type-level vectors of the selected context words are combined into one: in the studies described in this volume, they are added, but see Mitchell and Lapata (2008) for alternatives. If weighting is used, the type-level vector is first multiplied by the association score (ppmi or log-likelihood ratio) between itself and the target type. In such models, the higher the association strength between a context word and the target type, the larger its contribution to the vectorial representation of the token. If weighting is not used, all context words contribute equally to the representation of a token.

Once we have combined type-level vectors to create denser token-level vectors, we obtain a token-by-context matrix in which each row represents a token and each column corresponds to a second-order feature, that is, a column of the type-level matrix. The following steps mostly pertain to forms of post-processing, mainly clustering and visualization, and can also be applied to context-predicting models. Such models, based on neural networks, produce low-dimensional, dense vectors from the outset (see Jurafsky and Martin 2023, Chapter 6 for an overview of approaches to dense vector creation). Although dense vectors have shown to be better at a range of natural language processing tasks, the downside is that, whatever the dimensionality reduction approach (neural networks or singular value decomposition), they do not allow for a straightforward semantic interpretation in terms of shared contexts making two words or two occurrences similar.

3.4 From vector space to token clouds

With or without dimensionality reduction, a token-by-context matrix is a distributional model and its columns constitute the dimensions of the 'semantic space'. If instead of a count-based model, neural word embeddings are used, the output is also a token-by-context matrix, where the dimensions come from the weights of a layer of the neural network. These hundreds or thousands of dimensions are the ones being referred to when talking about the multidimensionality of the model and the interpretative challenge it presents. In order to capture and describe the patterns hidden in such massive representations, we make use of specific techniques: clustering techniques, which will be described in this section, and dimensionality reduction techniques for visualization purposes, which will be discussed in Chapter 4. The former return the cluster membership of the tokens, whereas the latter try to translate the relative distances between the tokens in the multidimensional space to a low-dimensional space that can be graphically plotted. There are yet other ways in which computational linguistics can use distance matrices, such as ranking nearest neighbours or representing analogies based on the differences between pairs of vectors, but here we focus on the detection of semasiological structure in the form of groups of semantically similar tokens.

To build a distance matrix, we need to compute the pairwise distances between tokens, and this implies a choice between various possible distance metrics (see Weeds, Weir, and McCarthy 2004 for an overview and comparison). In the studies of this volume, we have only included the cosine metric, as it is the most widely used metric in distributional semantics. By itself, the cosine is a measure of *similarity* (not distance) between two vectors *vv* and *ww*. Cosine similarity can be seen as a generalization of the two-dimensional case taught in secondary school trigonometry (i.e. the ratio of the adjacent side to the hypotenuse in a right triangle) to a high-dimensional space, where the dimensions are the context features of the token vectors and the pmi values are the coordinates of the tokens on these dimensions. In other words, cosine similarity measures the relative position of two tokens vis-á-vis each other as the cosine of the angle between the two token vectors in the high-dimensional space of context features. Mathematically, cosine similarity is defined as the normalized dot product of the two vectors (see Jurafsky and Martin 2023 for formal details), but here we will focus on the properties that are relevant for the linguistic interpretation. Firstly, the cosine similarity ranges between −1 and 1: it will be 1 between identical vectors, 0 for orthogonal vectors, and −1 for vectors that are completely opposite from each other. When we only use positive pmi (ppmi), the cosine similarity ranges between 0 and 1. Maximally dissimilar vectors that do not share any non-zero dimensions, like study$_{(3.1)}$ and study$_{(3.3)}$ in Table 3.2, will then have a cosine similarity of 0. A second important property is that cosine similarity is sensitive to the *angle* between the vectors but not to their magnitude: the similarity between study$_1$ and a vector created by multiplying the ppmi values in all the cells of study$_{(3.1)}$ by any constant will still be 1, even though the association strength between the token and each of context features would be much stronger. Despite this insensitivity to magnitude, cosine similarity is the most commonly used metric in distributional models (Jurafsky and Martin 2023) and it has been shown to outperform other measures, especially when combined with ppmi (Bullinaria and Levy 2007; Kiela and Clark 2014; Lapesa and Evert 2014, who also suggest the use of a correlation similarity metric). As mentioned above, most applications in distributional semantics require a *distance* matrix rather than a similarity matrix. Therefore, cosine similarities are usually transformed to distances by inverting the scale ($cosine_{dist}$ = 1 – $cosine_{sim}$), so that identical vectors—and each vector to itself—have a cosine distance of 0 and orthogonal vectors have a cosine distance of 1, as shown in Table 3.3.

Before applying clustering algorithms, the cosine distances are transformed with the aim of giving more weight to short distances, that is, nearest neighbours, and decreasing the impact of long distances. For each token vector *v* with *n* dimensions, we define the transformed vector $v_{transformed}$ as $v_{transformed_i} = log(1 + \log(rank(v)_i))$ for each i, with $1 \leq i \leq n$, and where $rank(v)_i$ is the similarity rank of the ith value in v. For example, if originally we have the distances $v = [0, 0.2, 0.8, 0.3]$, the rank transformation returns $rank(v) = [1, 2, 4, 3]$, which after the first logarithm transformation becomes $[0, 0.693, 1.39, 1.099]$, and after the second transformation, $v_{transformed} = [0, 0.52, 0.86, 0.74]$. On the one hand, the magnitude of the distance is not as important as its ranking among the nearest neighbours. On the other hand, the lower the ranking, the smaller the impact: the difference between the final values for ranks 1 and 2 is larger than between ranks 2 and 3. The new matrix, where each row v has been replaced by its $v_{transformed}$, is converted to Euclidean distances.

To recapitulate, the first set of choices results in a token-by-context matrix, but then we obtain a token-by-token matrix registering how different each token is from every other token. The rest of the steps we take to study the models are based on these (transformed) distances: clustering the tokens, the visualization, and even comparing the models to each other. The distance matrix therefore represents each token in its relationship to the rest of tokens in the sample, that is, indicating which other tokens it is more similar to or more different from.

In word sense disambiguation tasks, the vectorial representations of different attestations are clustered into groups of similar tokens (see Chapter 5; Montes, Franco, and Heylen 2021). There are a variety of clustering algorithms appropriate for different kinds of data and structures. We will not offer an overview of the options (see Navigli 2012; Nasiruddin 2013), but only describe the main clustering technique used in this volume: HDBSCAN, the algorithm that returns the coloured clusters in Figure 3.1. Note that in Chapter 6 traditional hierarchical clustering is used instead.

Hierarchical density-based spatial clustering of applications with noise or HDBSCAN (Campello, Moulavi, and Sander 2013; McInnes, Healy and Astels 2016) belongs to the family of density-based clustering approaches that look for dense areas in a space as possible clusters. In our case, we look for dense areas with tokens in a distributional semantic space. Density-based clusters consist of a dense core surrounded by a less dense periphery and are separated from each other by sparsely populated regions in the (semantic) space. This corresponds well to a prototype-based view of concepts, as introduced in Chapter 2. Unlike better-known hierarchical clustering algorithms, HDBSCAN does not try to place all the items in the sample in different groups, but instead assumes that the dataset might be noisy and that the items may have various degrees of membership to their respective clusters. In practice, it tries to discriminate between dense areas, that is, groups of elements that are very similar to each other, and sparse areas, where there are larger distances between the elements. This could be roughly compared to isolating groups of people at a party: there will be clear cores but mostly with fuzzy boundaries and some wandering guests still finding their crowd. Depending on the threshold we choose to define a group (as opposed to two or three people wandering together), the algorithm will return different clusters. In technical terms, the density of the area in which we find a point *a* is estimated by calculating its core distance $core_k(a)$ which is the distance to its *k* nearest neighbour, *k* being a parameter *minPts* – 1 and *minPts* the threshold to count a set of points as a group. This measure is at the base of the mutual reachability distance, which is used to compute a new distance matrix between the items. The mutual reachability distance between two points is the maximum between the core distance of each point and the original distances between the points. Afterwards, a single-linkage hierarchical clustering algorithm is applied on this new distance matrix. As a result, the items are organized in a hierarchical tree, from which clusters are selected based on the *minPts* requirement and their densities. A related notion to $core_k(a)$ is ε, which is defined as the radius around a point in which *minPts* – 1 can be found.

In R, the algorithm can be implemented with dbscan: :hdbscan() (Hahsler and Piekenbrock 2021; Hahsler, Piekenbrock, and Doran 2019). Its input can be an item-by-feature matrix or, like in this case, a distance matrix. The output includes, among other things, the cluster assignment, with noise points assigned to a cluster 0, membership probability values, which are core distances normalized per cluster, and ε values, which can be used as an estimate of density.

Finally, the approach taken in most chapters of this volume is to vary model parameters and generate many distance matrices for the same sample of tokens. In some cases, we might want to compare models to each other. We do so by comparing the profiles of the tokens in each pair of models, that is, the ranking of their nearest neighbours. The resulting distance matrix facilitates plotting the models (see Chapter 4) and clustering them, as is done in Chapter 5 with the Partitioning Around Medoids algorithm, or PAM (Kaufman and Rousseeuw 1990). This is a clustering technique that, given a predetermined number of clusters *k*, not only returns the membership of each element to a particular cluster but also a *medoid*, that is, a representative member of each cluster. This medoid is defined as the member that is most similar to all the other members of its cluster, additionally considering that every point is closer to its medoid than to any of the medoids of the other clusters. This technique allows us to explore the variation across hundreds of models by only inspecting the most representative ones, each of which stands for a number of very similar models.

3.5 Overview of implemented settings

To conclude this chapter, we will provide a short overview of all the parameter settings implemented in the different case studies included in this volume. For the sake of clarity, they are grouped in three categories: first, bag-of-word parameter settings that affect the selection of first-order context words; second, parameters based on association strength; and third, second-order parameter settings. First-order dependency-based parameters, as described above, are only implemented in the studies of Chapter 5. The most relevant settings are the distinction between REL models and PATH models: the former are those that select context words based on specific syntactic relationships to the target, whereas the latter filter them based on the distance in terms of syntactic paths. In the plots of Section 5.1, they are contrasted with bag-of-words models (BOW). For a more detailed description of the types of dependency models, the reader is directed to Chapter 2 of Montes (2021a). Table 3.4 gives an overview of the settings and their distribution over the chapters. It also specifies the code that will be the basis for a shorthand characterization of settings and models that we will occasionally use, in particular in the captions of figures or tables. Such a shorthand name for models will take the form of a straightforward concatenation of relevant codes, with dots separating the individual codes. Because the shorthand names are primarily used to distinguish models within the separate case studies of Chapters 5, 6, 9, and 10, settings that are constant within one chapter do not receive a code in Table 3.4. Note that the studies in Chapter 6 are based in different parameter spaces; here they will be referred to by the sections in which they are discussed, namely Section 6.3 and 6.4.

The first group of parameter settings includes the window size, frequency, and part-of-speech filters involved in the selection of first-order features, and the option to exclude items occurring outside the sentence of the target token. The first parameter setting, bag-of-words window size, can be implemented on any corpus and is thus relevant in all case studies. Chapter 5 uses models with symmetric windows of 3, 5, and 10 tokens to either side; Chapter 6 focuses on a model with 10 tokens to either side; Chapters 9 and 10, on the other hand, use models with windows of 5, 10, and 15 tokens to either side of the target. The part-of-speech filter is only applicable when the corpus includes such information, which renders this setting irrelevant in Section 6.4 and Chapter 9. In the rest of the chapters and in Section 6.3, instead, a distinction is made between models with no part-of-speech filter and those where first-order features were selected based on their part-of-speech: only common nouns, adjectives, and verbs for NAV models; common nouns, adjectives, verbs, and adverbs for LEX models; and common and proper nouns, adjectives, verbs, adverbs, and prepositions for NAV-NAP models. Chapters 5 and 10 use LEX models. While a window size filter selects items based on the occurrence in each token, the part-of-speech filter relies on a property of the context words. Another such property is frequency. (It does not receive a code in Table 3.4 because it is always kept constant within each case study.) Chapter 5 and the study in Section 6.3 only uses lemmas with a minimum relative frequency

Table 3.4 Overview of parameter dimensions and values

DIMENSION	VALUE	CODE	5	6	9	10
First-order window size	three tokens on either side	3-3	x	-	-	-
	five tokens on either side	5-5	x	-	x	x
	ten tokens on either side	10-10	x	x	x	x
	fifteen tokens on either side	15-15	-	x	x	x
First-order part-of-speech filter	common nouns, adjectives, verbs	NAV	x	-	-	x
	common nouns, adjectives, verbs, adverbs	LEX	x	-	-	x
	nouns (common and proper), adjectives, verbs, adverbs, prepositions	NAV-NAP	-	x	-	x
	no filter	ALL	x	x	x	x
First-order frequency filter	word forms with frequency > 10		-	x	-	-
	lemmas with frequency > 3		-	-	x	-
	lemmas with frequency > 5		-	-	-	x
	lemmas with relative frequency > 1 in 2M		x	x	-	-
First-order sentence boundary filter	only words in same sentence as target	BOUND	x	x	-	-
	no sentence boundary filter	NOBOUND	x	x	x	x
Dependency-based filter	based on specific syntactic patterns	REL	x	-	-	-
	based on distances in terms of syntactic paths	PATH	x	-	-	-
Association measure	positive pointwise mutual information	PPMI	x	x	x	x
	log-likelihood ratio	LLR	-	-	x	-
Association measure threshold	pmi > 0		x	x	x	x
	pmi >2		-	x	-	-
	llr > 1		-	-	x	-
Association measure filter	no use of filtering	ASSOCNO	x	-	x	x
	filtering context words below threshold	SELECTION	x		x	x
	weighting context words by association strength	WEIGHT	x	x	x	x
Second-order window size	four tokens either side		x	x	x	x
Second-order item	word forms		-	x	x	-
	lemmatized forms		x	x	-	x

DIMENSION	VALUE	CODE	5	6	9	10
Second-order part-of-speech filter	nouns, adjectives, verbs	SOCNAV	x	x	-	-
	no filter	SOCALL	x	x	x	x
Second-order frequency filter	excluding the 100 most frequent word forms		-	x	-	-
	excluding word forms with frequency < 400		-	-	x	-
Number of second-order features	union of first-order context words	FOC	x	-	x	x
	5000 most frequent items	5000	x	x	-	x
	all items with frequency above 400	MIN400	-	-	x	-
	200 dimensions as returned by SVD	SVD	-	-	x	-

of 1 in 2 million, Section 6.4 word forms with an absolute frequency larger than 10; Chapter 9 lemmas with a minimum co-occurrence frequency of 3 with the target type; and Chapter 10, lemmas with an absolute frequency larger than 5. These frequency thresholds are typically also implemented for the second-order features (with the exception of Chapter 9, see below). Finally, the sentence boundary parameter setting refers to the possibility to exclude context words that occur outside of the sentence boundaries. This can only be implemented on corpora with sentence delimiters, so only Chapter 5 and 6 make a difference between BOUND models (where those context words are excluded) and NOBOUND models (in which sentence delimiters are ignored). By default, all models in the rest of the chapters are NOBOUND.

The second category of parameter settings refers to the use of association measures between context features and target types for the purposes of filtering and/or weighting. They also affect the selection of first-order features, but they are relevant for both bag-of-words models and dependency-based models. Moreover, they can also affect the representation of first-order features. A first aspect to consider is which association measure will be used. As mentioned above, all chapters rely on ppmi, but Chapter 9 also uses models that rely on log-likelihood ratio instead. The second aspect is how the threshold is defined, which is kept constant in each study. Chapters 5, 9, and 10 and Section 6.3 only require pmi to be positive, whereas Section 6.4 sets a higher threshold, excluding word forms with a pmi lower than 2. When the association measure is log-likelihood ratio in Chapter 9, context words must have a log-likelihood ratio larger than 1 to be considered. Finally, an important parameter setting concerns how the association strength is used: whether it only selects context words (SELECTION), weights their contribution to the token-level vector (WEIGHT) or is ignored (ASSOCNO).

The third category of parameter settings concerns the selection and definition of second-order features. The symmetric window size of four tokens to either side, relevant to the computation of the ppmi values of the type-level vectors, is kept constant across all case studies. In addition, part-of-speech filters can be applied to the second-order dimensions. Chapter 5 uses both models with NAV filter and with no filter, and the rest of the chapters apply no part-of-speech filter to these features. Notice that in Chapters 5 and 10 and Section 6.3 the second-order features are lemmas, whereas in Section 6.4 and Chapter 9 they are word forms. Furthermore, Section 6.4 applies an additional frequency filter, excluding the 100 most frequent word forms, while Chapter 9 only includes word forms with a frequency higher than 400. The other relevant second-order parameter setting is the length of the vector, which can be defined in different ways. Chapters 5 and 10 use two values for this parameter: FOC and 5000. As described above, FOC refers to taking the final list of selected first-order context words as second-order features, while the alternative selects the 5000 most frequent types in the corpus (after applying other filters, if relevant). Chapter 6 only uses the latter value. In contrast, Chapter 9 uses three different values for the length of the second-order vector: FOC, MIN400, and SVD. In the case of FOC, it takes the list of selected first-order context words in the model and applies a second frequency threshold, only keeping the context words that occur at least three times in the context of the target type in the corpus. The final list is used as second order-dimensions. When the length of the second-order vector is MIN400, instead, all the types with a minimum absolute frequency of 400 are included as second-order dimensions, with no further filter applied. Finally, the SVD value refers to models in which singular value decomposition was applied to the type-level matrix, which has as rows all the types that co-occur with any concept in a symmetric window of 15 tokens to either side, and as columns the types with a frequency of at least 400. The SVD models then have only 200 dimensions, as explained in the previous section.

The bottom line

- Token-level vectors are built by combining the type-level vectors of context features in their immediate environment.
- The construction of a token-level model requires a number of choices; different decisions result in different models.

- Some of the decisions involved in distributional modelling are linguistically informed; for instance, window spans, use of syntactic information, and so on.
- Other decisions are mathematically informed, based on the properties of vectors, matrices, distance metrics, and so on.

4

Visual analytics for token-based distributional semantics

The modelling workflow described in Chapter 3 produces token-level distance matrices: one matrix per model, each indicating the pairwise dissimilarity between the occurrences of a certain word in a sample, according to that model. However, because of the large number of tokens in the sample and the diversity of models produced by multiple parameters, such output is challenging to interpret. In this chapter we will describe the steps followed to process the distance matrices and obtain a more manageable format, as well as a visual analytics tool designed to explore the results. Section 4.1 will introduce two dimensionality reduction algorithms that map the distances to a 2D-space, so that each token can be represented as a point in a scatterplot. Section 4.2 will address the issue of multiple models and suggest a clustering algorithm as a way of selecting representative models. Afterwards, Section 4.3 will describe a visual analytics tool, NephoVis, originally developed by Thomas Wielfaert (Wielfaert, Heylen, Speelman, and Geeraerts 2019), and then continued by Mariana Montes and Anthe Sevenants (Sevenants, Montes, and Wielfaert 2022). Finally, Section 4.4 will introduce a ShinyApp (Chang, Cheng, Allaire, Sievert, Schloerke, Xie, Allen, McPherson, Dipert, and Borges 2022) that expands the functionalities of the NephoVis tool. For readers who do not intend to explore or apply the tools described in Section 4.3 and 4.4, these parts of the chapter will be of secondary importance. Section 4.1 and 4.2, in contrast, introduce a number of notions that are relevant for the way in which the models discussed in Chapter 5 were analysed (via the NephoVis tool) and how the token-level plots in Chapters 6, 9, and 10 can be interpreted.

4.1 Dimensionality reduction for visualization

We can mentally picture or even draw positions, vectors, and angles in up to three dimensions, but distributional models have hundreds if not thousands of dimensions. Dimensionality reduction algorithms try to reduce the number of dimensions of a high-dimensional entity while retaining as much information as possible. We already surveyed some of these algorithms in Chapter 3, where we discussed ways of condensing vectors. In this section, instead, we will focus on

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0004

methods that try to project the distances between items in the multidimensional space to Euclidean distances in a low-dimensional space that we can visualize. The different implementations could take the token-by-feature matrix as input, but as they will not typically compute cosine distances between the items, we provide the distance matrix as input instead. The literature tends to go for either multidimensional scaling (MDS) or t-stochastic neighbour embeddings (t-SNE).

The first option, multidimensional scaling, is an ordination technique, like principal components analysis (PCA). It has been used for decades in multiple areas (see Cox and Cox 2008); its non-metric application was developed by Kruskal (1964). It tries out different low-dimensional configurations aiming to maximize the correlation between the pairwise distances in the high-dimensional space and those in the low-dimensional space: items that are close together in one space should stay close together in the other, and items that are far apart in one space should stay far apart in the other. The output from multidimensional scaling can be evaluated by means of the stress level, that is, the complement of the correlation coefficient: the smaller the stress, the better the correlation between the original distances and the reduced-dimensionality distances. Unlike principal components analysis, however, the new dimensions are not meaningful per se; two different runs of multidimensional scaling may result in plots that mirror each other while representing the same thing. Nonetheless, the R implementation vegan: :metaMDS() (Oksanen, Simpson, Blanchet, et al. 2022) rotates the plot so that the horizontal axis represents the maximum variation. In the cognitive linguistic literature both metric (Koptjevskaja-Tamm and Sahlgren 2014; Hilpert and Correia Saavedra 2017; Hilpert and Flach 2020) and non-metric multidimensional scaling (Heylen, Speelman, and Geeraerts 2012; Heylen, Wielfaert, Speelman, and Geeraerts 2015; Perek 2016; De Pascale 2019) have been used.

The second technique, t-SNE (van der Maaten and Hinton 2008; van der Maaten 2014), has also been incorporated in cognitive distributional semantics (Perek 2018; De Pascale 2019). It is also popular in computational linguistics (Smilkov, Thorat, Nicholson, Reif, Viégas, and Wattenberg 2016; Jurafsky and Martin 2023); in R, it can be implemented with Rtsne: :Rtsne() (Krijthe 2018). The algorithm is quite different from multidimensional scaling: it transforms distances into probability distributions and relies on different functions to approximate them. Moreover, it prioritizes preserving local similarity structure rather than the global structure: items that are close together in the high-dimensional space should stay close together in the low-dimensional space, but those that are far apart in the high-dimensional space may be even farther apart in low-dimensional space. Compared to multidimensional scaling, we obtain nicer, tighter clusters of tokens (see Figure 4.1), but the distance between them is less interpretable: even if we trust that tokens that are very close to each other are also similar to each other in the high-dimensional space, we cannot extract meaningful information from the distance *between* these groups. In addition, it would seem that points that are far away in a high-dimensional space might show up close together in the low-dimensional space (Oskolkov 2021). In contrast, uniform manifold approximation and projection, or UMAP (McInnes, Healy, and Melville 2020; Konopka 2022), penalizes this sort of discrepancy. This visualization technique is included in Figure 4.1 for comparison, but has not been used in the case studies in this book.

Unlike multidimensional scaling, t-SNE requires setting a parameter called *perplexity*, which roughly indicates how many neighbours the preserved local structure should cover. Low values of perplexity lead to numerous small groups of items, while higher values of perplexity return more uniform, round configurations (Wattenberg, Viégas, and Johnson 2016). Unless specified otherwise, the token-level plots included in this volume correspond to the default values of the R implementation, which has proved to be the most stable and meaningful in our datasets.

For both multidimensional scaling and t-SNE we need to state the desired number of dimensions before running the algorithm. For visualization purposes, the

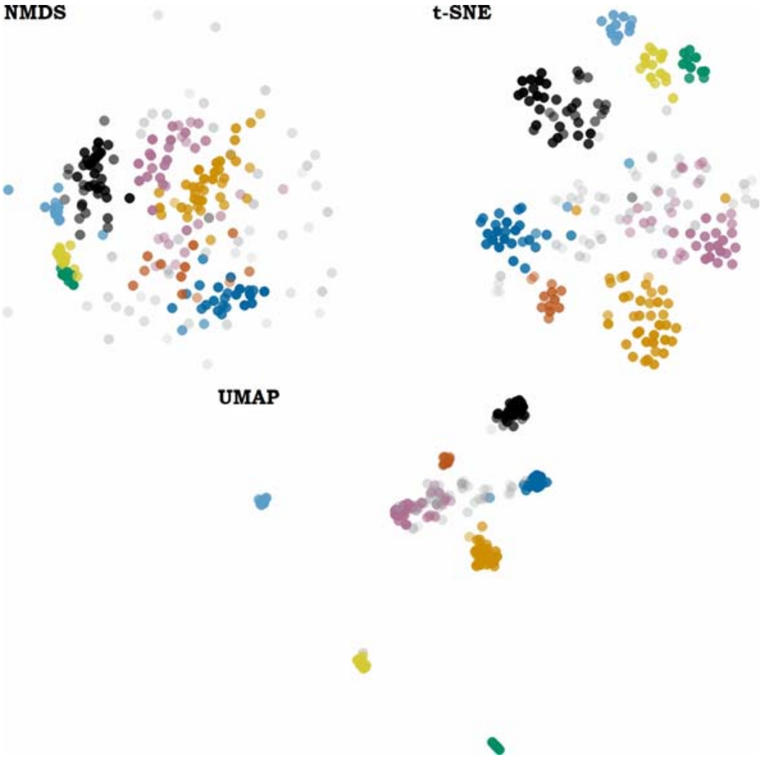


Figure 4.1 Two 2D representations of the same model of Dutch *hachelijk* 'dangerous/critical'. Non-metric MDS to the left, t-SNE in the centre, and UMAP to the right. Colours indicate HDBSCAN clusters

most useful choice is two; three dimensions are difficult to interpret if projected on a 2D space, such as a screen or paper (Card, Mackinlay, and Shneiderman 1999; Wielfaert, Heylen, Speelman, and Geeraerts 2019). For UMAP, instead, we can simply select the first two dimensions of the output. As we mentioned before, the dimensions themselves are meaningless: there is nothing to be interpreted in the actual values of each dimension, but only in the Euclidean distances between the plotted points. Hence, no axes or axis tick marks will be included in the plots. However, the scales of both co-ordinates are kept fixed: given three points $a = (1, 1.5)$, $b = (1, 0.5)$ and $c = (0, 1.5)$, the distance between a and b (one unit along the x-axis) will be the same as the distance between a and c (one unit along the y-axis).

4.2 Selecting representative models

The combination of the multiple variable parameters discussed in Chapter 3 return tens or hundreds of models. In order to explore and understand their diversity, we can compute similarities or distances between them. A distance matrix of models opens up further processing techniques: like with tokens, we can represent the similarities in 2D via multidimensional scaling or t-SNE, and we can apply clustering in order to identify groups of similar models. In this section we will describe the distance measure used to represent dissimilarities between models and a clustering algorithm that also identifies representative models.

While cosine distances are used to measure the similarity between token-level vectors, Euclidean distances will be used to compare two vectors of the same token across models, and thus compare models to each other. Concretely, let's say we have two matrices, A and B, which are two models of the same sample of tokens, built with different parameter settings, and we want to know how similar they are to each other, that is, how much of a difference those parameter settings make. Their values are cosine distances transformed according to the procedure described in Section 3.4. A given token i has a vector $\mathrm{a_i}$ in matrix A and a vector $\mathrm{b_i}$ in matrix B. For example, i could be example (4.1), and its vector in A is based on the co-occurrence with *computational* and *linguistics*, while its vector in B is only based on *computational*.

(4.1) They **study** this in computational linguistics as well.

The Euclidean distance between $\mathrm{a_i}$ and $\mathrm{b_i}$ is computed with the following formula in (4.2).

(4.2) *Euclidean distance*

$$(\mathrm{a}_i, \mathrm{b}_i) = \sqrt{\sum_{i=j}^{n} \left(a_j - b_j\right)^2}$$

After running the same comparison for each of the tokens, the distance between A and B is then computed as the mean of those tokenwise distances across all the tokens modelled by both A and B. Alternatively, the distances between models could come from Procrustes analysis like Wielfaert, Heylen, Speelman, and Geeraerts (2019) do, which has the advantage of returning a value between 0 and 1. However, the method described here is much faster and returns comparable results.

The resulting distance matrix can be mapped to two dimensions via multidimensional scaling, resulting in the Level 1 plots discussed in the following section. Additionally, we can apply Partition Around Medoids (Kaufman and Rousseeuw 1990), implemented in R with cluster: :pam() (Maechler, Rousseeuw, Struyf, and Hubert 2022), in order to select representative models, called *medoids*. Unlike HDBSCAN and other clustering algorithms, it requires us to set a number of clusters beforehand, and then tries to find the organization that maximizes internal similarity within the cluster and distances between clusters. For our purposes, we have settled for eight medoids for each lemma. The number is not meant to achieve the best clustering solutions—no number could be applied to all the cases with equal success, given the variability in the differences between the models. The goal, instead, is to have a set of models that is small enough to visualize simultaneously (on a screen, in reasonable size) and big enough to cover the variation across models. For some lemmas, there might not be that much variation, and the medoids might be redundant with each other. However, as long as we can cover (most of) the visible variation across models and the medoids are reasonably good representatives of the models in their corresponding clusters, the method is fulfilling its goal.

Although this is a clustering algorithm, we will avoid referring to the clusters of models as 'clusters', in order to avoid confusion with the clusters of tokens that are our main focus. The preferred phrase will be 'the models represented by the medoid'. Given that the clustering algorithms used on the tokens are HDBSCAN and hierarchical clustering, *medoid* will always refer to a representative model.

4.3 The NephoVis visualization tool

The visualization tool described here, *NephoVis*, was written in Javascript, making heavy use of the D3.js library, which was designed for web-based data-driven visualization (Bostock, Ogievetsky, and Heer 2011). The D3 library allows the designer to link elements on the page, such as circles in a graphic element of the webpage, dropdown buttons and titles, to data structures such as arrays and data frames, and manipulate the visual elements based on the values of the linked data items. In addition, it offers handy functions for scaling and mapping, that is, to fit the relatively arbitrary ranges of the coordinates to pixels on a screen, or to map a colour palette to a set of categorical values. While D3 offers a variety of useful colour palettes, the visualization currently relies on a (slightly adapted) colour-blind-friendly scale by Okabe and Ito (2002).

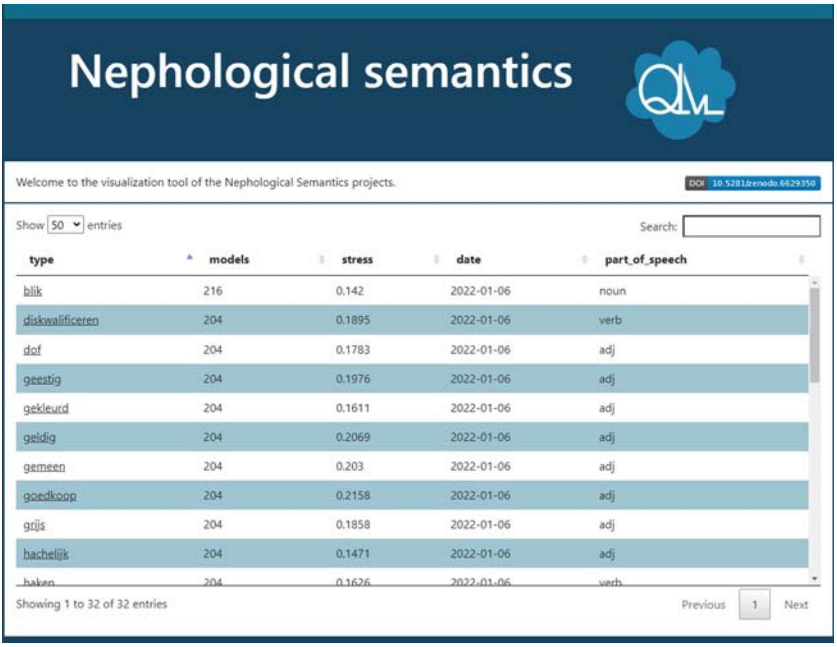
Section 3.4 discussed a procedure to measure the distance between tokens, Section 4.1 introduced the dimensionality reductions that can be applied to the resulting distance matrices, and Section 4.2 presented the technique used to measure the distance between models and select representative models, or medoids. Via these procedures and some additional processing, we can have access to the following datasets for each of the lemmas (in semasiological studies) or concepts (in onomasiological studies):

- a distance matrix between models;
- a data frame with one row per model, the 2D coordinates based on the distance matrix, and columns coding the different variable parameters or other pieces of useful information, such as the number of modelled tokens;
- a data frame with one row per token, 2D coordinates for each of their models and other information such as sense annotation, register, selection of context words, and concordance line;
- a data frame with one row per first-order context word and useful frequency information.

The R package semcloud (Montes 2021c) provides functionalities to generate these datasets based on the output of the nephosem workflow (QLVL 2021). In practice, the data frame for the tokens is split in multiple data frames with coordinates corresponding to different dimensionality reduction algorithms, such as multidimensional scaling, t-SNE with different perplexity values and uniform manifold approximation and projection (UMAP), and another data frame for the variables that are common to all models. In addition, one of the features of the visualization tool includes the possibility to compare an individual token-level model with the representation of the type-level modelling of its first-order context words. This feature is also part of the ShinyApp extension presented in Section 4.4. Crucially, the visualization tool works both for count-based models and prediction-based models. The main difference in terms of application is the way that context words are captured and mapped for certain features. On the one hand, the Python library NephoNeural (https://github.com/AntheSevenants/NephoNeural) can be used to generate NephoVis-compatible datasets from prediction-based models. On the other hand, missing data such as absent context word frequencies are dealt with by the tool seamlessly, simply deactivating the features that require them. (See also the *Software resources* section for an overview of the tools developed for the research presented in this monograph.)

In order to facilitate the exploration of all this information, NephoVis is organized in three levels, following Shneiderman's Visual Information Seeking Mantra: 'Overview first, zoom and filter, then details-on-demand' (1996: 97). The core of the tool is the interactive, zoomable scatterplot, but its goal and functionality are adapted to each of the three levels. In Level 1 the scatterplot represents the full set of models and allows the user to explore the quantitative effect of different parameter settings and to select a small number of models for detailed exploration in Level 2. This second level shows multiple token-level scatterplots—one for each of the selected models—and therefore offers the possibility to compare the shape and organization of the groups of tokens across different models. By selecting one of these models, the user can examine it in Level 3, which focuses on only one at a time. Shneiderman's (1996) mantra underlies both the flow across levels and the features within them: each level is a zoomed in, filtered version of the level before it; the individual plots in Levels 1 and 3 are literally zoomable; and in all cases it is possible to select items for more detailed inspection. Finally, a number of features—tooltips and pop-up tables—show details on demand, such as the names of the models in Level 1 and the context of the tokens in the other two levels.

As of August 2022, https://qlvl.github.io/NephoVis/ hosts the portal shown in Figure 4.2, which displays a list of lemmas for which there are visualizations available. The names of each lemma are hyperlinks to their respective Level 1



Nephological semantics

Welcome to the visualization tool of the Nephological Semantics projects.

DOI 10.5281/zenodo.6629350

Show 50 entries

Search:

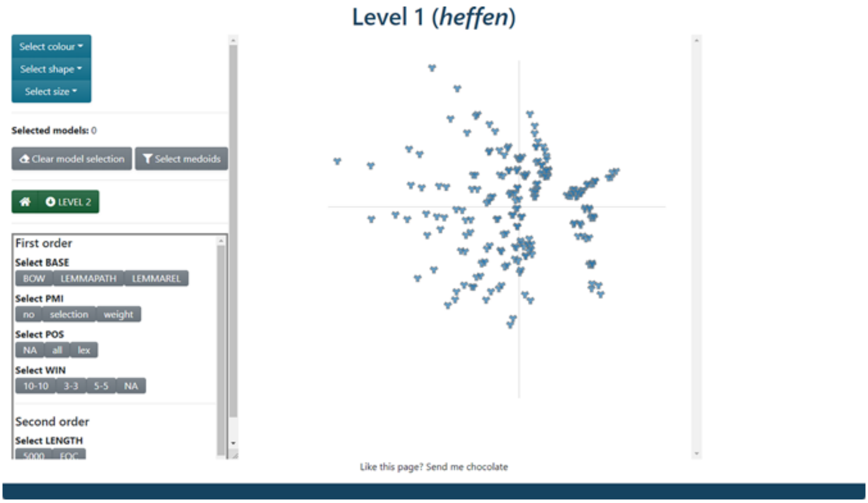
type	models	stress	date	part_of_speech
blik	216	0.142	2022-01-06	noun
diskwalificeren	204	0.1895	2022-01-06	verb
dof	204	0.1783	2022-01-06	adj
geestig	204	0.1976	2022-01-06	adj
gekleurd	204	0.1611	2022-01-06	adj
geldig	204	0.2069	2022-01-06	adj
gemeen	204	0.203	2022-01-06	adj
goedkoop	204	0.2158	2022-01-06	adj
grijs	204	0.1858	2022-01-06	adj
hachelijk	204	0.1471	2022-01-06	adj
haken	204	0.1626	2022-01-06	verb

Showing 1 to 32 of 32 entries

Figure 4.2 Portal of https://qlvl.github.io/NephoVis/ as of August 2022

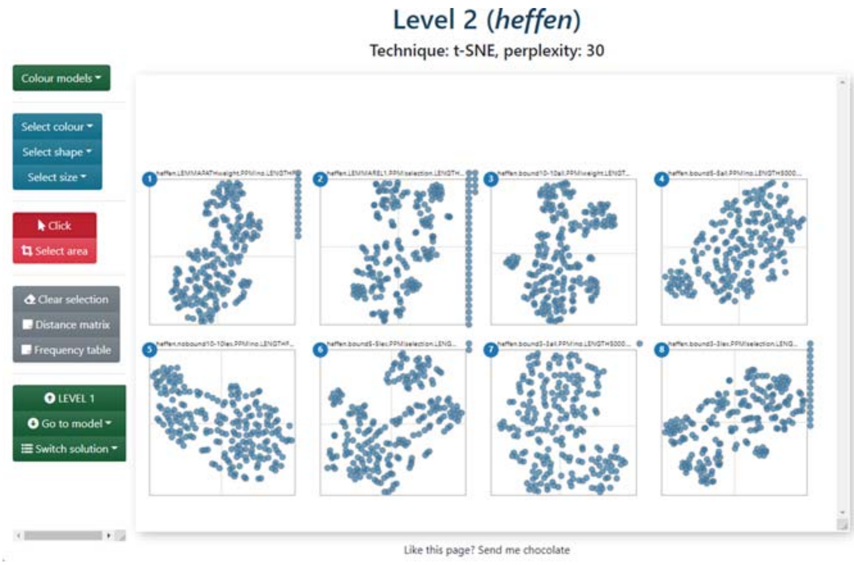
pages, shown in Figure 4.3. By exploring the scatterplot of models, the user can look for structure in the distribution of the parameters on the plot. For example, colour coding may reveal that models with nouns, adjectives, verbs, and adverbs as first-order context words are very different from those without strong filters for part-of-speech, because mapping these values to colours reveals distinct groups in the plot. In contrast, mapping the sentence boundaries restriction might result in a mix of dots of different colours with no obvious organization, meaning that the parameter makes little difference: models that only differ along the sentence boundary parameter are very similar to each other. Depending on whether the user wants to compare models similar or different to each other, or which parameters they would like to keep fixed, they will use individual selection or the buttons to choose models for Level 2. The *Select medoids* button quickly identifies the predefined medoids. By clicking on the *Level 2* button, Level 2 is opened in a new tab, as shown in Figure 4.4.

In Level 2, the user can already compare the shapes that the models take in their respective plots, the distribution of categories like sense labels, and the number of lost tokens. If multiple dimensionality reduction techniques have been used, the *Switch solution* button allows the user to select one and watch the models readjust to the new coordinates in a short animation. In addition, the *Distance matrix* button offers a heatmap of the pairwise distances between the selected models. Either by clicking on the name of a model or through the *Go to model* dropdown menu, the user can access Level 3 and explore the scatterplot of an individual model. As we will see below, Level 2 and Level 3, both built around token-level scatterplots, share several features. The difference lies in the possibility



Selected models: 0

Figure 4.3 Level 1 for *heffen* ‘to levy/to lift’



Level 2 (*heffen*)

Technique: t-SNE, perplexity: 30

Colour models
Select colour
Select shape
Select size
Click
Select area
Clear selection
Distance matrix
Frequency table
LEVEL 1
Go to model
Switch solution
1
heffen.LEMMAPATHweight.PPMIno.LENGTHFO
2
heffen.LEMMAREL1.PPMIselection.LENGTH...
3
heffen.bound10-10all.PPMIweight.LENGT...
4
heffen.bound5-5all.PPMIno.LENGTH5000...
5
heffen.nobound10-10lex.PPMIno.LENGTHF...
6
heffen.bound5-5lex.PPMIselection.LENG...
7
heffen.bound3-3all.PPMIno.LENGTH5000...
8
heffen.bound3-3lex.PPMIselection.LENG...

Like this page? Send me chocolate

Figure 4.4 Level 2 for the medoids of *heffen* ‘to levy/to lift’

of examining model-specific information, such as reading annotated concordance lines which highlight information captured by the model or selecting tokens based on the words that co-occur with it. In practice, the user would switch back and forth between Level 2 and Level 3: between examining single models and comparing many of them.

Before going into the detailed description of each level, a note is in order. As already mentioned in Section 4.1, the values of the dimensions of the plots are not meaningful: all that matters is the distances between the points. In consequence, there are no axes or axis ticks in the plots. However, the units are kept constant within each plot: one unit on the x-axis has the same length in pixels as one unit on a y-axis within the same scatterplot—this equality is not maintained across plots.

The main element of Level 1 is an interactive zoomable scatterplot where each glyph, by default a steel blue wye (the Y-like sign), represents one model. This scatterplot aims to represent the similarity between models as coded by the multidimensional scaling output and allows the user to select the models to inspect according to different criteria. Categorical variables (for instance, whether sentence boundaries are used) can be mapped to colours and shapes, as shown in Figure 4.5, and numerical variables (such as number of tokens in the model) can be mapped to size.

A selection of buttons on the left panel, as well as the legends for colour and shape, can be used to filter models with a certain parameter setting. These options

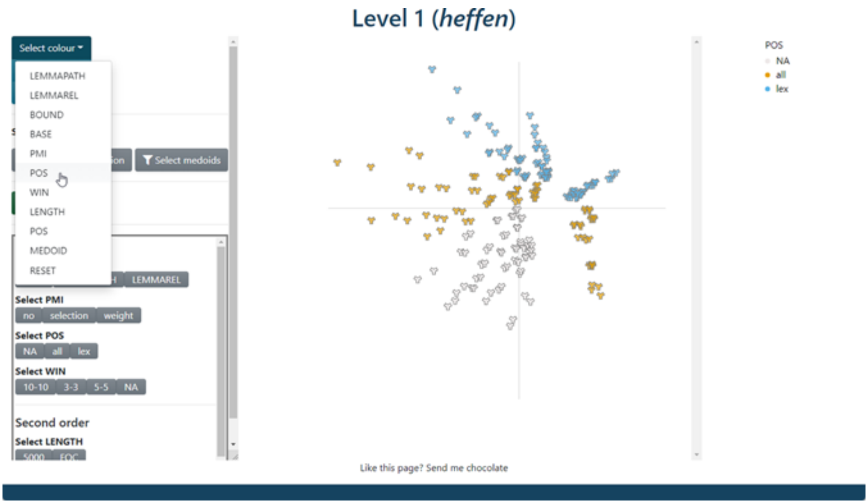


Figure 4.5 Level 1 for *heffen* 'to levy/to lift'. The plot is colour-coded with first-order part-of-speech settings; NA stands for missing data, in this case the dependency-based models

are generated automatically by reading the columns in the data frame of models and interpreting column names starting with *foc_* as representing first-order parameter settings, and those starting with *soc_* as second-order parameter settings. Different settings of the same parameter interact with an OR relationship, since they are mutually exclusive, while settings of different parameters combine in an AND relationship. For example, by clicking on the grey *bound* and *lex* buttons on the bottom left, only bag-of-words models with part-of-speech filter and sentence boundary limits will be selected. By clicking on both *lex* and *all*, all bag-of-words models are selected, regardless of the part-of-speech filter, but dependency-based models (for which part-of-speech is not relevant) are excluded. A counter above keeps track of the number of selected items, since Level 2 only allows up to nine models for comparison. This procedure is meant to aid a selection based on relevant parameters, as a manual alternative to selection by medoids. In Figure 4.6, instead, the *Select medoids* button was used to quickly capture the medoids obtained from Partitioning Around Medoids. Models can also be manually selected by clicking on the glyphs that represent them.

Level 2 shows an array of small scatterplots, each of which represents a token-level model. The glyphs, by default steel blue circles, stand for individual tokens, that is, attestations of the chosen lemma in a given sample. The dropdown menus on the sidebar (see Figures 4.4 and 4.7) read the columns in the data frame of variables, which can include any sort of information for each of the tokens, such as sense annotation, sources, number, and list of context words in a model,

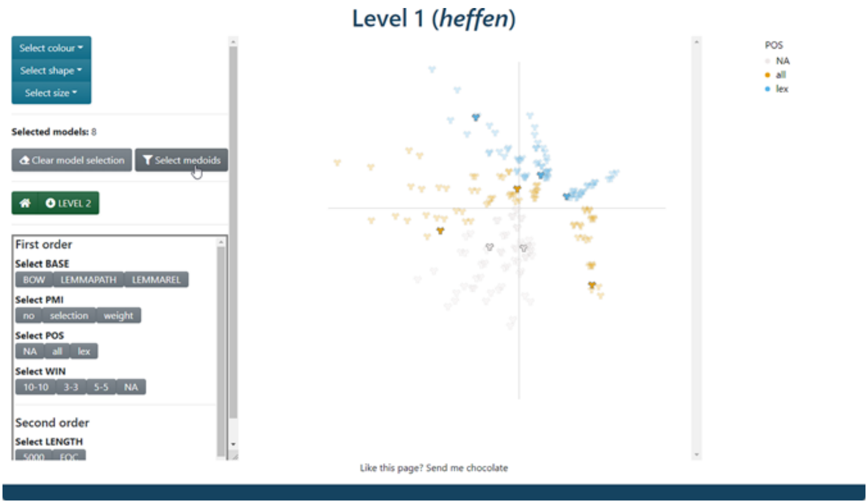
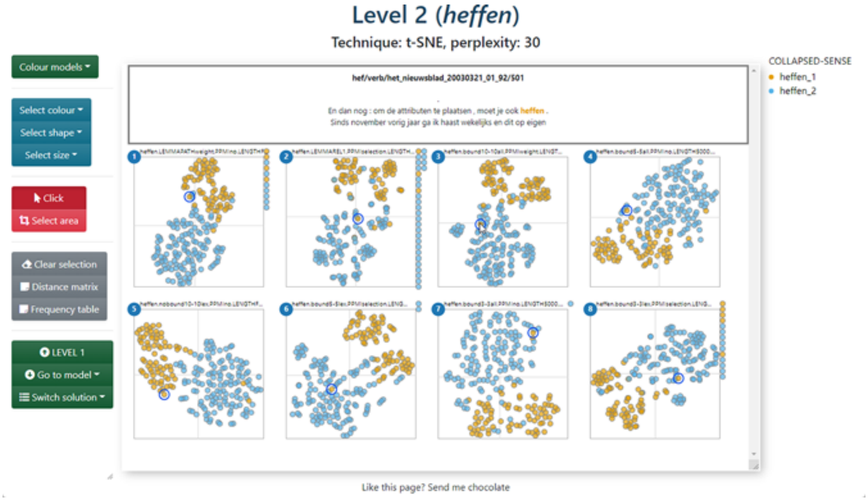


Figure 4.6 Level 1 for *heffen* 'to levy/to lift' with medoids highlighted


Level 2 (*heffen*)
Technique: t-SNE, perplexity: 30
Colour models
Select colour
Select shape
Select size
Click
Select area
Clear selection
Distance matrix
Frequency table
LEVEL 1
Go to model
Switch solution
COLLAPSED-SENSE
heffen_1
heffen_2
1 heffen.LEMMAPATHweight.PPMIno.LENGTHF...
2 heffen.LEMMAREL1.PPMIselection.LENGTH...
3 heffen.bound10-10all.PPMIweight.LENGT...
4 heffen.bound5-5all.PPMIno.LENGTH5000...
5 heffen.nobound10-10lex.PPMIno.LENGTHF...
6 heffen.bound5-5lex.PPMIselection.LENG...
7 heffen.bound3-3all.PPMIno.LENGTH5000...
8 heffen.bound3-3lex.PPMIselection.LENG...
Like this page? Send me chocolate

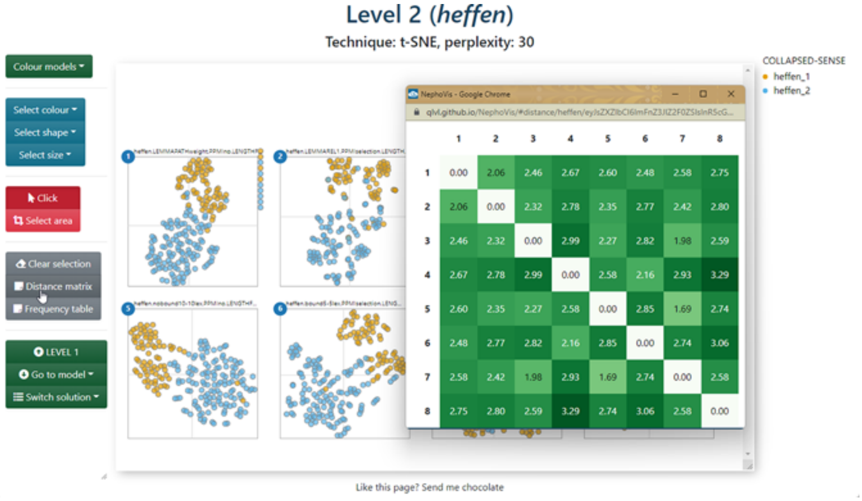
hef/verb/het_nieuwsblad_20030321_01_92/501

En dan nog : om de attributen te plaatsen , moet je ook heffen .
Sinds november vorig jaar ga ik haast wekelijks en dit op eigen

Figure 4.7 Level 2 for the medoids of *heffen* 'to levy/to lift', colour-coded with categories from manual annotation. Hovering over a token shows its concordance line

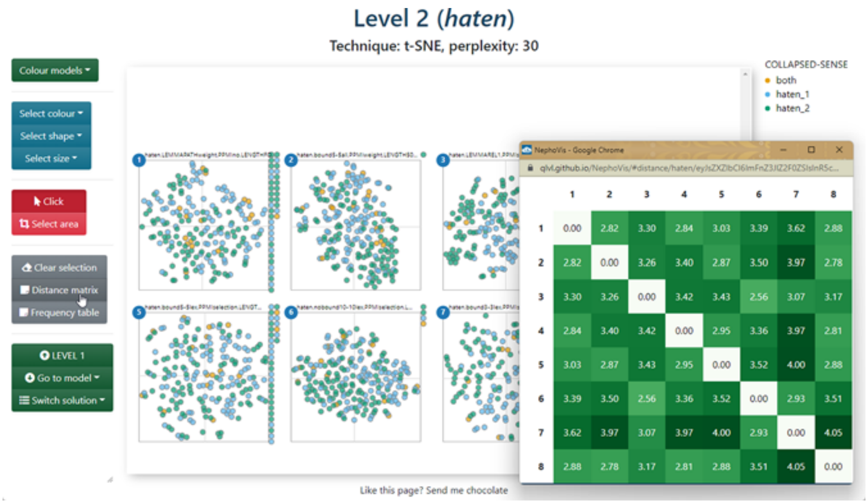
concordance lines, and so on. Categorical variables can be used for colour- and shape-coding, as shown in Figure 4.7, where the senses of the chosen lemma are mapped to colours; numerical variables, such as the number of context words selected by a given lemma, can be mapped to size. Note that the mapping will be applied equally to all the selected models: the current code does not allow for variables—other than the coordinates themselves—to adapt to the specific model in each scatterplot. That is the purview of Level 3.

Before further examining the scatterplots, a small note should be made about the distance matrix mentioned above. The heatmap corresponding to the medoids of *heffen* 'to levy/to lift' is shown in Figure 4.8. The multidimensional scaling representation in Level 1 tried to find patterns and keep the relative distances between the models as faithful to their original positions as possible, but such a transformation always loses information. Given a restricted selection of models, however, the actual distances can be examined and compared more easily. A heatmap maps the range of values to the intensity of the colours, making patterns of similar/different objects easier to identify. For example, Figure 4.8 shows that the fourth and sixth medoids are quite different from medoids seven and eight. Especially when the model selection followed a criterion based on strong parameter settings (for instance, keeping ppmi constant to look at the interaction between window size and part-of-speech filters), such a heatmap could reveal patterns that are slightly distorted by the dimensionality reduction in Level 1 and even hard to pinpoint from visually comparing the scatterplots. But even with the medoid selection, which aims to find representatives that are maximally different from each other (or that at least are the core elements of maximally different groups), the heatmap can show whether some medoids are drastically *more* different from or similar to others. As a reference, the heatmap is particularly useful to check hypotheses about the visual similarity of models. For example, unlike with *heffen* 'to levy/to lift' in Figure 4.7, if we colour-code the medoids of *haten*



	1	2	3	4	5	6	7	8
1	0.00	2.06	2.46	2.67	2.60	2.48	2.58	2.75
2	2.06	0.00	2.32	2.78	2.35	2.77	2.42	2.80
3	2.46	2.32	0.00	2.99	2.27	2.82	1.98	2.59
4	2.67	2.78	2.99	0.00	2.58	2.16	2.93	3.29
5	2.60	2.35	2.27	2.58	0.00	2.85	1.69	2.74
6	2.48	2.77	2.82	2.16	2.85	0.00	2.74	3.06
7	2.58	2.42	1.98	2.93	1.69	2.74	0.00	2.58
8	2.75	2.80	2.59	3.29	2.74	3.06	2.58	0.00

Figure 4.8 Heatmap of distances between medoids of *heffen* 'to levy/to lift' against the backdrop of Level 2

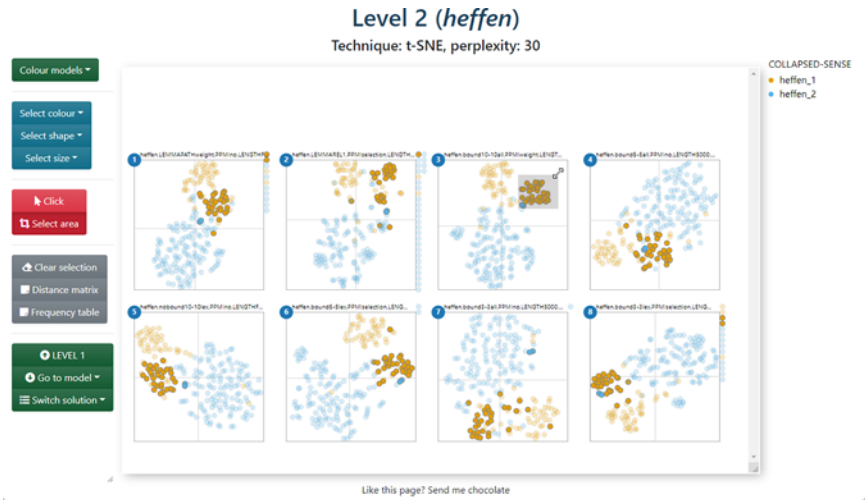


	1	2	3	4	5	6	7	8
1	0.00	2.82	3.30	2.84	3.03	3.39	3.62	2.88
2	2.82	0.00	3.26	3.40	2.87	3.50	3.97	2.78
3	3.30	3.26	0.00	3.42	3.43	2.56	3.07	3.17
4	2.84	3.40	3.42	0.00	2.95	3.36	3.97	2.81
5	3.03	2.87	3.43	2.95	0.00	3.52	4.00	2.88
6	3.39	3.50	2.56	3.36	3.52	0.00	2.93	3.51
7	3.62	3.97	3.07	3.97	4.00	2.93	0.00	4.05
8	2.88	2.78	3.17	2.81	2.88	3.51	4.05	0.00

Figure 4.9 Heatmap of distances between medoids of *haten* 'to hate' against the backdrop of Level 2

'to hate' with the manual annotation (Figure 4.9), all the models look equally messy. As we will see below, we can brush over sections of the plot to see if, at least, the tokens that are close together in one medoid are also close together in another. The heatmap of distances confirms that not all models are equally different from each other; for example, the seventh model is very different from all the others.

Next to the colour coding, Figure 4.7 also illustrates how hovering over a token indicates its location in other plots by surrounding their glyph with a circle and prints the corresponding identifier and concordance line. Figure 4.10, on the other hand, showcases the brush-and-link functionality. By brushing over a specific model, the tokens found in that area are highlighted and the rest are made more transparent. Such a functionality is also available in Level 3, but not in Level 1. Level 2 enhances the power of this feature by highlighting the same selection of tokens across the different models, whatever area they occupy. Thus, we can see whether tokens that are close together in one model are still close together in a different model, which is specially handy in more uniform plots, like the one for *haten* 'to hate' in Figure 4.9. Figure 4.10 reveals that the tokens selected in the third medoid are, indeed, quite well grouped in the rest of the medoids, with different degrees of compactness. It also highlights two glyphs on the right margin of the bottom right plot, where we find the tokens lost by that model due to lack of context words. In this case the tokens were lost by the eighth medoid, which has the most selective combination of parameter settings, so that no context words could be captured around those tokens.


Level 2 (*heffen*)

Technique: t-SNE, perplexity: 30

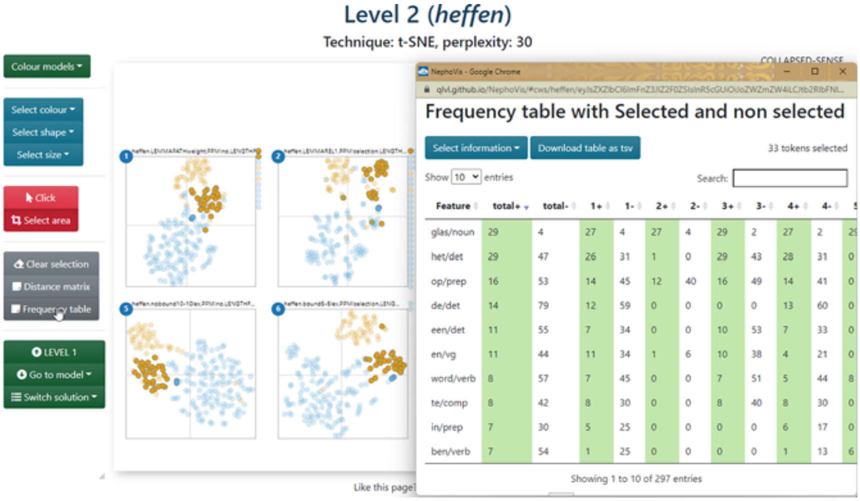
Colour models ▾
Select colour ▾
Select shape ▾
Select size ▾
Click
Select area
Clear selection
Distance matrix
Frequency table
LEVEL 1
Go to model ▾
Switch solution ▾
COLLAPSED-SENSE
● heffen_1
● heffen_2
1 heffen.LEMMAPATHweight.PPMIno.LENGTHF...
2 heffen.LEMMAREL1.PPMIselection.LENGTH...
3 heffen.bound10-10all.PPMIweight.LENGT...
4 heffen.bound5-5all.PPMIno.LENGTH5000...
5 heffen.nobound10-10lex.PPMIno.LENGTHF...
6 heffen.bound5-5lex.PPMIselection.LENG...
7 heffen.bound3-3all.PPMIno.LENGTH5000...
8 heffen.bound3-3lex.PPMIselection.LENG...
Like this page? Send me chocolate

Figure 4.10 Level 2 for the medoids of *heffen* 'to levy/to lift', colour-coded with categories from manual annotation. Brushing over an area in a plot selects the tokens in that area and their positions in other models

In any given model, we expect tokens to be close together because they share a context word, and/or because their context words are distributionally similar to each other: their type-level vectors are near neighbours. Therefore, when inspecting a model, we might want to know which context word(s) pull certain tokens together, or why tokens that we expect to be together are far apart instead. For individual models, this can be best achieved via the ShinyApp described in Section 4.4, but NephoVis also includes features to explore the effect of context words, such as frequency tables.

In Level 2, while comparing different models, the frequency table has one row per context word and one or two columns per selected model (such as the medoids). Such a table is shown in Figure 4.11. The columns in this table are all computed by NephoVis itself based on lists of context words per token per model. Next to the column with the name of the context word, the default table shows two columns called *total* and two columns per model, each headed by the corresponding number (seen next to the model name in the small scatterplot) and either a plus or a minus sign. The plus (+) columns indicate how many *of the selected tokens* in that model co-occur with the word in the row; the minus (-) columns indicate the number of non-selected tokens that co-occur with the word. The *total* columns indicate, respectively, the number of selected or non-selected tokens for which that context word was captured by at least one model. Here it is crucial to understand that, when it comes to distributional modelling, a context word is not simply any word that can be found in the concordance line of the token, but an item captured by the parameter settings of a given model. Therefore, a word can be a context word in a model but be excluded by a different model with stricter filters or different criteria.

For example, the screenshot in Figure 4.11 gives us a glimpse of the frequency table corresponding to the tokens selected already in Figure 4.10. On the top right corner, we are informed that the selection contains 33 tokens. The first row of the table indicates that the most frequent context word of these tokens is the noun *glas* 'glass', which is used in expressions such as *een glas heffen op iemand* 'to toast for someone, lit. to lift a glass on someone'. It co-occurs with at most 29 of the selected tokens and four non-selected tokens. Concretely, the third medoid captures *glas* in the 29 tokens, but only in two of the tokens outside of the selection; in contrast, the first and second medoid capture *glas* only in 27 of the captured tokens, but in the four non-selected ones. According to the names of the models, the third medoid is a bag-of-words model with a large window span and only ppmi weighting, whereas the other two models are dependency-based and have no ppmi weighting. The fourth medoid, like medoids 1 and 2, captures *glas* in only 27 of the selected tokens—and it does not use ppmi filters. However, like medoid 3, it only captures *glas* in two of the non-selected tokens—and it is a bag-of-words model. In other words, a large number of tokens co-occurring with *glas* are brought together by different models, but a few of them depend on the parameter settings. A context word that is far in the text but syntactically related will only be captured by syntactic models, whereas a context word closer in the text but farther in the syntactic



Frequency table with Selected and non selected

33 tokens selected

Feature	total+	total-	1+	1-	2+	2-	3+	3-	4+	4-	5
glas/noun	29	4	27	4	27	4	29	2	27	2	29
het/det	29	47	26	31	1	0	29	43	28	31	0
op/prep	16	53	14	45	12	40	16	49	14	41	0
de/det	14	79	12	59	0	0	0	0	13	60	0
een/det	11	55	7	34	0	0	10	53	7	33	0
en/vg	11	44	11	34	1	6	10	38	4	21	0
word/verb	8	57	7	45	0	0	7	51	5	44	8
te/comp	8	42	8	30	0	0	8	40	8	30	0
in/prep	7	30	5	25	0	0	0	0	6	17	0
ben/verb	7	54	1	25	0	0	0	0	1	13	6

Showing 1 to 10 of 297 entries

Figure 4.11 Frequency table of context words of selected tokens against the backdrop of Level 2 (medoids of *heffen* 'to levy/to lift')

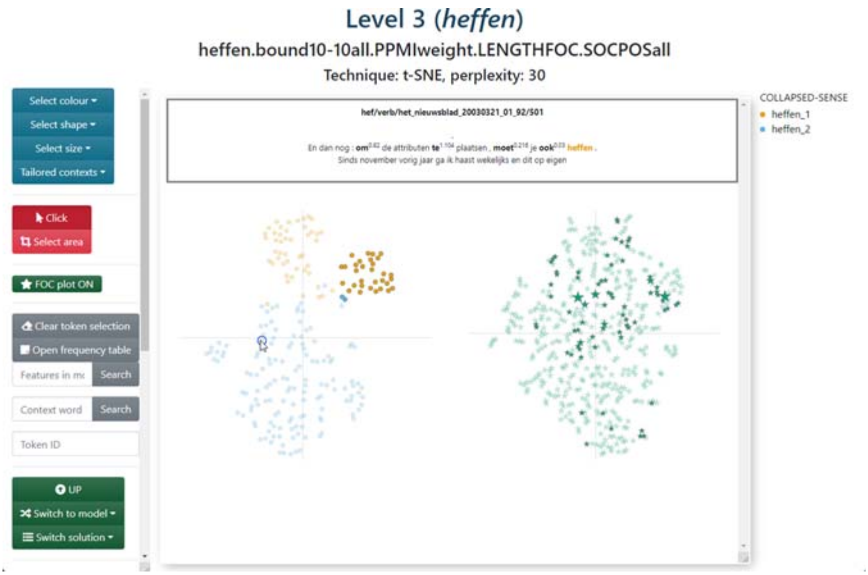
tree may require a bag-of-words model. This useful frequency information is available for all the context words that are captured by at least one model in any of the selected tokens. In addition, the *select information* dropdown menu gives access to a range of transformations based on these frequencies, such as odds ratio, Fisher Exact and cue validity.

The layout of Level 2, showing multiple plots at the same time and linking the tokens across models, is a fruitful source of information, but it has its limits. To exploit more model-specific information, we go one level down. Level 3 of the visualization tool shows a zoomable, interactive scatterplot in which each glyph, by default a steel blue circle, represents a token, that is, an attestation of the target lexical item. An optional second plot has been added to the right, in which each glyph, by default a steel blue star, represents a first-order context word, and the coordinates derive from applying the same dimensionality reduction technique on the type-level cosine distances between the context words. The name of the model, coding the parameter settings, is indicated on the top, followed by information on the dimensionality reduction technique. Like in the other two levels, it is possible to map colours and shapes to categorical variables, such as sense labels, and sizes to numerical variables, such as the number of available context words, and the legends are clickable, allowing the user to quickly select the items with a given value.

Figure 4.12 shows what Level 3 looks like if we access it by clicking on the name of the third model in Figure 4.10. Colour coding and selection are transferred between the levels, so we can keep working on the same information if we wish to do so. Conversely, we could change the mappings and selections on Level 3, based on model-specific information, and then return to Level 2 to compare the result across models. For example, if we wanted to inspect all the tokens for which *glas* 'glass' was captured by the model, we could input *glas/noun* on the *features in model* field. Or maybe we would like to find the tokens in which *glaasje* 'small glass' occurs, but we are not sure how they are lemmatized, so we can input *glaasje* in the *context words* field to find the tokens that include this word form in the concordance line, regardless of whether its lemma was captured by the model.

In sum, (groups of) tokens can be selected in different ways, either by searching words, inputting the unique identifier of the token, clicking on the glyphs or brushing over the plots. Given such a selection, clicking on 'Open frequency table' will call a pop-up table with one row per context word, a column indicating in how many of the selected tokens it occurs, and more columns with pre-computed information such as pmi (see Figure 4.13). These values can be interesting if we would like to strengthen or weaken filters for a smarter selection of context words.

Like Level 2, Level 3 also offers the concordance line of a token when hovering over it. But unlike Level 2, the concordance can be tailored to the specific model on focus, as shown in Figure 4.12. The visualization tool itself does not generate a tailored concordance line for each model but finds a column in the data frame that



Level 3 (*heffen*)

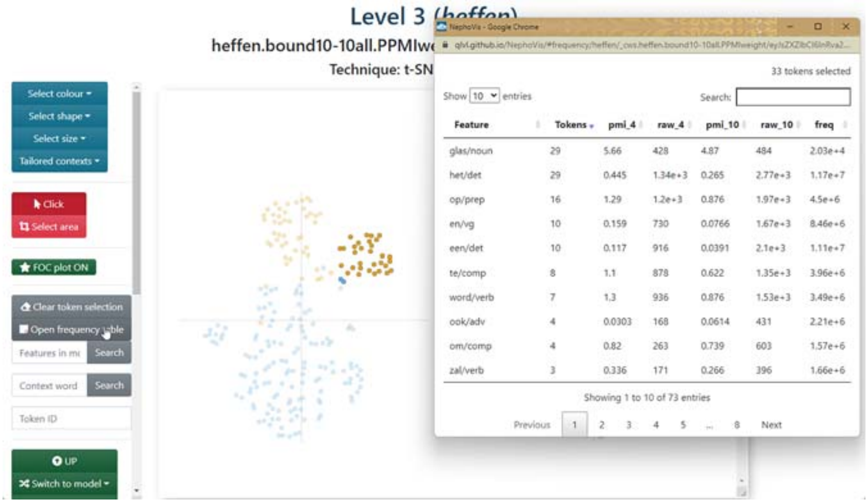
heffen.bound10-10all.PPMIweight.LENGTHFOC.SOCPOSall

Technique: t-SNE, perplexity: 30

hef/verb/het_nieuwsblad_20030321_01_92/501

En dan nog : **om**$^{0.82}$ de attributen **te**$^{1.104}$ plaatsen , **moet**$^{0.216}$ je **ook**$^{0.03}$ heffen .
Sinds november vorig jaar ga ik haast wekelijks en dit op eigen

Figure 4.12 Level 3 for the third medoid of *heffen* 'to levy/to lift', with parameters 10-10.ALL.BOUND.WEIGHT.SOCALL.FOC. Colours and selection of tokens have been transferred from Level 2



Level 3 (heffen)

heffen.bound10-10all.PPMIw

Technique: t-SN

NephoVis - Google Chrome

qlvl.github.io/NephoVis/#frequency/heffen/_cws.heffen.bound10-10all.PPMIweight/eyJsZXZlbCI6InRva2...

33 tokens selected

Show 10 entries

Search:

Feature	Tokens	pmi_4	raw_4	pmi_10	raw_10	freq
glas/noun	29	5.66	428	4.87	484	2.03e+4
het/det	29	0.445	1.34e+3	0.265	2.77e+3	1.17e+7
op/prep	16	1.29	1.2e+3	0.876	1.97e+3	4.5e+6
en/vg	10	0.159	730	0.0766	1.67e+3	8.46e+6
een/det	10	0.117	916	0.0391	2.1e+3	1.11e+7
te/comp	8	1.1	878	0.622	1.35e+3	3.96e+6
word/verb	7	1.3	936	0.876	1.53e+3	3.49e+6
ook/adv	4	0.0303	168	0.0614	431	2.21e+6
om/comp	4	0.82	263	0.739	603	1.57e+6
zal/verb	3	0.336	171	0.266	396	1.66e+6

Showing 1 to 10 of 73 entries

Figure 4.13 Level 3 for the second medoid of *heffen* 'to levy/to lift', with parameters 10-10.ALL.BOUND.WEIGHT.SOCALL.FOC. The frequency table gives additional information on the context words co-occurring with the selected tokens

starts with _*ctxt* and matches the beginning of the name of the model to identify the relevant format. A similar system is used to find the appropriate list of context words captured by the model for each token. For these models, the selected context words are shown in boldface and, for models with ppmi weighting such as the one shown in Figure 4.12, their ppmi values with the target are shown in superscript.

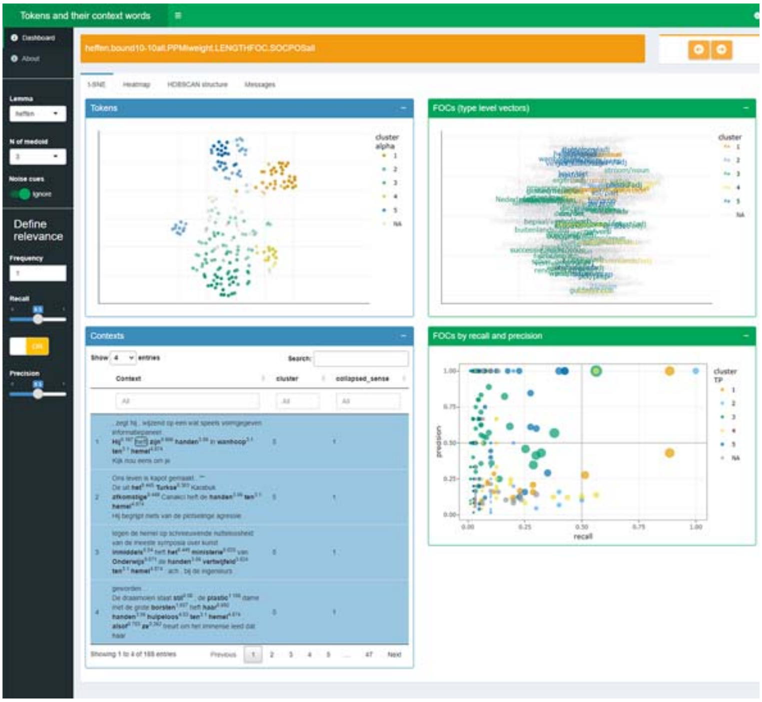
As we have seen throughout this chapter, the modelling pipeline returns a wealth of information that requires a complex visualization tool to make sense of it and exploit it efficiently. The Javascript tool described in this section, NephoVis, addresses many of the needs of this workflow, interconnecting the different levels of the analysis. The dashboard described in the next section, written in R instead, elaborates on some ideas originally conceived for NephoVis and particularly tailored to explore the relationship between the t-SNE solutions and the HDBSCAN clusters on individual medoids.

4.4 A ShinyApp extension for NephoVis

The visualization tool discussed in this section was written in R with the Shiny library (Chang, Cheng, Allaire, Sievert, Schloerke, Xie, Allen, McPherson, Dipert, and Borges 2022), which provides R functions that return HTML, CSS, and Javascript for interactive web-based interfaces. The interactive plots have been rendered with the R package plotly (Sievert, Parmer, Hocking, Chamberlain, Ram, Corvellec, and Despouy 2021). Unlike NephoVis, this dashboard requires an R server to run, so it is hosted on shinyapps.io instead of a static GitHub page. It takes the form of a dashboard, shown in Figure 4.14, with a few tabs, multiple boxes and dropdown menus to explore different lemmas and their medoids. All the functionalities are described in the About page of the dashboard, so here only the most relevant features will be described and illustrated.

The sidebar of the dashboard offers a range of controls. Next to the choice between viewing the dashboard and reading the documentation, two dropdown menus offer the available lemmas and their medoids, by number. By selecting one, the full dashboard adapts to return the appropriate information, including the name of the model in the orange header on top. The bottom half of the sidebar gives us control over the definition of relevant context words in terms of minimum frequency, recall, and precision, which will be explained below.

The main tab, *t-SNE*, contains four collapsible boxes: the blue ones focus on tokens while the green ones plot first-order context words. The top boxes (Figure 4.15) show t-SNE representations (perplexity 30) of tokens and their context words respectively, like we would find on Level 3 of NephoVis. However, there are crucial differences between the tools. First, the colours match pre-computed HDBSCAN clusters (*minPts* = 8) and cannot be changed; in addition, the transparency of the tokens reflects their ε (see Section 3.4). After all, the goal of this



Show 4 entries Search:

	Context	cluster	collapsed_sense
	All	All	All
1	... zegt hij , wijzend op een wat speels vormgegeven informatiepaneel . Hij heft zijn handen in wanhoop ten hemel Kijk nou eens om je	5	1
2	Ons leven is kapot gemaakt . " De uit het Turkse Kacabuk afkomstige Canakci heft de handen ten hemel Hij begrijpt niets van de plotselinge agressie .	5	1
3	tegen de hemel op schreeuwende nutteloosheid van de meeste symposia over kunst inmiddels heft het ministerie van Onderwijs de handen vertwijfeld ten hemel . ach , bij de ingenieurs	5	1
4	geworden . De draaimolen staat stil , de plastic dame met de grote borsten heft haar handen hulpeloos ten hemel alsof ze treurt om het immense leed dat haar	5	1

Showing 1 to 4 of 188 entries Previous 1 2 3 4 5 ... 47 Next

Figure 4.14 Starting view of the ShinyApp dashboard, extension of Level 3

dashboard is to combine the 2D visualization and the HDBSCAN clustering for a better understanding of the models. Second, the type-level plot does not use stars but the lemmas of the context words themselves. More importantly, they are matched to the HDBSCAN clusters based on the measures of frequency, precision, and recall. In short, only context words that can be deemed relevant for the definition or characterization of a cluster are clearly visible and assigned the colour of the cluster they represent best; the rest of the context words are faded in the background. A radio button on the sidebar offers the option to highlight context words that are relevant for the noise tokens as well. Third, the tooltips offer different information from NephoVis: the list of captured context words in the case of tokens, and the relevance measures as well as the nearest neighbours of the context word in the type-level plot.

For example, on the left-hand side of Figure 4.15 we see the same token-level model shown in Figure 4.12. Hovering over one of the tokens in the top right orange cluster, we can see the list of context words that the model captures for it: the same we could have seen in bold in the NephoVis rendering by hovering over the same token. Among them, *glas/noun* and the determiner *het* are highlighted, because they are the only ones that surpass the relevance thresholds we have set. On the right-hand side of the figure we can see the similarities between

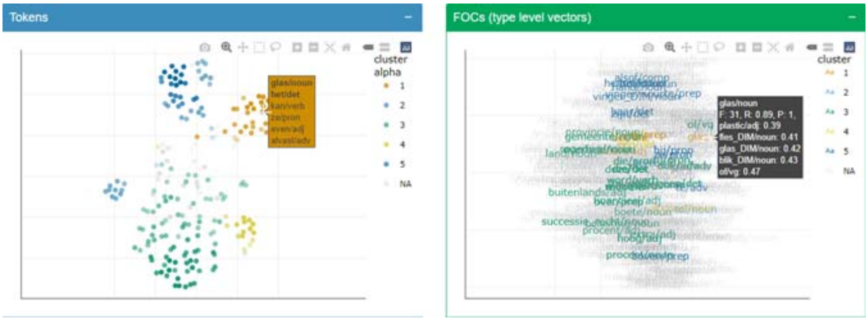


Figure 4.15 Top boxes of the 't-SNE' tab of the ShinyApp dashboard, with active tooltips

the context words that surpass these thresholds for any cluster. Hovering on one of them provides us with additional information. In the case of *glas/noun*, the first line reports that it represents 31 tokens in the orange HDBSCAN clusters, with a recall of 0.89 and precision of 1: it co-occurs with 89% of the tokens in the cluster and only with tokens in that cluster. Below we see a list of the nearest neighbours, that is, the context words most similar to it at type-level and their cosine similarity. For *glas* they are *plastic* and the diminutives of *glas*, *fles* 'bottle' and *blik* 'can'.

The two bottom boxes of the tab show, respectively, the concordance lines with highlighted context words and information on cluster and sense, and a scatterplot mapping each context word to its precision, recall, and frequency in each cluster. The darker lines inside the plot are a guide towards the threshold: in this case, relevant context words need to have minimum precision or recall of 0.5, but if the thresholds were modified the lines would move accordingly. The colours indicate the cluster the context word represents, and the size its frequency in it, also reported in the tooltip. Unlike in the type-level plot above, here we can see whether context words co-occur with tokens from different clusters. Figure 4.16 shows the right-side box next to the top token-level box. When one of its dots is clicked, the context words co-occurring with that context word—regardless of their cluster—will be highlighted in the token-level plot, and the table of concordance lines will be filtered to the same selection of tokens.

The first tab of this dashboard is an extremely useful tool to explore the HDBSCAN clusters, their (mis)match with the t-SNE representation, and the role of the context words. In addition, the *HDBSCAN structure* tab provides information on the proportion of noise per medoid and the relationship between ε and sense distribution in each cluster. Finally, the *Heatmap* tab illustrates the type-level distances between the relevant context words, ordered and coloured by cluster, as shown in Figure 4.17. In some cases, it confirms the patterns found in the type-level plot; in others it might add nuance to the relationships we thought we had found.

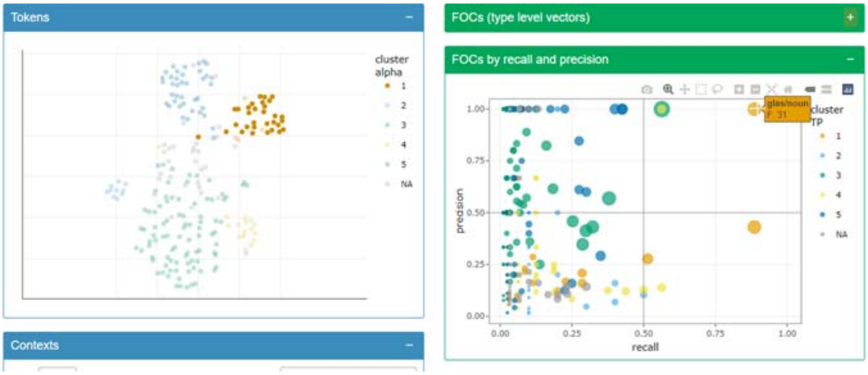


Figure 4.16 Token-level plot and bottom first-order context words plot of the 't-SNE' tab of the ShinyApp dashboard, with one context word selected

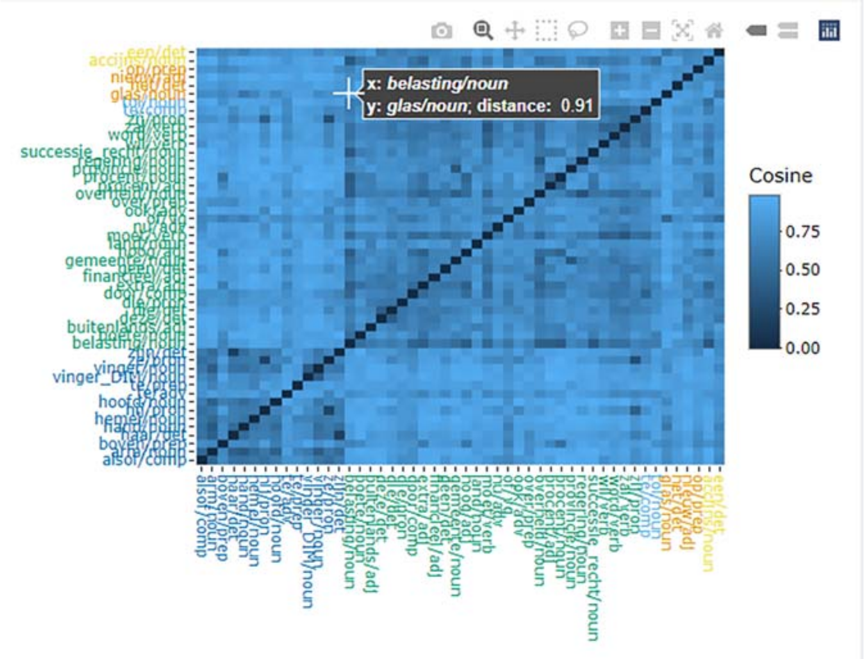


Figure 4.17 Heatmap of type-level distances between relevant context words in the ShinyApp dashboard

The bottom line

- Distances between tokens can be mapped to 2D via dimensionality reduction algorithms such as multidimensional scaling or t-SNE.
- Models can be compared by computing distances between them, which allows us to visualize them and to cluster them. The Partitioning Around Medoids clustering algorithm can then select representative models to examine in detail.
- A suite of software tools has been developed to implement the research reported on in this monograph. These tools are publicly available; an overview is found in the *Software resources* section at the end of the book.
- In particular, the NephoVis tool lets us explore the data, interactively, at different levels: from the distances between models through detail comparison of models to examination of individual models. A ShinyApp can be used to explore individual models in detail, combining information from t-SNE, HDBSCAN, and the context words selected by a model.

PART III

SEMASIOLOGICAL AND ONOMASIOLOGICAL EXPLORATIONS

Chapters 5 and 6 are the descriptive counterpart to Chapters 3 and 4. While the latter described the distributional method from a technical and a software point of view, we now turn to actual applications of the method—applications in which we will visualize token spaces and explore what the clusters that we can identify in them tell us about linguistic phenomena. In Chapter 5, that phenomenon is lexical meaning in its most direct, semasiological form: how can we use vector space models to identify word senses? Chapter 6 adds an onomasiological perspective: how can we use vector space models to describe the semantic relationship between various words? And how can it be used to get a grip on lexical variation and change?

5

Making sense of distributional semantics

What do clusters in distributional models look like and what do they mean? If we approach token-level models with the expectation of finding relatively well delineated clusters that match lexicographic senses, we will be thoroughly disappointed. ('Lexicographic senses' as meant here are meaning descriptions as one would find them in standard desk dictionaries in the form of a list of definitions.) As we will see in the following pages, distributional models take a variety of shapes, based on the frequency and distinctiveness of the context words that the models capture. There is no straightforward mapping between parameter settings and the resulting model, since the output depends mostly on the strength of the collocational patterns co-occurring with each target lemma. First, these patterns range from extremely resistant to changes in the parameter settings to constant shapeshifters. Second, they may or may not correspond to lexicographic senses or, more generally, to the kind of semantic distinctions a lexicographer might be interested in. As a consequence, distributional models are not reliable methods to identify such lexicographic senses, but instead may offer other insights into the collocational and semantic behaviour of the lemmas.

In this chapter we will look at the semantics in distributional semantics, as resulting from analyses performed on 32 Dutch nouns, adjectives, and verbs from a synchronic, semasiological perspective. The analyses presented in this chapter were performed on a corpus of Dutch and Flemish newspapers that we will refer to as the *QLVLNewsCorpus*. It combines parts of the *Twente News Corpus of Netherlandic Dutch* (Ordelman, De Jong, Van Hessen, and Hondorp 2007) with the yet unpublished *Leuven News Corpus*. It comprises articles published between 1999 and 2004, in equal proportion for both regions (the Netherlands and Flanders, the Dutch-speaking part of Belgium). The newspapers include *Het Laatste Nieuws*, *Het Nieuwsblad*, *De Standaard*, and *De Morgen* as Flemish sources and *Algemeen Dagblad*, *Het Parool*, *NRC Handelsblad*, and *De Volkskrant* as Netherlandic sources. The corpus amounts to a total of 520 million tokens, including punctuation. The corpus was lemmatized and tagged with part-of-speech and dependency relations with Alpino (Van Noord 2006).

For each of the lemmas, 240–360 attestations were extracted from the *QLVLNewsCorpus*, manually annotated for lexicographic senses and modelled with the parameter settings described in Chapter 3. Several examples from these analyses will illustrate and support the arguments made above, that is, first, that there is no configuration that returns an optimal solution across the board (see also Montes

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0005

2021b) and second, that the token clouds, that is, the clusters within the models, do not match lexicographic senses. Section 5.1 will offer an overview of the quantitative and qualitative effects of parameter settings across models. First, it will show how different parameter settings make the most difference in the distance between models as well as in the accuracy of the models in terms of lexicographic senses. In addition, relying on Section 7.2 of Montes (2021a), it will illustrate how the same parameter settings result in very different pictures across lemmas.

The following sections, based on Chapters 5 and 6 of Montes (2021a), analyse the types of information offered by the clouds, both at the syntagmatic and the paradigmatic level. At the syntagmatic level—the relationship between target types and their context words—they instantiate cases of collocation, colligation, semantic preference, or even tendencies towards the open-choice principle, in terms of Sinclair (1991, 1998). The paradigmatic level, on the other hand, codes the relationship between the clusters and lexicographic senses, from heterogeneous clusters to those that represent (proto)typical contexts of a sense. The combination of these dimensions results in a complex, rich picture of which lexicographic senses only cover a section. Section 5.2 will present the different linguistic phenomena that can be identified in the plots, from both a syntagmatic and a paradigmatic perspective. Sections 5.3 through 5.6 will focus on the different levels of the paradigmatic or semantic dimension, that is, the semantic interpretation that we can give the clouds, exploring how each of them combines with the different levels of the syntagmatic dimension.

All the models shown in this chapter result from a t-SNE dimensionality reduction with default values, ran with Rtsne: :Rtsne() in R (Krijthe 2018), and are colour-coded with clusters from HDBSCAN, computed with dbscan: :hdbscan() (Hahsler and Piekenbrock 2021; see Chapter 3 for a technical explanation). Grey indicates that the tokens were discarded as noise by the clustering procedure. The shapes of the glyphs correspond to different senses from the manual annotation. For a broader variety of case studies and models and a deeper analysis of the results, the reader is directed to Montes (2021a).

5.1 No single optimal solution

Parameter settings do not have an equal effect across all models, neither in the relative similarities or distances between the models, nor in terms of accuracy. In order to illustrate this, we will rely on conditional inference trees and random forests (Zeileis, Hothorn, and Hornik 2008; Hothorn and Zeileis 2015), run with the R package partykit (Hothorn and Zeileis 2021). Conditional inference trees try to predict a response or output (like the distances between models) based on a number of predictors or variables (such as the parameter settings in a model) by making binary decisions. Each decision in the tree tries to group observations with similar values in the response. Unlike parametric regression models, conditional inference trees can deal with large numbers of variables in small samples and are robust to multiple covariates, that is, correlating variables (Hothorn, Hornik, and Zeileis 2006). Note that this technique is typically used to learn the structure of a sample and predict the values in new data, but that is not how we will use it in this section. Instead, the goal is to illustrate how the set of choices that best describes the relations between the models of a lemma is different from the set of choices that describes a different lemma.

If we lay the focus on the ranking of parameter settings, that is, which makes the greatest difference in predicting the distances between models, we can run conditional random forests. This technique consists of running a large number of conditional inference trees combined with sampling techniques to avoid overfitting the prediction of a tree to the given sample and set of variables. From its output we can obtain variable importances, that is, the impact that each variable, in this case parameter setting, has on the result of a model.

Figure 5.1 illustrates the variable importances predicting the distances between models. The different lemmas are grouped by part-of-speech and the parameter settings are coded with colours. The farther to the right is the point, the higher the importance of the variable, that is, the higher the impact of that parameter setting on the pairwise distance between models. For example, for the verb *herinneren* 'to remember, to remind' the rightmost point is orange, which represents the part-of-speech filter. This means that when models have different part-of-speech filters, they tend to be very different from each other, more so than when they differ in some other dimension.

Note that, because some parameters are specific to bag-of-words models, viz. first-order window and part-of-speech, and always take the same value for dependency-based models, their partitions can be redundant with a distinction between bag-of-words and dependency models. For this reason, Figure 5.1 plots values based on a model that ignores these bag-of-words specific parameters against a model (in lower opacity) that does take them into account. First-order window and part-of-speech (in green and orange respectively) do indeed have a high variable importance when taking all parameters into account, with a higher impact of window size than of part-of-speech for the adjectives. However, the importance of the distinction between bag-of-words and different dependency formats greatly increases when the bag-of-words-specific parameter settings are ignored.

Figure 5.2 also plots variable importances, but with accuracy as the response. Accuracy was measured as the weighted proportion of k nearest neighbours of a token (with $k = 10$) that belong to the same lexicographic sense based on our manual annotation. For this purpose we used semvar::clusterqualkNN(); see Speelman and Heylen (2017). If a token has kNN = 1, its ten nearest neighbours belong to the same sense; if it has kNN = 0, none of its ten nearest neighbours belong to the same sense. The kNN of a sense is the average kNN of the points belonging to it; the values predicted here are the mean kNN across senses, in each model. In short, we want to see which parameters make a greater effect in increasing or decreasing the mean kNN value of the senses (as a proxy to accuracy).

As in Figure 5.1, we see that first-order part-of-speech is the most powerful parameter for *herinneren* 'to remember/to remind'. However, the range of values is different, as very few lemmas have a parameter with a variable importance larger than 0.001 (the vertical line). Sentence boundaries and the second-order parameter settings are consistently the least important.

The key insight from the variable importance ranking is that the relative importance of parameters varies across lemmas: the strongest parameter for *herinneren* 'to remember/to remind' is only second in the ranking for *harden* 'to become/make hard, to tolerate' and even lower for *geldig* 'valid'. What variable importances cannot show is the specific effect of each parameter setting on the response value. For example, given two lemmas for which the first-order part-of-speech setting is the most important predictor of accuracy, are the situations actually comparable? Could the same value of the same parameter have opposite effects in different lemmas?

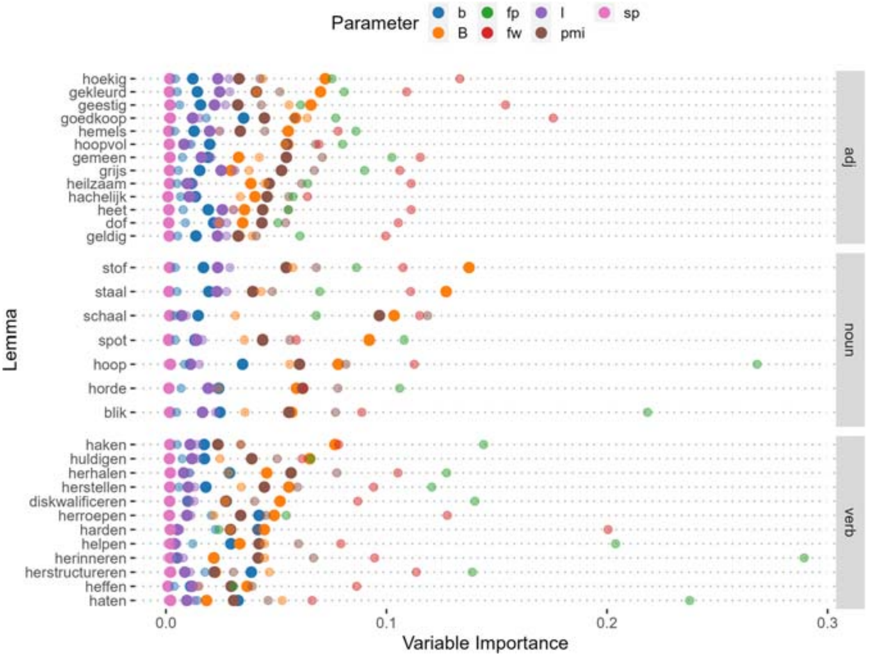


Figure 5.1 Variable importance predicting distances between all models. More transparent dots correspond to conditional random forests with all parameters as predictors, while highlighted dots come from models excluding parameters specific to bag-of-words models

Indeed, that is the case, and conditional inference trees are the evidence. Figures 5.3 and 5.4 show the conditional trees predicting mean kNN across senses for *herinneren* 'to remember/to remind' and *huldigen* 'to believe/to honour' respectively. In both cases, the first-order part-of-speech is the most important variable, with bag-of-words models that only include nouns, adjectives, verbs, and adverbs ('lex') exhibiting a different behaviour from both the rest of the bag-of-words models and the dependency-based models. However, in the case of *herinneren*, where the senses are characterized by the use of pronouns and prepositions (see Section 5.5), the models that implement the part-of-speech filter perform the worst; in contrast, in the case of *huldigen*, where the direct object and other lexical items are better cues for the senses (see Section 5.4), these models perform the best. It should be noted that the range of kNN values is rather similar in both lemmas, but that is not the case for all the case studies.

In short, conditional random forests show that no parameter setting is consistently the most important in defining either the similarity between models or their accuracy; conditional inference trees, on the other hand, show how even when two lemmas are more sensitive to the same parameter setting, the effect is not necessarily the same. These two observations pertain to measurable properties, namely the distance/similarity between models and the accuracy in terms of lexicographic senses. At the same time, we could ask whether there is a qualitative relationship

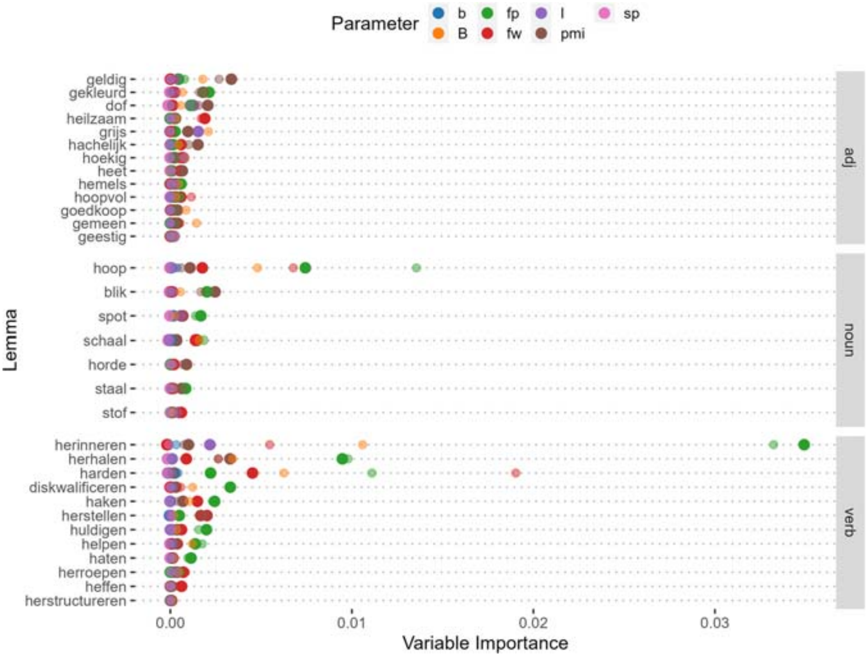


Figure 5.2 Variable importance predicting accuracy of models

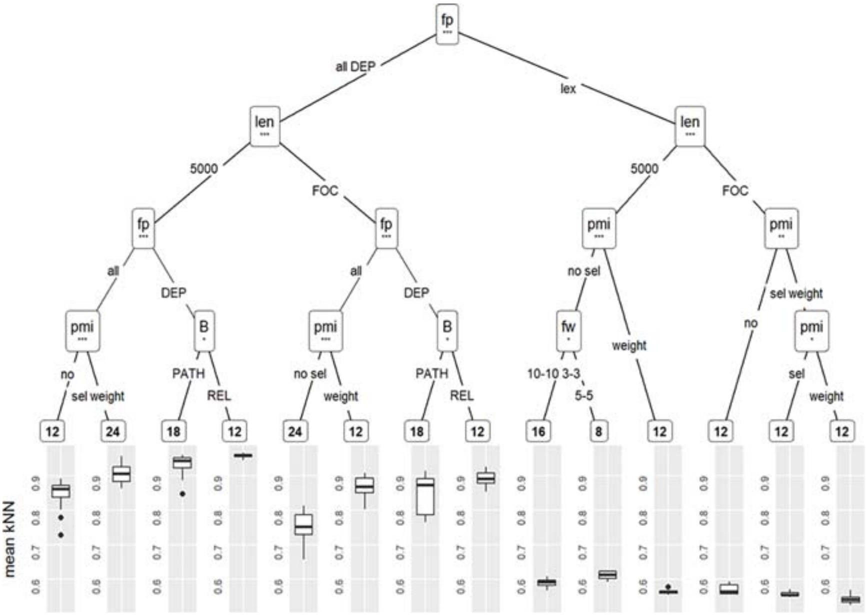


Figure 5.3 Conditional tree predicting the accuracy of *herinneren* 'to remember/to remind' models as kNN. Abbreviations: fp = first-order part-of-speech, B = distinction between bag-of-words and dependency models, fw = first-order window, pmi = ppmi weighting, len = vector length

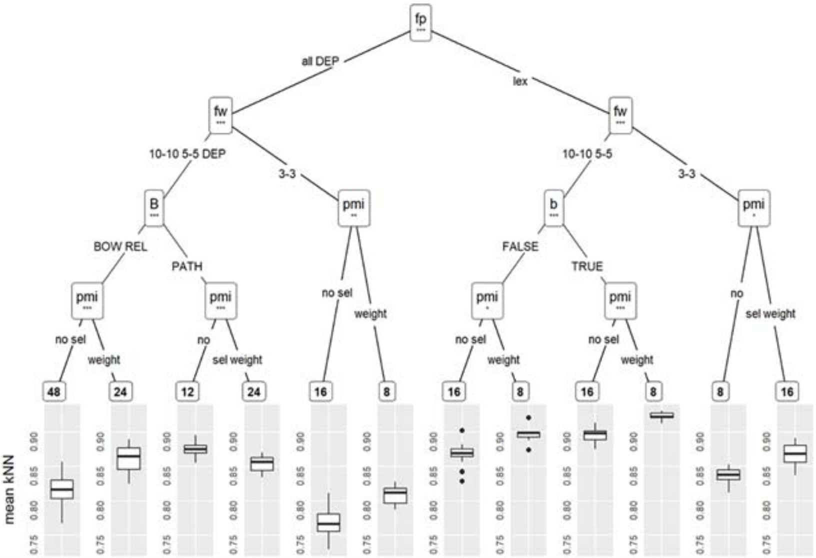


Figure 5.4 Conditional tree predicting the accuracy of *huldigen* 'to believe/to honour' models as kNN. Abbreviations: fp = first-order part-of-speech, B = distinction between bag-of-words and dependency models, fw = first-order window, pmi = ppmi weighting

between parameter settings and the shapes of the models. For example, certain parameter settings may favour tighter clusters, or a larger number of clusters. Yet again, this depends on the specific collocational patterns from the sample, which does not correlate with either the part-of-speech of the target or with the semantic phenomena we could expect.

For example, let us inspect Figures 5.5, 5.6, and 5.7, which show 2D representations of six models of different lemmas, built with the same parameter settings: symmetric bag-of-words window size of 5, but respecting sentence boundaries, only including nouns, verbs, adjectives and adverbs with a positive pmi with the target lemma; the second-order features coincide with the first-order features.

Each figure shows two plots that are quite similar to each other but different from the plots in the other figures. In Figure 5.5, we see models corresponding to *heet* 'hot' and *stof* 'substance, dust...'. The model of *heet* has 12 clusters with different degrees of tightness and distinctiveness. Most of them are dominated by individual context words that co-occur with most of the members of the cluster and represent typical patterns within a sense, as will be discussed in Section 5.5. Other clusters are characterized by groups of infrequent but semantically similar context words, as shown in Section 5.6. Finally, a few clusters are semantically heterogeneous and have no clear collocational pattern that characterizes them (see Section 5.3). The *stof* model has seven relatively homogeneous clusters. The three distinct clouds on the upper left corner are dominated by specific, individual context words and represent typical uses of the 'substance' sense (represented with circles). The red cloud also represents a typical use of this sense but is characterized by multiple similar context words instead. The rest of the clusters are lexicographically more heterogeneous. In other words, Figure 5.5 shows an

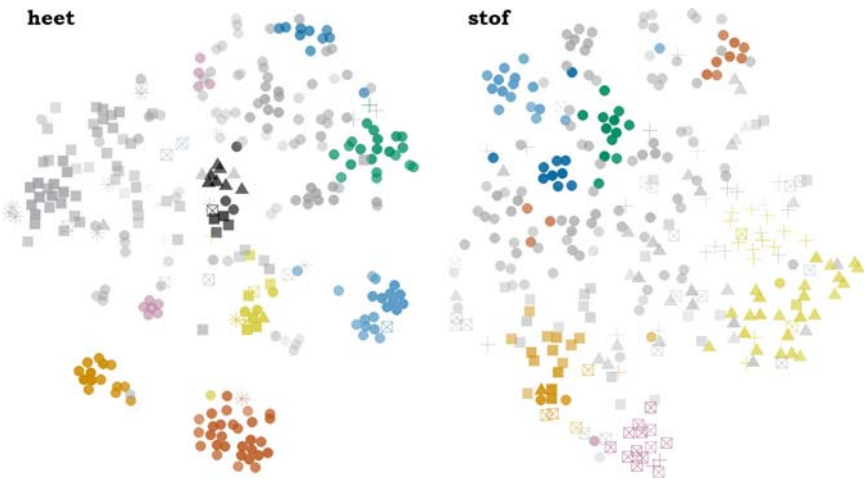


Figure 5.5 Models of heet '*hot*' and *stof* 'substance, dust...' with parameters 5-5.LEX.BOUND.SELECTION.SOCALL.FOC

adjective and a noun with similar shapes and similar interpretations: a number of tight clusters dominated by individual context words and a few dominated by semantically similar words, all of these representing typical uses of a single sense, and then some heterogeneous clusters. However, deeper inspection reveals that, while the homogeneous clouds of *stof* 'substance' represent typical uses that also profile different dimensions of the sense (for example, dangerous substances, harmful substances, or poisonous substances, as discussed in Section 5.6), the typical patterns within *heet* constitute idiomatic expressions, such as *hete aardappel* 'hot potato'.

The models shown in Figure 5.6 correspond to *dof* 'dull' and *huldigen* 'to believe/to honour'. Note that the model of the adjective *dof* is visually more similar to that of the verb *huldigen* than to the model of the adjective *heet* 'hot' in Figure 5.5. This is not only remarkable given their shared part-of-speech but because other models of *dof*, that is, with other parameter settings, do exhibit multiple clusters characterized by collocations with different types of sounds (among others, *klap* 'clap', *knal* 'bang', *dreun* 'thump') like the model of *heet* in Figure 5.5. In this model of *dof*, the metaphorical sense represented by the collocation with *ellende* 'misery' forms a tight, distinctive orange cloud on one side; the different context words referring to sounds give rise to the homogeneous, broader light blue cloud below, and the rest of the tokens, both those related to the visual sense ('dull eyes') and the rest of the metaphorical ones, gather in the green cloud. As we will see in Section 5.4, *huldigen* also has some strong collocates, but in this model the tokens of the 'to believe' sense, led by *principe* 'principle', *opvatting* 'opinion', and *standpunt* 'point of view', are part of the orange cloud, while most of the 'to honour' sense is covered by the large light blue cloud. In other words, these are lemmas with

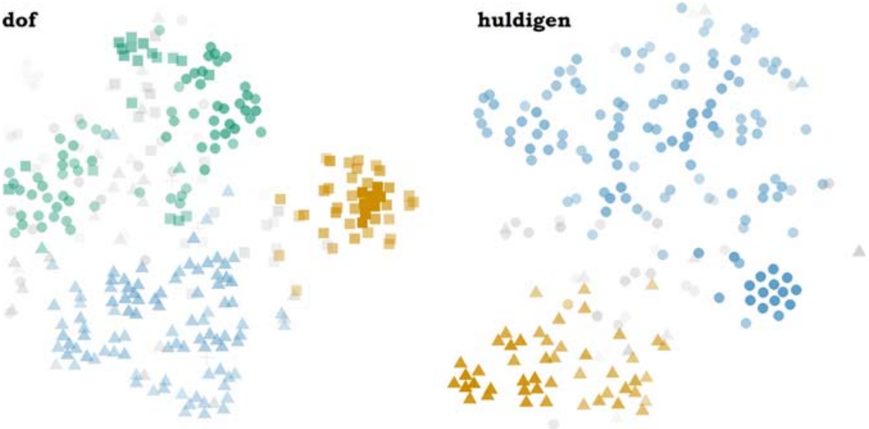


Figure 5.6 Models of *dof* 'dull' and *huldigen* 'to believe/to honour' with parameters 5-5.LEX.BOUND.SELECTION.SOCALL.FOC

some context words that can be strong enough to generate isolated clusters as seen in Figure 5.5, but not with the parameter settings chosen here. Moreover, for one of the lemmas both clusters happen to match lexicographic senses, whereas for the other lemma two clusters are semantically homogeneous and the other one is not.

Finally, the models shown in Figure 5.7, corresponding to *haten* 'to hate' and *hoop* 'hope/heap', show yet another configuration generated by the same parameter settings. Except for the green cloud in *haten*, roughly dominated by the co-occurrence with *mens* 'human, people', the rest of the clouds are small, diffuse, heterogeneous, and characterized by many different words. Moreover, most of the tokens of these models are excluded by the HDBSCAN algorithm as noise. This picture is endemic to all of the models of *haten*, which cannot find stronger collocational patterns than the co-occurrence with *mens*. In the case of *hoop*, on the other hand, dependency-informed models can—unlike this model—distinguish the two homonyms.

To sum up this section, we have seen (1) that the relative importance of the parameter settings is different for the different lemmas; (2) that even if the same parameter is important for two lemmas, its effect can be different in each of them; and (3) that the same parameter settings can return drastically diverging results when applied to different lemmas. This is tightly related to the particular configuration of the distributional patterns of that lemma, that is, how frequent and distinctive its context words are. The following sections offer some theoretical tools for the semantic interpretation of such configurations. However, it must be remembered that they interact with the parameter settings in complicated ways, as shown by the fact that both *heet* 'hot' and *dof* 'dull' can exhibit tight clusters dominated by specific collocations but they are not necessarily invoked by the same parameter settings.

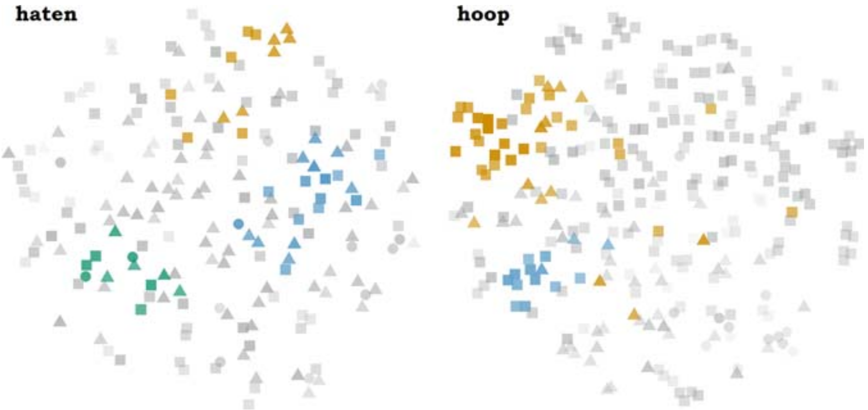


Figure 5.7 Models of *haten* 'to hate' and *hoop* 'hope, heap' with parameters 5-5.LEX.BOUND.SELECTION.SOCALL.FOC

5.2 Types of information

The linguistic interpretation of the clusters can be understood from a paradigmatic perspective—in relation to lexicographic senses—and from a syntagmatic perspective—in terms of co-occurrence patterns of different kinds. As already hinted at in the introduction to this chapter, both dimensions interact, resulting in a number of specific phenomena that we may encounter. The relationship is summarized in Table 5.1, which brings together a number of examples, most of which will be discussed in detail in the later sections.

The paradigmatic perspective refers in this case to the relationship between the HDBSCAN clusters and the manually annotated lexicographic senses. From this perspective we can initially distinguish between heterogeneous clusters, that is, those that do not exhibit a clear preference for one sense, and homogeneous clusters. Secondly, the homogeneous clusters may cover all the (clustered) tokens of a given sense or only some of them, effectively representing a typical (or even prototypical) context of the sense. Finally, said typical context may additionally highlight a certain aspect or dimension of the meaning of the target, different from that highlighted by a different context. As a result, the semantic dimension covers four different types of situations. The first one, that is, heterogeneous clusters or clusters with multiple senses, will be described in Section 5.3. If we considered the senses a gold standard and the target of our models, these clusters would be understood as bad or failed modelling. The second type of situation, that is, clusters that perfectly match senses, is the ideal result and what we could naively expect from distributional models. However, it is quite rare and often indicative of fixed expressions or very particular circumstances, as we will see in Section 5.4. Instead, rather than full senses, contextual patterns tend to represent typical contexts of a sense, which are the focus of Sections 5.5 and 5.6.

As already described in Chapter 1 (Section 1.2), the notion of prototypicality in cognitive semantics is related to the principle that categories need not be discrete and uniform and to its application to the semasiological structure of lemmas and their meanings (Geeraerts 1988, 1997). Distributional models in particular can give us insight into the prototypical structure at the extensional level: prototypical instances/contexts of a lemma, of a particular sense, or of a dimension of a sense. Section 5.5 will focus on the general cases where the tokens co-occurring with a specific contextual pattern represent a typical usage of a sense. Meanwhile, Section 5.6 will describe specific situations in which various contextual cues point to a specific semantic dimension of the sense that correlates with the contextual pattern. In other words, each typical usage within a sense highlights a different aspect of that sense.

From the syntagmatic or collocational perspective, clusters can be interpreted in terms of Sinclair's (1991) classification of collocations, already introduced in Chapter 2 (Section 2.2). The framework includes, next to the node, that is, our targets, four key components: one obligatory—semantic prosody, which will not be discussed here—and three more that will help us make sense of the observed output of the clouds: collocation, colligation, and semantic preference. In their simplest form, collocations are defined as the co-occurrence of two words within a certain span (Firth 1957: 13; Sinclair 1991: 170; 1998: 15; Stubbs 2009: 124). They might be further filtered by the statistical significance of their co-occurrence frequency or by their strength of attraction, such as pmi (see McEnery and Hardie 2012: 122–33 for a discussion). Even though a collocational relationship is asymmetric, that is, the co-occurrence with a more frequent word B may be more important for the less frequent word A than for B, the measures used to describe it are most often symmetrical (Gries 2013). When it comes to the interpretation of clouds, collocation takes a different form and is definitely asymmetric. Considering models built around a target term or node, frequent, distinct context words are bound to make the tokens that co-occur with them similar to each other and different from the rest: they will generate clusters, *aka* clouds. Such context words do tend to have a high pmi with the target, but, crucially, they stand out because they are a salient feature among the occurrences of the target, independently from how salient the target would be when modelling the collocate. In contrast to collocational studies, which tend to focus on lists of words that co-occur (significantly) frequently with a target node, vector space models also show whether these context words co-occur with each other in the context of the target or are mutually exclusive instead. In fact, this might reveal more complex collocational patterns that recruit multiple co-occurring context words. At a more technical level, clusters characterized by these strong, lexical collocational patterns tend to be very well defined: the distance between the tokens is small, the density of the semantic region is much higher than around the noise tokens, and they do not overlap with other clusters.

Whereas collocation is understood as a relationship between words (and, traditionally, as a relationship between word forms), colligation is defined as a relationship between a word and grammatical categories or syntactic patterns (Firth 1957: 14; Sinclair 1998: 15; Stubbs 2009: 124). In order to capture proper colligations as clusters, we would need models in which parts of speech or maybe dependency patterns are used as features, which is not the case in these studies. However, by rejecting a strict separation between syntax and lexis, we can make a grammatically oriented interpretation of collocations with function words, such as frequent prepositions or the passive auxiliary. We will therefore talk about lexically instantiated colligations when we encounter clusters dominated by items that indicate a specific grammatical function.

Finally, semantic preference is defined as the relationship between a word and semantically similar words (Sinclair 1998: 16; Stubbs 2009: 125; McEnery and Hardie 2012: 138–40). Within traditional collocational studies, this implies grouping collocates, that is, already frequently co-occurring items, based on semantic similarity, much as colligation can be the result of grouping collocates based on their grammatical categories. In the explanation of individual clusters, semantic preference appears in clusters that are not dominated by a single collocate or group of co-occurring collocates, but are instead defined by a group of infrequent context words with similar type-level vectors and for which we can give a semantic interpretation. This is a key contribution of token-level distributional models that is hardly achievable in traditional collocational studies: next to powerful collocates that group virtually identical occurrences, we can identify patterns in which the context words are not exactly the same but are similar enough to emulate a single, stronger collocate. These two last groups, that is, lexically instantiated colligations and semantic preference, tend to form larger, less compact clusters than lexical collocations. They cover somewhat less dense areas, sometimes with more overlap with other clusters.

The three notions described above assume identifiable patterns: occurrences that are similar enough to a substantial number of other occurrences, and different enough from other occurrences, to generate a cluster. Going back to Sinclair's founding notions (1991), we are assuming the domination of the idiom principle: '… a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments' (Sinclair 1991: 110). The opposite situation would be given by the open-choice principle: 'At each point where a unit is completed (a word or a phrase or a clause), a large range of choice opens up and the only restraint is grammaticalness' (Sinclair 1991: 109). The idiom principle and the open-choice principle are supposed to organize the lexicon and the production of utterances. But if, instead, they are understood as poles in the continuum of collocational behaviour, they can help us interpret the variety of shapes that we encounter within and across lemmas. Lemmas in which we tend to find identifiable clusters, with strong collocations, lexically instantiated colligations or sets with semantic preference, can be said to respond to the idiom principle. In contrast, lemmas that exhibit large proportions of noise tokens, as well as small, diffuse clusters, can be said to approximate the open-choice principle (see for example the plots in Figure 5.7). They don't necessarily lack structure, but whatever structure they have is less clear than for other lemmas, and harder to capture in those particular models. With this reasoning, next to the three categories described above, we include 'near-open choice' as a fourth category, meant to include the clouds that do not conform to either of the clearer formats.

As we can see in Table 5.1, the interaction between the four levels of each dimension results in a 4×4 table with all but two cells filled with at least one example. Naturally, not all the combinations are equally frequent or interesting; the most salient one is certainly the collocation that indicates the prototypical context of a sense. But this does not mean that the rest of the

Table 5.1 Examples of syntagmatic (columns) and paradigmatic (rows) perspectives on the linguistic interpretation of clouds

SEMANTIC INTERPRETATION	SINGLE COLLOCATION	LEXICALLY INSTANTIATED COLLIGATION	SEMANTIC PREFERENCE	NEAR-OPEN CHOICE
Heterogeneous clusters	*heilzaam* ‘healthy/beneficial’ + *werking* ‘effect’ (and near-synonyms)	*herstructureren* ‘to restructure’ + passive aux. *word* (part of the two transitive senses); *helpen* ‘to help’ + *om* & *te* ‘in order to’	*geestig* ‘witty’ + *wijze/manier* ‘manner’/various adverbs; *grijs* ‘grey’ + colours and clothes; *herroepen* ‘to recant/to void’ + *uitspraak* ‘statement/verdict’ & juridical field	*blik* ‘gaze/tin’—*werpen* ‘to throw’, *richten* ‘to aim’
Dictionary clusters	*staal* ‘sample’ + *representatief* ‘representative’; *schaal* ‘dish of a scale’ + *gewicht* ‘weight’; *schaal* ‘scale’ + *Richter*	*herhalen* ‘to repeat’ + *zich* ‘itself’; *hoop* ‘hope/heap’, in the one model that gets the senses right	*haken* ‘to make trip/to crochet’ + sports terms or hobby terms; *schaal* ‘scale’ + earthquake-topic or kitchen-topic	*huldigen* ‘to honour’
(Proto)typical context	*heffen* ‘to levy/to lift’ and all its collocates (except for *hand/arm*); *hachelijk* ‘dangerous/critical’ and its collocates	*diskwalificeren* ‘to disqualify’ + passive aux. *word*; *helpen* ‘to help’ + different pronouns/prepositions (*bij*, *aan*) as only remaining context words; *herinneren* ‘to remember/to remind’ + *(er)aan* ‘of (it)’, *ik* ‘I’ & reflexive pron. *me*, *zich*	*grijs* ‘grey’ + cars; *heet* ‘hot’ + food; *hemels* ‘heavenly’ + music; *dof* ‘dull’ + sounds	-Not relevant-
(Proto)typical context with profiling	*stof* ‘substance’ and its adjectives; *horde* ‘horde’	*horde* ‘horde’ + *journalist* & *door* ‘by’	*geldig* ‘valid’ + tickets & dates/identity documents & *voorleggen* ‘submit’ /*bezitten* ‘possess’; *staal* ‘steel’ + *ton* & *milioen* ‘million’/materials	-Not relevant-

phenomena should be ignored: we can still find interesting and useful information with other shapes of clouds, other contextual patterns, other semantic structure.

In the following sections, we will look in detail at examples of each attested combination. Each section will focus on one level of the semantic dimension and will be internally structured by the levels of the collocational dimension. The examples will be illustrated with scatterplots in which the colours represent HDBSCAN clusters and the shapes indicate manually annotated dictionary senses. The senses are not specified in the legends but in the captions, whereas the clusters will be named with the context word that represents it best followed by its *F*-score in relation to the cluster. The *F*-score is the harmonic mean of precision and recall. In this case, *precision* is the proportion of tokens co-occurring with a context word that are accounted for in a given cluster. Conversely, *recall* is the proportion of tokens of a cluster that co-occur with a context word. An *F*-score close to 1 for a context word in relation to a cluster means that the tokens in that cluster (almost) always co-occur with that context word, and that (almost) only tokens from that clusters co-occur with that context word. Textual reproductions of some tokens will also be offered; in all cases the target will be in boldface and the context words captured by the relevant model in italics. The source information, that is, the name of the newspaper, the date of publication, and the number of the article in the corpus, will accompany the original text. The source information is followed by an English translation between simple quotation marks.

5.3 Semantic heterogeneity

In many cases, clusters include tokens from different senses, without showing a strong preference for any of them. That is mostly the case in the near-open choice type of clusters but can also be identified in clusters with clear collocational patterns. In this section we will look at examples of heterogeneous clouds with different syntagmatic interpretation, from lexical collocations to near-open choice.

First, in some rare cases, collocations transcend senses, that is, they can be frequent and even distinctive of a lemma without showing a preference for a specific sense. The clearest example is found in *heilzaam* ‘healthy/beneficial’, which was annotated with two senses: one referring literally to health and one metaphorical, applied to a variety of domains. We expected to encounter mostly expressions such as *heilzame kruiden* ‘healthy spices’ or (5.1), in which the semantic domain of the noun that the adjective is applied to points more or less unambiguously to the lexicographical sense. At most, we expected some ambiguity in situations where the entity itself could either be literally healthy or more generally beneficial depending on the context. Instead, the construction in which the adjective tends to occur seems to stand in the way: *heilzaam* is mostly applied to nouns such as *werking* 'effect', *effect* and *invloed* 'influence', which do not discriminate between the two senses the annotation aimed for. Some examples are shown in (5.2) and (5.3) for the 'healthy' sense and in (5.4) and (5.5) for the 'beneficial' sense. As a result, clusters tend to be dominated by one of these context words.

(5.1) Versterking *van de* politieke controle *op de Commissie kan* **heilzaam** *zijn maar de huidige* ongenuanceerde *discussie is gevaarlijk voor* Europees beleid en besluitvorming. (*De Morgen*, 1999-03-18, Art. 45) 'Reinforcement *of the* political control *at the Commission can be* **beneficial**, *but the current* unnuanced *discussion is dangerous for* European policy and decision-making.'

(5.2) Het lypoceen, een *bestanddeel* dat bijdraagt aan *de* rode *kleur*, *zou een* **heilzame** *werking hebben op de* prostaat. (*De Volkskrant*, 2003-11-08, Art. 14) 'Lypocene, a *component* that contributes to *the* red *colour*, *would have a* **healing** *power on the* prostate.'

(5.3) Pierik *beschrijft de* **heilzame** *effecten van alcoholgebruik op de* bloedvaten en *de* bloeddruk, op mogelijke beroerten, galstenen, lichaamsgewicht, vruchtbaarheid, zwangerschap, botontkalking, kanker, verkoudheid, suikerziekte en seniele dementie. (*NRC Handelsblad*, 1999-11-27, Art. 148) 'Pierik *describes the* **healing** *powers of alcohol consumption on* [*the*] blood vessels and [*the*] blood pressure, on potential strokes, gallstones, body weight, fertility, pregnancy, osteoporosis, cancer, the cold, diabetes and senile dementia.'

(5.4) Voor politici met dadendrang een gruwel, maar als men de casus van de Betuwelijn nog *voor de geest* haalt dan *zou het* advocatensysteem zijn **heilzame** *werking hebben kunnen bewijzen*. (*De Volkskrant*, 2002-03-29, Art. 79) 'For politicians with thirst for action it is an abomination, but when one recalls (lit. 'brings *to the spirit*') the case of the Betuwe line then *the* lawyer system *would* have been *able* to *prove* its **beneficial** *effect*.'

(5.5) *De kwestie heeft alvast* één **heilzaam** *effect*: *het profiel van* commerciële boekenprijzen staat opnieuw ter discussie. (*De Standaard*, 1999-03-27, Art. 133) '*The matter certainly has* a **beneficial** *effect*: *the profile of* commercial book prizes is again under discussion.'

The model is shown in Figure 5.8: the clusters dominated by *werking* 'effect', *effect*, and *invloed* 'influence' are shown in yellow, light blue, and green, respectively, and the manually annotated senses are mapped to the shapes: the literal 'healthy' sense

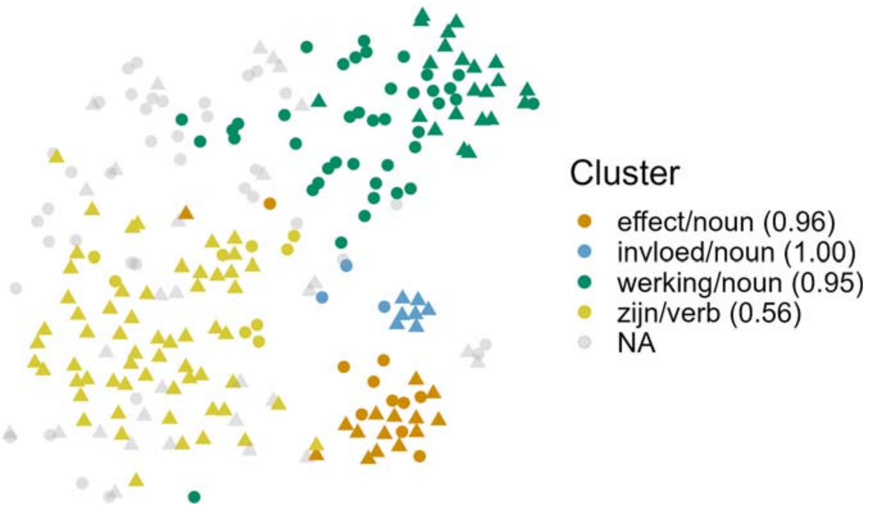


Figure 5.8 Model of *heilzaam* with parameters 10-10.ALL.BOUND.WEIGHT.SOCALL.FOC. Circles are 'healthy, healing', triangles are 'beneficial' in general

is coded in circles, and the metaphorical sense in triangles. Within the *werking* cluster, the literal tokens, like (5.2), are the majority and tend towards the left side of the cloud, whereas the metaphorical ones like (5.4) tend towards the right side. While there is a preference for the literal sense, especially considering that across the full sample the metaphorical sense is more frequent, it is far from homogeneous. The distribution is even more balanced within the *effect* cluster. Such a picture is pervasive across multiple models of *heilzaam*. The vague organization within the *werking* cluster suggests that models might actually be able to capture words representative of 'physical health', but these context words have to compete with the most salient context words, which are not discriminative of these two senses.

This is an issue if we come to distributional semantics expecting the nouns modified by adjectives in particular, or lexical collocates in general, to unequivocally represent different dictionary senses. On the other hand, *zijn* 'to be' and *werken* 'to work, to have an effect' (of which *werking* is a nominalization), exemplified in (5.1) and (5.6), co-occur almost exclusively with the tokens in the orange cluster, dominated by the metaphorical sense. In other words, the most frequent nouns modified by *heilzaam* tend to occur in attributive constructions (particularly *een heilzame werking hebben* 'to have a beneficial/healing effect/power' and *de heilzame werking van* 'the beneficial/healing effect/power of') and for either sense, whereas the predicative constructions present a wider variety of nouns and a stronger tendency towards the metaphorical sense.

(5.6) Ten slotte nog één fundamentele bedenking: ook *de* permanente actualiteit *van de* thematiek in *de media werkt* **heilzaam** *op de weggebruikers.* (*De Morgen*, 2001-02-28, Art. 107) 'To conclude, one final fundamental thought: *the* permanent presence *of the* topic in *the media* has a **beneficial** effect (lit. '*works* beneficially') *on road users*.'

The models of *heilzaam* 'healthy/beneficial' show that we cannot take for granted that collocations will be representative of senses. What is more, these words have both a high pmi with *heilzaam* and were often selected as informative cues by the annotators, which places doubt on the reliability of both methods as sources of semantic distinctiveness. When it comes to pmi, it is understandable: the measure is meant to indicate how distinctive a context word is of the type as a whole, in comparison to other types. It does not consider how distinctive it is of a group of occurrences against another group of occurrences of the same type. When it comes to cueness annotation, however, we could have expected a more reliable selection, but apparently the salience of these context words is too high for the annotators to notice that it is not discriminative of the different senses.

Just like lexical collocations, lexically instantiated colligations can also highlight dimensions that do not correspond to lexicographic senses. One example is the case of *herstructureren* 'to restructure', annotated with three sense tags emerging from a combination of specialization, that is, whether it's specifically applied to companies, and argument structure, distinguishing between transitive and intransitive forms. The intransitive sense is always specific—companies restructure, undergo a process of restructuring. Models of *herstructureren* are typically not very successful at disentangling any of these three senses. Instead, the clusters that emerge tend to highlight either the semantic or the syntactic dimension, disregarding the other one. The lexical items that most frequently dominate clusters of *herstructureren* are the passive auxiliary *worden*, *bedrijf* 'company', *grondig* 'thorough(ly)', and the pair of prepositions *om te* 'in order to', illustrated in (5.7) through (5.9).

(5.7) OK-score deelt bedrijven op in tien klassen; klasse 1 blaakt van gezondheid, klasse 10 is op sterven na dood, ofwel, staat op de rand van faillissement en *moet grondig worden* **geherstructureerd.** (*Het Parool*, 2003-04-16, Art. 69) 'The OK-score divides companies into ten classes: class 1 is brimming with health, class 10 is as good as dead, or rather, stands on the edge of bankruptcy and *must be thoroughly* **restructured**.'

(5.8) *Ze* **herstructureerden** *het bedrijf en* loodsten het de internationale groep Taylor Nelson Sofres (TNS) binnen. (*De Standaard*, 2004-01-06, Art. 59) '*They* **restructured** *the company and* steered it towards the Taylor Nelson Sofres (TNS) international group.'

(5.9) Uiteindelijk is dat de regering, want toen de crisis uitbrak nam de overheid een belang in de banken *om ze te* **herstructureren** *en uiteindelijk weer* te verkopen. (*NRC Handelsblad*, 2000-11-07, Art. 11)
'In the end that is the government, because when the crisis hit the authorities took an interest in the banks *in order to* **restructure** *them and eventually* sell them *again*.'

The two lexical collocates, *grondig* 'thoroughly' and *bedrijf* 'company', never co-occur with each other (even though they don't fulfil the same function), and only occasionally co-occur with *worden* or *om te*, which themselves co-occur with each other a few times. While they are both good cues for the company-specific senses, they may occur in either transitive or intransitive constructions. In contrast, *worden* is a good cue for transitive (specifically, passive) constructions, but may occur with either the company-specific or the general sense. Finally, *om te* may be attested in either of the three senses. The stark separation of the clusters in Figure 5.9 would seem to suggest opposite poles, but that is not the case at the semantic level. In fact, the clusters are merely slightly denser areas in a rather uniform, noisy mass of tokens and would be very hard for the naked human eye to capture without HDBSCAN clustering. Instead, each cluster indicates a pole of contextual behaviour which itself may code a semantic dimension, in the case

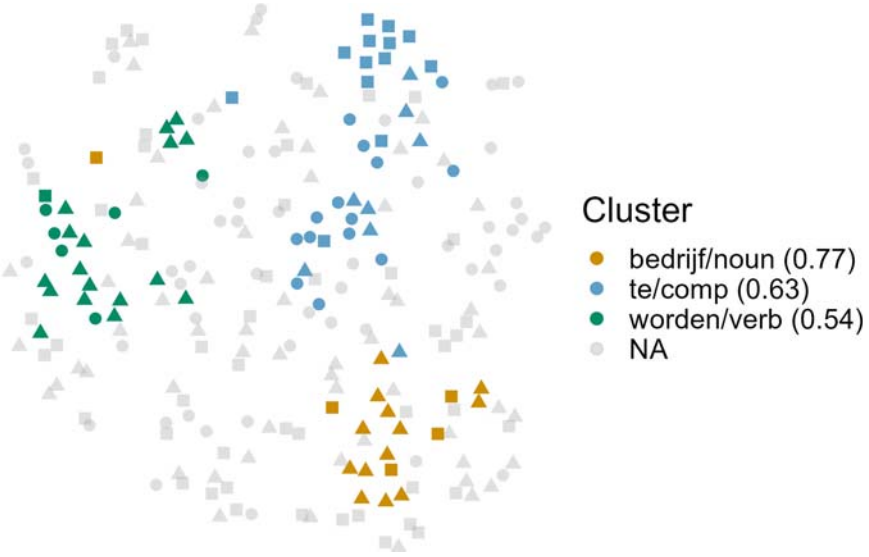


Figure 5.9 Model of *herstructureren* with parameters 3-3.ALL.BOUND.SELECTION.SOCALL.FOC. Circles indicate the transitive, general sense; triangles, the transitive companies-specific sense; and squares, the intransitive (companies-specific) sense

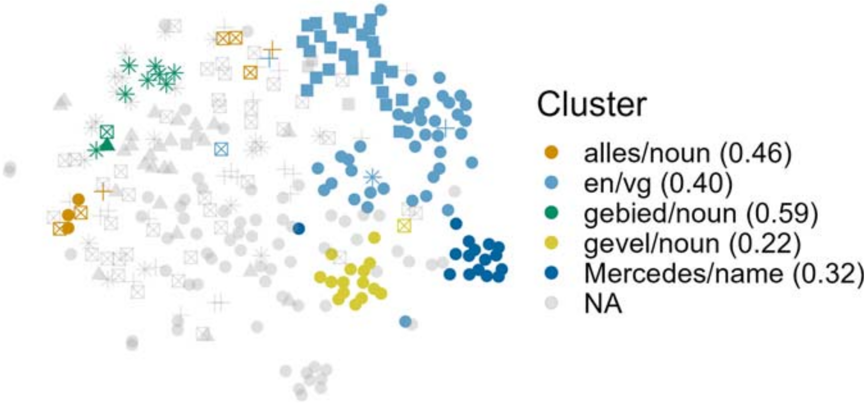


Figure 5.10 Model of *grijs* with parameters 5-5.ALL.BOUND.ASSOCNO.SOCALL.FOC. Circles represent the literal sense; triangles, 'overcast'; squares and crosses, to applications to hair and white-haired people respectively; crossed squares, 'boring'; and asterisks, 'half legal'

of the *bedrijf* cluster, or a syntactic one, as in the lexically instantiated colligation clusters.

Semantic preference clusters, that is, clusters dominated by multiple similar context words, can also be at odds with the manual semantic annotation. This is the case of *grijs* 'grey' tokens co-occurring with names of other colours and with clothing terms, which in a model like the one shown in Figure 5.10 includes tokens co-occurring with *haar* 'hair'. As a result, *grijs* tokens referring to concrete grey objects in general and, specifically, to grey/white hair (a specific sense, since the hair is not necessarily grey), form the light blue cloud on the top right of the figure. Note that, visually, the two senses occupy opposite halves of this cluster: the *haar* tokens (squares) occupy their own space, but the type-level similarity of the context word to the names of colours and clothing terms co-occurring with the light blue circles makes the two groups indistinguishable to HDBSCAN.

A second example is the set of juridical terms in *herroepen*, which means 'to recant' when the object is a statement or opinion, and 'to annul, to void' when it is a law or decision. In our newspaper corpus, it is often used in a broad legal or juridical context. One of the most frequent collocates of *herroepen* within this field is *uitspraak*, which can either mean 'verdict', therefore invoking the 'to void' sense like in (5.10), or 'statement', to which 'to recant' applies, like in (5.11). Unfortunately, the broader context is not clear enough for the models to disambiguate the appropriate meaning of *uitspraak herroepen* in each instance. At the type-level, *uitspraak* is very close to a number of context words of the juridical field, namely *rechtbank* 'court', *vonnis* 'sentence', *veroordeling* 'conviction', and so on. Together, they constitute the semantic preference of the light blue cloud in Figure 5.11, which, similar to the *grijs haar* ‘grey/white hair’ situation above, is visually split between the tokens co-occurring with *uitspraak* and those co-occurring with the rest of the juridical terms.

(5.10) *Het* beroepscomité **herriep** *gisteren de uitspraak* van de licentiecommissie en besliste om KV Mechelen toch zijn licentie te geven. (*De Standaard*, 2002-05-04, Art. 95) ‘*Yesterday the* court of appeal **voided** *the verdict* from the licencing committee and instead decided to grant KV Mechelen a licence.’

(5.11) Onder druk van Commissievoorzitter Prodi heeft Nielson verklaard dat hij verkeerd is geïnterpreteerd, maar hij heeft *zijn uitspraak niet* **herroepen**. (*NRC Handelsblad*, 2001-10-04, Art. 79) ‘Under pressure from committee chairman Prodi, Nielson declared that he had been misinterpreted, but he did *not* **recant** *his statement*.’

The result makes sense: the context words co-occurring with the tokens in the light blue cluster belong to a semantically coherent set and are distributional near neighbours. The problem is that, in the sample, the sense of *uitspraak* that occurs the most is not the juridical one shown in (5.10) but ‘statement’, like in (5.11), representing a different sense of *herroepen* than its juridical siblings. In some models, the two groups are split as different clusters, but in those like the one shown in Figure 5.11, they form a heterogeneous cluster generated by semantic preference. Interestingly, *verklaring* ‘statement’ and *bekentenis* ‘confession’ could be considered part of the same ‘juridical’ semantic field as well, in broad terms. However, they belong to a different frame within the field of legal action—a different stage of the process—and, correspondingly, their type-level vectors are different and tend to represent distinct, homogeneous clusters (the green cloud in the figure).

Finally, the most common situation for semantically heterogeneous clusters is the lack of any dominant context word or even semantic preference. In particular, they emerge as massive clouds in models where a small number of tokens that are very similar to each other—typically idiomatic expressions, but not necessarily—stand out as a cluster, and everything else either belongs to the same massive cluster or is excluded as noise. In many cases there is barely any noise left, while in others HDBSCAN does seem to find a difference between the many, varied tokens in the clouds and those that are left as noise. One such example is the orange cloud of *blik* in Figure 5.12. The small clouds to either side are represented by the co-occurrence of *werpen* ‘to throw’ and *richten* ‘to aim’, which indicate prototypical instances of *blik* ‘gaze’, as shown in (5.12) and (5.13) respectively. Very few tokens are excluded as noise—the patterns they form seem to be too different from the clustered tokens to merge with them, but too infrequent to qualify as a cluster on their own.

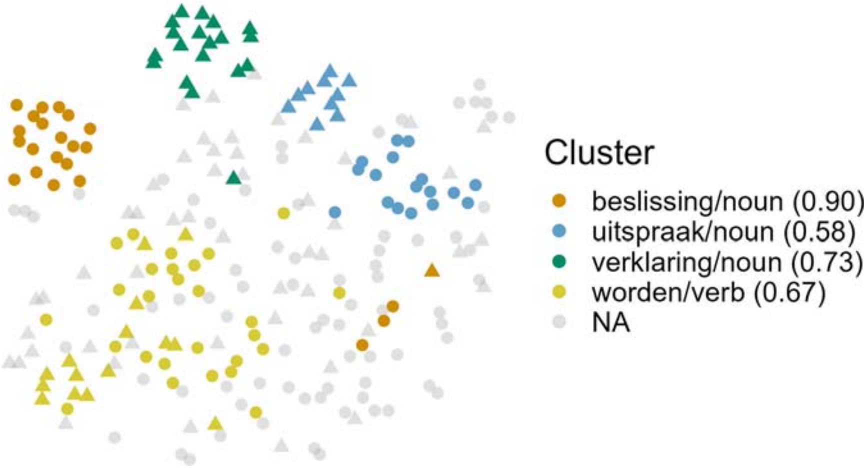


Figure 5.11 Model of *herroepen* with parameters 3-3.ALL.BOUND.SELECTION. SOCALL.FOC. Circles represent 'to void'; triangles, 'to recant'

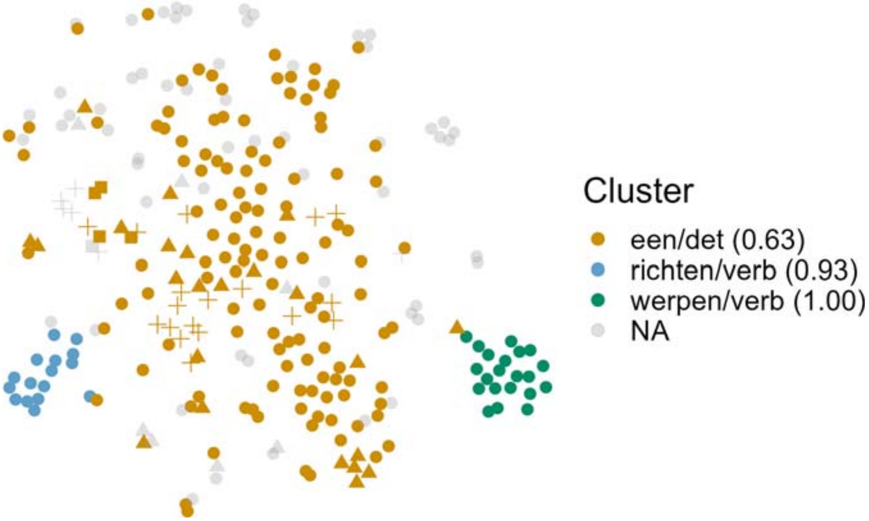


Figure 5.12 Model of *blik* with parameters 5-5.ALL.BOUND.WEIGHT.SOCNAV.5000. For the first homonym, circles represent 'gaze' and triangles, 'view, perspective'; for the second, squares represent 'tin' and crosses, 'made of tin' or 'canned food'

(5.12) Op zaterdag 27 april zwaait de lokale politie van de zone Kortrijk-Kuurne-Lendelede de deuren wijd *open* voor *al wie een* **blik** wil *werpen achter de schermen* van het politiewerk. (*Het Laatste Nieuws*, 2002-04-23, Art. 54) 'On Saturday 27 April the local police of the Kortrijk-Kuurne-Lendelede zone *opens* their doors wide for *all those who* want to *have a* **look** *behind the scenes* of police work.'

(5.13) Maar wat is goed genoeg, zo lijkt Staelens zich *af* te vragen, *haar* **blik** *strak naar beneden gericht*. (*De Volkskrant*, 2003-09-27, Art. 170) 'But what is good enough, Staelens seems to wonder, *her* **gaze** *looking straight down*.'

The orange cluster may seem homogeneous because of the predominance of the circles, but that is simply an effect of the large frequency of the 'gaze' sense, which can also occur in contexts like (5.14). The other sense of the 'gaze' homonym, 'perspective', as shown in (5.15), and of the 'tin' homonym illustrated in (5.16), are also part of this massive heterogeneous cluster. If anything brings these tokens together, other than the fact that they normally do not match the patterns in (5.12) and (5.13), it is that they typically co-occur with *een* 'a, an', *de* 'the', *met* 'with', *op* 'on', and other frequent prepositions, or with more than one at the same time. These frequent, partially overlapping, and not so meaningful patterns bring all those tokens together and, to a degree, set them apart.

(5.14) Totdat *Walsh met een droevige* **blik** *in zijn ogen* vertelt dat hij het moeilijk heeft. (*Het Parool*, 2004-03-02, Art. 121) 'Until *Walsh*, *with a sad* **look** *in his eyes*, says that he's having a hard time.'

(5.15) IMF en Wereldbank liggen al jaren onder vuur wegens *hun* vermeend *eenzijdige* **blik** *op de* ontwikkelingsproblemen *van Afrika*. (*Algemeen Dagblad*, 2001-02-20, Art. 129) 'The IMF and the World Bank have been under attack for years because of *their* allegedly *unilateral* **view** *on the* development issues *in Africa*.'

(5.16) Zijn vader had *een* fabriek waar *voedsel in* **blik** werd gemaakt. (*NRC Handelsblad*, 2003-12-05, Art. 120) 'His father had *a* factory where canned food (lit. '*food in* **tin cans**') was made.'

5.4 One cloud, one sense

The ideal but not frequently attested output of a model is clusters that equal lexicographic senses. Curiously enough, in spite of their rarity, they can be found in clusters with different kinds of collocational patterns. In a few cases we can see clusters characterized by one dominant context word that perfectly matches a sense, or at least its clustered tokens. These are normally fixed expressions, at least to a degree—cases where the collocational pattern might be provided in a dictionary entry for that sense, such as *representatieve staal* 'representative sample'. An interesting example of this phenomenon is shown in Figure 5.13, a model of the noun *schaal* 'scale/dish'. In the plot, the 'scale' homonym is represented by circles ('a range of values; for instance, the scale of Richter, a scale from 1 to 5'), squares ('magnitude; for instance, on a large scale'), and a few

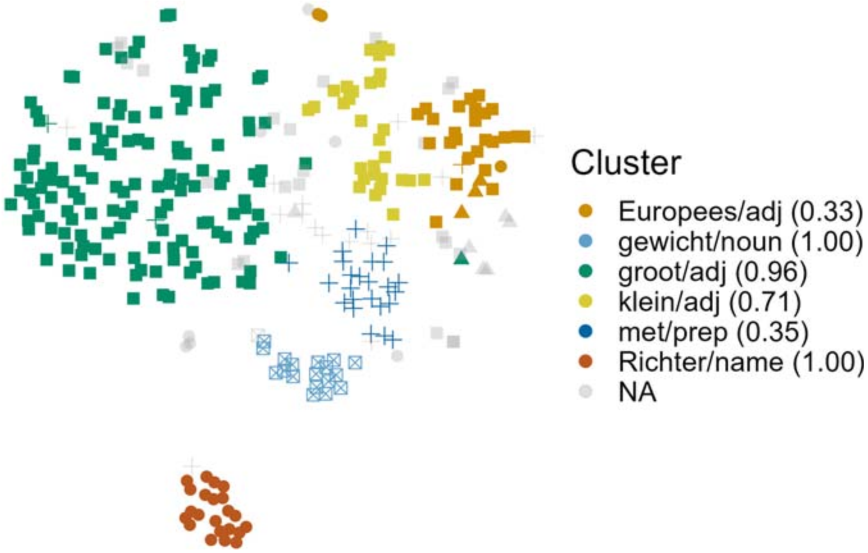


Figure 5.13 Model of *schaal* with parameters 5-5.ALL.NOBOUND.WEIGHT.SOCALL.FOC. Within the 'scale' homonym, circles are 'range'; triangles, 'ratio'; and squares, 'magnitude'; for the 'dish' homonym, crosses represent 'dish' and crossed squares, 'dish of a weighting instrument'

triangles ('ratio; for instance, a scale of 1:20'), whereas the 'dish' homonym is represented by crosses ('shallow wide dish') and crossed squares ('plate of a weighting instrument').

Both the 'range' and the 'dish of a weighting instrument' senses, exemplified in (5.17) and (5.18), have a perfect match (or almost) with an HDBSCAN cluster, represented by a context word with perfect *F*-score, that is, it co-occurs with all the tokens in the cluster and only with them. All the *schaal* tokens co-occurring with *Richter* are grouped in the red cloud and cover almost the full range of attestations of the 'range' sense, and all the tokens co-occurring with *gewicht* 'weight' are grouped in the light blue cloud and cover all the attestations of the 'dish of a weighting instrument' sense. The blue cloud of crosses is also homogeneously dedicated to the 'shallow wide dish' sense, but not dominated by a collocate, and the rest are variably homogeneous clouds representing parts of the 'magnitude' sense.

(5.17) Wenen, Beneden-Oostenrijk en Burgenland zijn dinsdagochtend opgeschrikt door een *aardschok van 4,8 op de* **schaal** *van Richter*. (*Het Nieuwsblad*, 2000-07-12, Art. 4) 'Vienna, Lower Austria and Burgenland have been scared up on Tuesday morning by an *earthquake of 4.8 on the Richter* **scale**.'

(5.18) Daarom is het van belang dat Nederland zich deze week achter de VS heeft geschaard, ook al legt ons land natuurlijk minder *gewicht in de* **schaal** dan *Duitsland in* het *Europese* debat over de al dan niet noodzakelijke toestemming van de Veiligheidsraad voor militaire actie tegen Irak. (*NRC Handelsblad*, 2002-09-07, Art. 160) 'Therefore it is important that the Netherlands has united behind the US this week, even though our country has of course less influence (lit. 'places less *weight on the* **dish of the scale**') than *Germany in* the *European* debate on the potentially necessary permission of the Security Council for military action against Iraq.'

In a way, the phenomenon indicates a fixed, idiomatic expression: a combination of two or more words that fully represents a sense. However, the picture is more nuanced. First, the 'range' sense could potentially occur with context words other than *Richter*. In fact, one of the examples given to the annotators along with the definitions of the different sense tags is *schaal van Celsius* 'Celsius scale', as well as a pattern like the one found in (5.19), one of the orange circles at the top of Figure 5.13. However, in the corpus used for these studies, *Celsius* does not co-occur with *schaal* in a symmetric window of four; moreover, of the 32 tokens of this sense attested in this model, 22 co-occur with *Richter*, three follow the pattern from (5.19), and the rest exhibit less fixed patterns or the infrequent *glijdende schaal* 'slippery slope' construction. The few matching the pattern in (5.19) are more readily clustered with other tokens co-occurring with the preposition *op* 'on', such as (5.20). In other words, in the register of newspapers, the 'range' sense of *schaal* is almost completely exhausted in the *schaal van Richter* 'Richter scale' expression.

(5.19) 'Misschien deelt de computer mij op grond *van statistische* analyses *op een* **schaal** *van 1 tot 12 in* categorie 3', zegt woordvoerder B. Crouwers van de registratiekamer. (*NRC Handelsblad*, 1999-01-09, Art. 10) '"Maybe the computer on the basis *of statistical* analyses *on a* **scale** *of 1 to 12* puts me *in* category 3", says spokesperson B. Crouwers of the registration chamber.'

(5.20) Die stad vormde de opmaat tot de latere *collectieve regelingen op nationale* **schaal**, stellen *de* auteurs, in navolging van socioloog prof. dr. Abram de Swaan. (*De Volkskrant*, 2003-05-03, Art. 253) 'That city was the prelude to the later *collective arrangements at national* level (lit. '*on* a *national* **scale**'), state *the* authors, in accordance with sociologist Prof. Dr. Abram de Swaan.'

Second, the 'dish of a weighting instrument' sense need not be used in the metaphorical expression illustrated in (5.18), but that is indeed the case in our data. Next to *gewicht* 'weight', these tokens also mostly co-occur with *leggen* 'to lie, to place' or, to a lesser degree, with *werpen* 'to throw'. Even in other models, this cluster tends to be built around the co-occurrence with *gewicht*, normally excluding tokens that only co-occur with *leggen*, which do not belong to the same sense in any case. These examples don't disprove the possibility of clouds dominated by a collocate perfectly covering a sense, as long as we keep in mind the characteristics and limitations of the corpus we are studying and the difference between describing 'how a sense is used' and 'how a sense is used *in this particular corpus*'.

Although even more rare, we might be able to find a cluster dominated by a grammatical pattern that matches a lexicographic sense. One clear case is the reflexive sense of *herhalen* 'to repeat', characterized by its co-occurrence with *zich* 'itself' in bag-of-words models without part-of-speech filters and in dependency-based models that select specific syntactic patterns (see Section 3.2), especially if the context words are weighted in function of their pmi. Dependency models that filter based on the length of the syntactic path instead also capture *zich*, but somehow don't build clusters around it. The reflexive sense is represented in the clearest cluster in Figure 5.14, the red cloud of squares at the bottom. In this figure, the code 'REL1' in the name of the model refers to a dependency-based model that uses the

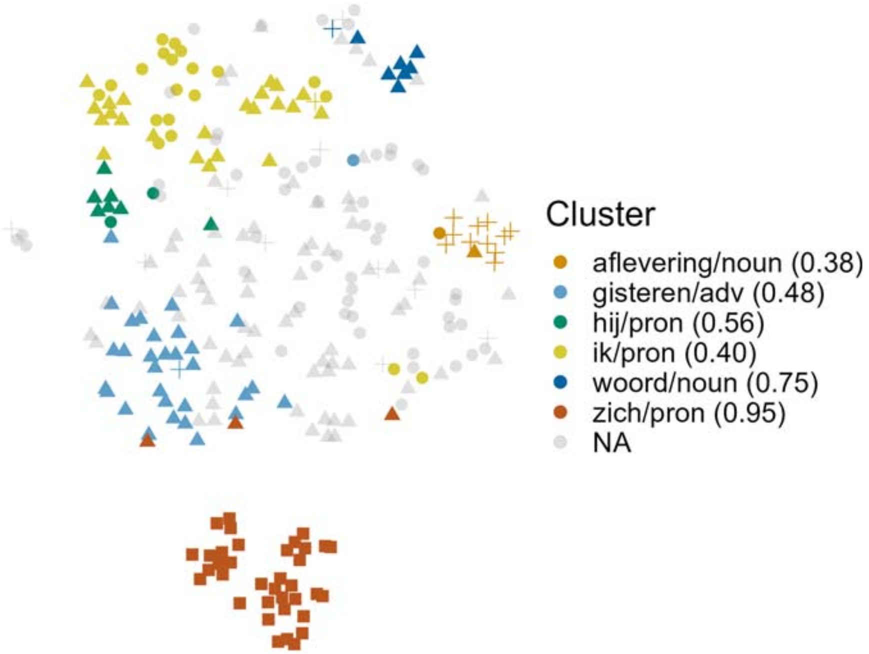


Figure 5.14 Model of *herhalen* with parameters REL1.SELECTION.SOCALL.FOC. Circles are 'to do again'; triangles, 'to say again'; squares, '(reflexive) to happen again'; and crosses, 'to broadcast again'

most restrictive filter based on syntactic patterns: for a verb, this means that context words are limited to 'direct objects, active and passive subjects (with up to two modals for the active one); reflexive complements, and prepositions depending directly on the target' (Montes 2021a: 33). Looking closely, we can see that it is made of two separated 'halves': a small one on the left, in which the tokens also co-occur with *geschiedenis* 'history', and a bigger one on the right, where they do not. This particular model is very restrictive: it normally captures only one or two context words per token, which is all that we need to capture this particular sense. We expected this kind of output in other lemmas with purely reflexive senses as well, but it is not easy to achieve. For example, *diskwalificeren* 'to disqualify' was annotated with three senses: two transitive ones and a reflexive one. One of the transitive senses is restricted to the contexts of sports, whereas the other two senses apply to any other context, typically politics. In this situation, the reflexive sense is very infrequent and mostly absorbed within the transitive sense that matches it semantically, that is, the non-sports-related sense.

In some cases, senses can also be completely clustered by groups of similar context words. One of these cases was already discussed in the context of *schaal* tokens: in models that exclude *Richter* because of its part-of-speech *name* (i.e. proper noun), the tokens co-occurring with it can alternatively be grouped by *kracht* 'power', *aardbeving* 'earthquake', and related context words. As in the case of *Richter* as a dominating collocate, the semantic field of earthquakes is not part of the definition of the 'range' sense of *schaal*, but it is the dominating semantic pattern within the corpus under study. Another example is found in *haken*, where the 'to make someone trip' sense is characterized by a variety of football-related terms (*strafschop* 'penalty kick', *penalty*, *scheidsrechter* 'referee', etc.), and the very infrequent 'crochet' sense, by *breien* 'to knit', *naaien* 'to sew', *hobby*, and similar words. They are represented as a cloud of dark blue squares and one of light blue crossed squares in Figure 5.15 respectively. As indicated by the name of the dark blue cluster, the passive auxiliary *worden* is also characteristic of the 'to make someone trip' cluster and very rarely occurs outside of it: here, lexically instantiated colligation is working together with the clear semantic preference of the cloud.

Finally, we might not expect to find near-open choice clusters that perfectly match meanings, but they do occur. Such is the case of the model of *huldigen* shown in Figure 5.16. Like with other transitive verbs, the senses of this lemma are characterized by the kinds of direct objects they can take. When the direct object of *huldigen* is an idea or opinion, it means 'to hold, to believe': in our sample, typical cases include *principe* 'principle', *standpunt* 'point of view', and *opvatting* 'opinion' (see examples (5.21) through (5.23)). The three of them are near neighbours at type level, but frequent enough to generate their own clouds in most models, like in Figure 5.16 (compare Figure 5.6, where they work together to form one larger cloud). In other contexts, *huldigen* means 'to honour, to pay homage', and the role

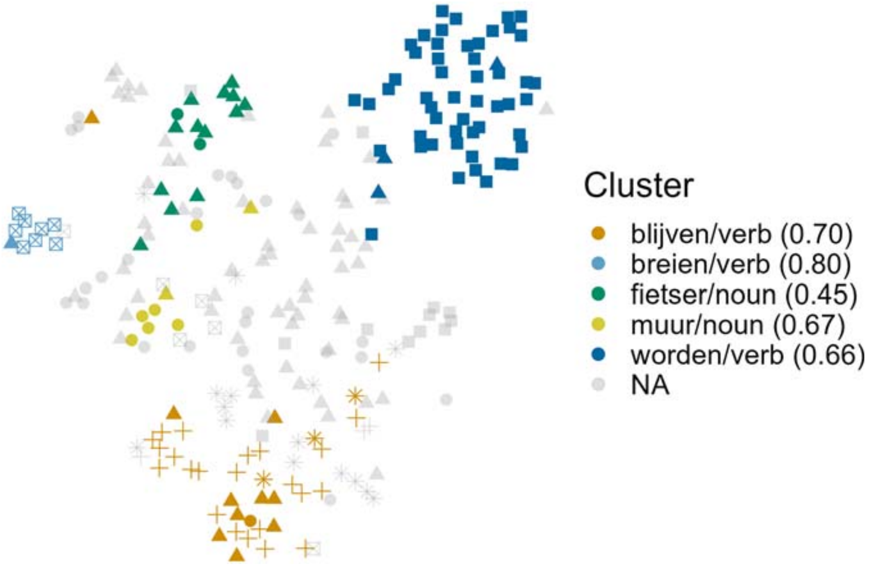


Figure 5.15 Model of *haken* with parameters 10-10.LEX.BOUND.SELECTION.SOCNAV.FOC. Circles and triangles represent the transitive and intransitive literal 'to hook'; crosses represent the figurative (intransitive) sense; filled squares represent 'to make someone trip'; crossed squares, 'to crochet'; and asterisks, 'to strive for' (with *naar* 'towards')

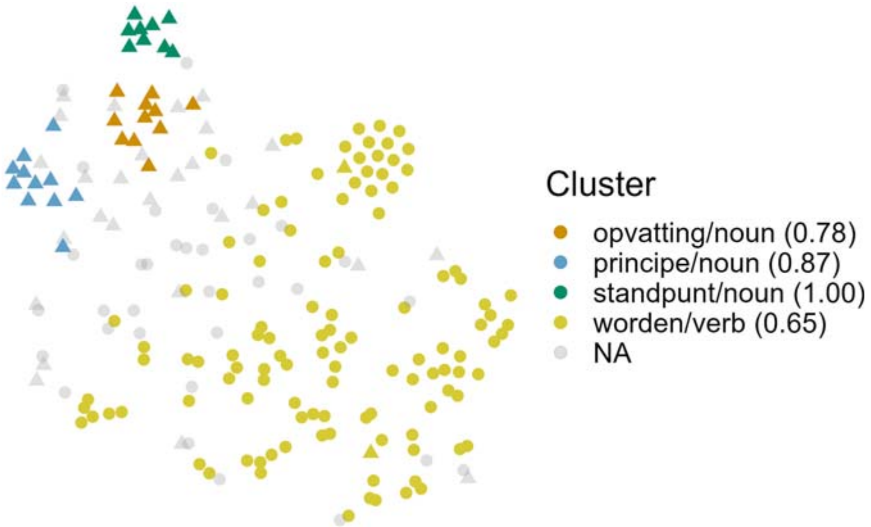


Figure 5.16 Model of *huldigen* with parameters 3-3.LEX.NOBOUND.SELECTION.SOCALL.FOC. Circles represent 'to believe, to hold (an opinion)'; triangles, 'to honour'

of the patient is normally filled by human beings (see examples (5.24) and (5.25)). In practice, the variety of nouns that can take this place is much larger than for 'to believe', and as a result, the clusters that cover 'to honour' are less compact and defined than the clusters representing the 'to hold, to believe' sense. And yet, the cluster shown in yellow in Figure 5.16 almost perfectly represents the 'to honour' sense. How is that possible?

(5.21) Jacques: 'Voor het *eerst* **huldigen** we het *principe* dat de vervuiler betaalt.' (*De Morgen*, 1999-03-10, Art. 12) 'Jacques: "For the *first* time we **uphold** the *principle* that polluters must pay."'

(5.22) De regering in Washington **huldigt** het *standpunt* dat volgens Amerikaans recht de vader beslist over het domicilie van zijn minderjarige zoon. (*NRC Handelsblad*, 2000-04-03, Art. 97) 'The government in Washington **holds** the *view* that according to American law fathers decide on the primary residence of their underage sons.'

(5.23) …de objectieve stand van zaken in de buitenwereld zou kunnen *weerspiegelen*. Rorty **huldigde** *voortaan* de *opvatting* dat waarheid synoniem is voor wat goed is voor ons. (*De Standaard*, 2003-01-09, Art. 93) '… would *reflect* the objective state of affairs in the outside world. *Ever since* Rorty has **held** the *opinion* that the truth is a synonym for what is good for us.'

(5.24) 'Elk *jaar* **huldigen** wij onze *kampioenen* en sinds enkele jaren richten we een jeugdkampioenschap in', zegt voorzitter Eddy Vermoortele. (*Het Laatste Nieuws*, 2003-04-15, Art. 121) '"Every *year* we **honour** our *champions* and for a few years we've been organizing a youth championship," says chairman Eddy Vermoortele.'

(5.25) Langs de versierde straten zijn we naar de kerk gereden en na de plechtigheid hebben we Karel nog **gehuldigd** in *feestzaal* Santro. Hij is nog een heel kranige man. (*Het Laatste Nieuws*, 2003-07-18, Art. 256) 'We drove through the ornate streets towards the church and after the ceremony we **honoured** Karel at the *party hall* Santro. He is still a spry man.'

One of the factors playing a role in the layout of this model is that the co-occurrences with *principe* 'principle', *standpunt* 'point of view', and *opvatting* 'opinion' exhaust about half the attestations of the 'to believe' sense. The rest of the tokens of this sense are too varied and typically fall into noise. The variety within the 'to honour' sense cannot compete against the stark differences between these clusters and everything else. Nonetheless, there is some form of structure within the sense that differentiates it from the equally varied remaining tokens of 'to believe', and that is a family resemblance structure. No single semantic field suffices to cover the variety of contexts in which *huldigen* 'to honour' occurs in our sample: instead, we find different aspects and variations of its prototypical reference, namely ceremonies organized by sports- and city organizations in public places, in honour of successful athletes. In order to get a better picture of the syntagmatic relationships between the context words within the cluster, we can represent them in a network, shown in Figure 5.17. Each node represents one of the 150 most frequent context words co-occurring with tokens from the yellow cloud in Figure 5.16, and is connected to each of the context words with which it co-occurs in a token of that cluster. The thickness of the edges represents the frequency with which the context words co-occur within the sample; the size of the nodes summarizes that frequency, and the size of the label roughly represents the frequency of the context word among the tokens in the cluster.

The most frequent context word is the passive auxiliary *worden*: it is the only context word captured in the tokens of the dense core on the upper-right corner of the cloud, and co-occurs with about half the tokens of this cluster. A number of different, less frequent context words partially co-occur with it, such as *kampioen* 'champion', *stadhuis* 'city hall', and *sportsraad* 'sports council'. They subsequently generate their own productive branches in the family resemblance network. Semantically and distributionally, the context words plotted in this network belong to different, loosely related fields, such as sports (*kampioen*, *winnaar* 'winner', *sportsraad*), town administration (*stadsbestuur* or *gemeentebestuur* 'city administration'), and temporal expressions (*jaar* 'year', *weekend*). Each of them corresponds to an aspect of the situation that *huldigen* 'to honour' refers to, and

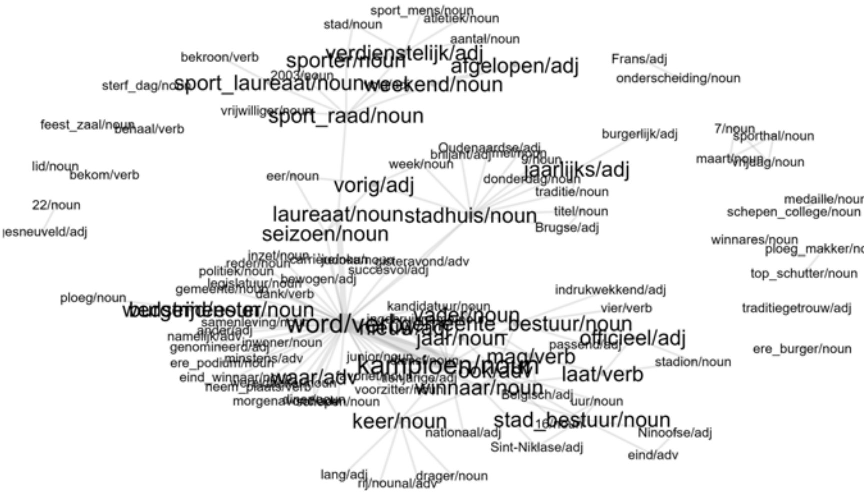


Figure 5.17 Network of context words of the *huldigen* 'to honour' cluster

each token might make one or two of them explicit. In short, the predominance of the passive auxiliary *worden* (lexically instantiated colligation), the presence of unified semantic fields (multiple semantic preferences), and the family resemblance among tokens, resulting from an intricate network of co-occurrences, work together to model the subtle, complex semantic structure of *huldigen* 'to honour'.

5.5 Prototypical contexts

The most frequent phenomenon among the most compact and well-defined clouds is a cluster dominated by one context word or group of co-occurring context words that represents a typical context of a sense. It may be the prototypical context, if the rest of the sense is discarded as noise or spread around less clear clusters, but we might also find multiple clusters representing different typical contexts of the same sense. Neither t-SNE nor HDBSCAN can tell whether one of these contexts is more central than the other, at least not in the same way we would expect from prototype theory. Denser areas of tokens, as perceived by HDBSCAN, are those where many tokens are very similar to each other. The more tokens are similar, and the more similar they are, the denser the area, but only the first point is relevant when talking of prototypicality in psycholinguistic terms. Typical clusters can be characterized by lexical collocations, lexically instantiated colligations, or semantic preference, but no such cases were found among near-open choice clusters.

One of the clearest examples of this phenomenon in collocation clusters is found in *heffen* 'to levy/to lift', whose typical objects are also characteristic of its two main senses (see Figure 5.18). On the one hand, the 'to levy' sense occurs mostly with *belasting* 'tax', *tol* 'toll' (typical of the Netherlandic sources, since tolls are not levied in Flanders), and *accijns* 'excise', as shown in (5.26) through (5.28). Their frequencies are large enough to form three distinct clusters, which tend to merge in the higher levels of the HDBSCAN hierarchy, that is, they are closer to each other than to the clusters of the other sense. On the other hand, the 'to lift' sense occurs with *glas* 'glass', where the final expression *een glas(je) heffen op* takes the metonymical meaning 'to give a toast to' (see (5.29)), and with the body-parts *hand*, *arm*, and *vinger* 'finger', which might take other metonymical meanings. The latter group does not correspond to the category of collocation clusters but is instead defined by semantic preference.

(5.26) Op het inkomen boven *die* drie miljoen *gulden wil* De *Waal honderd procent belasting* **heffen**. (*Het Parool*, 2001-05-02, Art. 99) 'De *Waal wants* to **levy** a *one hundred percent tax* on all incomes above *those* three million *guilders*.'

(5.27) Mobiliteitsproblemen, rekeningrijden, op een andere manier het *gebruik* van de *weg belasten*, kilometers *tellen*, *tol* **heffen**—de *mogelijkheden om* de ingebouwde chip *te* benutten zijn vrijwel onbeperkt. (*NRC Handelsblad*, 1999-10-02, Art. 31) 'Mobility problems, road pricing, *taxing* the *use* of *roads* in a different way, *counting* kilometres, **levying** *taxes*—the *possibilities to* utilize the built-in chip are almost unlimited.'

(5.28) ... in landen als Groot-Brittannië (waar de accijnzen op 742 euro per 1.000 liter liggen), Italië *en Duitsland* (*die beide accijnzen boven* de *400 euro* **heffen**) komt de *harmonisering* ten goede van de transportsector. (*De Morgen*, 2002-07-25, Art. 104) '... in countries like Great Britain (where excise duties are at 742 euros per 1,000 liters), Italy *and Germany* (*both of which* **levy** *excise duties above 400 euros*) the transport sector benefits from the *harmonization*.'

(5.29) Nog *twaalf* andere deelnemers *konden maandagavond het glas* **heffen** *op* de *hoogste winst*. (*De Standaard*, 2004-10-20, Art. 150) '*On Monday night* another *twelve* participants *could* **raise** *their glasses to* the *highest profit*.'

As we can see in Figure 5.18, the model is very successful at separating the two senses and the clusters are semantically homogeneous: the most relevant collocates of *heffen* are distinctive of one or the other of its senses. Crucially, no single cluster is even close to covering a full sense; instead, each of them represents a prototypical pattern that stands out due to its frequency, internal coherence, and distinctiveness. It seems reasonable to map the clusters to prototypical patterns because of their frequency and distinctiveness, but we should be careful about how we apply the results of the modelling to this kind of semantic analysis. From the perspective of prototype theory, a feature of a category is more central if it is more frequent, that is, it is shared by more members, while a member is more central if it exhibits more of the defining features of the categories. As such, within the 'to levy' sense, the *belasting heffen* 'to levy taxes' pattern is the most central, and tokens instantiating such a pattern will be more central. In contrast, HDBSCAN prioritizes dense areas, that is, groups of tokens that are very similar to each other. Thus, membership probabilities, which we might be tempted to use as proxy for centrality, indicate internal consistency, lack of variation. From such a perspective, given that *belasting heffen* 'to levy taxes' is more frequent and applies to a wider variety of contexts than the other two patterns of 'to levy', its area is less dense, and its tokens have lower membership probabilities within a compound of 'to levy' clusters. In other words, the models can offer us typical patterns of a lemma and of its senses

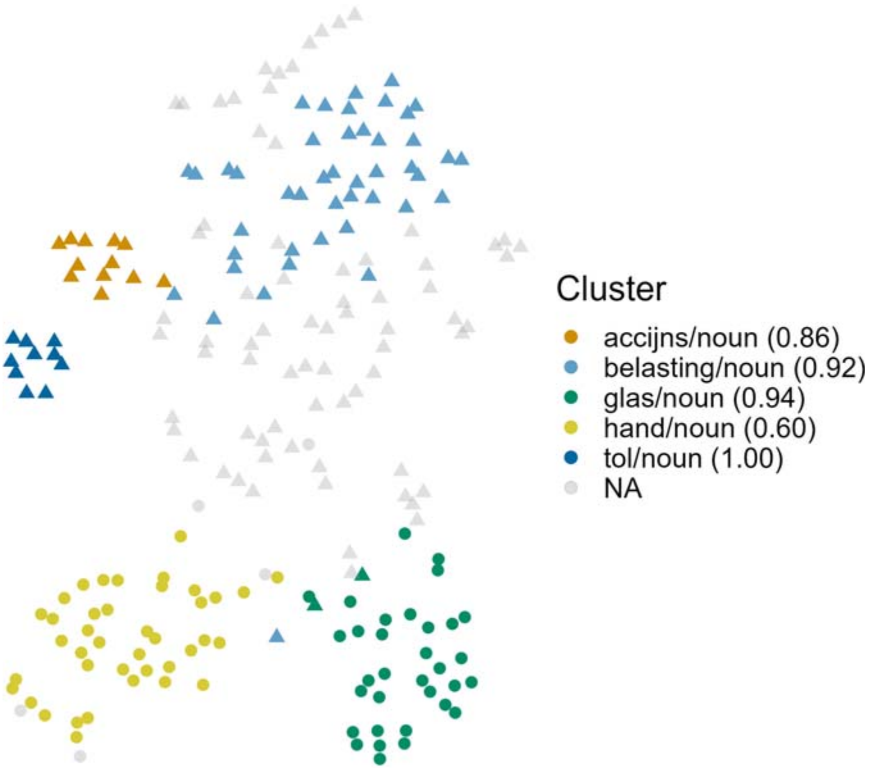


Figure 5.18 Model of *heffen* with parameters 10-10.ALL.BOUND.WEIGHT.SOCNAV.FOC. Circles are 'to lift'; triangles are 'to levy'

and tell us how distinctive they are from each other and how much internal variation they present. Beyond this information, they don't map in a straightforward manner to our understanding of prototypicality.

It must be noted that clusters defined by collocations may not be just characterized by one single context word, but by multiple partially co-occurring context words. A clear example is *hachelijk* 'dangerous/critical', where both senses are characterized by prototypical contexts, exemplified in (5.30) through (5.35): *onderneming* 'undertaking', *zaak* 'business', and *avontuur* 'adventure' for the 'dangerous, risky' sense, *moment*, *situatie* 'situation', and *positie* 'position' for the 'critical, hazardous' sense. A model is shown in Figure 5.19. These six frequent context words are paradigmatic alternatives of each other, all taking the slot of the modified noun, that is, the entity characterized as dangerous or critical. However, unlike its very near type-level neighbour *situatie*, *positie* also co-occurs with *bevrijden* 'to free' (and *uit* 'from') and, additionally, with *brandweer* 'firefighter', typically in Belgian contexts. The frequency of these co-occurrences in the sample, next to the type-level dissimilarity between these three lexical items, splits the co-occurrences with *positie* in three clusters based on these combinations. Concretely, the green cluster in Figure 5.19 groups the occurrences where *positie* co-occurs with both *bevrijden* and *brandweer*, the light blue one groups those where it occurs with *bevrijden* but not *brandweer*, and the red one those where neither *bevrijden* nor *brandweer* occur.

(5.30) Het is geen gewaagde stelling dat de deelname van de LPF aan de *regering een* **hachelijke** *onderneming blijft*. (*De Volkskrant*, 2002-08-05, Art. 46) 'It is not a bold statement that the participation of the LPF in the *government remains a* **risky** *undertaking*.'

(5.31) Daar baseerden de media zich op slechts één bron, en elke journalist weet dat *dat een* **hachelijke** *zaak is*. (*De Volkskrant*, 2004-05-05, Art. 42) 'The media relied on only one source, and every journalist knows that *that is a* **dangerous** *thing to do*.'

(5.32) ... met storm opzij is het *inhalen* van *een vrachtwagen een* **hachelijk** *avontuur*... (*Het Parool*, 2000-03-17, Art. 34) '... under sidewind conditions *overtaking a truck* is *a* **risky** *adventure*...'

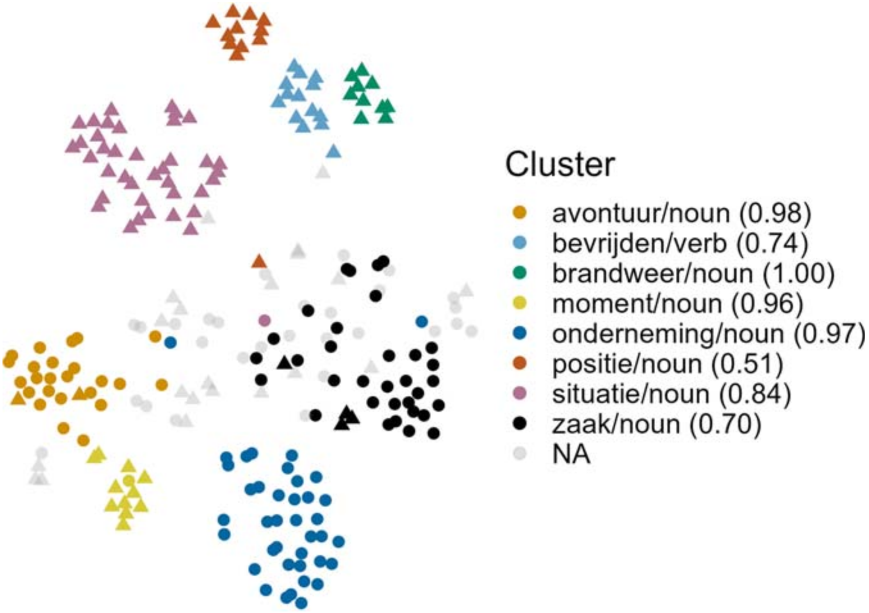


Figure 5.19 Model of *hachelijk* with parameters 5-5.ALL.BOUND.WEIGHT.SOCALL.FOC. Circles are 'dangerous, risky'; triangles are 'critical, hazardous'

(5.33) *Kortrijk beleefde enkele* **hachelijke** *momenten* tegen *Brussels*, dat in zijn ondiep bad bewees zijn vierde plaats in de play-offs waard te zijn. (*Het Laatste Nieuws*, 2001-05-14, Art. 375) '*Kortrijk experienced some* **critical** *moments* against *Brussels*, who in their shallow pool proved to be worthy of their fourth place in the play-offs.'

(5.34) Kort maar krachtig staat er: 'De **hachelijke** *situatie* van *Palestina is* vooral een interne aangelegenheid, hoewel de bezetting en de confrontatie met Israël er de context voor schept.' (*De Standaard*, 2004-10-02, Art. 162) 'Short but powerful, it reads: "The **critical** *situation* in *Palestine is* mostly an internal matter, even though the occupation and the confrontation with Israel create the context for it."'

(5.35) Zij toont knappe filmpjes, *opgenomen vanuit* de **hachelijke** *positie* van *een* deltavlieger... (*De Morgen*, 1999-06-07, Art. 126) 'She shows outstanding videos, *taken from* the **hazardous** *position* of *a* hang glider...'

The model does not give us information about the relative centrality of the three *positie* clusters. They result from the combination of three features, and each cluster exhibits a different degree of membership based on how many of these overlapping features it co-occurs with. At the same time, they have a distinctive regional distribution. Based on this data, we might say that a prototypical context of *hachelijke positie* in Flanders is a situation in which fire-fighters free someone/something from such a position, while this core is not present, or at least not nearly as relevant, in the Netherlandic data. We might also say that the same situation is not typical of *hachelijke situatie*, and this therefore presents a (local) distributional difference between two types that otherwise, at corpus level, are near neighbours.

Prototypical contexts can also be identified by means of lexically instantiated colligations, as in the case of *herinneren*. This verb has two main senses characterized by well-defined constructions: either an intransitive construction co-occurring with the preposition *aan*, meaning 'to remind', or a reflexive construction meaning 'to remember'; a third, transitive sense is also attested but very infrequently. This lemma is sometimes rendered as three equally sized clouds, as shown in Figure 5.20: the orange cluster is characterized by the preposition *aan* (see (5.36)), the green one by the subject and reflexive first person pronouns *ik* and *me* (see (5.37)), and the yellow one by the third person reflexive pronoun *zich* (see (5.38)). A smaller group of tokens co-occurring with *eraan*, a compound of the particle *er* and *aan* (see example (5.39), where it works as a placeholder to connect the preposition to a subordinate clause), may form its own cloud, like the light blue one in Figure 5.20, or be absorbed by one of the larger ones.

(5.36) Vinocur **herinnert** *aan* een tekening van Plantu in L'Express. (*Het Parool*, 2002-05-18, Art. 101) 'Vinocur **reminds** [the spectator] *of* a drawing by Plantu in L'Express.'

(5.37) *Ik* **herinner** *me* een *concert waarop hij* hevig gesticulerend applaus in ontvangst kwam nemen. (*Het Parool*, 2003-11-14, Art. 79) '*I* **remember** a *concert in which he* received a round of overwhelming applause.'

(5.38) 'Het was *die dag bloedheet*', **herinnert** de *atlete uit Sint-Andries zich nog levendig*. (*Het Nieuwsblad*, 2001-08-08, Art. 192) '"It was *scorching hot that day*", **remembers** the *athlete from Sint-Andries vividly*.'

(5.39) In *zijn voorwoord* **herinnert** Manara *eraan dat* deze *meisjes* in hun *tijd vaak* met toegeknepen oogjes werden aanschouwd. (*De Morgen*, 2001-11-10, Art. 40) 'In *his preface* Manara **reminds** [the reader] *that* back in their *time* these *girls* were *often* looked at with squinted eyes.'

As the shape coding in the plot indicates, the clusters are semantically homogeneous, with the exception of three tokens in the first person cluster also co-occurring with *aan*, and one instantiating *ik zal herinnerd worden als* 'I will be remembered as'. Such colligation-driven homogeneity is possible because these

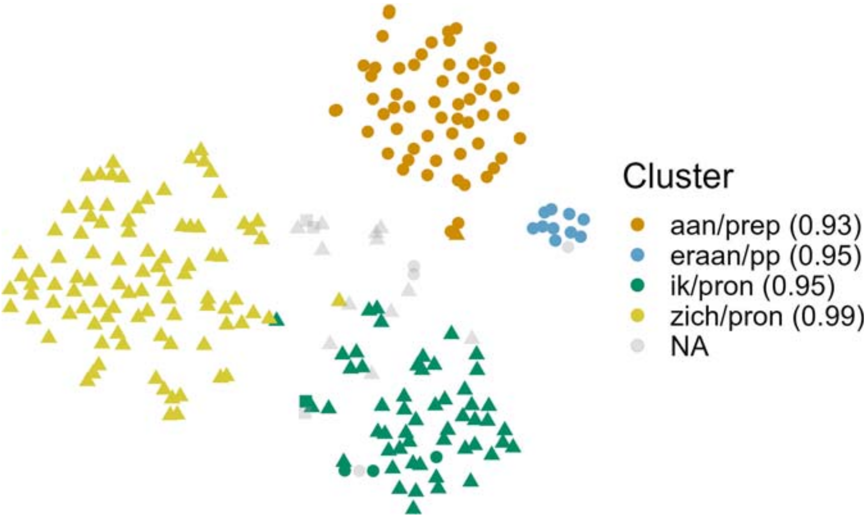


Figure 5.20 Model of *herinneren* with parameters 10-10.ALL.BOUND.WEIGHT.SOCNAV.5000. Circles indicate 'to remind' (with aan); triangles, '(reflexive) to remember'; and (the very few) squares, '(trans.) to remember'

function words are perfect cues for the senses. The rest of the co-occurring context words do not make a difference: they are not strong enough, in the face of these pronouns and prepositions, to originate further salient structure. Nonetheless, both the *aan* and *eraan* clusters belong to the 'to remind' sense, and both pronoun-based clusters belong to 'to remember'. Thus, what these lexically instantiated colligation clusters represent is a typical or salient pattern within each sense.

Prototypical clusters can also be defined by semantically similar infrequent context words, that is, semantic preference. In Figure 5.10, for example, the dark blue cloud is represented by cars, mostly indicated by the co-occurrence with *Mercedes* and *Opel*, among other brands. A typical semantic group attested in different lemmas is culinary, as in the cases of *schaal* 'dish'—the blue cloud of crosses in Figure 5.13—and *heet* 'hot', the red cloud of mostly circles in Figure 5.21. In the case of *heet*, almost all the tokens co-occurring in this cluster refer to literally hot foods and drinks, although the full expression might be idiomatic, like in (5.40), and only a few of them belong to the much less frequent sense 'spicy'. In other models, the tokens co-occurring with *soep* 'soup' and/or those co-occurring with *water* form separate clusters.

(5.40) Hoogstwaarschijnlijk zal Poetin Ruslands afgeknapte westerse partners discreet laten weten dat zodra hij eenmaal in het Kremlin *zit*, *de soep minder* **heet** *gegeten zal worden*. (*De Volkskrant*, 1999-12-21, Art. 22) 'Most probably Putin will discretely let Russia's former western allies know that as soon as he *is* in the Kremlin, things will look up (lit. "*the soup will be eaten less* **hot**").'

In addition, *aardappel* 'potato' is at type-level a near neighbour of the context words in the semantic group of food, but it still tends to form its own cluster, like the orange cloud in Figure 5.21. This is due both to its frequency and the distinctiveness of its larger context, like, for instance, the co-occurrence with *doorschuiven* 'to pass on'. Like other expressions annotated with the 'hot to the touch' sense (circles in the figure), including *hete hangijzer* 'hot issue, lit. hot iron pot hanger' in yellow and *hete adem (in de nek)* 'hot breath (on the neck)' in light blue, *hete aardappel* 'hot potato' is used metaphorically. In the strict combination of adjective and noun, the meaning of *heet* proper is still 'hot to the touch': it is the combination itself that is then metaphorized (for a discussion see Geeraerts 2003). The context words themselves are frequent and distinctive enough to generate clusters of their own with the tokens that co-occur with them, but *aardappel* 'potato' tends to stick close to the culinary cluster or even merge with it.

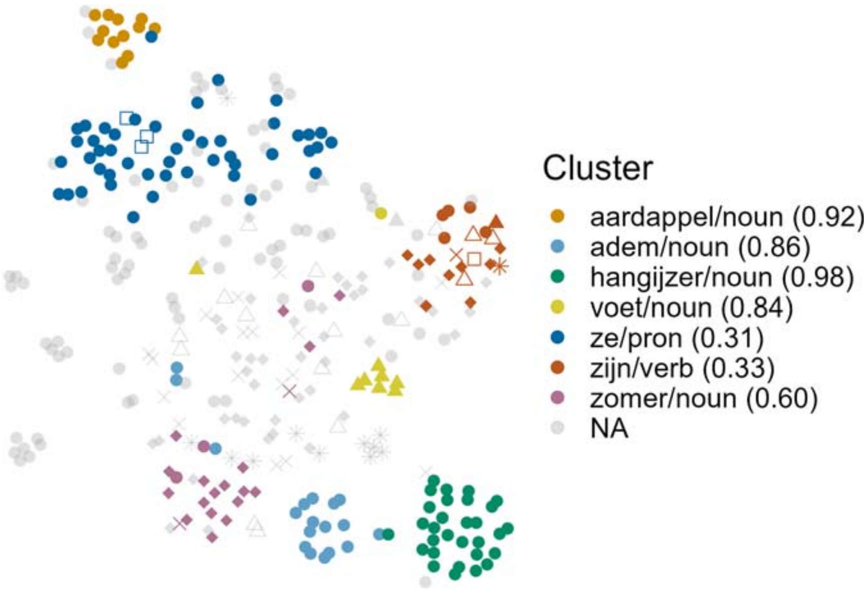


Figure 5.21 Model of *heet* with parameters 5-5.ALL.BOUND.ASSOCNO.SOCALL.FOC. Among the literal senses, circles, filled triangles, and filled diamonds represent tactile, weather, and body senses; empty squares and triangles represent 'spicy' and 'attractive' respectively; crosses represent 'conflictive'; and asterisks, 'popular or new'

5.6 Semantic profiling

Clusters that represent typical contexts might also be profiling specific dimensions of the senses. This is not extremely frequent and requires an extra layer of interpretation, but it is an additional explanation to some of the clustering solutions. As in the case of general prototypical contexts, it can occur as a collocation cluster, as lexically instantiated colligation or as semantic preference, but has not been attested in near-open choice clusters.

One example defined by a collocation is given by the 'substance' meaning of *stof*, represented as circles in Figure 5.22. Within this sense, we tend to find clusters dominated by *gevaarlijk* 'dangerous', *schadelijk* 'harmful' (which also attracts *kankerwekkend* 'carcinogenic'), and *giftig* 'poisonous' (which often attracts *chemisch* 'chemical'). These dominant context words are nearest neighbours at type-level, and the clusters they govern belong to the same branch in the HDBSCAN hierarchy. However, we can find additional information among the context words that co-occur with them, suggesting that frequency is not the only factor separating the clusters. Concretely, the tokens in the cluster dominated by *schadelijk* tend to focus on the environment and composition of substances, as indicated by the co-occurrence with *uitstoot* 'emissions', *lucht* 'air', *stank* 'stench',

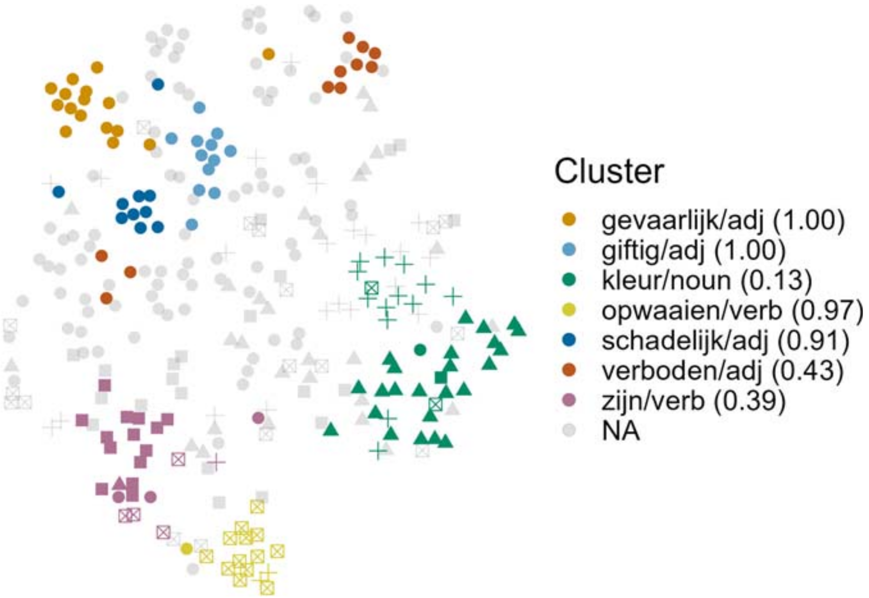


Figure 5.22 Model of *stof* with parameters 5-5.LEX.BOUND.SELECTION.SOCALL.FOC. Within the first homonym, circles are 'substance'; triangles, 'fabric'; filled squares, 'topic, material'. For the second, crosses are literal 'dust' and crossed squares idiomatic expressions

and *bevatten* 'to contain'; meanwhile, those in the cluster dominated by *giftig* focus on the context of drugs or profile the liberation of substances, with context words such as *vormen* 'to form', *vrijkomen* 'to be released', and *drugsgebruik* 'drug use'. The clusters are not distinguished by their meaning as they would be coded in a dictionary entry, but by semantic dimensions that are highlighted in some contexts and hidden in others, yet always latent. This effect of the less frequent context words is one of the consequences of less restrictive models: at some levels of analysis, one word (*gevaarlijk*, *schadelijk*...) might be enough to disambiguate the target, but this extra information added by the less frequent context words enriches our understanding of how the words are actually used. It is also contextualized information: not just about how *stof* 'substance' is used, but how it is used when in combination with certain frequent collocates.

Clusters defined by lexically instantiated colligations can also represent a typical context that highlights a specific dimension of the sense of the target. One such case is found in the 'horde' sense of *horde*, whose most salient collocates in this sample are *toerist* 'tourist' and *journalist*. The two collocates are quite similar to each other at type-level, but the rest of the context words in their clusters point towards a different dimension of the 'horde' sense: hordes of journalists, photographers, and fans (other nouns present in the same cluster) will surround and follow celebrities, as suggested by the co-occurrence of *omringen* 'to surround', *opwachten* 'to wait', and *achtervolgen* 'to chase', among others. In contrast, hordes of tourists will instead flood and move around in the city, with words such as *toestromen* 'to flood' and *stad* 'city'. As it stands, the situation is equivalent to the case of *stof* 'substance' described above. However, in the models that capture function words like the one shown in Figure 5.23, the profiling in these clusters is strengthened by lexically instantiated colligations. The *journalist* cluster (in orange) is dominated by the preposition *door*, which signals explicit agents in passive constructions; the passive auxiliary *worden* also occurs, albeit less frequently. Meanwhile, the *toerist* cluster (in red) includes tokens co-occurring with *naar* 'towards'. The prepositions are coherent with the dimensions of 'horde' highlighted by each of the clusters, that is, aggressivity and flow respectively. Interestingly, they don't co-occur with all the tokens that also co-occur with *journalist* and *toerist* respectively; instead, the nouns and prepositions complement each other.

Finally, profiling prototypical clusters can also be syntagmatically defined by semantic preference, as in the case of *geldig* 'valid'. This adjective can relate to a legal or regulated acceptability, which is its most frequent sense in the sample, or may have a broader application, to entities like *redenering* 'reasoning'. By definition, and like for most of the lemmas studied here, each sense matches some form of semantic preference. In addition, models of this lemma reveal semantic

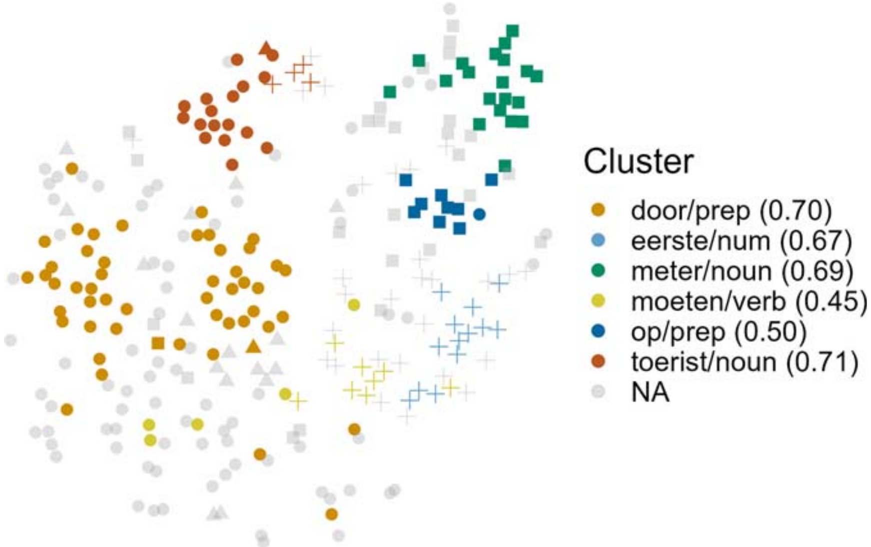


Figure 5.23 Model of *horde* with parameters 5-5.all.bound.selection.socall.foc. Within the 'horde' homonym, circles indicate human members and triangles, nonhuman members; within the 'hurdle' homonym, squares show the literal sense and crosses the metaphorical one

preference patterns within the frequent, specific sense, each of which, in turns, highlights a different dimension of this sense. These patterns may be only identified as areas in the t-SNE plots or, in models like the one shown in Figure 5.24, as clouds. The green cloud is characterized by context words such as *rijbewijs* 'driving licence', *paspoort* 'passport', and other forms of identification, as well as verbs like *voorleggen* 'to present', *hebben* 'to have', and *bezitten* 'to possess'. In other words, it represents contexts in which someone has to demonstrate possession of a valid identification document, as shown in (5.41). The light blue and the yellow clouds, on the other hand, co-occur with other kinds of documents (*ticket*, *abonnement* 'subscription'), *euro*, the preposition *tot* 'until', and times (*maand* 'month', *jaar* 'year', numbers, etc.). In this case, the price of the documents and the duration of their validity are more salient, as illustrated in (5.42).

(5.41) Aan de incheckbalie *kon* de *Somaliër echter* geen **geldige** *papieren voorleggen*. (*Het Laatste Nieuws*, 2001-08-24, Art. 64) '*But* the *Somali could* not *show* any **valid** *papers* at the check-in desk.'

(5.42) Klanten van Kunst In Huis zijn bovendien zeker van variatie: wie lid is, kan elke maand een ander werk uitkiezen, het *abonnement blijft* een leven *lang* **geldig** en de *maandelijkse huurprijs* van 250 *frank is ook niet* bepaald hoog te noemen. (*De Standaard*, 1999-05-29, Art. 41) 'Moreover, customers of Kunst In Huis (lit. 'Art At Home') are guaranteed variation: members can choose a different work each month; the *subscription remains* **valid** for a life*time* and the *monthly fee* of 250 *franks is not* particularly high *either*.'

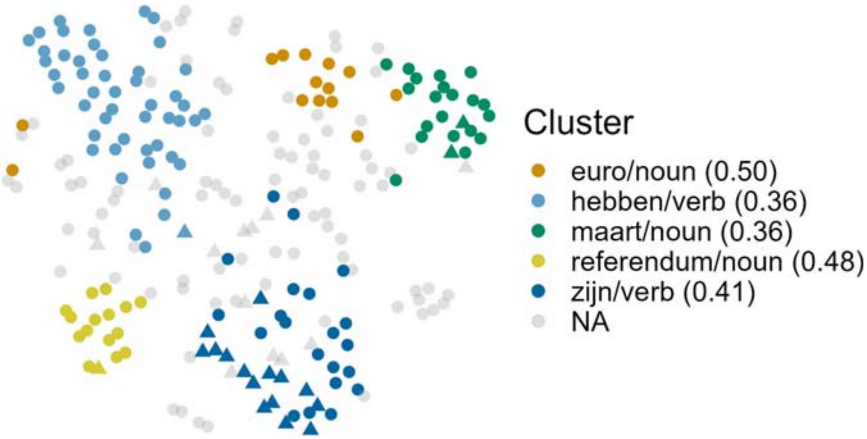


Figure 5.24 Model of *geldig* with parameters 10-10.LEX.BOUND.SELECTION.SOCALL.FOC. Circles represent the specific sense and triangles the general one

These different dimensions of *geldig* are always part of this sense: it is applied to a document for a given span of time, it is typically paid for and it must be shown to an authority. However, different aspects are more relevant, and thus foregrounded, to different kinds of documents, which is reflected in the various patterns of usage. Some documents, such as a driving license and a passport, are talked about in contexts in which individuals have to show possession of the document, and that document must be valid in that moment: both their price and their temporal dimension are left implicit. Other documents, such as tickets and subscriptions, are talked about in terms of their purchase and highlighting the period in which they are valid, which does not need to include the time of speaking. In terms of the definitions used for manual annotation, all these situations are covered by the same sense of *geldig*, but in practice, contextual patterns correlate with the foregrounding of specific dimensions of that sense.

The bottom line

- There is no parameter settings configuration that works optimally across the board.
- Clouds do not necessarily match lexicographic senses. Instead, they match co-occurrence patterns that may or may not coincide with lexicographic senses.
- Clouds may take different shapes, depending on the frequency and distinctiveness of the context words. These shapes correlate with collocational phenomena and can, in turn, overlap to different degrees with lexicographic senses.
- Next to the most typical result of strong collocations pointing to prototypical contexts, we encounter a variety of phenomena combining syntagmatic and paradigmatic aspects. Along with collocations, we find colligation and semantic preference as motors behind most of the clusters, but also a number of cases where no clear distributional pattern can be found. At the paradigmatic level, next to clusters that represent typical contexts, we find heterogeneous clusters and some that match senses completely. In addition, typical contexts may include richer information regarding different semantic dimensions of a sense that are highlighted in certain contexts, that is, that are prototypical of that contextual pattern.

6
The interplay of semasiology and onomasiology

In the previous chapter, we saw that there is no single distributional model that yields optimal results across a wide range of lexical items. Two solutions then suggest themselves: either a choice is made for a specific model on the basis of additional or external evidence, or a variety of models is included in the analysis and the stability of descriptive results is investigated across that set of models. In this chapter, we illustrate the former approach. The latter will be the basis for the lectometric studies in Chapters 9 and 10. Empirically speaking, the present chapter explores the use of distributional models for analysing the evolution over time of the nearly synonymous Dutch verbs *vernielen* and *vernietigen*, 'to destroy'. Compared to the preceding chapter, the descriptive perspective is different in two ways. First, rather than analysing semasiological structure as such, the focus shifts to an onomasiological point of view. But evidently, semasiology will not be absent: onomasiologically comparing the semantic relations between two near-synonyms involves simultaneously mapping out their semasiological structure. Second, next to a purely synchronic perspective, which forms the focus of the first case study included in this chapter (Section 6.3), the second case study (Section 6.4) introduces a diachronic approach. Compared to the previous chapter, then, we add a lectal dimension, and that dimension is a diachronic, or if one wishes 'chronolectal' one. With this perspective, we will demonstrate that semasiological and onomasiological variation and change are not independent of each other: changes in the onomasiological structure of a language can go hand in hand with semasiological changes.

6.1 Onomasiology and token clouds

In contrast with the previous chapter, where the semasiological structure of a lexical item was modelled with a distributional approach, in this study, these models are used to analyse an onomasiological pair. Two types of information that are relevant for an onomasiological perspective are introduced into the visualization of a token model. On the one hand, the onomasiological relationship between a pair of (nearly) synonymous lexical items can be visualized by considering the distance

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0006

between individual tokens of the variants. On the other hand, the lectal stratification of a linguistic variable can be introduced as an independent variable in the distributional modelling procedure. As a prelude to the analysis of *vernielen* and *vernietigen*, the present section illustrates how these two types of information can be made visible in token spaces.

As an example of the first type of information, consider Figure 6.1. In this figure, a distributional model is used to analyse the onomasiological structure of a pair of near-synonyms in Dutch, viz. *woedend* and *laaiend*, both meaning 'furious'. *Woedend* is an adjective that etymologically stems from the present participle of the verb *woeden*, meaning 'to rage'. All the senses of the adjective in present-day Dutch can be roughly translated as 'furious', though the Van Dale dictionary also mentions a figurative use 'zeer onstuimig', that is, 'acting in a wild or violent way', which we did not find in the sample of tokens from our corpus that were manually disambiguated. *Laaiend* can mean 'furious', but it additionally occurs in a number of other usage contexts. Etymologically, it is the present participle of the verb *laaien*, meaning 'to be in flames, to burn heavily', and it can still be used adjectivally with this sense in present-day Dutch, as in example (6.1). Several senses related to 'burning' have evolved from there, not only 'furious' (6.2) but also the intensifier 'very', which most often occurs in the expression *laaiend enthousiast* 'very enthusiastic', see (6.3). A few other less frequent senses are metaphorically related to the high intensity of the burning fire expressed in the original meaning of the verb; see (6.4)–(6.6).

(6.1) **Laaiend** vuur verwoest drie gebouwen. (*Het Nieuwsblad*, 2004-06-18, Art. 370)
'**Burning** fire destroys three buildings.'

(6.2) Een **laaiend**e burgemeester [...] stapte meteen op en sloot zich op in haar kantoor. (*Het Nieuwsblad*, 2003-05-10, Art. 78) 'The **angry** mayor [...] quit immediately and locked herself in her office.'

(6.3) De critici, zeker de jongeren onder hen, zijn in elk geval **laaiend** enthousiast. (*Het Parool*, 2001-11-05, Art. 22)
'The critics, especially the younger ones among them, are **very** enthusiastic in any case.'

(6.4) De levensduur van tijdschriften is door de **laaiend**e concurrentie en het enorme tempo waarin er bladen bijkomen, een stuk korter geworden. (*Het Parool*, 2001-01-05, Art. 68)
'The lifespan of magazines has become a lot shorter due to the **intense** competition and the high pace at which new magazines appear.'

(6.5) In eigen land kreeg Susheela Raman **laaiende** kritieken, maar de plaat verkocht voor geen meter. (*De Standaard*, 2003-06-27, Art. 111)
'In his own country, Susheela Raman received **very positive** reviews, but his record was not sold often.'

(6.6) En de eerste reacties zijn echt **laaiend**. Mensen op straat zeggen: 'Jezus Christus, wat hebben we gelachen, wat ben jíí gestoord.' (*Het Parool*, 1999-07-03, Art. 51)
'And the first reactions are really **very positive**. People in the streets are saying: Jesus Christ, that was hilarious, you are so funny.'

In the bottom panel of Figure 6.1, a distributional model for 600 tokens for *laaiend* and *woedend* is shown. For each verb, 300 tokens were sampled from the *QLVLNewsCorpus* (see Chapter 5). The model relies on five tokens to each side of the target as first-order context words, only considering nouns, adjectives, verbs, proper names, adverbs, and prepositions. These first-order context words are weighted by their log-likelihood ratio with the target lemmas *laaiend* or *woedend*. As second-order context words, 5000 dimensions are considered, restricted to the most frequent nouns, adjectives, and verbs that occur in both regiolects in the corpus (see De Pascale 2019). The values of these dimensions are their ppmi with the first-order context words in a window of four to each side. Clusters are identified by means of a non-density based hierarchical clustering algorithm, hierarchical agglomerative clustering. As explained in Section 3.4, this is a more traditional clustering algorithm than the HDBSCAN algorithm that is used in most chapters of the book. The procedure classifies the tokens into three clusters. In the top panel of the plot, colours correspond to manual disambiguation according to the senses described above. (Strictly speaking, the tokens for *woedend* were not manually disambiguated on a token-by-token basis: a 15% sample of them was manually checked but all of these tokens referred to the prototypical meaning of 'furious'; this result was extrapolated.) The plot shows that the tokens with the meaning 'very' (in blue in the top panel), prototypically occurring in the construction *laaiend enthusiast*, are clearly distinguished by the distributional model from the other tokens: they are contained in cluster 1 (in green) in the bottom panel. Thus, using distributional modelling can help identify the relationship between onomasiological synonyms, in the sense that it can distinguish contexts where the variants are not interchangeable. However, at the same time, the figure also confirms that further consideration is needed because some other meanings of *laaiend*, not related to angry, are located close to the tokens of *woedend*.

Relevant information can also be retrieved by mapping metadata onto the token spaces. In the following example, we look at a case in which a difference in language use shows up by projecting lectal information (i.e. the geographic distribution of text sources) onto the visual representation of the model. As it happens, the example also reveals an imprecision in the lemmatization of the corpus—which

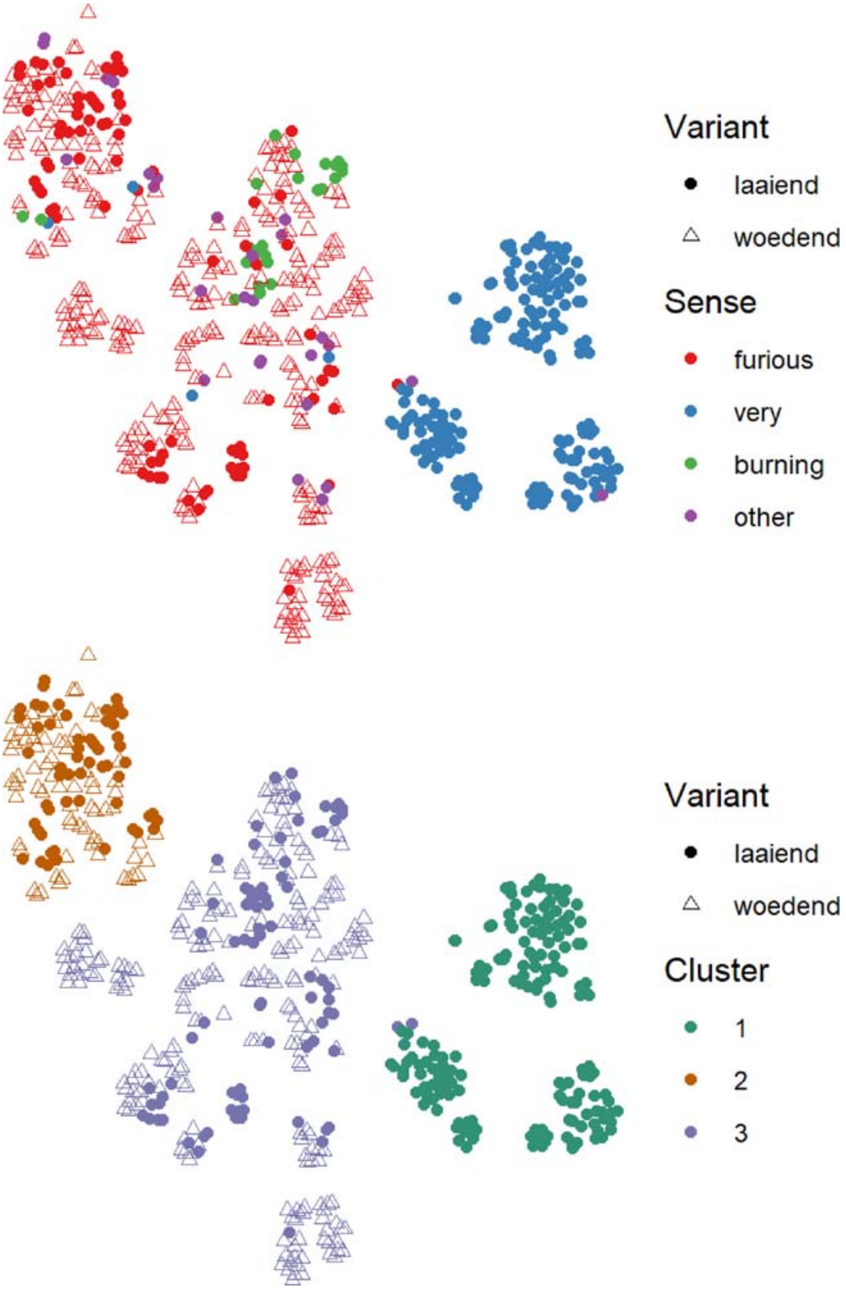


Figure 6.1 Models for *woedend* and *laaiend*. Colour coding in the bottom panel represents a clustering solution with three clusters. Colour coding in the top panel shows the senses that occur in the data based on a manual coding of the tokens

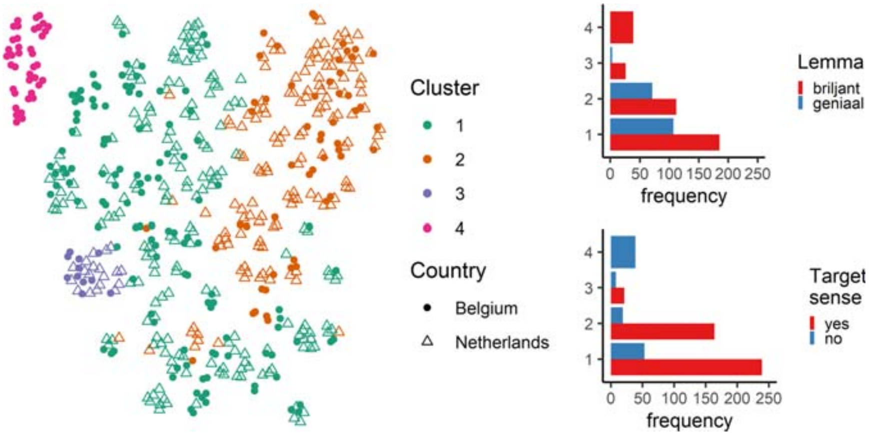


Figure 6.2 Models for *briljant* and *geniaal*. The left panel shows the result of the model, visualized with t-SNE. Clusters (from hierarchical clustering) are shown in colours; lects are shown with shapes. The right panel shows the frequency of the lemmas (top) and of the target sense 'brilliant, genius' (bottom) per cluster

is yet another kind of information one would like to be aware of, even if it may complicate the procedure. We focus on two near-synonyms in Dutch: *briljant* and *geniaal* 'brilliant, genius' (see Figure 6.2). The left panel of the figure shows a t-SNE visualization of a model for these adjectives. The parameter settings are identical to those used in the model underlying Figure 6.1. In total, 549 tokens were sampled, with N = 365 tokens for *briljant* and N = 184 tokens for *geniaal*. The model was analysed with hierarchical clustering in four clusters, which are shown with colours. The lect of the token (here: country, i.e. Belgium or the Netherlands) is shown with shapes. The right panel of the plot shows, at the top, the frequency of each lemma per cluster: red for *briljant*; blue for *geniaal*. The bottom of the right panel contains the frequency of the target sense 'brilliant, genius' for each cluster: red for tokens that have the target sense, that is, in-concept tokens; blue for tokens with another sense, that is, out-of-concept tokens.

The top right panel of the figure shows that the near-synonyms are not equally frequent in the four clusters: in cluster 1, *briljant* is more frequent than *geniaal*, whereas in cluster 2, the adjectives occur with approximately equal frequency. In cluster 3 there are hardly any tokens of *geniaal* and in cluster 4 there are none. Additionally, the left panel shows that cluster 4 (in pink) only contains Belgian Dutch tokens while in the other clusters, tokens from both lects occur. Finally, the bottom right panel reveals that this lectal distribution in cluster 4 is also related to the fact that only tokens without the target sense occur in the cluster. It seems that cluster 4 represents a usage of *briljant* that only occurs in Belgian Dutch.

By inspecting the tokens, it becomes clear that the usage of *briljant* in cluster 4 is very specific. This cluster only contains tokens with collocations like *briljanten (huwelijks)jubileum* '(lit.) brilliant (wedding) jubilee', *briljant*en *bruiloft* '(lit.) brilliant wedding', *briljant*en *huwelijksverjaardag* '(lit.) brilliant wedding anniversary', where the lexical item lemmatized as *briljant* refers to the stone (cut diamond) and stands for a couple's 65th wedding anniversary. Our data suggest that this usage does not occur in Netherlandic Dutch, or at least, not in the corpus sample examined here. At the same time, the analysis reveals a lemmatization mistake in the newspaper data: on closer inspection it turns out that rather than *briljant*, the examples in cluster 4 actually contain the lexeme *briljant*en, an adjective derived from the noun *briljant* 'cut diamond'. The morphological pattern behind the adjective *briljant*en is a regular one by which the suffix *-en* is attached to a noun denoting a material. The adjectival meaning is 'made of the said material', as in *houten* 'wooden' from *hout* 'wood'.

6.2 Verbs of destruction in Dutch

The case studies in this chapter focus on two verbs meaning 'to destroy' in Dutch, *vernielen* and *vernietigen*. A classical analysis of these verbs is presented in Geeraerts (1997), a monograph that argues that a prototype-theoretical approach to lexicology is relevant for diachronic semantics, because 'differences of cognitive salience get an integrated and natural place in the semantic structure ascribed to lexical categories' (1997: 200). The analyses in Geeraerts (1985, 1988, 1997) show that these differences of salience play a crucial role in the meaning structure of the (near) synonyms *vernielen* and *vernietigen*. The description is based on 19th-century data extracted from the citations corpus of the largest historical dictionary of Dutch, the *Woordenboek der Nederlandsche taal* 'Dictionary of the Dutch language'. Etymologically, *vernielen* is a verb formed with a verbalizing prefix *ver-* and an obsolete Dutch adjective *niel* that roughly translates to 'down to the ground'. The literal meaning of the verb *vernielen* is then 'to throw down to the ground, to tear down'. *Vernietigen*, in contrast, is based on the same verbalizing prefix *ver-* with the adjective *nietig* which itself comes from *niet* 'not, nothing' plus an adjectival suffix *-ig*. The literal meaning of *vernietigen* is then 'to annihilate, to bring to naught'. It can be shown that in spite of these divergent etymologies, the near-synonyms can be used in similar contexts in the 19th-century material. For instance, in examples (6.7) and (6.8) (reproduced from Geeraerts 1988: 30–1), both *vernielen* and *vernietigen* occur in the context of the destruction of a material artefact, viz. a (part of a) building.

(6.7) Dat huis was ··· evennmin als de naburige tegen de verwoestende veeten dier tijd bestand. Reeds onder den zoon en opvolger des stichters werd het ··· tot den grond toe **vernield** (Veegens, Hist. Stud. 2, 282, 1869).
'Like the neighboring one, this house was not able to stand up against the destructive quarrels of the age. Already under the son of the founder, it was **demolished** down to the ground.'

(6.8) Alleen zijn de vroegere kruisvensters door vensterramen van nieuweren trant vervangen en hebben de vrijheidsmannen van 1795 ··· het wapen des stichters in den voorgevel met ruwe hand **vernietigd** (Veegens, Hist. Stud. 1, 125, 1864).
'Only, the earlier cross-windows have been replaced by windows in a newer style, and in 1795, the freedom fighters **demolished** the founder's arms in the facade with their rough hands.'

Geeraerts discusses many more examples that clearly show that the verbs are interchangeable in 19th-century Dutch. Overall, three semantic groups of uses for *vernielen* and *vernietigen* can be distinguished in the 19th-century data: concrete uses, abstract uses, and personal uses (see Geeraerts 1997: 191–2, reproduced in the list below):

- three large groups of usage contexts with a concrete item as the patient can be distinguished: to demolish (parts of) buildings, to destroy other human artefacts, to destroy natural objects;
- two large groups of usage contexts with an abstract item as the patient can be distinguished: to annihilate existing situations, characteristics, and so on, to prevent the execution of plans, intentions, and so on;
- four large groups of usage contexts with a person as the patient can be distinguished: to kill someone, to undermine someone's physical health, to undermine someone's psychological wellbeing, to defeat groups of armed men or armies.

While the verbs are found in similar contexts, the prototypical cores of the verbs crucially differ. More specifically, *vernielen* prototypically occurs with concrete uses, such as destroying parts of buildings. *Vernietigen* prototypically occurs in abstract contexts such as the complete annihilation of existing situations or plans. It is also noted that, while both verbs can occur with instances of partial or complete destruction, *vernielen* is prototypically used in the partial destruction sense (for instance, when a building is destroyed by a fire, parts of the structure and ashes from the fire remain), whereas *vernietigen* often implies complete annihilation to naught (for instance, when a plan is destroyed, nothing remains). In addition, the difference between the concrete and abstract uses is also visible in the context of the 'destruction' of people: while *vernielen* occurs more frequently with the more concrete sense of killing someone, *vernietigen* is more often found in the more abstract contexts where someone's physical or mental health is affected. Finally, the prototypical cores of the verbs correlate with their divergent etymologies. On the one hand, *vernietigen* is mostly used in abstract contexts, which correspond to the abstract adjective *nietig* on which it is formed. The destruction of buildings with *vernielen*, on the other hand, corresponds with its literal meaning 'to tear down'. In this way, the synchronic distribution in the 19th-century data, both on the semasiological and on the onomasiological level, reflects the diachronic, etymological background of the verbs.

6.3 Destruction in contemporary Dutch

The distribution of *vernielen* and *vernietigen* in the 19th century is intriguing: they are etymologically distinct and have different prototypical cores but are still largely interchangeable. The question then arises what the onomasiological pair looks like in contemporary Dutch. Could a tendency towards isomorphism be at work, strengthening the prototypical cores to become even stronger and weakening the interchangeability? We will now analyse this question with the distributional semantic models that form the focus of this book. (This section is based on Montes, Franco, and Heylen 2021.)

We extracted data from the *QLVLNewsCorpus* introduced in Chapter 5, lectally balanced for the *Twente News Corpus of Netherlandic Dutch* and *Leuven News Corpus* sections of the corpus; see also De Pascale (2019: 30). In practice, we sampled 186 tokens for *vernielen* and 300 tokens for *vernietigen*, to ensure that their relative frequency distribution in our sample is the same as in the full corpus (7507 and 12 128 respectively). Next, we constructed a set of 100 token-level semantic vector space models that differ with regard to their parameter settings. We explored how similar these models were by calculating the Euclidean distance between each pair of models and analysing these distances with the Partitioning Around Medoids algorithm (see Chapter 4). The models can be inspected at https://qlvl.github.io/NephoVis/#model/destroy. Analysing these models showed that they are quite similar, in the sense that most of the nine medoids that we inspected in detail contained a clear cluster with *vernielen* tokens and a cluster with *vernietigen* tokens with similar semantic properties. Below, we only show a detailed analysis of one of these models, viz. the one that models the semantic structure in the data best according to our manual inspection. In particular, the model we discuss below was, according to our manual analysis, the best at distinguishing the variants under analysis and contains the most semantically homogeneous clusters. This model has the following parameters:

- it is a bag-of-words model, with a window size of 15 context words to each side of the target word (*vernielen/vernietigen*);
- a part-of-speech filter is applied, only including nouns, verbs, adjectives, adverbs, prepositions, and proper names;
- the vectors of the context words are weighted by their ppmi value with the target;
- words used to gauge the similarity between context words come from a list of the 5000 most frequent nouns, adjectives, and verbs that occur in both regiolects in the corpus, as selected by De Pascale (2019). The values of these second-order vectors are based on their ppmi value.

To analyse the semantic structure of this model, we coded the tokens for two explanatory variables: lect (Netherlandic versus Belgian Dutch) and newspaper quality (quality or popular newspaper). In addition, we manually coded a subset of the tokens under investigation (N = 77) for agent and patient type, as well as for agent and patient expression according to the coding scheme in Table 6.1. While patients were already central to the argument laid out in Geeraerts' work on *vernielen* and *vernietigen* in the 19th-century data, coding for agents as well is a novel addition. In the visualizations below, we use a scatterplot (generated with t-SNE with default settings, see Chapter 3) in which four clusters are indicated by colours. The clusters are based on a hierarchical agglomerative clustering algorithm (Ward method). The analysis was conducted in R (R Core Team 2022).

The scatterplot is shown in Figure 6.3, with shape coding for the variants, whose organization is clear enough to be visible across clusters. The *vernielen* tokens (triangles) take up the upper half of the plot, while the *vernietigen* tokens (circles) occupy the lower half. In addition, the frequency of each variant in each cluster is shown in the last column of Table 6.2. Here we can see that clusters 1 and 4 consist nearly exclusively of tokens of *vernietigen* (1) or *vernielen* (4). In contrast, clusters 2 and 3 have around three times as many tokens of *vernietigen* as of *vernielen*, while the overall proportion is closer to 2:1. Clusters 1 and 4 are also located on opposite sides of the plot in Figure 6.3 and, as we will see, represent the prototypical contexts of each target. The remaining question is what kinds of contexts occupy the central area: do they represent shared semantic features or underdetermined contexts?

In order to interpret the clusters, we consider the context words that the model has picked up for its members. The procedure we use is based on the ShinyApp tool described in Section 4.4. In Table 6.2, we only consider context words for which at least 50% of their occurrences are within the cluster of interest. Moreover, the table only shows the ten most frequent context words per cluster.

Clusters 1 and 4 are characterized by context words with completely different semantic associations. For the former, with only tokens of *vernietigen*, *Raad van*

Table 6.1 Coding schema for agent and patient expression

VARIABLE	TOP LEVEL	MEDIUM LEVEL	BOTTOM LEVEL	EXPLANATION	PLOTTED LABEL
agent	persons	individuals	criminals	individuals breaking legal or social norms	criminal
			professional	individuals acting in their professional capacity	other_persons
			other		other_persons
		organizations	(legal) authorities	any legally sanctioned authority, including individuals acting under this authority	authority
			other		other_persons
		indirect	instruments	instrument of destruction controlled by humans	other_persons
	natural phenomena		fire	including explosions	fire
			other	including animals and biological phenomena	other_inanimate
	circumstances		circumstances	situations or events construed as agents	other_inanimate
patient	concrete	human artefacts	(parts/groups) of buildings		building
			utilitarian artefacts	vehicles, instruments, objects with a practical function in everyday life	utilitary
			with military use	utilitarian artefacts used for military purposes	other_concrete
			commercial goods	agricultural or other goods intended for economic transactions	other_concrete
		natural objects	natural objects	including biological phenomena	other_concrete
	abstract		(legal) decisions		abstract
			organizations		abstract
	persons		wellbeing	both physical and psychological	abstract

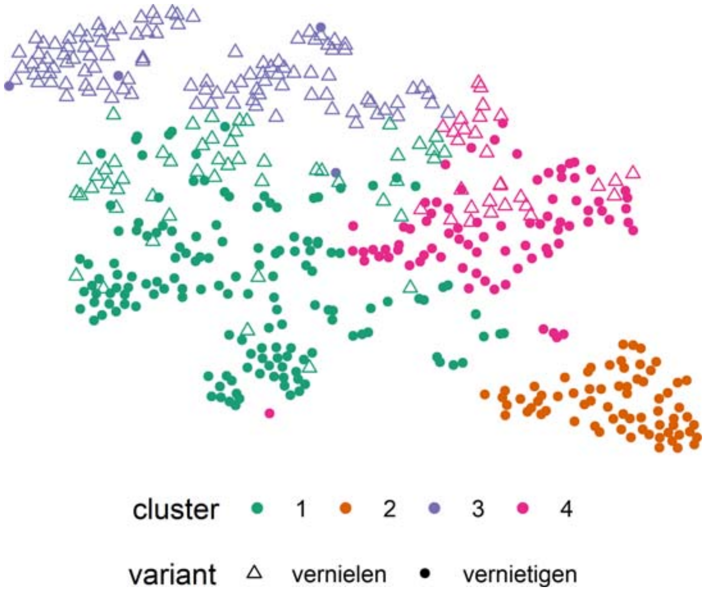


Figure 6.3 Scatterplot of t-SNE visualization of one model of *vernielen* and *vernietigen*, coloured by four clusters and shape-coded by variant

Table 6.2 Clusters in the model for *vernielen* and *vernietigen* in contemporary data

CLUSTER NUMBER	CONTEXT WORDS	VARIANTS
1 authority destroying decisions, plans, etc.	*Raad* (24); *van* (22); *State* (20) ('Council, of, State'); *beslissing* 'decision' (11); *rechter* 'judge' (9); *vergunning* 'permit' (8); *bouwvergunning* 'building permit' (8); *beroep* 'appeal' (7); *vonnis* 'verdict' (7); *Vlaams* 'Flemish' (6)	*vernielen*: 0 (0) ***vernietigen:*** 1 (69)
2 (diverse, destruction of (sick) livestock)	*bedrijf* 'company' (10); *rund* 'beef' (5); *ziek* 'sick' (4); *kost* 'cost (n)/to cost (v)'; *stuk* 'piece'; *oogst* 'harvest'; *besmet* 'infected'; *Enron* (3); *bewijsmateriaal* 'evidence' (3); *schilderij* 'painting' (3)	*vernielen*: 0.27 (29) *vernietigen*: 0.73 (80)
3 (diverse, auxiliary and modal verbs)	*word* 'passive auxiliary' (114); *door* 'by (for passive agents)' (36); *zal* 'will' (27); *kan* 'can' (22); *moet* 'must' (21); *volgens* 'according to' (17); *politie* 'police' (15); *stad* 'city'(12); *wapen* 'weapon' (11)	*vernielen*: 0.26 (51) *vernietigen*: 0.74 (147)
4 fire destroying material artefacts	*van* 'of, from' (57); *brand* 'fire (event)' (23); *volledig* 'completely' (18); *woning* 'house, residence' (16); *vuur* 'fire' (15); *ook* 'also' (13); *auto* 'car' (12); *huis* 'house' (11); *brandweer* 'fire department' (11); *vandaal* 'vandal' (10)	*vernielen*: 0.96 (106) ***vernietigen:*** 0.04 (4)

State 'Council of State' and similar names referring to government authorities represent typical agents, with *beslissing/besluit* 'decision' and *vergunning* 'permit' as the typical patients. In contrast, cluster 4, with overwhelmingly *vernielen* tokens, includes *vuur/brand* 'fire' and *vandaal* 'vandal' as typical agents, and *woning/huis* 'house' and *auto* 'car' as patients. Another interesting context word in cluster 4 is *volledig* 'completely': its high association with *vernielen*—or rather, its repulsion with regard to *vernietigen*—suggests that the notion of complete annihilation is implicit in *vernietigen*, rendering such a combination redundant.

The third cluster mostly contains auxiliaries and modal verbs, and a few lexical context words such as *politie* 'police', *stad* 'city', and *wapen* 'weapon'. Only the first two co-occur evenly with both variants. Context words in the second cluster, finally, refer to sick or contaminated produce/livestock: *ziek* 'sick', *besmet* 'infected', *rund* 'beef', *oogst* 'harvest', and *bedrijf* 'company' (to which the produce belongs). However, these context words occur exclusively with *vernietigen*. In other words, characteristic context words of the diverse clusters are more frequent with the frequent variant, but they are less distinctive than those in the clusters with the prototypical usages of the verbs (one for *vernietigen* and four for *vernielen*).

As a final step, we compared the results of the model to our own manual coding of a subset of the tokens. Figure 6.4 visualizes the model, but the colour coding now

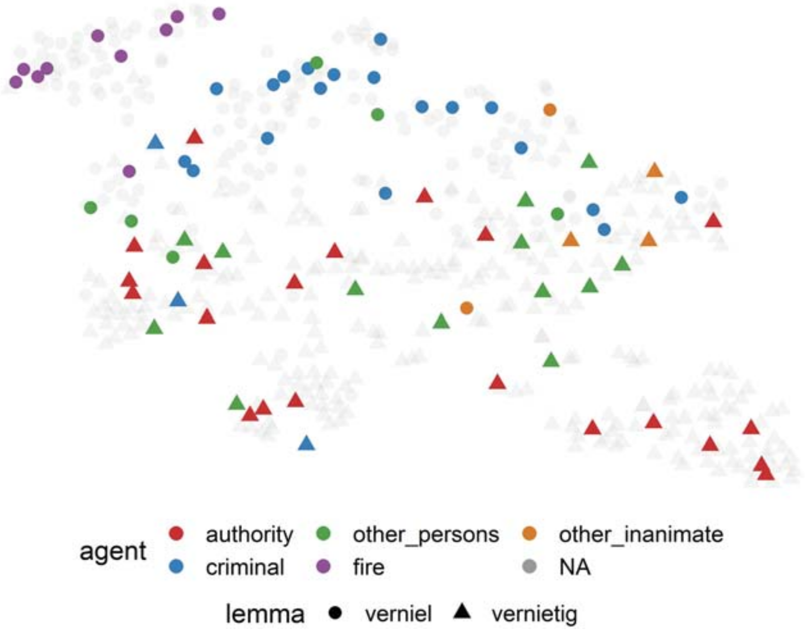


Figure 6.4 Scatterplot with colours for manual coding of agent type and shapes for variants

shows the manual coding for agent type. The top half of the plot mostly contains tokens of *vernielen*, whereas the bottom half predominantly has *vernietigen* tokens. The figure indicates that the model is successful at distinguishing agent types, especially the prototypical ones. There are two clear agent types that occur in the top half (mostly *vernielen*): *fire* (purple) and *criminals* (blue). These agents were also clearly distinguished in cluster 4 discussed above, with context words like 'fire' and 'vandal'. In the bottom half of the plot (mostly *vernietigen*), a wider variety of agents occurs. Only in the bottom right—the region that coincides with cluster 1 discussed above—there is a clear cluster of (legal) authority agents (red), although these authorities actually occur throughout the lower half of the plot.

In sum, the manual coding of the agents reveals that the automatic modelling procedure outlined above is very good at distinguishing the prototypical cores of each verb: fire and criminals for *vernielen*; (legal) authorities for *vernietigen*. However, in the centre of the plot, the picture is more diverse. In this region, agent type alone is not sufficient to distinguish *vernielen* and *vernietigen*, as both verbs occur with all the other agent types (although *vernietigen* is in general the most frequent verb). This may be an explanation for the diffuse semantic nature of the context words in clusters 2 and 3.

Figure 6.5 again visualizes the model, but the colour now reflects a manual coding for patient type. The plot shows that there are two types of patients

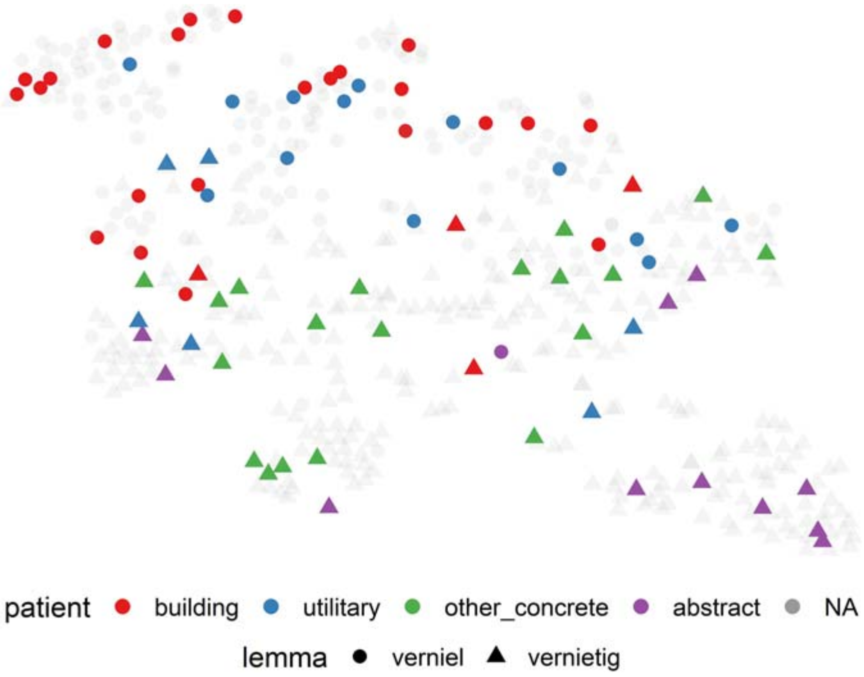


Figure 6.5 Scatterplot with colours for manual coding of patient type and shapes for variants

that occur most often with *vernielen* (in red and blue): (parts) of buildings and utilitarian artefacts, like vehicles, instruments or other objects with a practical function in everyday life. In contrast, abstract patients, like decisions and organizations (in purple), typically occur with *vernietigen*, as in the 19th-century data. These results are mostly in line with the older work that we referenced in Section 6.2, which showed that in the 19th century *vernielen* prototypically occurs with material objects. However, for the contemporary data, the plot shows that only two specific types of material patients are very frequent with this verb: buildings and utilitarian artefacts. Other concrete items, such as goods, military artefacts, or natural objects, occur more in the bottom half of the plot (in green), with the *vernietigen* tokens. However, most of these tokens are semantically quite specific. Although the patients in these tokens are concrete objects, the sense expressed by the verb is that of an authority ordering the destruction of the objects. This is the case for every token that was manually coded as 'other' and these patients also exclusively occur with *vernietigen*; see (6.9). Crucially, the sense expressed in these tokens is closely related to what is the most frequent reference of *vernietigen* in these data, viz. a situation in which an authority, usually featuring in the agent role, orders the elimination of something abstract, such as a decision or an existing situation. Though we only have synchronic data at our disposal in this analysis, we hypothesize that *vernietigen* has diachronically evolved through modulations of its prototypical core: from the destruction of an abstract concept (the most prototypical sense in Geeraerts' 19th-century data), to the destruction of an abstract concept on behalf of an authority (the most frequent sense in these data), and, in the final stage, to the destruction of a concrete object on behalf of an authority.

However, one puzzling fact is that natural objects occur exclusively with *vernietigen* as well. Although this may have to do with the fact that in the tokens where natural objects occur the objects are completely annihilated (just like decisions or existing situations are completely annihilated or cancelled with *vernietigen*), an authoritative figure is never involved. See (6.10) for an example.

(6.9) Zweedse boeren die genetisch gemanipuleerde koolzaadplanten telen, moeten hun oogst voor 7 juli **vernietigen**. Dat heeft de Zweedse Raad voor de Landbouw woensdag besloten. (*De Volkskrant*, 2000-05-25, Art. 136) 'Swedish farmers who grow genetically modified rapes are required to **destroy** their crops by July 7. This was decided by the Swedish Council of Agriculture on Wednesday.'

(6.10) [Het] is al gebleken dat deze afweercellen in een reageerbuis kankercellen van mensen en muizen kunnen **vernietigen**. (*Algemeen Dagblad*, 2000-08-31, Art. 111) 'It has already become clear in vitro that these immune cells can **destroy** cancer cells of people and mice.'

In sum, the analysis of our distributional semantic model shows that *vernielen* and *vernietigen* have diverged in the 21st-century data in comparison to Geeraerts' 19th-century dictionary citations: the verbs are no longer easily interchangeable in every context. A highly prototypical context for *vernielen* in the 21st-century data is the destruction of (parts of) buildings by fire, and *vernietigen* no longer occurs in this context. In contrast, for *vernietigen*, the cancellation of decisions or ideas by a governmental body makes up a large portion of the tokens in the corpus, and *vernielen* is no longer possible there. Semasiologically, the prototypical core that was already there for both variants in the 19th century has become stronger and changes have only occurred in the periphery. For instance, uses like those pertaining to human patients have been lost. At least one new usage has also come into existence: to destroy livestock, crops, and so on, with *vernietigen*, where a concrete patient is used instead of the prototypical abstract patients like ideas or decisions. The evolution likely stems from the fact that the agent, a governmental body, is typically similar to the agent in the most frequent (prototypical) usage contexts of *vernietigen*, that is, contexts where the verb is used to refer to the cancellation of decisions, ideas, plans, and so on. Nonetheless, it is also important to consider the fact that a different type of dataset is used in the current analysis (newspaper material) compared to the older manual study (dictionary citations). The register under analysis may also affect the semantic patterns that are found through the distributional semantic analysis. In any event, comparing the manual analysis from Geeraerts (1997) with a distributional analysis of contemporary data, we find evidence of a conceptual reorganization. Would we get similar results if we track the diachronic developments on a strictly distributional basis? In the following section, we will distributionally follow *vernielen* and *vernietigen* over a timespan of five centuries. (An earlier version of this study was published as Franco, Montes, and Heylen 2022.)

6.4 Destruction across the centuries

In the diachronic analysis, we use a corpus of prose texts from DBNL, the *Digitale Bibliotheek voor de Nederlandse Letteren* 'Digital library for Dutch language and literature'. Some information about the corpus can be found in Depuydt and Brugman (2019), though the corpus is not publicly available at this point. We extracted all corpus texts tagged as prose in the metadata from the 16th, 17th, 18th, 19th, and 20th centuries. Due to data sparseness, we combine the subcorpora for the 16th and 17th centuries in the analysis.

In contrast with the work in most chapters of this book, the diachronic corpus that we use here consists of raw xml files without any further linguistic tagging (no tokenization, no lemmatization, no part-of-speech-tagging etc.). As no high-quality lemmatizers or part-of-speech taggers are as of yet available for historical

Dutch, the only pre-processing we applied to the corpus was to transform the entire corpus to lower case and to automatically indicate word and sentence boundaries using the pretrained nltk tokenizers (Loper and Bird 2002). However, as will become apparent below, this means that a number of parameters settings considered in the previous sections of this chapter and in the rest of this book cannot be applied to the dataset used here.

After pre-processing, we needed to extract the tokens for *vernielen* and *vernietigen* (including inflected forms and spelling variants) from the four subcorpora. However, spelling variation is abundant in historical Dutch—a standardized spelling only became widely used at the beginning of the 20th century. Table 6.3, for instance, shows all the spelling variants and their inflected forms that occur in the corpus for the Dutch verbs under analysis in this study. In total, ten unique word forms occur for *vernielen* and 13 for *vernietigen*. To collect these inflected forms and spelling variants of *vernielen* and *vernietigen*, we searched for all word forms starting with *verniel* and *vernyel* for the verb *vernielen*, and all forms starting with the forms *vernietig* and *vernietich* for *vernietigen* in the entire corpus (16th–20th centuries). Next, we manually cleaned up these forms in order to only consider spellings that can be used for the verbal uses of *vernielen* and *vernietigen*.

Next, we collected all tokens containing one of the items of this cleaned up list of forms. Table 6.4 provides an overview of the number of tokens per century for each verb, and for the total corpus size per century. The table already shows that *vernielen* decreases in (relative) frequency (from 61% to 19% of all tokens for the concept 'to destroy'), whereas *vernietigen* becomes more popular over time (from 39% to 81% of all tokens for the concept 'to destroy'). Finally, we took a random sample of N = 400 tokens for each subcorpus which will be analysed further below.

Next, we constructed a single vector space model for the tokens in each subcorpus using the procedure outlined in Chapter 3. The parameters that we used are largely based on the best model described above for the analysis of *vernielen* and *vernietigen* in 21st-century newspaper data. However, due to the lack of part-of-speech tagging and lemmatization, most parameters needed to be amended

Table 6.3 Inflected forms and spelling variants occurring for *vernielen* and *vernietigen* in the diachronic corpus

VERB	INFLECTED FORMS AND SPELLING VARIANTS FOUND
vernielen	vernielen, vernielt, vernield, vernielde, verniele, verniel, vernielden, vernieldt, vernyelt, vernyelen
vernietigen	vernietigen, vernietight, vernietigt, vernietigd, vernietig, vernietige, vernietigden, vernietigde, vernietighen, vernietighd, vernietighde, vernieticht, vernietichde

Table 6.4 Frequency of *vernielen* and *vernietigen*, and total number of tokens per century

CENTURY	NUMBER OF TOKENS FOR *vernielen* (PROPORTION)	NUMBER OF TOKENS FOR *vernietigen* (PROPORTION)	TOTAL NUMBER OF TOKENS
16th & 17th	610 (0.61)	394 (0.39)	20 369 255
18th	2175 (0.35)	4127 (0.65)	123 542 380
19th	3589 (0.29)	8900 (0.71)	317 140 193
20th	108 (0.19)	446 (0.81)	19 849 314

(Table 6.5). On the one hand, some parameter settings are not applicable to the diachronic corpus. More specifically, we could not apply any part-of-speech filtering and all analyses are based on word forms rather than on lemmas. In addition, it is not possible to obtain a list of the 5000 most frequent words across the different lects considered (here: the four chronolects) due to the lack of lemmatization and the abundant spelling variation: very few words (if any) are shared between the oldest and most recent subcorpus. On the other hand, we decided to decrease the window size from 15 to 10 words to the left and right of the target token for the first-order context words because preliminary analyses revealed that in models with a broader window, too many irrelevant or noisy context features were included. Similarly, we only considered first-order context words with a relatively high pmi value (pmi > 2) with the target verbs because preliminary analyses revealed that context words with a weaker association with the target items tend to be noisy and irrelevant. This is probably also related to the fact that part-of-speech tagging and lemmatization are not available, resulting in less relevant words being included in the context window. The models were constructed according to the procedure described in Chapter 3. As in the synchronic contemporary model, we used hierarchical clustering (Ward method), distinguishing four clusters. There are two reasons why we did not use the HDBSCAN algorithm. First, we wanted to keep the analysis as similar as possible across both case studies. Second, for all four subcorpora, HDBSCAN classifies between 77.6% (19th-century data) and 93.9% (20th-century data) of the modelled tokens as noise tokens. We assume this is related to the fact that no lemmatization was employed, though it is striking that the largest proportion of noise tokens occurs in the most recent data, where spelling variation is not as abundant. We also visualize the data with t-SNE setting perplexity to 20. Finally, we analyse the clusters using the same procedure as in the contemporary study.

Figure 6.6 shows the visualizations of the models, with one panel per subcorpus. Plot symbols show the variants (*vernielen* versus *vernietigen*) and colours indicate the clusters. Note that the order of the clusters is fully based on the hierarchical

Table 6.5 Parameter settings in the contemporary study compared to parameter settings in the diachronic study

	PARAMETERS CONTEMPORARY STUDY (SECTION 6.3)	PARALLELS IN DIACHRONIC STUDY (SECTION 6.4)
Context window	Window size of 15 (first-order) context words to each side of the target lexical item.	Window size of 10 (first-order) context words to each side of the target lexical item.
Considered first-order context words	Part-of-speech filter, only considering nouns, verbs, adjectives, adverbs, prepositions, and proper names.	All word forms [w+] with a frequency of at least 10 in the subcorpus. For all first-order context words, a single pmi value with the target verbs *vernielen* and *vernietigen* is calculated. The context words are subsequently filtered by their pmi value with the target token: only wordforms with pmi > 2 are considered.
Weighting of first-order context words	The vectors of first-order context words (lemmas) are weighted by their ppmi value with the target lexical item (lemmas).	The vectors of first-order context words (wordforms) are weighted by their pmi value with the target lexical item (lemma-level). This ppmi value, calculated separately per subcorpus, is based on the combined frequency of each context word over all wordforms occurring for *vernielen* or *vernietigen*.
Second-order context words	Considered second-order context words come from a list of the 5000 most frequent noun, adjective, and verb lemmas that occur in both lects available in the newspaper corpus (Netherlandic and Belgian Dutch), as selected by De Pascale (2019). The values of these vectors are based on their ppmi value (lemmas).	5000 most frequent word forms [w+] per subcorpus, excluding the first 100 word forms, as these are usually function words rather than content words and therefore do not contribute a lot of semantic information. The values of these vectors are based on their ppmi value (word forms).

clustering method. Therefore, there is no relationship, semantic or other, between the cluster numbers in subsequent centuries. Cluster 1 in the first subcorpus, for example, does not necessarily contain similar tokens as cluster 1 in the second

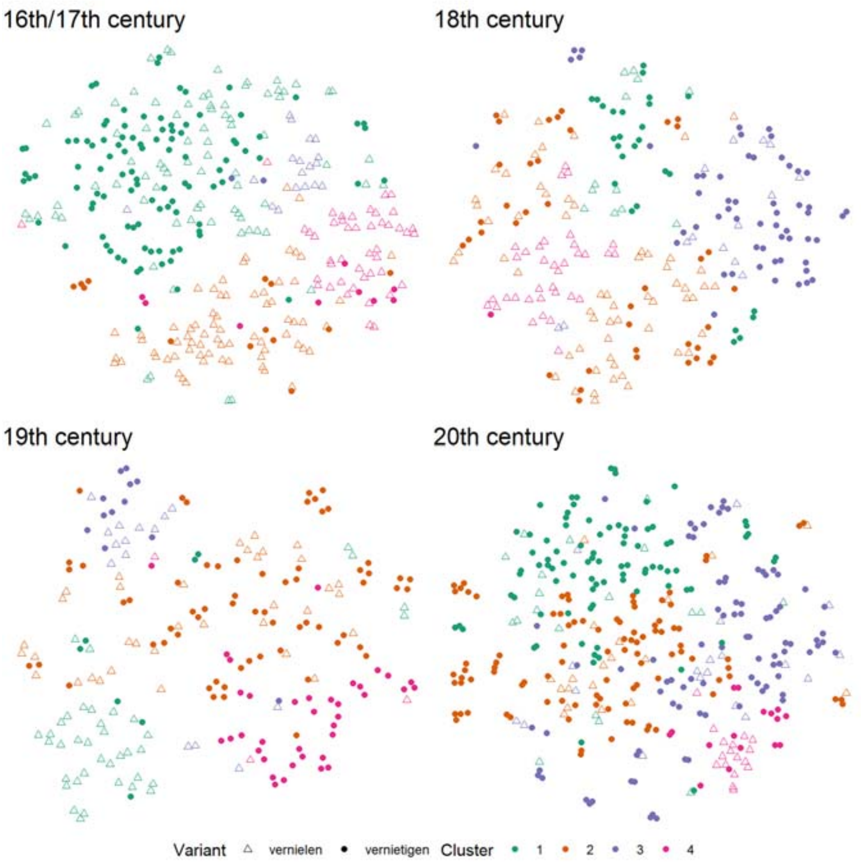


Figure 6.6 Cluster analyses of *vernielen* and *vernietigen* in the four diachronic subcorpora

subcorpus and so on. The figure shows that over time, *vernielen* and *vernietigen* are distinguished more clearly by the models. In the 16th/17th century, there is still quite some overlap between the variants, in the sense that there is one large cluster (cluster 1 in green), which clearly contains tokens for both variants. This indicates that they are still interchangeable in a large number of contexts at this point in time. However, over time, the overlap becomes smaller: clusters where both variants are possible become much smaller in size. In the 18th century *vernielen* mostly occurs at the bottom left of the plot and *vernietigen* at the top right. In the 19th century, *vernielen* is found in the left side of the plot and *vernietigen* mostly in the bottom right. By the 20th century, there is only one small cluster (cluster 4 in pink) where both variants still occur with comparable frequency, while in the other clusters, *vernietigen* is always the most frequent variant. Moreover, by the 20th century, the variant *vernielen* had decreased dramatically in frequency and *vernietigen* takes up most of the figure.

Table 6.6 Clusters in the model for *vernielen* and *vernietigen* in the 16th and 17th centuries

CLUSTER NUMBER	CONTEXT WORDS	VARIANTS
1 (diverse)	7: *alles* 'everything', *geheel* 'completely' 4: *natuur* 'nature', *geluk* 'luck', *duizend* 'thousand', *veranderingen* 'changes', *zonder* 'without', *werden* 'became (pl.)', *schulden* 'debts', *werd* 'became (sg.)' 3: *gramschap* 'wrath', *beeld* 'statue, picture', *vorsten* 'monarchs', *oogenblik* 'moment', *kunt* 'can', *word* 'become (sg.)', *plantagien* 'plantations', *nieuwe 'new'*, *compagnie* 'company', *dezelve* 'itself'	*vernielen*: 0.44 (80) *vernietigen*: 0.56 (102)
2 to kill persons	10: *dese* 'this' 5: *doot* 'death' 4: *desen* 'this', t 'it' 3: *wet* 'law', *sulcke* 'this', *selve* 'self', *vyanden* 'enemies', *Christi* '(of) Christ', *dooden* 'to kill', *omme* 'in order to', *macht* 'power', *verlaten* 'to leave', *sonde* 'sin'	***vernielen:*** 0.84 (69) *vernietigen*: 0.16 (13)
3 natural objects	/	***vernielen:*** 0.90 (18) *vernietigen*: 0.10 (2)
4 concrete objects	5: *schepen* 'ships' 4: *vernielen* 'to destroy', *steden* 'cities', *vloot* 'fleet' 3: *zwaert* 'sword', *bergen* 'mountains'	***vernielen:*** 0.83 (45) *vernietigen*: 0.17 (9)

The following Tables (6.6–6.9) provide an overview of the most important context words per period and per cluster. In the second column, only context words that occur in at least 50% of the cases within the cluster and with a frequency of more than 2 within the cluster are shown, to avoid that infrequent words get too much weight in the interpretation. The numbers in this column indicate the frequency of the context words. The first column also shows a semantic interpretation of each cluster. The final column indicates the relative and absolute proportion of each variant in the cluster. If one of the variants takes up more than 80% of the tokens in a cluster, we consider it the major variant and we assume that the context is non-interchangeable. Major variants are marked in bold in Tables 6.6–6.9.

The context words for the first subcorpus (16th–17th century) are shown in Table 6.6. There are three clusters where *vernielen* clearly is the major variant (clusters 2, 3, and 4 in orange, purple, and pink in Figure 6.6). It occurs in clusters with context words related to killing persons (such as *doot* 'death', or *vyanden* 'enemies'), in a small cluster with natural objects (where there are no context words with frequency > 2) and in a cluster with context words referring to concrete objects like *schepen* 'ships' and *steden* 'cities'. The first and largest cluster

(N = 182 tokens) is still quite diverse and *vernielen* and *vernietigen* are both possible. Thus, in the 16th and 17th century, *vernielen* and *vernietigen* are still mostly interchangeable, although there are already a few contexts where *vernielen* is preferred.

Table 6.7 shows the results for the 18th century. In this second subcorpus, there are two clusters where *vernielen* is more frequent (clusters 2 and 4 in orange and pink in Figure 6.6) and two clusters where *vernietigen* takes over (clusters 1 and 3 in green and purple), though each variant is the major variant only in a single cluster (covering at least 80% of the tokens). Following the hypotheses outlined above, *vernielen* mostly occurs with concrete objects like *kerken* 'churches' and *huizen* 'houses' which are destroyed by *brand* 'fire' (cluster 4). In addition, in cluster 2 it seems to occur in passive tokens (with *werden* 'became (pl.)') where persons are destroyed (*vijand* 'enemy', *troepen* 'troups', *leger* 'army', and some lexemes related to people, such as *hunne* 'their' and *elkaâr* 'each other'). In contrast, *vernietigen* occurs in tokens with abstract objects like *invloed* 'influence' and *kracht* 'strength' (cluster 3). *Vernietigen* is also the most frequent variant in the first cluster, which does not show a clear semantic picture. In sum, while in the 18th century the variants already start to occur in their prototypical contexts, they are still interchangeable in most contexts.

Table 6.8 shows the results for the 19th century. In this century, which coincides with the data analysed in Geeraerts (1985, 1988, 1997), *vernielen* remains the most frequent variant in contexts of the destruction of (parts of) buildings by fire, in cluster 1 (green). In contrast with the 18th-century subcorpus, *vernietigen* is now

Table 6.7 Clusters in the model for *vernielen* and *vernietigen* in the 18th century

CLUSTER NUMBER	CONTEXT WORDS	VARIANTS
1 (diverse)	5: *daardoor* 'because of' 3: *worde* 'become (pl.)', *gansch* 'completely', *hoop* 'hope'	*vernielen*: 0.31 (12) *vernietigen*: 0.69 (27)
2 to kill persons + war	12: *werden* 'became (pl.)' 4: *hunne* 'their' 3: *elkaâr* 'each other', *vijand* 'enemy', *troepen* 'troups', *leger* 'army', *gebroken* 'broken', *slag* 'battle', *vloot* 'fleet', *oogst* 'harvest'	*vernielen*: 0.61 (60) *vernietigen*: 0.39 (39)
3 abstract objects	5: *invloed* 'influence' 4: *zedelijk* 'virtuous', *kracht* 'strength', *macht* 'power', *revolutie* 'revolution', *bestaan* 'existence, to exist', *vrijheid* 'freedom'	*vernielen*: 0.20 (14) ***vernietigen:*** 0.80 (56)
4 (parts of) buildings (fire)	6: *brand* 'fire' 4: *stad* 'city' 3: *kerken* 'churches', *huizen* 'houses', *steden* 'cities	***vernielen:*** 0.97 (35) *vernietigen*: 0.03 (1)

Table 6.8 Clusters in the model for *vernielen* and *vernietigen* in the 19th century

CLUSTER NUMBER	CONTEXT WORDS	VARIANTS
1 (parts of) buildings (fire)	5: *huis* 'house' 4: *brand* 'fire' 3: *grond* 'ground', *boel* 'things', *vlammen* 'flames'	***vernielen:*** 0.84 (38) *vernietigen*: 0.16 (7)
2 to kill persons + war	6: *zichzelf* 'himself/herself/themselves', *volkomen* 'completely', *steden* 'cities' 5: *werden* 'became (pl.)', *leger* 'army', *vloot* 'fleet', *schepen* 'ships' 3: *zorgvuldig* 'carefully', *gedeeltelijk* 'partly', *volledig* 'completely', *willen* 'to want', *brieven* 'letters'	*vernielen*: 0.36 (35) *vernietigen*: 0.64 (63)
3 abstract objects?	3: *waan* 'delusion'	*vernielen*: 0.56 (15) *vernietigen*: 0.44 (12)
4 abstract objects	3: *vrijheid* 'freedom'	*vernielen*: 0.08 (3) ***vernietigen:*** 0.92 (37)

more frequent in contexts related to the destruction of persons, including armies (cluster 2 in orange), but *vernielen* is still possible. In this cluster *zichzelf* 'himself/herself/themselves' is among the most frequent context words. This frequent use of the reflexive pronoun may indicate that the patient role for *vernietigen* in the 19th century is often the subject itself, or that it at least plays a major role. Another highly frequent context word is *volkomen* 'completely', indicating the complete destruction of the patient of the verb. In Geeraerts' work, complete destruction was typically associated with *vernietigen* as well. Like in the 18th century, *vernietigen* also still occurs the most with abstract lexemes such as *vrijheid* 'freedom' (cluster 4 in pink). Cluster 3 (in purple) only has one important context word, *waan* 'delusion', and both variants are possible in this cluster. The interpretation is not as clear as for the other clusters. Overall then, there are two prototypical contexts where one variant takes over: *vernielen* for the destruction of (parts of) buildings by fire, *vernietigen* for the destruction of abstract objects. In the other clusters, the variants are still interchangeable.

Finally, in the subcorpus for the 20th century, *vernietigen* is much more frequent than *vernielen*. Only 108 tokens for *vernielen* occur in the complete 20th century subcorpus, but 446 occur for *vernietigen*. This may indicate that *vernielen* is on its way out, or that it is retreating to very specific contexts. Recall that *vernietigen* was also approximately twice as frequent as *vernielen* in the contemporary newspaper data. The cluster analysis confirms that *vernielen* is on the decline

(Table 6.9): there are no more clear contexts where only *vernielen* is possible. In contrast, there are clear clusters in which *vernietigen* is the preferred variant, though the interpretation of the cluster is not always straightforward. In cluster 3 (in purple), some function words are present (for example, *uiteindelijk* 'eventually', *waarna* 'after which'), as well as lexical items referring to complete annihilation (*alles* 'everything', *niets* 'nothing'). In this third cluster, there are also some lexical items related to persons (*zichzelf* 'himself/herself/itself', *mens* 'human'). Further, *vernietigen* is the most frequent variant in cluster 1 (in green), which is also a diverse cluster, with the most frequent context word related to complete destruction (*geheel* 'completely'). This cluster also contains other types of lexical items such as abstract concepts (*bestaan* 'existence, to exist', *rede* 'reason', *schoonheid* 'beauty') and function words (*daardoor* 'because of', *zulke* 'such').

Table 6.9 Clusters in the model for *vernielen* and *vernietigen* in the 20th century

CLUSTER NUMBER	CONTEXT WORDS	VARIANTS
1 (diverse)	5: *geheel* 'completely' 4: *bestaan* 'existence, to exist' 3: *daardoor* 'because of', *groepen* 'groups', *rede* 'reason', *zulke* 'such', *schoonheid* 'beauty', *natuur* 'nature', *waarde* 'value', *dreigt* 'threatens'	*vernielen*: 0.18 (19) ***vernietigen:*** 0.82 (87)
2 to kill persons + war	8: *oorlog* 'war' 5: *werden* 'became (pl.)', *nadat* 'after', *hele* 'whole', *gehele* 'whole', *oplage* 'edition' 4: *documenten* 'documents', *moesten* 'had to', *volk* 'people' 3: *recht* 'right', *kaart* 'map', *zouden* 'would', *joodse* 'jewish', *steden* 'cities', *exemplaren* 'samples', *geworden* 'become (participle)', *goden* 'gods', *zestig* 'sixty', *wereldoorlog* 'world war', *europese* 'european'	*vernielen*: 0.16 (19) ***vernietigen:*** 0.84 (97)
3 (diverse)	14: *alles* 'everything' 10: *zelfs* 'even' 9: *zichzelf* 'himself/herself/itself' 7: *niets* 'nothing' 6: *uiteindelijk* 'eventually' 4: *mens* 'human' 3: *waarna* 'after which', *god* 'god', *erbij* 'near it', *erop* 'on it', *definitief* 'definitive', *jezelf* 'yourself', *onmogeliijk* 'impossible'	*vernielen*: 0.18 (19) ***vernietigen:*** 0.82 (88)
4 (parts of) buildings	4: *huis* 'house', *aarde* 'earth' 3: *muren* 'walls', *stenen* 'stones'	*vernielen*: 0.52 (17) *vernietigen*: 0.48 (16)

In cluster 2 (in orange), *vernietigen* is the most frequent variant as well. This is a semantic cluster with many people- and war-related lexical items (*volk* 'people', *oorlog* 'war'), although it also contains other lexical items such as function words (*werden* 'became', *nadat* 'after'), lexemes related to complete annihilation (*hele* 'whole', *gehele* 'whole') and concrete objects (*documenten* 'documents', *kaart* 'map'). In cluster 4 (in pink), *vernielen* and *vernietigen* are interchangeable. The context words are mostly related to destroying (parts of) buildings. This is a clear shift compared to the earlier data, where the destruction of parts of a building was strongly associated with the use of *vernielen*. However, the context words in this cluster have quite a low frequency so possibly not all tokens are related to the destruction of (parts of) buildings. In addition, the cluster only contains 33 tokens in total. Perhaps *vernielen* has become so infrequent in the 20th century that even this prototypical use is not salient enough any more to be distinguished by the model and clustering procedure.

6.5 The evolution of onomasiological sets

The first five columns of Table 6.10 summarize the results across the four diachronic subcorpora. The table shows in how many clusters each variant is the preferred one per century (i.e. occurring with a relative frequency of 80% or more), as well as in how many clusters both variants are interchangeable (i.e. both variants take up less than 80% of the tokens in the cluster). At the bottom of the table, the total number of tokens in interchangeable clusters is shown in absolute numbers and in terms of the proportion of all tokens modelled—remember that the total number of tokens modelled per century is always N = 400 in the diachronic data.

For *vernielen*, we see that over time, there are fewer and fewer clusters in which the variant is the preferred one. For *vernietigen*, we see the opposite trend: the number of contexts where this variant is the preferred one increases over time. With regard to the interchangeability of the variants, we see that in the first subcorpus (16th and 17th century), there is one cluster where both variants are clearly

Table 6.10 Summary of the cluster analyses of *vernielen* and *vernietigen* in the subcorpora

	16/17th	**18th**	**19th**	**20th**	**21st**
vernielen	3 clusters	1 cluster	1 cluster	/	1 cluster
vernietigen	/	1 cluster	1 cluster	3 clusters	1 cluster
interchangeable?	1 cluster N = 182 (0.46)	2 clusters N = 138 (0.35)	2 clusters N = 125 (0.31)	1 cluster N = 33 (0.08)	2 clusters (N = 307) (0.63)

interchangeable. This cluster is also quite large in size, consisting of N = 182 tokens out of the 400 (46%) that were modelled per century. Since *vernielen* occurs as the preferred variant in the other three clusters for this subcorpus, we may assume that *vernietigen* does not yet have any clear semantic contexts in which it prototypically occurs. Instead, in the contexts where *vernietigen* is possible, *vernielen* can still be used as well. One century later, the variants are interchangeable in two out of the four clusters, indicating that they are probably still highly synonymous. However, in the following century, which coincides with the data analysed by Geeraerts (1985, 1988, 1997), the verbs start retreating to their prototypical cores, and the size of the clusters with interchangeable tokens begins to decrease. By the 20th century, *vernielen* has become very infrequent in the data. Only one cluster remains in which this variant occurs and in this cluster, *vernietigen* is also possible. In tandem, the size of the interchangeable cluster also decreases: fewer and fewer tokens occur in semantic contexts where both variants are possible, resulting in only 8% of tokens in an interchangeable cluster in the 20th century. Instead, each variant gets more strongly associated with a particular meaning over time.

A direct comparison with the 21st-century data needs to be approached with caution due to the larger sample size in the contemporary data (N = 486) and, more importantly, due to register differences (newspaper rather than prose). Still, the last column of Table 6.10 shows that in the 21st-century data, there is one cluster where *vernielen* is the prototypical variant (the destroying material artefacts by fire cluster) and one where *vernietigen* takes over (the cluster with authorities destroying abstract things like decisions, plans etc.). The other clusters are semantically more diverse and both variants are possible, although *vernietigen* is also much more frequent in both clusters (taking up 73%–74% of the tokens).

On the one hand, these patterns seem to confirm the trend that is also visible in the diachronic data: *vernietigen* increases in frequency at the expense of *vernielen* over time and *vernielen* is almost restricted to one prototypical usage context, that of the destruction of material artefacts by fire. *Vernietigen* becomes prototypically associated with an authority destroying abstract things and there are indications of it becoming associated with the destruction of concrete objects like livestock or crops by authorities too. On the other hand, one exception to the pattern in the diachronic data is that the interchangeable clusters are quite large in size (N = 307 tokens in total). Even though *vernietigen* takes up nearly 75% of these datapoints, this indicates that *vernielen* is not yet on its way out. Here, it is important to take into account that the contemporary data comes from a different register than the diachronic data (prose versus newspaper): it is likely that things being destroyed (for instance, in fires or through vandalism) are newsworthy topics rather than literary themes.

Thus, the models for the different centuries show how the relationship between the near-synonyms *vernielen* and *vernietigen* has changed over time. Semasiologically, *vernielen* was the major variant in the 16th and 17th centuries, occurring in tokens related to the death of persons and concrete, natural objects. Over the course of the 18th century, it developed its prototypical meaning related to the destruction of (parts of) buildings, often by fire, and this meaning remained its core usage in the 19th century. By the 20th century, the verb had decreased in frequency and its prototypical core was no longer distinguishable in the prose data. In the 21st-century newspaper data, the prototypical core remains strong.

Vernietigen, in contrast, was the less frequent variant in the 16th and 17th centuries and at that time, there were no clear contexts yet where the verb occurred. It was mostly found in a semantically diverse cluster where its near-synonym *vernielen* was possible as well. From the 18th century onwards, the verb started to increase in frequency, and it developed its prototypical sense of the destruction of abstract objects. In the 19th century, it also started to invade contexts where *vernielen* was preferred before (specifically related to the death of persons and to war). In the 20th-century data, *vernietigen* was by far the most frequent variant, taking over two diverse clusters, as well as a cluster related to killing persons and war. In the 21st-century data, it is also the most frequent (but not sole) variant in a cluster related to the destruction of livestock.

Onomasiologically, the analysis showcases how the nuances in the concept 'to destroy' evolve over time and have become more outspoken. For instance, the cluster related to the destruction of parts of buildings is not yet visible in the oldest data, but this is an important cluster in the more recent subcorpora. Similarly, the cluster with abstract objects is not yet distinguished by the analysis for the 16th and 17th centuries, but these objects form a cluster on their own in the 18th-, 19th-, and 21st-century data. Moreover, the analysis also showed how these particular nuances of meaning are typically expressed by a particular verb. In the visualization (Figure 6.6), for instance, there is clearly less overlap (or interchangeability) between the verbs in the later periods (except in the 20th-century data, where *vernielen* is infrequent). Thus, this is a clear case study of how semasiological changes in a verb's meaning interact with changes in an onomasiological pair.

Methodologically, our usage of distributional models combined with a cluster analysis and an identification of representative context words allowed us to show how both verbs changed semantically over time. As an example of a corpus-based exercise in diachronic semantics with a distributional approach to meaning, the links up with the growing interest in quantitative and computational diachronic semantics; see Jenset and McGillivray (2017), Tahmasebi, Borin, Jatowt, Xu, and Hengchen (2021). In terms of the methodological options charted in Table 2.5, our description of the diachronic development of near-synonyms may be compared to Petterson-Traba (2022): whereas we illustrated a full-fledged vector space approach, Petterson-Traba details the interaction of semasiological and onomasiological changes with a behavioural profile approach that makes use of the manual annotation of context features.

In contrast with the previous chapter, where multiple parameter settings were considered, the procedure employed here was quite straightforward. For the contemporary data, we also constructed a set of 100 distributional models with varying parameters, but we only analysed the best model, that is, the model that visualizes the semantics of the verbs the best. In the diachronic data, we relied on this model's parameter settings and adapted them to the diachronic corpus to model tokens from four centuries. While this method proved useful to track semasiological change and to investigate how this interacts with onomasiological variation over time in a single onomasiological pair, it cannot be applied on a dataset consisting of a larger set of linguistic variables. More specifically, with a larger set of linguistic variables, it becomes unfeasible for the linguist to analyse a large set of visualizations and to interpret the context words that occur in each cluster for each variant to obtain a picture of how interchangeable two or more variants are. In addition, while in this chapter we could build on the earlier synchronic work to get an idea of the parameter settings that would work well for the diachronic analysis, this previous knowledge is not available for (near) synonyms that have not yet been analysed with a distributional methodology. Accordingly, in some circumstances it is necessary to choose the other option that we mentioned at the very start of this chapter, viz. to include a variety of models in the analysis and to investigate the stability of descriptive results across this set of models. In the lectometric chapters of the book, to which we will now turn, we will illustrate how this could work.

The lectometric studies that we will showcase differ in yet another dimension from the materials presented in this chapter. Our story of *vernielen* and *vernietigen* involves an interaction of semasiological and onomasiological change, and as such, it deals with what we may call a conceptual reorganization of the lexicon. Such a reorganization is characteristic for the evolution of near-synonyms. But we may narrow the perspective to strict synonyms, where any lexical variation occurs against a background of conceptual stability, that is, where we look at lexical changes while keeping the meaning constant. Treating lexical variation as a sociolinguistic variable, as we will in the lectometric chapters, assumes exactly that perspective.

The bottom line

- A joint distributional analysis of semantically related lexical items, like near-synonyms, allows to identify shared senses as common clusters in a cluster analysis.
- The absence of a model configuration that works optimally across the board suggests a choice between two alternatives: select a specific model on the basis of additional or external evidence, or include a variety of models

in the analysis and study the stability of descriptive results across that set of models.

- The method of this chapter illustrates the first alternative: as a first step, a manageable number of medoids for a set of models with variable parameter settings in one corpus is selected; as a second step, one model for consecutive centuries is manually chosen.
- The diachronic evolution of near-synonyms can be described distributionally as an interaction of semasiological and onomasiological changes.

PART IV

LECTOMETRIC METHODOLOGY

With Chapters 7 and 8, we turn to the lectometric part of the study. As was the case with the semasiological and onomasiological perspective in Chapters 3, 4, 5, and 6, the lectometric perspective is represented by two sets of chapters: a methodological one in Chapters 7 and 8, and a descriptive one in Chapters 9 and 10. Chapter 7, then, introduces the formulae that use lexical variation to quantify the relationship between language varieties. Chapter 8 specifies how a token-based distributional method may be used to identify the sets of synonymous expressions on which such quantification is built.

7
Quantifying lectal structure and change

In the preceding four chapters, we looked at the lexeme-lection-lect triangle (as we called it in Chapter 1) from the most common, most traditional perspective: we described the varying associations between forms and readings, and investigated how that semasiological-onomasiological variation might be influenced by lectal variation, including 'chronolectal' variation, that is, diachronic change. In a second group of four chapters, we will now reverse the perspective. Instead of treating semasiological-onomasiological variation as a variable dependent on lectal factors, we take variation in form-meaning pairings as an input variable for determining lectal structure. More specifically, by focusing on formal onomasiological variation, we can treat lexical variation as a sociolinguistic variable in the Labovian sense, and then aggregate over such variables to find out how various lects relate to each other, both synchronically and diachronically. As in the preceding four chapters, the first pair of the set have a methodological orientation, while the second pair serve purposes of description and illustration. The division of labour between the two methodological Chapters 7 and 8 corresponds roughly to the essential components of our lectometric approach, which applies aggregate-level distance calculations to onomasiological profiles. The present chapter, then, expands on the lectometric measures that we outlined in Section 1.3. To flesh out the formulae, we will rely on data from previous research that does not yet use vector-based modelling. The following chapter will integrate this quantitative perspective on lectal structure with a distributional workflow. Specifically, it will focus on how to derive onomasiological profiles from the clustering of tokens that we explored in Chapters 3–6.

Within the field of (dia)lectometry, our approach is most closely related to the corpus linguistic lexical variation studies of Grieve and his associates (Grieve, Asnaghi, and Ruette 2013; Grieve 2016; Grieve, Nini, and Guo 2018; Grieve, Montgomery, Nini, Murakami, and Guo 2019). Two differences stand out, though. First, relating to the procedures we will describe in Chapter 8, Grieve's methods do not rely on distributional semantics to identify relevant lexical variants and their precise area of overlap (the 'envelope of variation' as it is called in the tradition of sociolinguistics). Second, relating to the lectometric measures presented in the current chapter, we try to cover a broader range of lectal dimensions. By including chronological periods or social stratification in the distance calculations, our focus departs more than Grieve's from the traditional geolinguistic interests of dialectometry.

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0007

7.1 Measuring lectal distances

In Section 1.3, we introduced formulae (1.1), (1.2), and (1.3) as basic measurements for investigating lectal distances on the basis of onomasiological profiles. For ease of reference, the formulae are repeated here as (7.1), (7.2), and (7.3).

(7.1) *Uniformity for a single concept*

$$U_Z(Y_1, Y_2) = \sum_{i=1}^{n} \min\left(F_{Z,Y_1}(x_i), F_{Z,Y_2}(x_i)\right)$$

(7.2) *Average uniformity for a set of concepts*

$$U(Y_1, Y_2) = \tfrac{1}{n} \sum_{i=1}^{n} U_{Z_i}(Y_1, Y_2)$$

(7.3) *Weighted average uniformity for a set of concepts*

$$U'(Y_1, Y_2) = \sum_{i=1}^{n} U_{Z_i}(Y_1, Y_2) \cdot G_{Z_i}(Y_1 \cup Y_2)$$

As a reminder, Y_1 and Y_2 refer to the lectal datasets we intend to compare. Z is a concept that may be expressed by n competing expressions in an onomasiological profile, from x_1 to x_n. The (relative) frequency of an expression x_i in naming Z in the dataset Y_1 is represented by $\mathrm{F}_{\mathrm{Z},\mathrm{Y}_1}(\mathrm{x}_\mathrm{i})$, while $\min(\mathrm{F}_{\mathrm{Z},\mathrm{Y}_1}(\mathrm{x}_\mathrm{i}), \mathrm{F}_{\mathrm{Z},\mathrm{Y}_2}(\mathrm{x}_\mathrm{i}))$ indicates the minimum value of the frequencies of x_i for Z in Y_1 and Y_2. To calculate a uniformity index for a single concept, the minimum values for all n items are summed. If more than one concept is investigated, the uniformity index is defined as an average of the uniformity indexes of the separate concepts. The average can be computed as a mean, or as a weighted average, with the relative frequency of each concept in the combined datasets, expressed as $\mathrm{G}_{\mathrm{Z}_\mathrm{i}}(\mathrm{Y}_1 \cup \mathrm{Y}_2)$, used as a weighting factor for the uniformity index of each concept separately.

Before we turn to examples of how these formulae can be used, there are three technical topics to be discussed: the statistical nature of our uniformity index as a distance measure, the role of significance tests in the calculations, and alternatives for the weighting of clusters and concepts. First, distances between quantitatively characterized objects can be calculated in a number of ways. The distance measures in formulae (7.1), (7.2), and (7.3) are based on a City Block Distance measurement. To adequately characterize City Block Distance, it may be contrasted with Euclidean distance, which is probably the most intuitive and straightforward method for thinking geometrically about distances. The Euclidean distance between two points *a* and *b* is defined as the square root of the sum of the squares of the differences between the corresponding positions of the points on each of the n dimensions that constitute the space under consideration. The formula for Euclidean distance was mentioned earlier, as equation (4.2). In a two-dimensional space, so with n=2, the formula in (4.2) corresponds with the Pythagorean theorem that in a rectangular triangle, the sum of the squares of the rectangular sides equals the square of the hypotenuse. The formula merely generalizes this idea to an n-dimensional space. As such, it also applies to data like the ones in our toy example in Table 1.6: the 'points' characterized there are the lectal varieties Arp and Picabia expressing the concept NONSENSE, and the three dimensions with regard to which those points are characterized are the lexical items *tressli*, *bessli*, and *nebogen*. In this way, the linguistic similarities and differences that we are interested in can be measured in terms of geometrical distances in an n-dimensional space.

Against the background of the formula for Euclidean distance, City Block Distance can now be defined as in Formula (7.4). Here, the distance is the sum of the differences of the positions of *a* and *b* on each of the n dimensions of the space in which *a* and *b* are measured. Applied to a two-dimensional space, with an x-axis and a y-axis, this means that the distance between *a* and *b* is first measured along the x-axis and then along the y-axis; the total distance is the sum of these two distances. According to this method, you can only move along one dimension at the same time. Going diagonally, as in Euclidean distance, is not possible. Because City Block Distance is calculated as the distance in dimension x plus the distance in dimension y, it resembles the way one moves in a city: you have to walk around the block along one street and then turning the corner along the other instead of going diagonally through the buildings. This explains the name of this distance measure (and why it is also called *Manhattan distance*).

(7.4) *City Block Distance*

$$CBD\,(a,b) = \sum_{i=1}^{n} |a_i - b_i|$$

Although Euclidean distances are by definition smaller than or equal to City Block Distances (the shortest route between two points is a straight line), they yield similar results. We opt for a measurement in terms of City Block Distances, though, because it yields a straightforward quantitative translation of the idea that similarity of linguistic behaviour consists of an overlap in the expressive choices made by language users. As explained in Chapter 1, we take a usage-based perspective, and accordingly think of linguistic similarity as commonality of behaviour. That idea receives an intuitive translation in terms of City Block Distance: the difference between lects is the sum of the choices the speakers of those lects make with regard to the use of n functionally equivalent lexical items. Further, City Block Distance has the advantage of not amplifying the effect of a large difference in a single dimension. Since distances are not squared as they are in Euclidean distance, all dimensions contribute on an equal basis to the overall difference. But City Block Distance measures dissimilarity whereas our uniformity formulae (7.1), (7.2), and (7.3) capture similarity. The relationship between both perspectives is given by (7.5) which expresses that uniformity as defined earlier is the complement of a normalized City Block Distance—normalized in the sense that the division by two maps the distances on an interval [0,1].

(7.5) *Uniformity as a function of City Block Distance*

$$U\,(Y_1, Y_2) = 1 - \tfrac{1}{2}CBD\,(Y_1, Y_2)$$

Second, using a distance measure like City Block Distance presupposes that the relative frequencies of the items in the profiles are a reliable estimate of the relative frequencies in language usage at large. But with small samples, we cannot simply assume that the sample on which we measure the distances delivers a good picture of the overall linguistic behaviour: the lower the frequencies in the sample, the higher the danger that they may not be entirely representative. We may therefore introduce a double form of control. To begin with, we can work with a threshold for the frequency of the profile in the lectal varieties, by only considering profiles that are attested at least n times in each of the lectal varieties. In addition, we can apply a statistical test to judge how confident we can be that the differences in the sample reflect actual differences, rather than being a side-effect of the small sample size. A Fisher Exact test and a Log-likelihood Ratio test are statistical tests for significance that handle low-frequency samples well. The Log-likelihood Ratio test (which is used in the examples of Section 7.2) yields a value for the log-likelihood test statistic *–2 log* λ. On the basis of this log likelihood statistic a p-value can be calculated for the chance that the underlying distribution is the same for both lectal profiles, in spite of the observed frequency differences in the sample: if $p > 0.05$, it is considered unlikely that the frequency differences in the sample reproduce real differences of behaviour in the lectal varieties. In practice, we use the p-value of the Log-likelihood Ratio test as a restriction on the distances measured on the sample. If $p > 0.05$, we assume that the differences in frequency across the lectal varieties are due to chance rather than being representative for actual differences in behaviour, and we then set the City Block Distance to 0 and the U-value to 1. Conversely, if $p < 0.05$ in the Log-likelihood Ratio test, the profile-based distance measurements are considered to be trustworthy.

By default, concepts with complete uniformity—more precisely, without attested significant variation—are maintained in the aggregate calculations. If you are interested in the overall distance relations between lects, the concepts for which they do not differ in their linguistic habits are arguably just as important as the concepts for which their lexicalization preferences differ. However, if you want to zoom in on the cases where significant differences show up, the concepts without demonstrable variation may be omitted from the aggregate calculations. An example of this approach is included in Chapter 10.

Third, we may point to alternative approaches with regard to the weighting of clusters and concepts as incorporated in (7.1), (7.2), and (7.3). This discussion is not immediately relevant for the remainder of the book, though. We will not apply any of the alternatives, but we include the discussion for the sake of completeness, and to point to further perspectives that may be pursued in the development of the lectometric research programme. Two kinds of modulation on the calculations are worth mentioning, one that involves the weighting of clusters within a concept, and one that involves the weighting of concepts.

In the first place, notice that (7.1) does not take into account the internal semantic structure of the concept, that is, the fact that the distributional approach identifies meaningful groups of tokens within concepts. As we saw in Chapter 6, an onomasiological analysis along distributional lines may reveal the area of overlap between lexical items, and that is the area in which we want to study the alternation between the items; clusters outside the area of overlap are discarded with a procedure that will be specified in Chapter 8. But we also saw in Chapter 6 that even within that area of overlap, there may be distributionally defined clusters of tokens. Treating such clusters differently in the calculation of uniformity measures across lects could be relevant when the onomasiological profiles for different clusters are different, that is, when the choice between two synonymous items depends on the specific usage contexts. The procedure that can then be followed is fully parallel with the treatment of entire concepts described above: first, to establish whether the cluster-based lexical profiles of a given concept are significantly different and to treat non-significant clusters as exhibiting hundred per cent uniformity while treating significantly different clusters as exhibiting 'real' differences; second, to aggregate over the uniformity measures of different clusters weighted by the relative frequency of clusters.

In the second place, we may consider an asymmetrical weighting of concepts. In (7.3), the weighting of concepts is based on the reasoning that contexts that are communicatively less important should play a lesser role in aggregate calculations of distances. The onomasiological profiles for concept x in lects A and B tell us something about the lexical overlap between an A-lecter and a B-lecter when talking about x. Next, when aggregating over concepts x and y, we take into account that the probability that they might talk about x is lower than the probability they might talk about y, and that probability is a combination of the inclination of an A-lecter to talk about x and the inclination of a B-lecter to talk about x. In this perspective, it doesn't matter all that much that these inclinations might differ: what we're interested in is the overall communicative incidence of talk about x compared to talk about y. Still, when the relative frequencies of x in A and B differ, an asymmetric approach may be considered, in which the weighting of concepts is based on their frequency distribution in A alone or in B alone. The intended construct could then be paraphrased as: 'If we pretend that B-lecters talk about x (and y and z etc.) as frequently as A-lecters, how would that affect the overall uniformity between A and B? And conversely from the point of view of B?' One may think about this in terms of discourse practices. If a communicative culture is defined in terms of the relative importance with which certain topics are discussed, an asymmetric weighting of concepts is a way of examining the effect of communicative culture on linguistic uniformity: if the two lects had the same discursive culture, either the A one or the B one, how would that influence the uniformity measures?

Thinking further along these lines suggests yet another weighting schema. It again makes use of the concept frequency of just one lect, but instead of systematically taking the perspective of either one of the lects, it takes into account the concept frequency of the smallest lect in the comparison, whichever lect that is.

For a given concept, the weighting factor is the fraction of, as numerator, the onomasiological profile with the smallest number of tokens among the ones compared, and as denominator, the sum of these smallest lect concept frequencies. From a communicative perspective, choosing the frequency of the smallest lect as weighting factor can then be thought of as focusing on the minimal number of events in the observed communicative space in which speakers of the A-lect and speakers of the B-lect talk about the same things.

7.2 Standardization and informalization

We can illustrate the measures introduced in (7.1), (7.2), and (7.3) with the case study described in Geeraerts (2018a), which investigates aspects of the recent evolution of Dutch (more specifically, the developing relationship between Netherlandic Dutch and Belgian Dutch) against the background of contemporary views on the evolution of standard languages in Europe. Two ideas are dominant in the current theory formation: on the one hand, Auer's typology of dialect and standard language constellations (Auer 2005, 2011); on the other, the notions of 'destandardization' and 'demotization' introduced by the SLICE (Standard Language Ideology in Contemporary Europe) network. Disregarding many subtypes and variations painstakingly described by Auer, the former involves the idea that the languages of Europe tend to follow a long-term evolution from exoglossic diglossia in the medieval period to endoglossic diglossia in Early Modern times, followed by an evolution to a diaglossic situation—a fully fleshed out stratigraphic spectrum between standard language and base dialects—in the Modern period. In some cases, dialect loss in the contemporary period may further lead to a shrinking of the spectrum: the original dialects disappear, in the situations in which they were used a colloquial version of the standard language takes over, and the range of stratigraphic variation becomes narrower. In the framework developed by the SLICE network (Coupland and Kristiansen 2011; Kristiansen 2016), contemporary changes at the top of the stratigraphic spectrum are considered. Specifically, an increasing, postmodern tolerance for variation is supposed to take shape in two different forms: either as 'demotization' or as 'destandardization'. Demotization (a terminological reference to the *demotisierung* introduced by Mattheier 1997) involves cases in which more variation enters into the standard language but in which the standard language ideal as such is not affected: the 'best language' becomes more open to variation, but the normative concept of a best language as such is not weakened. Destandardization by contrast involves changes through which established standard languages lose their exclusive status as 'best language' and a broader range of speech varieties is accepted within the public sphere. The distinction between destandardization and demotization has triggered a lot of debate, not least so because they were not introduced with a clear operational definition.

To see how a lectometric approach may shed light on the issue of destandardization, the conical representation of stratigraphic spectra used by Auer (2005) provides a fruitful starting point. The conical visualization assumes an essentially two-dimensional structure of variation. The vertical dimension represents a hierarchical ordering along a situational dimension: the higher a context of use is situated in the stratificational cone, the more standard language use will be expected. Informative media language, for instance, whether written or spoken, will generally be expected to conform to the standard language norm, regardless of how internally varied that norm may be. Casual conversations in an informal context, by contrast, will generally come with less outspoken expectations with regard to standard language use. The vertical dimension, in other words, assumes different contexts of use, but also an attitudinal ordering among those contexts. The horizontal dimension, conversely, involves the variation that exists at any of the stratigraphic layers. It may primarily be thought of in terms of geographic variation: to the extent that dialect differences exist, they will show up more readily in situations with less stringent standard language expectations. But the geographical dimension would obviously not be the only one to be considered; at least social features (such as the speaker characteristics of sociolinguistics) and thematic differences (as for instance in Language for Special Purposes) would need to be added to get a more complete picture of the variation. The conical representation, in other words, is a simplified model of a multidimensional variational structure, but precisely as a simplified model, it can help us to think analytically about the dynamics of standardization—always keeping in mind that more complicated approaches may need to be introduced later to accommodate the multidimensionality of variation.

For explanatory purposes, we will assume a minimal conical structure with two layers. In terms of Auer's typology, we make no assumptions about the real-world nature of the layers or their specific position within the stratigraphic spectrum. They could for instance involve a configuration with at one end a supraregional written standard language, and traditional local dialects at the other end, or they could involve the written standard in relation to the spoken standard. And there would be other options: in a diaglossic linguistic situation with many intermediate levels between the top and the bottom layer, the relationship between any two levels could be represented by a minimal cone of the kind we will be using. Given a simplified two-layered configuration, then, when could we talk about an increasing or decreasing standardization? Three types of structural change may be identified as possible definitions of destandardization.

In the first place, if we think of standardization as the final stage described by Auer (the lower levels move up towards the top register), destandardization would be the opposite: the linguistic distance between the upper and the lower level increases, as a result of changes in the upper level or changes in the lower level, or both. Figure 7.1 graphically represents such a process of decreasing stratigraphic standardization. The dotted line represents the original situation, while the solid line depicts a situation in which the upper level has moved further away from the base level. Terminologically, we will refer to this type of destandardization as *hierarchical destandardization*. In the situation pictured in Figure 7.1, both levels in themselves maintain their original degree of variation, as represented by the surface of the ellipse. This is not necessarily the case, though: increasing distances may go hand in hand with changes in the internal variation of the levels. Below, in relation to Figure 7.3, we will suggest how such internal changes can be quantified.

In the second place, a decrease in the distance between the two levels might still be considered a form of destandardization if the movement between the two levels is rather from the top level to the bottom level rather than the other way around. When not just the degree of rapprochement but also the direction of the process is taken into account, a distinction can be made between the type of standardization that fits the traditional conception of standardization, and developments in the other direction. In the former, the features of the hierarchically superior level trickle down towards the inferior one, as when colloquial language use loses most of the original dialect features. In the latter, opposite process, what used to be informal language percolates into the formal, upper-level situations, thus bringing qualitative change in the substance of the standard language. This second process is a type of destandardization to the extent that the old standard norm gives way to a new one that is influenced by the initial informal, colloquial, less valued forms of language use. To distinguish this development from the previous type, we will identify it as *informalization*. It is pictured in Figure 7.2.

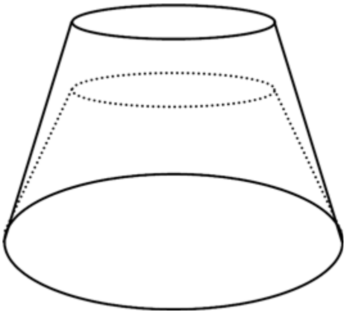


Figure 7.1 Hierarchical destandardization as increasing distance between strata

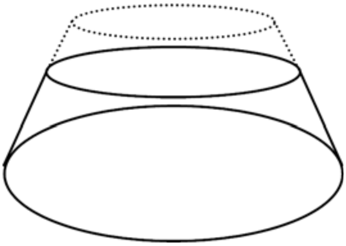


Figure 7.2 Informalization as top-down decreasing distance between strata

In the third place, destandardization may take the form of increasing variation within the highest level, regardless of whether this growing variation correlates with changes in the relationship with regard to the other level. Again, various terminological alternatives can be considered for identifying such a process: it might be called dehomogenization, but heterogenization or internal destandardization could also be considered. Figure 7.3 graphically represents such a process of internal destandardization.

If these three configurations may help us to understand better what might be going on in terms of destandardization processes, the next step is to provide operational definitions of the three configurations: to distinguish between hierarchical destandardization, informalization, and dehomogenization, we would like to measure what is happening, and that is where lectometry can help us. But before we proceed to a quantitative, lectometric definition of the three developments, let us consider how the destandardization/demotization framework (the SLICE point of view) relates to the three potential processes. Demotization, with its emphasis on the relaxation of existing standard norms, is probably best conceived of in terms of the third process: more variation enters into the standard language, but the position of the standard with regard to other levels of language use remains roughly the same. The SLICE notion of destandardization, on the other hand, seems to relate primarily to the second process, to the extent that higher level language use grows closer to lower level language use, though not in the bottom-up way that is expected by traditional standard language ideologies but rather in a top-down way: hierarchical differences are levelled out, but they are levelled out in favour of an initially subordinate level rather than the other way around.

To provide lectometric definitions of the three types of destandardization that we distinguished, one extra type of measurement needs to be introduced. Formula (7.6) provides a measure for what we refer to as the internal uniformity of lexical usage in a given dataset, based on the assumption that a language situation could be considered more uniform to the extent that there are less competing forms for expressing a given concept, and to the extent that dominant forms exist within that set of alternatives. In (7.7) and (7.8), the internal uniformity measure is aggregated over a set of n concepts, respectively without and with weighting. Against the background of the classification of perspectives in Figure 1.1, these are not strictly

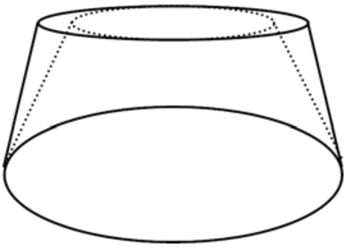


Figure 7.3 Dehomogenization as increasing variation within one stratum

speaking lectometric measurements, because they do not capture the relationship between different lects. In themselves, they only describe properties of onomasiological profiles within a single lect, but they may obviously be compared across lects, and that is how they will be used here. (In Section 7.3 we will come back to the internal uniformity measure and situate it in a wider context.)

(7.6) *Internal uniformity for a single concept*

$$I_Z(Y) = \sum_{i=1}^{n} F_{Z,Y}(x_i)^2$$

(7.7) *Average internal uniformity for a set of concepts*

$$I(Y) = \tfrac{1}{n} \sum_{i=1}^{n} I_{Z_i}(Y)$$

(7.8) *Weighted average internal uniformity for a set of concepts*

$$I'(Y) = \sum_{i=1}^{n} I_{Z_i}(Y) \cdot G_{Z_i}(Y)$$

The definitions of the three types of destandardization now follow in a straightforward fashion. We consider four lects, differentiated by stratificational position and chronology. *H* represents the stratificationally higher situation, where we may expect language use that is representative of or at least closer to standard language use (to the extent that standardization exists at all in the linguistic situation at hand), and *L* a lower-ranking situation. If t_1 and t_2 represent an earlier and a later point in time, then the three types of destandardization (in the order in which they were introduced above) are defined as follows.

(7.9) Hierarchical destandardization occurs
if $U(H_{t_1}, L_{t_1}) > U(H_{t_2}, L_{t_2})$
or if $U'(H_{t_1}, L_{t_1}) > U'(H_{t_2}, L_{t_2})$

(7.10) Informalization occurs
if $U(L_{t_1}, H_{t_2}) > U(H_{t_1}, L_{t_2})$
or if $U'(L_{t_1}, H_{t_2}) > U'(H_{t_1}, L_{t_2})$

(7.11) Dehomogenization occurs
if $I(H_{t_1}) > I(H_{t_2})$
or if $I'(H_{t_1}) > I'(H_{t_2})$

Corresponding to the three processes introduced above, the formulae will be self-evident, except perhaps in the second case. Formula (7.10) measures the direction of change by comparing the similarity between, on the one hand, the lower level at time t_1 and the higher level at t_2, and on the other, that between the higher level at time t_1 and the lower level at t_2. If the former is bigger than the latter, the attraction exerted by the originally lower level is stronger than the attraction of the higher level in the initial stage, or in other words, the change is from bottom to top rather than from top to bottom.

To illustrate the formulae and the phenomena they capture, we use a longitudinal study on the lexical development of Dutch in the lexical field of clothing terms. Although, as we shall see, the results can be plausibly interpreted in the light of the recent evolution of Dutch, it will be clear that a single lexical field is not enough to yield general conclusions about the evolution of Dutch. We would need to know more about other parts of the vocabulary and other levels of linguistic structure for a comprehensive picture. In this sense, the results are primarily meant to illustrate the method rather than to support far-reaching descriptive statements. The study drawn on here is a replication of Geeraerts, Grondelaers, and Speelman (1999), in which clothing terms and football terms were followed from 1950 over 1970 to 1990 in Netherlandic Dutch and Belgian Dutch sources. These sources primarily comprised supraregional written data from national newspapers and magazines, with the addition of shop window materials for the 1990 clothing terms. These 'shop window materials' took the form of price tags in local shops, with the exclusion of national or international chain stores. In this way, a second situational layer is added to the dataset: if naming practices differ in less formalized contexts, this is one communicative situation in which less formal usage may be found. The shop window data were collected in two Dutch and two Flemish towns with similar characteristics: the centrally located university towns Leiden and Leuven, and the peripheral towns Maastricht and Kortrijk, each with a smaller university. The replication study of 2012 (see Daems, Heylen, and Geeraerts 2015 for an extended description) repeated the 1990 clothing terms study, so that we now have real time data for two stratigraphic levels at two points in time—a crucial condition for applying the definitions in (7.9)–(7.11). In quantitative terms, the dataset contains 8797 observations for Belgian Dutch in 1990, and 3761 for 2012. For Netherlandic Dutch, the figures are 6205 and 5255 respectively.

The 14 concepts included in the analysis are the following: SHIRT$_{\text{M}}$, SHIRT$_{\text{F}}$, T-SHIRT$_{\text{MF}}$, SWEATER$_{\text{MF}}$, CARDIGAN$_{\text{MF}}$, TROUSERS$_{\text{MF}}$, JEANS$_{\text{MF}}$, LEGGINGS$_{\text{F}}$, SKIRT$_{\text{F}}$, DRESS$_{\text{F}}$, SUIT JACKET$_{\text{M}}$, SUIT JACKET$_{\text{F}}$, JACKET$_{\text{MF}}$, SUIT$_{\text{MF}}$. The subscripts indicate whether the item of clothing is meant for women or men. This could either mean that the clothing type is gender-specific (like SKIRT) or that the same type receives different names when worn by men or women (as in a jacket as part of a suit, which is often called *colbert* in the case of men, but hardly ever so in the case of women). If the gender distinction does not correlate with differences of naming pattern, the concept is considered gender neutral. The lexical alternatives involve synonyms like *jeans, jeansbroek, spijkerbroek*. Only in the case of SKIRT no alternatives emerge: skirts are always called *rok* (but because we want to have an aggregate-level view of the lectometric relations, concepts with little or no lexical variation are retained as part of the calculations). Overall, statistical significance is checked by applying a Log-likelihood Ratio test with a threshold of 5% to the naming patterns under comparison.

If we then collect the results for the Belgian Dutch dataset, the three types of possible destandardization captured by (7.9)–(7.11) appear as follows. Restricting the overview to non-weighted averages, the *B* figures refer to the higher-level stratum of national magazine data, while the *LeuKor* figures are based on the shop window materials in Leuven and Kortrijk.

(7.12) Hierarchical destandardization

$U(B_{90}, LeuKor_{90}) = 50.47$
$U(B_{12}, LeuKor_{12}) = 73.72$
$U(B_{90}, LeuKor_{90}) < U(B_{12}, LeuKor_{12})$

(7.13) Informalization

$U(LeuKor_{90}, B_{12}) = 60.25$
$U(B_{90}, LeuKor_{12}) = 53.12$
$U(LeuKor_{90}, B_{12}) > U(B_{90}, LeuKor_{12})$

(7.14) Dehomogenization

$I(B_{90}) = 69.21$
$I(B_{12}) = 74.96$
$I(B_{90}) < I(B_{12})$

For the Netherlandic Dutch dataset, the *N* figures refer to the higher-level stratum of national magazine data, while the *LeiMaa* figures are based on the shop window materials in Leiden and Maastricht.

(7.15) Hierarchical destandardization

$U(N_{90}, LeiMaa_{90}) = 69.07$
$U(N_{12}, LeiMaa_{12}) = 73.62$
$U(N_{90}, LeiMaa_{90}) < U(N_{12}, LeiMaa_{12})$

(7.16) Informalization

$U(LeiMaa_{90}, N_{12}) = 61.57$
$U(N_{90}, LeiMaa_{12}) = 84.93$
$U(LeiMaa_{90}, N_{12}) < U(N_{90}, LeiMaa_{12})$

(7.17) Dehomogenization

$I(N_{90}) = 68.48$
$I(N_{12}) = 71.06$
$I(N_{90}) < I(N_{12})$

The evolutions contained in these figures turn out to point to standardization, rather than destandardization. In both national varieties of Dutch, the distance between the stratigraphic layers diminishes and the internal uniformity of the upper layer increases. In the Netherlandic case, the directionality of the compression corresponds to a traditional conception of standardization: the lower level moves in the direction of the upper level. In the Belgian Dutch data, on the other hand, the opposite is the case, and this is the only example of 'destandardization' as defined above that may be found in the dataset. This destandardizing aspect of the Belgian Dutch development needs to be understood in a broader historical context. (For more background, see Geeraerts and Van de Velde 2013 or De Sutter 2017 for a comprehensive view of recent developments in Netherlandic and Belgian Dutch.)

In Flanders, the standardization process that started off, as in most European countries, in the Early Modern Period was slowed down as a result of Flanders' political separation from the Netherlands during the Eighty Years' War. Standard Dutch started to develop in the Netherlands in the course of the 17th century, but as Flanders was politically separated from the Netherlands, remaining under foreign rule, it did not link up with this process of standardization. Rather, French was used more and more as the language of government and high culture, a practice that received an important impulse after the birth of the Belgian state in 1830. Dutch then survived predominantly in the form of a range of Flemish dialects. However, as a result of a social and political struggle for the emancipation of Flanders and the Dutch-speaking part of the Belgian population, Dutch again gained ground as a standard language (the language of learning, government, and high culture) in Flanders. This process started somewhat hesitantly in the late 19th century as a typically romantic movement, gained momentum during the first half of the 20th century, and finally made a major leap after World War II and during the booming 1960s. Importantly, the official linguistic policy of Belgian Dutch during this process of standardization was based on a normative dependency on Netherlandic Dutch: when the use of Dutch as a language of higher education and culture spread, the existing Netherlandic Dutch norm was officially promoted, in educational practices and elsewhere, as the model to be taken over. This linguistic policy was successful: if we look at our dataset for the evolution of $U(B, N)$ over 60 years, we see a steady increase from 1950 over 1970 to 1990: U' figures rise from 69.21 over 77.50 to 86.50. From 1990 to 2012, however, the uniformity drops from 86.50 to 81.50. If this drop signals a growing independence of Belgian Dutch with regard to Netherlandic Dutch, then the 'destandardizing' directionality revealed in (7.13) makes sense. At the same time as looking away from (or at least looking less attentively at) Netherlandic Dutch as a norm to be adopted, Belgian Dutch makes more room for its own forms of linguistic usage. When all aspects of the evolution are taken into account, the 'destandardizing' change of Belgian Dutch does not signal an abandonment of the traditional model of standardization, but rather reveals that the Belgian Dutch standardization process has acquired a dynamic of its own, with more autonomy with regard to Netherlandic Dutch than used to be the case.

A case study like this shows the relevance of a quantitative approach to onomasiological variation and the relationship between lects: lexical variation can be successfully treated as a sociolinguistic variable, and the relevant phenomena of an onomasiological and lectometric kind can receive quantitative definitions that help to clarify what is going in the language. Evidently, a comprehensive picture of the standardization dynamics in the Low Countries should include a variety of speech situations. For colloquial Belgian Dutch, for instance, existing research includes contexts ranging from advertising (Van Gijsel, Speelman, and Geeraerts 2008), over social media (Hilte, Daelemans, and Vandekerckhove 2020) and tv series and films (Jaspers and Van Hoof 2015), to educators' language (Delarue and Ghyselen 2016) and dinner-table child-directed speech (Zenner and Van De Mieroop 2021). Within that range, the analysis of large text corpora has an obvious role to play, but at the same time, our case study points to the importance of methodological scaling up. It is based on a single lexical field, and not much more is feasible if the methodology relies to a large extent on the manual collection of data. There are major opportunities for lexical research in the growing availability of corpus materials, but exploiting those opportunities requires a corpus-oriented method that is maximally automated. So, this will be one of our dominant questions: can we overcome the limitations of a manual workflow by making use of distributional semantics?

7.3 Lexical diversity and lexical success

Measures of lexical distance and homogeneity do not stand on their own; they invite an interpretation and analysis. Interestingly, such a further scrutiny can take its starting point in each of the corners of the lexeme-lection-lect triangle. Obviously and inevitably, one may first look at the observed effects from a lectal point of view, that is, from the point of view of the language varieties themselves. This was the case in the previous section, where we investigated potential differences between Netherlandic Dutch and Belgian Dutch, and interpreted those differences in terms of the specific history of the varieties and their mutual relationship. The background questions go beyond the individual case, though: could there be any generalizations along the lines we discerned in the case study? If we look at other pluricentric languages with an initially asymmetric relationship, what are the patterns of their development, and under which conditions does a minority centre acquire independence? Or, to come back to the question that initiated the investigation in 7.2, is there indeed a postmodern informalization effect in the languages of Europe, how strong is it, and how does it manifest itself in different languages? Or still, on a different dimension, what factors influence rates of lectal change? Answering questions of this kind requires performing lectometric studies on a large scale—and here again, an automated or semi-automated distributional workflow might be very helpful.

But second, lectometric findings do not depend on the lects alone. Already in the original Geeraerts, Grondelaers, and Speelman (1999) study, we could notice differences between the two lexical fields under investigation, viz. that of football terms and that of clothing terms. Even though the overall trends between the two fields were similar, the internal dynamics of the fields were slightly different. A similar result is found in Daems (2022). A corpus-based calculation for the lexical fields of traffic, information technology, and emotion concepts yields the lowest degree of uniformity between Netherlandic Dutch and Belgian Dutch for the traffic terms, an effect that may be due to the existence of different official traffic codes for the two countries. Information technology and emotion have higher, roughly similar degrees of uniformity. In the former case, this seems to reflect the common pressure of English in this highly international domain. In the latter, the smaller distance between the varieties may be due to the fact that emotion is a basic field with an old and well-established vocabulary. The theoretical point then is that lexical field belongs in the 'lection' corner of the triangle; it is a factor relating to the meaning of the terms. This recognition leads to a more general question: what is the role of concept characteristics in shaping lectometric results? For instance, consider concept frequency. In formula (7.3), the overall frequency of a concept is used as a weighting factor: a weighted uniformity measure reflects the idea that less frequent concepts play a communicatively secondary role, and that their profiles may accordingly count for less in a calculation of the similarity in the communicative behaviour of two groups of people. But if we keep frequency out of the calculation, it can function as an analytic variable in its own right: are (unweighted) uniformity values similar across frequency ranges? This is but one example of a potentially relevant concept characteristic; others are mentioned below.

Third, the high uniformity values for information technology concepts in Daems (2022) point to the relevance of looking at lectometric measures from the point of view of the lexemes: the onomasiological profiles in the lexical field of information technology are largely filled with English terms, and so, given that Netherlandic Dutch and Belgian Dutch are in equal measure subjected to the influence of global English, the origin of the lexemes in a profile offers a relevant perspective for looking at lectometric results. But again, the observation that next to lectal and conceptual factors the characteristics of the lexemes play a role opens up a wider perspective: instead of using onomasiological profiles wholesale as input for aggregate-level lectometry, we can look inside the profiles, and study the lexemic innards of a profile as a topic in its own right. In so doing, we are again reversing the perspective in terms of Figure 1.1: we are looking at formal onomasiological variation—profiles and their lexical characteristics—as a phenomenon shaped by other factors, rather than as a factor giving shape to lectal relations and lectal changes.

There are two foci for such an internal examination of profiles: we can probe their internal lexical diversity, or the success of specific lexemes or groups of lexemes. The first perspective is one that we are already familiar with. The measures of internal uniformity defined by formulae (7.6) to (7.8) embody one specific way of quantifying lexical diversity (or more precisely, the absence thereof):

the homogeneity of a profile is considered to increase to the extent that there are less alternatives within a profile, and to the extent that one or more of the alternatives tends to dominate over the others. So, a profile with three competitors with a relative frequency of 50%, 25%, and 25% yields a lower value than one with two alternatives that each have a relative frequency of 50%, but in turn, the latter profile has a lower internal uniformity than one with two terms and a 75–25% distribution. This is only one method of measuring lexical diversity, though. A more traditional, system-oriented view of linguistic structure could be to look exclusively at the number of different lexical types in a profile, without considering their relative frequency in actual usage: how many different words are there to express a given meaning, regardless of their difference in frequency? Conversely, when that frequency distribution is taken into account, standard statistical measures of dispersion like variance or standard deviation could be used to gauge the variability within the profile. In view of these alternatives, it may be noted that internal uniformity as defined in (7.6) and following combines the essentials of these alternatives, that is, it is sensitive both to the number of types in a profile and to their frequency distribution.

The second angle for examining the internal lexical make-up of profiles links up with the influence of English in the terminology of information technology. To measure that influence, it is necessary to quantify the proportion of English terms in the profiles in question. Applying the template provided by the U and I measures, measures of lexical proportion can be specified as in formulae (7.18), (7.19), and (7.20). In (7.18), a proportion measure A is defined for a given concept Z in a given dataset Y, with x_1 to x_n referring to the alternative terms for Z, and with K representing the set of terms that share a given feature (like being an English loanword). Membership in that set may be graded, however, such as when a compound noun in Dutch may combine an English loan with an original Dutch word. Accordingly, $W_{x_i}(K)$ is a weighting factor ranging from 0 to 1 and specifying the degree to which feature K is present for any x_i. $A_{K,Z}(Y)$ is calculated by summing over the relative frequencies $F_{Z,Y}(x_i)$ of each of the alternative terms multiplied by their respective weighting factor. The next two formulae are then straightforward: $A_K(Y)$ in (7.19) represents the mean proportion of words with a specific feature K in the dataset, and $A'_K(Y)$ in (7.20) is the weighted alternative that takes into account the relative frequency of the concept.

(7.18) *Proportion for a single concept of terms with feature K*

$$A_{K,Z}(Y) = \sum_{i=1}^{n} F_{Z,Y}(x_i) \cdot W_{x_i}(K)$$

(7.19) *Average proportion for a set of concepts of terms with feature K*

$$A_K(Y) = \tfrac{1}{n} \sum_{i=1}^{n} A_{K,Z_i}(Y)$$

(7.20) *Weighted average proportion for a set of concepts of terms with feature K*

$$A'_K(Y) = \sum_{i=1}^{n} A_{K,Z_i}(Y) \cdot G_{Z_i}(Y)$$

In Geeraerts, Grondelaers, and Speelman (1999), lexical proportions are explored for various features of the lexemes: their French or English origin, their typicality for either Netherlandic or Belgian Dutch, and whether they are propagated (or advised against) in the reference works that bolstered the normative linguistic policy in Flanders. Tracking the evolution of the proportions over the three time periods under consideration sheds an interesting light on the convergence that may be observed between the two language varieties. The convergence is clearly due to changes on the Belgian Dutch side. For one thing, the proportion of terms that are initially typical for Belgian Dutch gradually diminishes in the Belgian Dutch data, but no similar effect is noted on the Netherlandic Dutch side. For another, the proportion of terms that are disapproved in the Belgian Dutch normative literature actually diminishes in that variety. So overall, we may conclude that at least in this period, the official Belgian Dutch policy of taking over the existing Netherlandic Dutch linguistic standards is reflected in the evolution of linguistic usage. English origin is a variable that also contributes to the convergence, in the sense that the impact of English increases in both varieties. French on the other hand works differently in the two varieties (although this only shows up in the clothing terms data, because borrowing from French plays no noteworthy role in the lexical field of football terms). Reflecting the fact that Dutch in Belgium had to establish its position as a standard language in competition with the traditional dominance of French, Belgian Dutch tends to shy away from French loans. In Netherlandic Dutch by contrast, no such reticence is observed.

Coming back to the point made earlier in this section, it will be noted that the factors contributing to lexical success are not just ones that relate to the lexemes themselves, like their origin or their status in reference works, but also include ones that relate to the other corners of the lexeme-lection-lect triangle, viz. the lexical field and the language variety to which they belong. This multidimensionality is further illustrated in Zenner's work on borrowing from English (2013; see also Zenner, Speelman, and Geeraerts 2012). As in the 1999 data, the distinction between the two main varieties of Dutch does not have an impact on the success of borrowed English person reference nouns. However, they are more successful, first, if the anglicism is the shortest lexicalization of the concept; second, if it is used to express a low-frequency concept; third, if the loanword is introduced in Dutch as a necessary loan for which a Dutch alternative was only coined later, which is especially true for younger loanwords; and fourth, if the loanword lexicalizes a concept from a lexical field closely related to or originating in Anglo-American culture. The first of these, economy of expression, involves the lexeme, while the other three are of a conceptual nature.

The focus of the following chapters does not lie on the non-lectometric onomasiological perspective embodied by the measures of internal uniformity and lexical proportion. Internal uniformity will be used as a supplement to the U-measures, as was done in the case study discussed in the previous section, but lexical proportion values will not receive separate attention.

The bottom line

- The 'onomasiological profile', defined as the set of synonymous expressions for a concept differentiated by their relative frequency, is the basis for quantifying onomasiological variation and lectometrically measuring the relationship between lects.
- When onomasiological variation is studied as a topic of investigation in its own right, the focus may fall either on lexical diversity or on lexical success. Lexical diversity involves the degree of variation in a profile, in terms of the number of lexical types, or in terms of their frequency distribution. Lexical success involves the strength or weakness in the profiles of specific lexemes or groups of lexemes with shared characteristics.
- When onomasiological variation is taken as input for an aggregate-level study of language varieties, onomasiological profiles provide a basis for quantifying the closeness or distance of language usage in different lects (including chronological periods).
- Such a lectometric approach sheds light on descriptive issues in sociolinguistic and diachronic linguistic research, like the relations that hold within pluricentric languages, or the convergence/divergence of varieties.

8

Lectometry step by step

The approach to the empirical study of lexical variation demonstrated in this book rests crucially on the concept of an *onomasiological profile* (see Chapters 1 and 7). For an onomasiological profile to be a valid operationalization of how a concept is linguistically expressed, the frequency counts used in the lectometric calculations should represent the target concept and only that concept. This is the lexical instantiation of the general notion of ‘envelope of variation’ as used in sociolinguistics (Labov 1972). A concept’s occurrence is assessed indirectly by means of the lexemes that are considered to lexicalize the concept, and it is the frequency of those lexemes that is recorded in the onomasiological profile. But clearly, as many lexemes are polysemous, we need to identify and retain only those occurrences of a lexeme that correspond to the target concept and discard the ones that we will call the ‘out-of-concept’ occurrences.

When looking at onomasiological variation in a single or a few concepts it is still feasible to take a substantial sample of tokens and manually check whether each token of each lexeme instantiates the intended sense. However, when the analysis is scaled up and the calculations are aggregated over potentially tens or hundreds of concepts, a manual disambiguation procedure becomes increasingly time-consuming. In addition, potential issues do not only emerge at the level of tokens, where the researcher has to decide whether or not a given token represents the target sense, but also in the preceding stage, when the researcher needs to determine at the level of types which lexemes populate the profile, that is, for which near-synonyms the envelope of variation will be identified.

This chapter introduces a workflow for the identification of onomasiological profiles that tackles the issue of semantic disambiguation and at the same time meets the requirements of scale, systematicity, and representativeness. The workflow capitalizes on what we saw in Chapter 6: when we analyse near-synonyms in the same vector space, we can identify the areas where their sense clusters overlap, that is, where they are interchangeable. More precisely, the workflow combines type-based and token-based vector space models. First, type-based models will be used for selecting concepts to be studied and for identifying the lexemes expressing those concepts. Token-based models will, as a next step, be employed for identifying and selecting observations in the corpus that instantiate these concepts. The successive sections of the chapter review the steps involved in the identification of onomasiological profiles and discuss the specific reasons and practical concerns that may decide between their alternative implementations: the selection of

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0008

near-synonyms (Section 8.1), the demarcation of the model space (Section 8.2), the fine-tuning of the onomasiological profiles (Section 8.3), the selection of the pruned models (Section 8.4), and the final lectometric measures (Section 8.5).

8.1 Selection of near-synonyms

In this section we review the pros and cons of two types of procedures for selecting near-synonyms: a bottom-up, semi-automatic, corpus-based one, and a top-down, manual, resource-based one. The latter, as illustrated among others by Geeraerts, Grondelaers, and Speelman (1999), Soares da Silva (2010, 2014), and Ruette, Speelman, and Geeraerts (2011, 2014), relies on existing information as may be found in dictionaries, thesauri, or lexical databases. Compared to a bottom-up approach, it makes use of a pre-established selection of lexemes as found in independently executed descriptions of the vocabulary, while the former performs a selection on the basis of the corpus evidence alone. Such a resource-based choice of synonyms may be advantageous from two points of view. First, starting from existing descriptions is likely to enhance the precision of the results. If the lexicographical resource is authoritative, the synonym sets may be considered semantically reliable: during the compilation of the resource, at least one expert has recognized the items as semantically equivalent. Second, relying on a pre-established collection is useful when a specific research question is at stake: the items may then be selected in function of that question. For instance, if one is interested in the success of English loanwords in a given language, it may be appropriate to pick out concepts from a limited number of lexical fields and examine if the degree of English influence is the same in these fields. Similarly, the impact of English may be affected by features like the recency of the concepts or the formal features of the existing native competitors (see Zenner, Speelman, and Geeraerts 2012 for an analysis along these lines). In such cases, building a balanced dataset of synonym sets may be the preferred option compared to a bottom-up corpus approach.

But both advantages of pre-established synonym sets come with their own disadvantages. First, there are limits to the reliability of lexical resources (as any lexicographer will confirm). Recent developments in the vocabulary, for instance, may not yet be adequately covered by existing descriptions—not just because it takes time to compile them, but also because changes in the vocabulary mostly need to have reached a certain level of conventionalization before they are included in dictionaries and similar resources. For linguists interested in the dynamics of the lexicon, the words that do not make it to the reference works are as interesting as the ones that do, and they will be missed by an approach that relies only on the latter. And second, the price to pay for a balanced design with respect to concept and lexeme characteristics is the size of the dataset: if you are interested in a large-scale coverage of the vocabulary, an automated or semi-automated approach will more easily allow you to scale up the investigation.

In addition, there is a third, methodological disadvantage to working with pre-established synonym sets. With hand-chosen synonym sets, the distributional semantic properties of the items are not known in advance. This is particularly relevant when entering the phase of modelling the tokens of the lexemes. If near-synonyms are primarily chosen because they seem useful for answering specific questions, but they vary widely with respect to the type and quality of vectors that are associated with them, the calculated lectometric results might be confusing at best and compromised at worst. If the overall frequency of a set of near-synonyms, or the individual frequency of one the alternating lexemes is not enough to build a sufficiently dense vector, maintaining the set or the item may skew the calculations, but discarding them may unbalance the design. Initially casting the net wider by means of a bottom-up corpus approach helps to avoid such problems.

So, how can such a bottom-up perspective be implemented? The corpus-based retrieval of near-synonymous lexemes relies on leveraging the semantic (vector) similarity of those lexemes. Generally, the input for this retrieval consists of the type-based vectors of the most frequent lexemes in a corpus. These vectors are assumed to be very reliable and high quality because the high frequency of the lexemes ensures that a large amount of contextual information is taken into account. In Ruette (2012), Ruette, Geeraerts, Peirsman, and Speelman (2014), and Ruette, Ehret, and Szmrecsanyi (2016) a dedicated algorithm for the retrieval of concepts and near-synonyms for lectometric analysis is developed, relying on the Clustering by Committee approach introduced by Pantel (2003). Pantel's model is a general-purpose clustering algorithm that measures cosine distance between type-based vector representations to cluster them into so-called committees. The method has been applied to document clustering (Pantel and Lin 2002b), word sense disambiguation (Pantel and Lin 2002a) and concept discovery (Lin and Pantel 2002). Although some of these tasks resemble our goal, the committees that are the output of Pantel's algorithm are semantically much broader than the sets of lexemes that we are looking for. Because Pantel's committees consist of all word types that are topically associated, they represent small lexical fields rather than sets of near-synonyms. In Ruette (2012), this problem was dealt with by extracting one specific part of the algorithm (Phase II) and adapting it slightly so as to produce much smaller sets of lexemes. In Ruette, Geeraerts, Peirsman, and Speelman (2014) vectors were constructed for the 10 000 most frequent nouns in a corpus of Belgian Dutch and Netherlandic Dutch. This amended version of the method yielded 2019 committees of usually two or three words. A manual clean-up phase was carried out to weed out committees that did not constitute an acceptable committee, either because its constitutive lexemes belonged to the same lexical field without being near-synonymous, or because they were known to be highly polysemous. Thus, only 224 committees, or about 11%, were retained for the lectometric analyses. Ruette (2012: 117) notices that although the sample of linguistic variables does not cover the complete vocabulary, the algorithm certainly increases the generalization power, at the cost of precision with the profiles.

In De Pascale (2019) further refinements were made to Ruette's final version. The most important change was made in the very first step. In this step, the algorithm loops over all the lexemes in the list and the space defined by their nearest neighbours. However, instead of selecting one single subcluster of near-synonyms with the highest similarity score, all subclusters that reached a certain similarity threshold (called the 'cutting height') were retained. It is this modified version of Pantel's algorithm that has been used for the retrieval of the large sets of concepts for the case studies in Chapters 9 and 10.

Like with the resource-based method, there are pros and cons to a semi-automatic corpus-based method for the selection of near-synonyms. To begin with, it solves some of the issues that come with a manual, resource-based selection. First, the different steps of the abovementioned algorithm rely on a single building block, namely the cosine similarity between type-based vectors. It is therefore completely unsupervised, refraining as much as possible from the use of manually compiled existing resources for training. Second, by relying almost solely on information intrinsic to the corpus, one can be confident that the discovered near-synonymy patterns are those that occur in actual usage and in the dataset. In other words, the operationalization of near-synonymy through cosine similarity meets the need for a usage-based conception of lexical-semantic relations. Third, with corpus-derived measures for the identification of near-synonyms one avoids the biases that might creep in during a manual selection, such as researcher's bias, but also the peculiarities of lexical resources, for instance, how up to date those resources are.

In addition, the semi-automatic method has the added benefit that distributional semantic properties extracted from a corpus offer rich and multidimensional information, that can be decomposed in many different quantitative indices which can all be individually manipulated and highlighted—this in a much more sophisticated and flexible way than a manual judgement and choice of a concept's relevance for a study can do. For example, the internal semantic tightness of a set of lexemes can be measured by calculating the distance of each lexeme to the centroid representation of that concept (as in Ruette, Geeraerts, Peirsman, and Speelman 2014), or one could compute the similarity between entire concepts to assess the structure of a whole lexical field. Lastly and most obviously, a semi-automatic technique is faster than a manual selection.

However, semi-automatic, vector-based algorithms also entail some risks. Whereas the increased recall (i.e. a diverse and perhaps complete set of concepts) is an evident benefit, this is not true regarding precision, which will most likely be higher when a concept has undergone the linguist's scrutiny. And even though the reliance on contextual properties of words might be an ecologically very sound way of modelling semantic relations, taking semantic similarity, expressed by the cosine metric, as the pivotal criterion of our algorithm, turns out to be a risky double-edged sword. In fact, in a semantic space of the whole vocabulary of a language one could say that *all words* are semantically similar to each other *to some extent*, since a cosine similarity value is a continuous measure that can in principle be computed between all words, and determining similarity cut-offs is an insidious task. Certainly, what can be assessed is whether words are *more* or *less* similar to each other, but still, there is no easy way to decide at which boundary words are (no longer) similar *enough* to each other. In sum, it is not straightforward to reconcile an operationalization of lexical relations functioning in a boundless, multidimensional continuum (i.e. a type-based semantic space) with the need to extract practicable, bounded units, that is, our sets of near-synonyms.

Finally, the output generated by the current algorithm has only been analysed partially (see De Pascale 2019: 183–204) and by consequence more research is necessary to assess whether the algorithm itself is biased towards certain types of concepts. In Chapter 9 we will report that the algorithm seems to identify more good candidate sets among nouns than among verbs, but this does not necessarily prove a design bias of the algorithm. It might reflect several unrelated, but so far understudied properties of vector representations: the surrounding textual context might be a better indicator of the semantics of nouns compared to that of other word classes; near-synonymy might be more easily found in the domain of objects, events, and ideas (typically encoded by nouns) than in that of actions and evaluations (usually encoded by other parts of speech), and so on. A further striking feature of the list generated in Chapter 9 is that almost all concepts are lexicalized by just two near-synonyms, with only three nominal concepts having three near-synonyms (JOB, MAGAZINE, and EXPERT). The limited number of near-synonyms per concept is not so much a reflection of the specific semantics of the concept (thesauri often report many more synonyms), but rather a consequence of the vector- and corpus-based retrieval procedure. The algorithm is particularly tailored to prioritize a high semantic similarity of the word forms in the set rather than the exhaustivity of a set. In general, the overall vector similarity of a larger set of items tends to be lower than the vector similarity of a smaller set of items.

We may conclude that both a bottom-up, semi-automatic, corpus-based procedure for selecting candidate synonyms, and a top-down, manual, resource-based one have advantages and disadvantages. In line with our overall approach, Chapters 9 and 10 rest firmly on a bottom-up procedure. But generally speaking, both approaches are clearly not mutually exclusive, and Chapter 10 will in fact illustrate how resource-based information can be used to fine-tune the results of a bottom-up procedure.

8.2 Demarcation of the model space

Once a list of near-synonyms has been compiled, at least one token-based vector space model needs to be created for each set of near-synonyms; such a model includes the token vectors for the lexemes belonging to a set. The fundamental conclusion of the previous chapters, and a leitmotif throughout this book, is that a single model per set of lexemes constitutes a limited, and perhaps biased view on the semantic structure of the modelled object. We formulated two strategies to deal with this insight. In Chapter 6, we selected a single model on the basis of available external information. In Chapters 9 and 10, we opt for the creation and aggregated analysis of a large range of models, all of them varying along the set of model hyperparameters that have been discussed and evaluated at length in the previous chapters. Choices that are specific to the case studies described in Chapters 9 and 10, and that derive from restrictions posed by the respective corpus materials or by their theoretical focus, will be explained in the respective chapters. Here, we discuss a general choice that may be made in onomasiological modelling with regard to the selection of first-order context words.

Two types of models may in fact be distinguished, namely 'union-based' models and 'intersection-based' ones. The intersection-based approach boils down to selecting only the first-order context words that are shared between the lexemes in the profile, that is, that occur in the intersection of the context words for all the lexemes. For instance, consider a corpus in which, across occurrences, we find the words *watch, show, movie*, and *national* in the context window of each of the three lexemes for the concept TELEVISION [*television, tv, tube*], but in which the words *network, broadcaster, cable*, and *toothpaste* occur only with one or two of them. In an intersection-based approach, we only retain the first set of context words for the construction of the token vector. In contrast, in the union-based approach all context words that occur for at least one target lexeme in the profile are considered for the computation of the token vector, that is, *watch, show, movie*, and *national* plus *network*, *broadcaster, cable*, and *toothpaste*.

At first sight, the intersection-approach is the technical implementation that seemingly stays most true to Labov's original interpretation of the 'envelope of variation' for a lexical-semantic application. If one is interested in the sociolinguistic distribution of alternating near-synonyms, then it is necessary to study the alternation in interchangeable contexts, that is, contexts that are semantically equivalent. A very strict way of defining interchangeability is to restrict the envelope of variation to context words that occur with each of the lexemes under scrutiny. There are five comments to be made about this choice. First, the fact that identical context words are the best proxy for the semantic equivalence of a specific set of target lexemes does not mean that observing identical context words between any pair of target lexemes automatically and generally entails that the pair is near-synonymous. The usefulness of the proxy appears primarily *after* one has already identified the alternating lexical variants. Second, the more lexemes take part in the onomasiological profile, the more complex the identification of identical context words becomes: should only full overlap count, or is partial overlap, only between a subset of the lexical items, already good enough? Should the overlap depend on the frequency of the lexemes? In fact, the extent to which the shared context words will suffice to model the tokens depends largely on the frequency of the lexemes: in onomasiological profiles with a very skewed lexeme distribution, it would be impractical if the candidate context words were only those in the intersection of the occurrences of those lexemes. Third, compared to a union-based approach, the number of context words will be reduced by definition, which might give rise to sparse token vector representation, that is, vectors based on just a few (shared) context words, instead of all context words in the window of that token. This might have consequences for the quality of the final token vector. Fourth, although the intersection-based approach might, on theoretical grounds, be the most appropriate procedure for either single-level or aggregate-level sociolinguistic studies which make use of Labovian variables, other types of questions could more easily benefit from the union-based approach. For instance, if one is interested in the semantic structure of a set of near-synonyms (as in Chapter 6), or if one wants to investigate how changes in one lexeme affect other lexemes in the set, a more exhaustive set of context words is needed, and the restriction to the shared context words would probably hide changes in the non-shared uses of a lexeme. Fifth, and crucially, theoretical arguments can be formulated against a very strict demarcation of the envelope of variation. In the example, *network* and *broadcaster* are near-synonyms, so from a semantic point of view, in terms of the concept they express, they *do* represent a shared context (and we assume that representing the context words *network* and *broadcaster* by their type vector will bring out their semantic similarity). The disadvantage of restricting the context words to those that are shared between the lexemes is that context words that are semantically similar but not attested with each lexeme in the profile will be overlooked. While an intersection-based approach might ultimately make the use of context vectors superfluous, the union-based approach leverages the similarity between context words. It increases the number of context words that will contribute to the creation of the token vector, potentially reaching a more enriched representation.

It follows that the choice between an intersection-based approach and a union-based approach is very much about how strictly one wants to define the envelope of variation. Choosing the former favours a severely restrictive variationist-sociolinguistic interpretation, where interchangeability of variants is driven by local-level lexical choices; choosing the latter highlights a more cognitive-conceptual take on near-synonym alternation, in which a higher level and more abstract conceptual structure influences the choice of either lexeme. The intersection-based and union-based approaches can thus be seen as choice points on a spectrum of context-selection strategies, in which a good balance needs to be found between being too restrictive with an intersection-based selection, because non-shared context words are weeded out, and not restrictive enough with a union-based selection, because one could potentially include context words that are neither lexically nor semantically shared between the near-synonyms. Rather than choosing a single position on this spectrum, we will deal flexibly with it. While in Chapter 10, only union-based models are considered, in Chapter 9 we encompass the spectrum as such, that is, both options will be included in the set of models to be discussed.

8.3 Fine-tuning profiles

After union-based and intersection-based models with varying hyperparameters have been constructed, the next phase in the workflow is to fine-tune these models further, that is, to (semi-)automatically select the tokens that represent the intended sense of the concept from the modelled data. As stated in the introduction, an important obstacle to this selection is the fact that the near-synonyms in the alternation are themselves polysemous.

We illustrate this issue with the concept SECONDARY with near-synonyms *secundair* and *middelbaar*. A peculiarity of this concept is that the lexical alternation is only found in Belgian Dutch, and in particular when denoting secondary education and schools, where the near-synonyms function as (substantivized) adjectives. This is, for instance, the case in *middelbaar/secundair onderwijs* 'secondary education', *leerlingen van het middelbaar/secundair* 'pupils from the secondary [school]'. In Netherlandic Dutch, only *middelbaar* is used for this type of education, although it is not excluded that *secundair* is attested in tokens which refer to the Flemish school system. Moreover, both words have other senses as well: *middelbaar* can more generally mean 'middle' or 'average, intermediate' as in *een man van middelbare leeftijd* 'a middle-aged man' and *secundair* is mostly used in the sense of 'minor' as in *van secundair belang* 'of minor importance'. In sum, the concept of SECONDARY features many semantic and lectal properties that are useful to explain the upcoming steps in fine-tuning the profiles for lectometric calculations.

Figure 8.1 shows two versions of the same token-based model for this concept. The model uses the union-based procedure and considers five context words to each side of the target with a log-likelihood ratio larger than 1 with the target. As second-order context features the union of the first-order context words is used and these dimensions are reduced to 200 with singular value decomposition. This model was not chosen randomly or because it is particularly representative of the full model space that was created for this concept, but it was selected because it satisfies a number of quantitative criteria, so as to maximize its utility and clarity as an example for the discussion. In particular, it has less than ten clusters, more

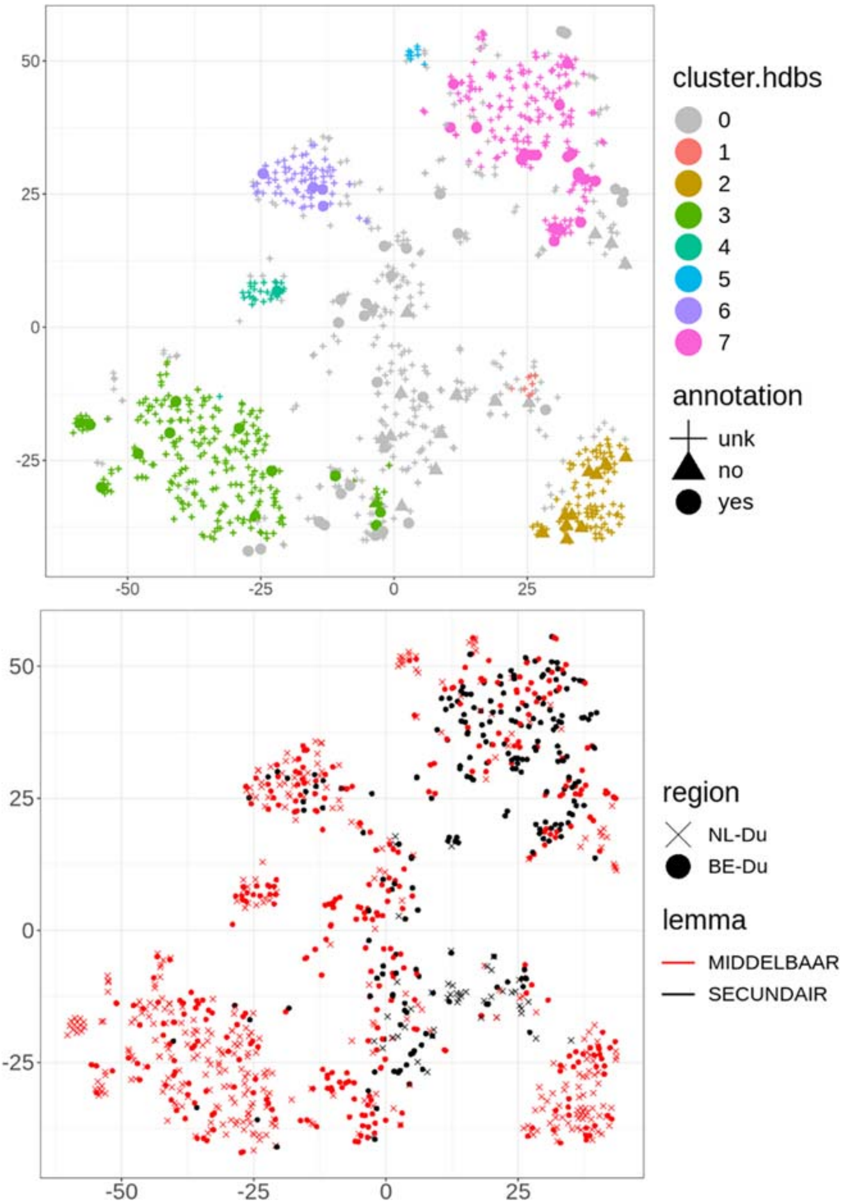


Figure 8.1 Visualization of a model for SECONDARY. Top panel: colour-coded by cluster (from HDBSCAN) and shape-coded by annotation. Bottom panel: colour-coded by variant and shape-coded by region

than five annotated (i.e. manually annotated) tokens per cluster, a simultaneous presence of several in-concept clusters and at least one out-of-concept cluster. On the one hand, the top panel of Figure 8.1 plots the token space by highlighting, first, its cluster structure with eight different colours, and, second, whether the token has been annotated or not and which annotation it receives, by means of shape coding. Annotated tokens are larger in size (• for in-concept tokens and ▲ for out-of-concept tokens), unannotated tokens are shown with plus signs (+). Grey tokens are tokens that ended up being noise (see Section 3.4). On the other hand, the bottom panel of Figure 8.1 plots the token space by highlighting the near-synonyms *middelbaar* and *secundair*, with respectively red and black colour, and the regiolect in which the tokens were found, that is, Belgian Dutch as signalled by dots (•) and Netherlandic Dutch by crosses (×).

The workflow for this fine-tuning phase consists of several steps. Each step removes a set of tokens from the data, relying on some specified criterion from the modelling procedure (step 1) or from the cluster analysis with HDBSCAN (steps 2–4). The aim of the procedure is to solely keep tokens that are in-concept, that is, that represent the intended sense of the concept. All successive steps are executed at the level of an individual model.

STEP 1. REMOVING UNMODELLED TOKENS AND APPLYING A CLUSTERING ALGORITHM

Tokens that have not been modelled by a given model are removed from the data. It is indeed possible that the initial number of tokens sampled from a corpus prior to the creation of a token space is larger than the final set of actually modelled tokens. This (slight) reduction is caused by the lack of context words that are necessary to create a token vector representation. For instance, when only context words with ppmi score higher than 0 are selected, it can happen that no context word is retained for a token. After removing unmodelled tokens, we apply a clustering technique on the token space to automatically get an overview of which tokens are the most similar to each other. When looking at the top panel of Figure 8.1, we can see that the clustering algorithm has identified seven clusters for our model of SECONDARY. Crucially, the clustering itself has an important role in scaling up the removal of out-of-concept tokens, which is the goal of the last step. As in most previous chapters, except for Chapter 6, we use HDBSCAN with the standard settings as implemented in the function hdbscan() from the R package dbscan (Hahsler and Piekenbrock 2021).

STEP 2. REMOVING NOISE TOKENS

In the next step of the workflow, the tokens considered noise by the HDBSCAN algorithm are removed from the data. As explained in Chapter 3, an advantage of using HDBSCAN is that tokens with vectors not similar enough to other tokens are considered noise. These noise tokens are always included in a cluster labelled '0' (grey in the top panel of Figure 8.1). We remove these tokens from the data because in a multiple-variant concept, noise tokens more likely represent senses or usage contexts where not all variants are frequently used: if multiple variants, or even a single variant, would have often occurred in these contexts, there probably would have been enough tokens with sufficient coherence to form a cluster of their own. Of course, the tokens that are considered noise by the algorithm also depend on the settings chosen for HDBSCAN (see Chapter 3 for a discussion of the relevant parameters). As explained in Section 3.4 a crucial hyperparameter of this clustering technique is setting a threshold for the minimum size of a cluster in terms of the number of tokens (i.e. minPts in the R implementation). More specifically, if a lower value for the minPts parameter is chosen, fewer tokens are likely to be categorized as noise, because the conditions for being recognized as a cluster become less stringent. In addition, choosing an appropriate value depends on the initial size of the token space. In Chapter 9 we set the minPts parameter to 8 while in Chapter 10 we choose the value 15 for minPts, as the input token spaces are much larger.

STEP 3. REMOVING TOKENS IN MONOLECTAL AND/OR MONOLEXICAL CLUSTERS

After removing the noise tokens from the data, we check if any of the clusters are monolexical and/or monolectal. Monolectal clusters are defined as clusters in which all tokens come from a single lect; here: from a single regiolect. (It may be noted that this definition of monolectality is very strict. Less strict implementations could be considered, with a more flexible bound to the proportion of tokens that come from a single regiolect, but we leave that as an option for future research.) The rationale for removing tokens occurring in monolectal clusters is that they likely represent usage contexts which only occur in just one of the regiolects under analysis. When in a lectometric study one is interested in measuring lexical variation while keeping conceptual content constant, such regiolect-specific contexts are by definition not available to the speakers of the other lect and cannot function as shared conceptual content. An example of such a monolectal set of tokens may be found in the *briljanten* cluster in Figure 6.2, which was attested in the Belgian Dutch but not the Netherlandic Dutch data under consideration.

A further option is to also remove tokens in monolexical clusters. Monolexical clusters are defined as clusters in which all the tokens are observations for only one of the lexical variants under scrutiny. The rationale for removing these tokens is that these contexts represent specific senses of the variant that do not belong to the target concept. However, removing monolexical clusters carries more risks than removing monolectal ones. In our example concept SECONDARY, for instance, clusters 2, 4, and 5 are monolexical (respectively the large ochre cluster, and the small emerald green and blue clusters in Figure 8.1). The three clusters all contain only *middelbaar* tokens, which is overall also the most frequent near-synonym (74% of the whole token space and 26% *secundair*). However, they represent distinct collocational patterns of *middelbaar*, some of which are in-concept, whereas others are out-of-concept. Cluster 2 brings together tokens where *middelbaar* is used in the out-of-concept sense 'of middle-age'. The other two clusters index an in-concept collocation of *middelbaar*, that is, *middelbare scholieren* 'pupils from secondary schools' in cluster 4 and *middelbaar beroepsonderwijs* 'secondary vocational training'. These two last usages pertain to the range of application of both *middelbaar* and *secundair*, even though the second variant does not seem to appear in that pattern in our sampled dataset. Thus, the example shows that an indiscriminate removal of monolexical clusters can lead to the removal of tokens that should actually participate in the lexical alternation under study. In unbalanced concepts, where one variant has a much higher frequency than the other(s), monolexical clusters may also occur simply because the alternative lexeme was not observed in the data in this context. Since we often have concepts like this in the data in the following chapters, the removal of monolexical clusters is not further explored. (Note that this problem can also be an issue for monolectal clusters if the data is biased towards one of the regiolects under analysis. This is not the case in the datasets analysed in this book and for this reason we do remove the monolectal clusters in the following chapters.) If monolexicality is indeed not a watertight cue for an out-of-concept cluster, we need another way of determining whether a cluster should be discarded or kept, which is explained in the last step of the procedure.

STEP 4. ANNOTATING AND CLUSTER PRUNING

Typically, with many model parameters that can be combined, tens of models are generated for each concept. The steps outlined above are carried out in a fully automatic fashion, as some tokens or clusters are removed because they are categorized as noise by HDBSCAN or because they are clustered into a monolectal and/or monolexical cluster. However, in line with what we saw in Chapter 5, the models are not perfect in recognizing and separating senses: tokens instantiating the same sense can be separated over multiple clusters and, what is worse, tokens instantiating the target sense and tokens instantiating another, out-of-concept sense may be lumped in the same cluster. This last case is particularly harmful, because we want to construct onomasiological profiles that are not polluted by tokens from another sense of the near-synonyms. The consequence of this impurity is that a human annotator may be enlisted to check the quality of the models, and as there are no standard annotation datasets available that cover the many concepts included in a lectometric study, a manual disambiguation may have to be conducted for each individual study. Crucially, opting for manual annotation is justified if the inaccuracies in the clustering are more than negligible, and if the effort to perform the annotation does not undermine the goal of scaling up the research.

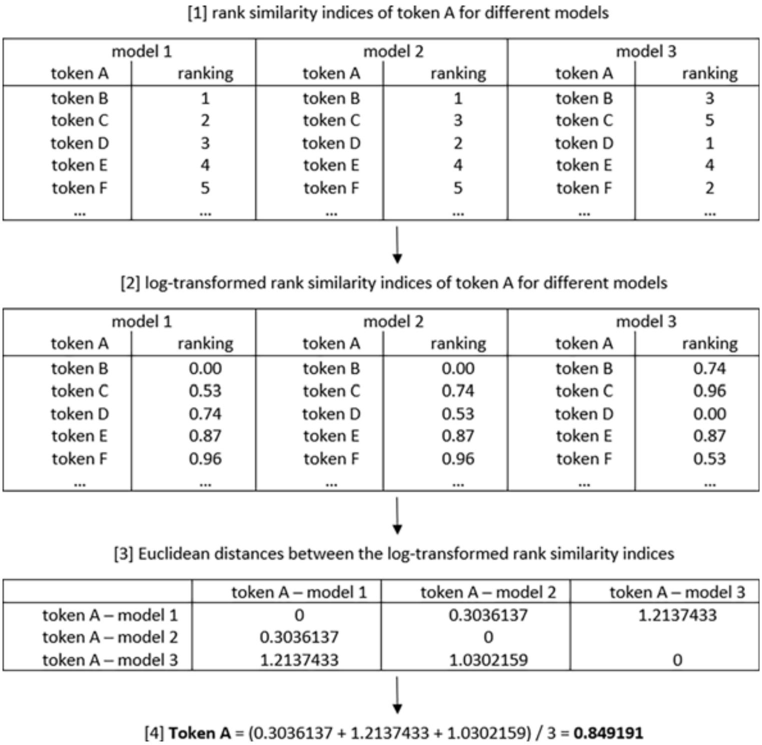
With regard to the first point, we may note that the model in Figure 8.1 succeeds quite well in separating in-concept tokens from out-of-concept tokens. That does not mean that each cluster is equally good, though. The large, green cluster 3 happens to contain 13 annotated tokens, of which 12 (92%) are in-concept and one

(8%) out-of-concept. The other clusters 2, 4, 6, and 7 fare even better, as these are 100% pure clusters (2 is fully out-of-concept, while 4, 6, and 7 fully in-concept). All in all, this token-based model achieves a high quality with respect to the separation of in-concept versus out-of-concept tokens, but many other models unfortunately show a much more noisy picture. Accordingly, both Chapter 9 and Chapter 10 include manual annotation. Further research aimed at optimizing parameter settings may eventually diminish the relevance of a manual intervention, but we do not seem to have reached that point yet.

With regard to the second point, we keep the manual effort low by sampling a subset of tokens, annotating them, and, on the basis of these annotations, deciding whether or not the full cluster likely contains mostly in-concept or out-of-concept tokens. The working assumption is that tokens within the same clusters signal a similar sense or contextual use, and inspecting a handful of tokens should suffice to identify the sense of the entire cluster. Clearly, because the procedure aims at determining the dominant sense of clusters and discarding the out-of-concept clusters, the impurities in the retained clusters will not be removed. We take this as the price to pay for keeping the annotation effort down.

A manual annotation and pruning procedure of this type involves three decisions: selecting the tokens to be annotated, choosing a specific annotation technique, and determining a threshold for discarding clusters. With regard to the first decision, a first option is to take a random sample of tokens from a specified size or proportion from each of the models, or from each cluster within each model for annotation. The benefit of this procedure is that it is relatively straightforward to implement in the workflow. However, a major disadvantage is that there is no control over which tokens are sampled and, again, that the size of the dataset to be annotated may increase greatly due to the fact that tokens that occur in one model need not be retained in another. For instance, tokens removed as noise in one model need not turn up as noise in another. To keep the number of tokens to annotate at a manageable size, a second strategy may be envisaged in which only tokens in a smaller set of models are analysed, for instance only tokens in the medoids selected by the Partitioning Around Medoids algorithm (see Section 4.2). While this procedure probably greatly reduces the amount of manual annotation necessary and, thus, the time investment needed compared to the first strategy, it has as a disadvantage that a much smaller number of models is considered. For this reason, we suggest a third strategy here which relies on distributional properties of the tokens themselves while at the same time taking into account all the distributional models that were constructed. This strategy will be applied in Chapters 9 and 10. The leading idea is that tokens with a high distributional stability across models are more interesting candidates for disambiguation. A token is considered to be 'distributionally stable' when its neighbouring tokens are highly similar in each model, across models.

We work out an example to show how such an index of distributional stability is calculated in Figure 8.2. For example, we take a token A and calculate, via cosine distance, the list of nearest neighbours and repeat this calculation for every generated model. Then we transform this nearest neighbour distance list into a nearest neighbour rank list, as shown in Figure 8.2, panel [1]. In model 1 token C is the second closest token to token A, but in model 3 that same token C is only the fifth closest token to A. Afterwards we apply a double log-transformation of these rankings (the rationale of it has already been explained in Chapter 3). These transformed vectors are shown in panel [2] of Figure 8.2. In Figure 8.2, panel [3] calculates the pairwise Euclidean distances indices between every such log-transformed list. In other words, we are dealing with a distance between rankings, which originally consisted of similarity values themselves. The result is a distance matrix for an individual token across all the available models: each cell shows the Euclidean distance between a neighbour rank list in one model compared to another. When taking the average of these Euclidean distances, as is done in Figure 8.2, panel [4], we arrive at our index of 'distributional stability'. The lower this number, the more alike the token neighbourhoods are across models; the higher this number, the



[1] rank similarity indices of token A for different models

model 1		model 2		model 3	
token A	ranking	token A	ranking	token A	ranking
token B	1	token B	1	token B	3
token C	2	token C	3	token C	5
token D	3	token D	2	token D	1
token E	4	token E	4	token E	4
token F	5	token F	5	token F	2
...	...	...	...	...	...

[2] log-transformed rank similarity indices of token A for different models

model 1		model 2		model 3	
token A	ranking	token A	ranking	token A	ranking
token B	0.00	token B	0.00	token B	0.74
token C	0.53	token C	0.74	token C	0.96
token D	0.74	token D	0.53	token D	0.00
token E	0.87	token E	0.87	token E	0.87
token F	0.96	token F	0.96	token F	0.53
...	...	...	...	...	...

[3] Euclidean distances between the log-transformed rank similarity indices

	token A – model 1	token A – model 2	token A – model 3
token A – model 1	0	0.3036137	1.2137433
token A – model 2	0.3036137	0	
token A – model 3	1.2137433	1.0302159	0

[4] **Token A** = (0.3036137 + 1.2137433 + 1.0302159) / 3 = **0.849191**

Figure 8.2 Example workflow for calculating distributional token stability

more they differ. Finally, we rank the tokens by their distributional stability index from smallest to largest, and manually annotate the 10% most stable tokens.

The assumption here is that tokens with high distributional (neighbourhood) stability likely have clear and robust semantic properties that are easily and consistently picked up by the distributional models, and that they are therefore good representatives of a particular sense. A further practical advantage of focussing on these tokens for annotation is that their high stability value indicates that they are kept in many models after the workflow so far (rather than being categorized as noise or in a monolexical/monolectal cluster), greatly reducing the amount of time and effort needed for annotation. In Chapter 9, for instance, the total amount of tokens to annotate using this procedure is 8744, which is less than 10% of the total amount of sampled tokens (883 000).

After tokens have been selected for annotation, a consistent and systematic methodology has to be chosen with regard to the annotation process. For an onomasiological lectometric study, the main goal is to only analyse tokens that display the target sense. Thus, for this kind of studies, it is mostly relevant to determine, for each token, whether it is an instance of this sense. In traditional variationist theory, interchangeability plays a key role here. The annotator can be asked to judge whether or not the meaning of the token under scrutiny remains stable if the target lexeme is replaced with one of each of the other lexical items included in the profile. The advantage of this method is that it is relatively fast. It will be employed in Chapter 9 to annotate large sets of onomasiological profiles in Dutch.

An alternative method is to have annotators conduct a more detailed lexical-semantic analysis of the tokens under scrutiny, indicating which sense of a list of possible senses determined beforehand (for instance, a list based on lexicographic reference works) the target lexeme in the token represents. While this type of more detailed lexical-semantic analysis requires more effort from the annotator, as well as from the researcher, who needs to construct a valid list of possible senses beforehand, it is quite important in studies where the main interest is to obtain a detailed distribution of the semasiological variation in a set of onomasiological profiles. This method was employed for the (semasiological) dataset analysed in Chapter 5 and a similar approach (without annotation) is showcased in Chapter 6. It also lies at the basis of the annotation approach in Chapter 10.

These two methods of annotation do not exhaust the possibilities (see also Schlechtweg, Tahmasebi, Hengchen, Dubossarsky, and McGillivray 2021). A further possible addition to the methodology of annotating tokens is to include a measure of certainty with regard to the annotation of the token. More specifically, semantic indeterminacy of the type explained in Chapter 2 may also play a role when it comes to the annotation of tokens: there may be cases where a human annotator cannot be sure about the intended meaning of the target lexeme in the token because some type of information (contextual, encyclopaedic) is not available. If a measure of certainty is included in token annotation, this information can be used to gauge whether tokens that are more straightforward or clearer are also better modelled by the distributional method described in this book. On the one hand, the annotator can be asked to indicate, on a Likert scale or by similar methods, how certain they are about their annotation. On the other hand, several annotators can be asked to annotate the same set of tokens and their mutual agreement can be calculated as a proxy for certainty. The latter represents the strategy taken in the large-scale annotation task carried out in Chapter 10. Further, additional questions, such as which context words helped the annotator choose a specific sense, can also be asked. However, asking for this type of additional information greatly increases the effort and time investment required from the annotator. If the main aim of the study is merely to employ distributional methods to conduct a lectometric study, it may therefore not be necessary to include this type of information in the design. In contrast, if, like in Chapter 5, the aim is to understand how different distributional models model aspects of semasiological meaning (which may also affect onomasiological pairs, like in Chapter 6), this type of information may be a valuable addition.

After the sampled tokens have been annotated, the third component of the procedure comes into play: using the annotation information to remove clusters from each model that likely do not represent the intended target sense. In the chapters that follow, the cut-off point is set at 80%, that is, at least 80% of the annotated tokens need to represent the intended sense. If fewer than 80% of the annotated tokens in a cluster represent the concept, the cluster is removed from the model. Obviously, this limit of 80% is arbitrary and other boundaries can be set.

8.4 Selection of pruned models

In the wake of Chapter 5, we mentioned two basic strategies for dealing with the indeterminacy of semasiological distributional modelling: either a choice is made for a specific model on the basis of additional or external evidence, or a variety of models is included in the analysis and the stability of descriptive results is investigated across that set of models. The first approach was illustrated in Chapter 6, the second forms the backbone of Chapters 9 and 10. But some possible refinements and alternative implementations with regard to the latter need to be introduced.

For starters, it is important to recognize that the procedure that we have described for identifying and removing out-of-concept clusters may also indirectly eliminate entire models. The result of the workflow described so far is a set of tokens that instantiate the target sense. Which tokens are retained depends on the outcome of the clustering algorithm, which in turn depends on the specific token-based model taken as input. After the fine tuning and pruning procedure, the number of tokens left can vary surprisingly widely: some models will lose just a few tokens, but crucially, other models will disappear in their entirety because every token belongs to either the noise class, a monolectal cluster, or a cluster considered out-of-concept. It is therefore not known in advance how many models are left after the procedure, but with our in-concept threshold of 80%, we can be confident that the remaining models are of sufficient quality. In the broader scheme of things, this is a refinement of the strategy of looking at a multitude of models: only a subset of models that passes a certain quality criterion is taken along to the actual lectometric calculations. The models that are not retained so to speak evaporate: even though with another concept they may be retained by the procedure, they disappear from the modelling of a given concept because they do not succeed in producing at least one 80% in-concept cluster for that concept.

We will apply that criterion in Chapter 9, but in Chapter 10, we will raise the bar by making the criterion more strict. Given that the fine-tuning and pruning procedure may result in the removal of entire models, some concepts will be more richly modelled than others, in the sense that they suffer less from model evaporation. In Chapter 10, this effect is included in the lectometric exploration. The results obtained by considering all concepts that pass the initial threshold will be compared to results obtained from concepts that are modelled by enough models, with the cut-off point set at 50% of all models per concept. In other words, instead of keeping only concepts for which at least one model passes the 80% in-concept cluster threshold, concepts only enter into the lectometric calculations if at least half of all the models for that concept comply with the criterion.

In principle, one could go even further and apply more sophisticated quantitative measures to gauge the models' performance and keep only models that perform well on these measures for the remainder of the analysis. And if we push that method to the extreme and select the very best performing model only, we meet the other basic strategy—focusing on a single model—at the other end of the spectrum. Two different approaches may be considered.

If a large set of tokens has been manually disambiguated by human annotators, it is possible to use standard evaluation measures for sense disambiguation (see Manning, Raghavan, and Schütze 2008 for an overview). Within the field of natural language processing, a common way to evaluate systems for their sense disambiguation performance is by comparing a clustering based on manual annotations to the solution from a cluster algorithm based on distributional models (Manandhar, Klapaftis, Dligach, and Pradhan 2010; Navigli and Vannella 2013). By means of a given cluster evaluation measure one can check how well the induced clusters, in terms of their number, composition, and size, mirror the distribution of the manually labelled tokens. Evaluation measures that have been devised specifically for this task are the *Rand Index*, *Normalized Mutual Information*, the *F-measure*, and the *V-measure* (Rosenberg and Hirschberg 2007; Manning, Raghavan, and Schütze 2008) and they have been extensively used in many word sense disambiguation tasks (see previous references). With these measures, we would be able to decide which model fits the data the best, and only continue working with the best model(s). A disadvantage of this approach is that by incorporating both a clustering technique and an evaluation measure, we introduce two new levels of potential bias and uncertainty related to the specific nature of the techniques and the measures, and researcher's degrees of freedom, which all render the workflow more opaque and prone to distortions that can percolate up to the lectometric calculations.

A partial solution would be to employ evaluation measures that do not need a prior clustering solution, such as the indices introduced in Speelman and Heylen (2017), specifically intended for their application in distributional semantic research, and in particular the exploration of token space models. Useful measures are the 'local' measures: the *k nearest neighbours index* (kNN) (already used in Chapter 5) and the *same class path index* (SCP). Another advantage of this approach is that by skipping the intermediate step of the cluster analysis, one is able to evaluate the quality of the token representations more directly than in a scenario with clustering, in which it would be harder to disentangle the merits of the underlying vector representations from those of the specific cluster algorithm itself. Furthermore, as said, the two indices are 'local' indices, in that they assess whether at a more fine-grained level in the token space, some, many or all smaller areas of tokens therein nicely coincide with either the 'in-concept class' or the 'out-of-concept class'. Given the noisy properties of natural language data, which make for notoriously hard test cases for classic statistical modelling techniques, we can expect the token spaces to exhibit properties that do not satisfy many assumptions on which many clustering techniques have been built (such as unequal sizes and density of regions, the absence of regular shapes, etc.). In addition, the local perspective allows us to detect small sense-homogeneous areas even in cases where those regions are interspersed across each other and/or glued together. The focus on the local patterns in the token space circumvents these problematic aspects that would instead greatly penalize cluster methods that optimize for the global structure in the data.

Neither of these alternatives will be pursued in the following pages. In contrast with the straightforward quality criteria that we implement in Chapters 9 and 10, the very sophistication of the alternatives makes them more difficult to interpret, and also, because the indeterminacy of meaning description has surfaced a number of times throughout our research, we hesitate to adopt an approach that boldly looks for a single best model. In Chapter 6, selecting a single model worked well because we could focus on a single concept, and because we only identified a distributional model that corresponded maximally with an existing manual analysis—without claiming that either the analysis or the model is the very best description per se. But what we did in Chapter 6 cannot be extrapolated directly to a larger set of concepts. More generally, we feel that in the current methodological stage of semantic research, care needs to be taken with promoting any source of information, including annotated data, to the status of 'gold standard' (see also Plank 2022). Accordingly, we prefer to stay with an evaluation criterion for models that is relatively rough but easy to interpret, and most importantly, with a method that increases stability by considering a range of solutions.

8.5 Lectometric measures

The workflow described so far selects the models and the in-concept clusters that constitute the basis for the lectometric calculations. The latter then apply the U-measures and I-measures described in Chapter 7, including weighting by concept frequency. Other types of weighting, like a cluster-based weighting procedure as discussed in Chapter 7, are reserved for future research. In Chapters 9 and 10, two further methodological features are added to the lectometric calculations: a method for examining stability across models, and methods for looking for patterns and underlying factors behind the lectal distances as such.

First, since a large number of models will be considered, the question remains how to analyse lectometric results across models. How can we integrate the fundamental instability and variability that comes with different model parameters and modelling strategies into the lectometric framework? We suggest using the variety of token-based models that results from parameter combinations as a way to quantify the uncertainty that might characterize the calculation of uniformity indices. The reasoning goes as follows: a single token-based model of a concept generates one onomasiological token space with particular semantic properties, and on the basis of that space, a uniformity value is calculated for a specific pair of lects represented by the tokens. In other words, in an individual model one lectal comparison will have one uniformity value. This does not mean, to be sure, that one model only yields a single uniformity value. The number of uniformity values that can be possibly calculated given the model's tokens depends on the number of lects in which one can categorize that set of tokens. For instance, if a token space contains formal Belgian Dutch and formal Netherlandic Dutch, from two time periods each, the number of possible pairwise comparisons would be (no less than) six, which in turns yields (at least) six distinct uniformity values.

Further, when we create different models for the same set of tokens, we will have as many uniformity values for a specific lectal comparison as there are models. By doing so we actually simulate the process of calculating uniformity indices on different samples, and this allows us to generate a distribution of uniformity values for the same pair of lects. On such a distribution we can then calculate measures of central tendency and dispersion. The idea is that the mean of a list of uniformity indices, calculated over many different models, will be a better estimate of the true lectometric distance than a lectometric value based on a single model. Or perhaps more appropriately, we are reluctant to assume that there is an ultimate 'true distance'. Rather, we think of different distributional parameter settings and workflow variants as perspectives that bring forward different semantic facets of the concepts at hand, and we are keen to see how lectal distances perform under a range of perspectives. In practice, Chapter 9 visually analyses the distribution of the I- and U-measures per model (and per concept and semantic field). Additionally, it considers the mean of means to obtain an aggregate lectometric result across models and across concepts, that is, the mean of the mean uniformity values per concept across models. Further, this figure is compared between concept-weighted and non-concept-weighted analyses.

Second, we can go beyond merely establishing lectal distances, and look for tools to better interpret the structure and background of those distances. One such tool is visualization. In Section 7.2 and Chapter 9, the distances between the lectal data points are presented in a numerical way only, but Chapter 10 adds a further layer by submitting the lectal distances to a multidimensional scaling analysis that allows for a visual representation of the lectal relations and that thus helps to better understand the pattern behind the distances. Another insightful tool, demonstrated in Chapter 9, is to submit the measures to a regression analysis, for instance by examining how model parameters, semantic fields, and lectal features contribute differently to the observed distances. Again, these tools do not exhaust the possibilities. Ruette and Speelman (2014) use individual differences scaling (INDSCAL) to gather information on how individual profiles contribute to the aggregate-level solution. Speelman (2021) works with Procrustes distances to establish the extent to which distance matrices based on individual profiles differ from each other, and accordingly, to determine subsets of similarly behaving profiles. Here too, a further expansion of the research programme may be envisaged.

The bottom line

- The area where near-synonyms overlap in a semantic vector space constitutes the envelope of variation for studying those lexemes as sociolinguistic variables, more specifically, for identifying onomasiological profiles as input for lectometric calculations.
- Determining such onomasiological profiles consists of a number of steps, involving both type-based and token-based vector representations: selecting an initial set of near-synonyms, setting the parameters that demarcate the range of models to be considered, pruning the models by removing out-of-concept clusters, and potentially also selecting models.
- The pruning process as we implement it is itself constituted of different components: removing unmodelled tokens, removing noise tokens, removing monolexical clusters, and removing out-of-context clusters on the basis of manually annotated samples of tokens.

- Our approach involves specific choices with regard to alternatives that present themselves at the successive stages of procedure, in particular, whether the selection of near-synonyms happens in an automated, corpus-based or in a manual, resource-based manner; whether an intersection-based or union-based selection of context words contributes to the demarcation of the model space, and how token samples are chosen and annotated for the identification and removal of out-of-concept clusters.
- The incorporation of manual annotation into the procedure is motivated by the recognition that token clusters are not always semantically homogeneous. The annotation method implemented here is designed to reduce time investment, so as not to endanger the goal of scaling up the lectometric workflow.

PART V

LECTOMETRIC EXPLORATIONS

Just as Chapters 5 and 6 are the descriptive counterpart to Chapters 3 and 4, Chapters 9 and 10 apply the methods presented in Chapters 7 and 8 to actual examples of linguistic variation: Chapter 9 looks diachronically at the evolution of Dutch, and Chapter 10 presents a synchronic view of international varieties of Spanish.

9

Dimensions of standardization

In Chapter 8 we gave an overview of the successive steps for setting up a lexical lectometric study. In this chapter we will put that workflow into practice, with both a descriptive and a methodological focus. The descriptive goal is to advance our knowledge of the standardization dynamics of the two national varieties of Dutch used in the Low Countries: Belgian Dutch and Netherlandic Dutch. Both have been examined in previous lectometric work, and as such, the present study is situated in the longstanding tradition of lectometric research that was initiated in Geeraerts, Grondelaers, and Speelman (1999). A survey of how the approach was descriptively expanded and methodologically refined was presented in Section 1.5. In Chapter 7 a first study of the lectometrically reinterpreted notions of 'hierarchical destandardization', 'informalization', and 'dehomogenization' was carried out, based on the lexical field of clothing terms. However, the results presented there rest on a limited set of concepts from a single lexical field. Drawing general conclusions from there would be dangerous, but at the same time, expanding the approach faces a methodological hurdle. The manually compiled dataset consisted of thousands of tokens for each region, and any broadening of the scope, be it in the number of concepts, time periods, or regions quickly clashes with the limitations of a fully manual workflow. In this chapter we pick up the three dimensions with which Chapter 7 characterized standardization processes and submit them to a large-scale study—the first on this scale devoted to the destandardization of written Dutch in Flanders and in the Netherlands. With the integration of distributional semantics in lectometric research as introduced in Chapter 8, we are ready to tackle the disambiguation bottleneck that comes with having to identify thousands of semantically equivalent tokens for the onomasiological profiles. The current chapter therefore serves as a replication of the study carried out in Chapter 7, and partially also in Daems, Heylen, and Geeraerts (2015): we intend to compare the trends observed in these studies with the ones generated in our analysis.

The second goal of the chapter is a methodological one. As we have seen in Chapter 8, the use of token-based models for measuring distances raises the issue of how to evaluate the influence of different types of modelling on the lectometric task. In Section 8.4 in particular, we surveyed various ways in which models can be selected for the measurement of lectal distances. In this chapter, then, we illustrate the strategy in which, per concept, all the available models are included in the analysis. Instead of comparing the semantic structure of each single model to

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0009

a reference semantic classification, we compare a variety of models among one another, without necessarily recurring to a gold standard, that is, we focus on the role of all available models in the quantification of uncertainty around lectometric measurements. In Section 9.1 we introduce the various corpora and set of concepts for our study on Dutch standardization. Section 9.2 reports on the specific choices that we have implemented regarding the construction of token spaces and onomasiological profiles. The next three sections (Section 9.3, 9.4, and 9.5) provide the analyses of the different dimensions of destandardization introduced above. Section 9.6 concludes the chapter with an overall evaluation of the integration of distributional semantics in the lectometric workflow. Do we get valuable insights by working with multiple models and distributions of uniformity values, instead of single measurements? Does the integration need small modifications only, if at all, or will future research have to address more fundamental shortcomings?

9.1 Corpora and concepts

For the analysis of the three dimensions of (de)standardization put forward in this project we first need (sub)corpora that represent the lects that are necessary to carry out such analyses. We will work with eight lectal strata, corresponding to eight measuring points: two situational lects or registers in two regions across two time periods. A first hurdle is that, at present, we do not have at our disposal one single corpus that covers the full lectal variation necessary for our study and that was compiled and processed according to a systematic set of instructions and tools. Therefore we brought together several separate corpora, whose composition is shown in Table 9.1. The total corpus size equals 161 million tokens and we have tried, whenever possible, to obtain balanced subcorpora for each of the eight measuring points. It is important to keep the size of the subcorpora as comparable as possible, in order to keep the size of concepts extracted from those corpus sections also comparable across sections. If, for instance, a concept is twice as large in one lect as opposed to the other, this can only be interpreted as a true difference in concept prevalence when the underlying data sample of both lects is similar in size. Ideally, we would have had subcorpora of about 20 million words for each measuring point, but we will see shortly that we could not always meet that requirement.

The materials for representing the formal lects are taken from quality newspaper issues, in both time periods. For Netherlandic Dutch, we choose issues from *NRC Handelsblad*. To cover the first time period, we sampled about 23 million words from issues published during the years 1999 and 2000. For Belgian Dutch we opted for the quality newspaper *De Standaard* and gathered a number of articles that would amount to circa 22 million words, for the same time window as the

Table 9.1 Corpus composition for the Dutch standardization study

REGISTER	PERIOD	BELGIUM	THE NETHERLANDS	TOTAL
formal	1999-2000	*De Standaard*: 22M	*NRC Handelsblad*: 23M	45M
informal	1999-2004	Usenet groups: 19M	Usenet groups: 20M	39M
formal	2017-2018	*De Standaard*: 20M	*NRC Handelsblad*: 20M	40M
informal	2017-2018	tweets: 10M	tweets: 27M	37M
TOTAL		71M	90M	161M

NRC Handelsblad articles. Issues from both newspapers for the first time frame are extracted from the larger *QLVLNewsCorpus* already introduced in Chapter 5. As a more recent counterpart of the older *De Standaard* and *NRC Handelsblad* issues, we sampled the same number of issues from these newspapers during the years 2017–2018 from the text collection available in the *Corpus Hedendaags Nederlands* (Corpus of Contemporary Dutch, CHN 2021), totalling 20 million words each.

As a representation of older informal language, we took the complete collection of Usenet discussion groups, active in Belgium and the Netherlands, already scraped and processed by Tom Ruette (see Ruette 2012 for a description and a first lectometric use of these discussion groups). Due to the overall smaller size of this collection for both countries (19 and 20 million words) we were compelled to consider the full temporal range of the extracted groups with the .be domain, so from 1999 to 2004.

The informal lect in the second time period was covered by tweets sampled for a separate study in Van de Cruys (2021); the dataset compiled for this study is used here by courtesy of the author. Tweets can be automatically geotagged, or the author can add a location tag in his bio, but often this information is missing and the location needs to be inferred by other means, such as machine learning classifiers. This is indeed the approach chosen in Van de Cruys (2021), who uses the Dutch contextualized word embeddings network RobBERT (Delobelle, Winters, and Berend 2020) to provide a regional label for each tweet, either 'Belgian Dutch' or 'Netherlandic Dutch'. The first consequence of this choice is that the regional attribution is only predicted, and to a certain extent uncertain, contrary to the higher certainty we have about the country affiliation of the other materials. The second consequence is, given the nature of these contextualized word embeddings, that we are not sure which features of a tweet help to reach a specific prediction. RobBERT achieves an F1-score of 79% for the regional classification of tweets on a test set, which means that 21% of the tweets might receive the wrong label. This finding should be kept in mind when interpreting the lectometric calculations that involve the informal lects in Belgian Dutch and

Netherlandic Dutch. Furthermore, when a random sample of 1 million tweets is sampled in each year (i.e. in 2017 and 2018), it turns out that two-thirds are written by Dutch Twitter users, and one-third by Belgian Twitter users. This fact is not surprising given the different sizes of the Dutch-speaking population in both countries (i.e. 17 million speakers in the Netherlands and about 6.5 million in Belgium, which nicely corresponds to the same rates found in the Twitter samples).

An important consequence of bringing together heterogeneous data sources is the diversity of pre-processing layers applied on them. The quality newspapers of the first time period have been lemmatized and part-of-speech tagged with different tools than the ones from the second time period, and even between the Twitter data and the recent newspaper issues there are discrepancies. We therefore decided to work with the most basic textual units, that is, word forms. This choice has the advantage of making the subcorpora directly comparable (as no major spelling reforms have been implemented between 1999 and 2018) but unfortunately increases formal redundancy and by consequence also 'vector redundancy', as word forms belonging to the same lemma will all receive different vectors. In the choice of parameter settings for our token-based models, we opt for the use of singular value decomposition on the context vectors to reduce the redundancy in the context vector representation (see Section 9.3). We refer to Section 3.3 of Chapter 3 for a more elaborate discussion on the relevance of this technique.

Once the lectal structure and the associated corpus parts are clarified, the task is to determine which and how many concepts will be sampled for the study. The research in Daems, Heylen, and Geeraerts (2015) and Chapter 7 extends the lectometric framework on a, respectively, descriptive and theoretical level, but both studies carry out their analyses on the original set of concepts of Geeraerts, Grondelaers, and Speelman (1999). The novelty of the present chapter is the expansion to many more and more diverse concepts, partially continuing the methodological innovation that was spearheaded by Ruette (2012). The approach for the selection of concepts in this chapter is corpus-based as introduced in Section 8.1, and makes use of the concept selection algorithm described in that same section and in De Pascale (2019: 160–206).

The concept retrieval algorithm received as input the type vectors of the most frequent 20 215 word forms, which corresponds to word forms with at least 400 occurrences across the whole corpus. As context dimensions we opted for that same list of target types, and frequency co-occurrences were recorded in a window span of four word forms left and right. The hyperparameters of the algorithm were taken from the evaluation study in De Pascale (2019): the number of nearest neighbours is set to 100, the recurring dendrograms are pruned at 10% cutting height, committees that reach a cosine similarity score of 0.8 are merged and residual target types are taken to the next iteration of the algorithm if their similarity to each other retrieved committee is lower than 0.4 cosine similarity. The generated list of committees is still very large, 17 458, but based on that same evaluation we further pruned the list to a smaller collection of 1357 committees with quantitative properties that would likely make them better candidates for inclusion: only committees with an average mutual pmi-association lower than 6, an average mutual cosine distance lower than 0.22 and an average mutual similarity rank lower than 3. If a committee had at least one of these properties, it was kept for a manual inspection. In this last collection we only looked at concepts with word forms that are unlikely to be ambiguous with respect to different parts-of-speech, as a way to avoid having to remove too many tokens of a different part-of-speech in the subsequent stepwise disambiguation procedure.

The final list has 85 concepts and is summarized in Table 9.2. Next to each concept label we show the dictionary form of each near-synonym. Notice that this dictionary form hides the fact that we applied the algorithm to all paradigmatic forms of a variant separately (for instance, all inflections of a verb, plurals and singulars for nouns, etc.). In the unstructured list that is generated by means of the algorithm we were able to group the concepts in different part-of-speech categories: 19 adjectival concepts, 6 adverbial concepts, 10 verbal concepts, and 50 nominal concepts. The classification in parts-of-speech is based on the word forms of the variants that appear in the generated committee, and in doing so it delimits the set of tokens that are going to be considered in-concept in the next steps. There are a couple of comments to be made regarding this part-of-speech distribution. First, the distinction between adjectives and adverbs is notoriously hard in the Dutch language, and one cannot identify the part-of-speech in which a potential adverbial or adjectival word form is used without taking into account the immediate sentential context. In principle, almost all adjectives can be used as adverbials, but not all adverbs can be used as adjectives. Therefore, we classify the former set as 'adjectival concepts' and the latter set, the so-called 'real' immutable adverbs, as 'adverbial concepts'. Second, within the group of nominal concepts we identified smaller clusters indexing different lexical fields: science, sports, economy, politics, abstract concepts, and a rest group of uncategorized concepts. As this grouping in lexical fields is the result of a post-hoc analysis, carried out manually, the groups are unequal in size and certainly subject to revision. Nevertheless it is one possible attempt to carve up this list of nominal concepts in relatively homogenous groups. A more data-driven approach to the detection of semantic fields among concepts was illustrated in De Pascale (2019), in which type-based vectors were first created for concepts themselves, by averaging over the vectors of the near-synonyms, and afterwards submitted to a cluster analysis. The resulting clusters of concepts would eventually be taken as the bottom-up derived semantic fields.

Table 9.2 List of concepts for the Dutch standardization study

	CONCEPT	NEAR-SYNONYMS	FREQUENCY
adjectival concepts (19)	OFTEN	*geregeld, regelmatig*	16 930
	NEXT	*komend, volgend*	73 951
	EXPLICIT	*expliciet, uitdrukkelijk*	5101
	MODERN	*hedendaags, modern*	16 854
	GRADUAL	*geleidelijk, langzaam*	8667
	REMARKABLE	*opmerkelijk, opvallend*	15 125
	SIMILAR	*gelijkaardig, soortgelijk*	4150
	WEIRD	*raar, vreemd*	23 497
	DRASTIC	*drastisch, ingrijpend*	4419
	DISAPPOINTED	*ontgoocheld, teleurgesteld*	3398
	SECONDARY	*middelbaar, secundair*	5152
	SUCCESSIVE	*achtereenvolgend, opeenvolgend*	1693
	ACCEPTABLE	*aanvaardbaar, acceptabel*	2112
	PROFITABLE	*rendabel, winstgevend*	2149
	NERVOUS	*nerveus, zenuwachtig*	1935
	RESOLUTE	*doelbewust, opzettelijk*	2044
	ALARMING	*verontrustend, zorgwekkend*	1732
	BRAVE	*dapper, moedig*	2497
	SLOW	*langzaam, traag*	9597
adverbial concepts (6)	MEANWHILE	*inmiddels, ondertussen*	44 377
	INDEED	*immers, namelijk*	33 995
	ANYWAY	*overigens, trouwens*	45 022
	(THIS) MORNING	*vanmorgen, vanochtend*	9888
	AFTERWARDS	*daarna, vervolgens*	40 612
	SUDDENLY	*ineens, opeens*	14 143
verbal concepts (10)	TO COME UP	*bedenken, verzinnen*	15 044
	TO GUARANTEE	*garanderen, waarborgen*	6426
	TO STIMULATE	*bevorderen, stimuleren*	5970
	TO ACCEPT	*accepteren, aanvaarden*	10 892
	TO PROTEST	*demonstreren, protesteren*	5885
	TO SEND	*versturen, verzenden*	4397
	TO SHOP	*shoppen, winkelen*	11 409
	TO DEBATE	*debatteren, discussiëren*	4122
	TO INTERROGATE	*ondervragen, verhoren*	3302
	TO INSULT	*beledigen, kwetsen*	3352
nominal concepts (50)			
science	LABORATORY	*laboratorium, lab*	3009
	RESEARCHER	*onderzoeker, wetenschapper*	15 866

	CONCEPT	NEAR-SYNONYMS	FREQUENCY
	SYSTEM	*system, stelsel*	25 201
	EXPERT	*expert, specialist, deskundige*	12 917
sports	CYCLIST	*renner, wielrenner*	4636
	FAN	*fan, supporter*	13 903
	WIN	*overwinning, zege*	11 650
	COACH	*coach, trainer*	14 597
	PENALTY	*penalty, strafschop*	3082
economy	TREASURY	*schatkist, staatskas*	1237
	CENT	*cent, eurocent*	5356
	TAX	*taks, heffing*	2526
	BRANCH	*filiaal, vestiging*	4407
	IMPORT	*import, invoer*	4315
	SCARCITY	*krapte, schaarste*	1337
	HOUSING MARKET	*huizenmarkt, woningmarkt*	1197
	LICENSE	*licentie, vergunning*	7527
	JOB	*job, baan, arbeidsplaats*	32 667
politics	DEMONSTRATION	*betoging, demonstratie*	5178
	COUP	*coup, staatsgreep*	2580
	REPRESSION	*onderdrukking, repressive*	2605
	ARREST	*arrestatie, aanhouding*	3855
	ABUSE	*misstand, wantoestand*	1582
	MISERY	*ellende, miserie*	4900
	INMATE	*gedetineerde, gevangene*	4554
	CRIME	*criminaliteit, misdaad*	10 047
	DISSATISFACTION	*ongenoegen, onvrede*	3038
	DEMONSTRATOR	*betoger, demonstrant*	3145
	TURN	*ommekeer, omwenteling*	1100
abstract	REALITY	*realiteit, werkelijkheid*	15 085
	INTENTION	*bedoeling, intentie*	13 820
	FUSS	*commotie, ophef*	2679
	CHARACTERISTIC	*eigenschap, kenmerk*	5792
	ISSUE	*kwestie, vraagstuk*	16 297
	COURAGE	*lef, moed*	4744
	CONCERN	*ongerustheid, bezorgheid*	1801
	EMOTION	*emotie, gevoel*	9311
	INTERFERENCE	*bemoeienis, bezorgdheid*	1599
	ANNOYANCE	*ergernis, irritatie*	2289
	DIVERSITY	*diversiteit, verscheidenheid*	3139
	FANTASY	*fantasie, verbeelding*	4560

	CONCEPT	NEAR-SYNONYMS	FREQUENCY
rest	ZOO	*dierentuin, zoo*	2110
	RACK	*rek, schap*	2433
	MEDICINE	*geneesmiddel, medicijn*	6150
	MAGAZINE	*blad, magazine, tijdschrift*	11 793
	TEACHER	*leekracht, leraar*	9909
	EMAIL	*e-mail, mail*	16 236
	AIRPLANE	*toestel, vliegtuig*	13 885
	TICKET	*ticket, kaart*	19 838
	WEBSITE	*site, website*	34 748

9.2 Modelling of token spaces and selection of profiles

We created 108 token-based vector space models per concept. A textual description of the parameters involved in the modelling is given in Section 3.5 of Chapter 3, together with a comparison with modelling choices in other chapters. In Table 9.3 we summarize again the relevant parameters used in this study. In Table 9.2 the absolute concept frequencies are listed in the last column. These frequencies are important as a weighting factor in the calculation of the weighted uniformity indices U’. As stated in Chapter 7, in a usage-based framework the frequency with which concepts appear in actual communication should have an impact on our assessment of the systemic relationship between lects, so that very frequent concepts play a larger role than less frequent concepts. Therefore, the relative frequency of each concept within a certain lectal comparison is used as a weighting factor for the uniformity index of each concept separately. The concepts under scrutiny in this study vary widely in frequency (from 1110 tokens for HOUSING MARKET up to 73 951 tokens for NEXT). However, in the token-based modelling procedure extremely frequent concepts—especially those above 20 000 tokens—lead to extremely large cosine distance matrices, causing the lectometric calculations to drastically slow down or in some case even break down because of insufficient computational power. The solution that we suggest for the study in this chapter is to adopt a graded downsampling scheme in order to keep the number of tokens to be modelled at a manageable size, while at the same time still taking into account the relative frequency differences between concepts. At least since Zipf (1945; Casas, Hernández-Fernández, Català, Ferrer-i-Cancho, and Baixeries 2019) it is in fact known that more frequent words are more polysemous and this graded downsampling scheme precisely applies this knowledge to our test concepts. Instead of taking a fixed amount of tokens irrespective of original concept size, we sample more tokens from highly frequent concepts, in view of their potentially larger semantic diversity. Such a scheme allows us to keep the modelling of

Table 9.3 Overview of parameters for the Dutch standardization study

PARAMETER	VALUES
First-order window size	5-5; 10-10;15-15
Association measure	PPMI (positive pmi); LLR (log-likelihood ratio)
Association measure filter	ASSOCNO (no user of filtering) SELECTION (filtering context words below threshold) WEIGHT (weighting context words by association strength)
Second-order item	word forms
Number of second-order features	MIN400 (all items with frequency above 400) FOC (union of first-order context words) SVD (200 dimensions as returned by SVD)
Type of context overlap	intersection-based; union-based

Table 9.4 Downsampling scheme for concept sizes

ORIGINAL CONCEPT SIZE (X)	DOWNSAMPLED CONCEPT SIZE
x ≤ 1000	keep original concept size
1000 < x < 15 000	1000
15 001 < x < 30 000	1100
30 001 < x < 45 000	1200
45 001 < x < 60 000	1300
60 001 < x < 75 000	1400
x > 75 001	1500

token-based spaces feasible while at the same time doing justice to the semantic properties in our dataset. The downsampling scheme is shown in Table 9.4.

When the weighted U'-index will be finally calculated, the original relative frequencies of the concepts will be taken into account again. At this point one might raise the following objection: why is such a graded scheme implemented if at the end the relative frequencies are used anyway? Could a fixed number of tokens, for instance 1000 for each concept, not be a more straightforward solution? The reason for this graded increase is mainly motivated by the manual disambiguation step in the procedure. As explained in Section 8.3, accurate lectometric measurements still require a set of manually disambiguated tokens, in contrast with a fully automatic workflow. Precisely because we know that large concepts might yield a higher degree of polysemy, we need to have sufficient tokens to annotate for the large concepts, but less for the small concepts.

The procedure for selecting stable and high-quality tokens for manual annotation has already been introduced in Section 8.3. Here it suffices to mention that by only sampling one fifth of the tokens in the 50%-top percentile of the most stable tokens, we end up with sets of tokens to disambiguate per concept in a range from 97 for a concept like INMATE to 137 tokens for the largest concept NEXT (and an average of 100 tokens per concept). In total, the number of tokens to annotate for the full list of concepts amounts to 8744 tokens. Although this is in absolute terms still a high number, it is on average only a tenth of the tokens modelled in a single space, and certainly a very small fraction of the original concept sizes.

For the creation of the onomasiological profiles we followed the stepwise procedure outlined in Section 8.3. This means that in the modelling of each concept, tokens were removed at different stages and because of different reasons. These removal steps take place within individual models, and therefore only affect the size and token composition of a single model. First, tokens that could not receive a token representation were deleted, so as to keep only the modelled tokens. Second, after HDBSCAN clustering, tokens that were classified in the noise category were removed. Third, tokens belonging to monolectal clusters were also removed. While in principle in this study we take three lectal dimensions into account (regiolect, register, and chronolect), a cluster is considered monolectal when it exclusively contains tokens from one regiolect, that is, Belgian Dutch or Netherlandic Dutch. We refrained from removing monolexical clusters, but in a final step did filter out remaining clusters in which more than 20% of the annotated tokens were out-of-concept. At the end of the workflow, we have 108 different onomasiological profiles for a given lect in a given concept. Given the operationalization of hierarchical (de)standardization and (in)formalization, which require the creation and comparison of eight onomasiological profiles in total, we arrive at 864 onomasiological profiles (i.e. individual measurement points) for each concept.

It goes without saying that with the stepwise filtering workflow, many individual token-based models that started out with 1000 to 1500 tokens, will sometimes be drastically reduced. The variation between concepts is relatively large: some of them, such as INTERROGATE, SHOP, and TICKET, only keep about 20 or 30 tokens, distributed over different lects, whereas other concepts, such as EXPLICIT, PENALTY, and MISERY, kept on average between 500 and 600 tokens. It is worth noting that even the concepts that are the least affected by the filtering procedure lose almost 50% of their initially modelled token set. It is possible that the final set of tokens has also changed in a more qualitative way: while every starting token space would contain tokens from the eight different lects under scrutiny, final sets might well only be made up of tokens from a more limited set of lects. By consequence not all lectal comparisons can be calculated, and the resulting uniformity value will therefore receive an NA value.

Recall that the initial token set present in the distance matrix was already the result of graded downsampling, which by definition leads to a disruption of the true relative weight of the concepts. In order to reconstruct the true size of the in-concept tokens of a regiolect, we work with the following formula:

(9.1) *Formula for reconstructing the true concept size per regiolect*

$$\frac{\text{final frequency of regiolect}}{\text{downsampled concept size}} \times \text{original concept size}$$

As an example, let us look at the concept ANYWAY, and in particular at a token space with ten words left and right, log-likelihood-based selection, and second order union of first order context words. In the full corpus ANYWAY has a frequency of 45 022, which according to the downsampling scheme is reduced to 1300 tokens to model. After the filtering procedure, there are 132 Netherlandic Dutch tokens and 119 Belgium-Dutch tokens left. This means that the proportion of useful Netherlandic Dutch tokens is 10% (132/1300), and that of Belgian Dutch tokens 9% (119/1300). By multiplying these proportions with the original concept size we now have the final and true number of in-concept tokens as if we had sampled them from the full corpus directly: 4121 Belgian Dutch tokens and 4571 Netherlandic Dutch tokens. These are the figures that are going to be used as lect-specific weights for the weighted uniformity indices.

9.3 Hierarchical standardization and destandardization

We can now turn to the actual analyses of the three operationalizations of the destandardization phenomena: hierarchical (de)standardization, (in)formalization, and (de)homogenization. For each operationalization the analyses will be conducted on three levels that vary from a methodological to a descriptive focus. First, we present bird's-eye view visualizations in which we try to show the consequence of varying and aggregating over different models. These plots concern more directly the issue of the (in)stability of lectometric results with respect to choosing one set of parameter values instead of another. Second, we carry out a similar analysis, but this time we want to grasp the variability within the chosen concepts, given the many different modelling solutions. Are all concepts equally affected by the variability of model parameters or not? In these plots it will also be possible to capture potential (de)standardization, (in)formalization, and (de)homogenization tendencies of individual concepts. In this way, these figures form a bridge with the third, most descriptive level of our analysis, where we present focused numerical comparisons on the semantic fields in the noun datasets, giving us more insight for the benefit of the descriptive objectives of this chapter.

Figures 9.1 and 9.2 present different types of information in a format that will recur in the rest of the analysis. Specifically, they show how building many token-based vector space models can help us derive more realistic and trustworthy estimates of the lectometric calculations. First, the three dimensions of destandardization are inherently comparative, that is, each time two uniformity values are compared. The three dimensions of destandardization in Section 7.2 of Chapter 7 are formulated as (mathematical) inequalities, for which it holds that hierarchical destandardization, informalization, and dehomogenization occur when $a > b$, whereas standardization, formalization, and homogenization occur when $a < b$. For the sake of visual and analytic simplicity we will rewrite these inequalities in the form $b - a > 0$ and $b - a < 0$. This reformulation has the advantage of a more intuitive rescaling of the dynamics: 0 is to be interpreted as no change in the stratificational configuration, values higher than 0 signal standardization, formalization, and homogenization, and values lower than 0 indicate destandardization, informalization, and dehomogenization. In Figures 9.1 and 9.2 the x-axis shows exactly this scale for the dimension of hierarchical (de)standardization, respectively for Belgian Dutch and Netherlandic Dutch.

The y-axis, on the other hand, shows the 108 models. These are ranked in descending order of the aggregate (or better: mean) hierarchical (de)standardization index explained above, weighed by the frequency of the Belgian Dutch tokens in Figure 9.1 and the frequency of the Netherlandic Dutch tokens in Figure 9.2 for each of the 85 concepts. This weighted average index is visualized by the red dot. The blue dot, on the other hand, is the unweighted average index, so without taking into account information about the relative weight of the concept within the list. In other words, for each of the 108 models we show the mean (de)standardization index within the distribution of concepts' individual (de)standardization indices. The dispersion around the weighted mean is plotted with error bars (one standard deviation above and under the mean value) and therefore captures the diversity of the individual concepts' values. The green vertical line is the reference line set at 0: if a mean or the entire distribution is to the right of the line, hierarchical standardization has taken place; if it is left of the line, hierarchical destandardization, the opposite, occurred. The red vertical line is the across-models mean of the within-models weighted mean, that is, the (unweighted) average of all the red dots. The blue vertical line is the across-models mean of the within-models unweighted mean, so again the unweighted average of all the blue dots.

Figures 9.1 and 9.2 are so-called 'caterpillar plots', that is, a ranking of the distributions of values, embedded in a certain level of a factor. The factor is in this case the modelling space, and the levels the individual models. On the y-axis of these plots we therefore find the 108 models, and the horizontal whiskers define the distribution of values for all the concepts modelled with a specific set of model parameters. We have omitted the full names of the models on the y-axis of these plots, both because they would take up too much space on the margins of the plots and because for our analysis it is not important to know exactly which model is at the top and which at the bottom. Since we are interested in a bird's-eye view perspective on the stability of the measurements, we look at the plot holistically and pay less attention to the distribution of indices within single models. Of course, models at the extremes of the ranking (i.e. the very top and the very bottom) tend to differ considerably in their mean aggregate hierarchical (de)standardization scores: in the Belgian Dutch data, the lowest value, −0.19, is reported by the model that uses the union-based procedure and considers 15 context words to each side of the target with a log-likelihood ratio larger than 1 with the target. As second-order context features the union of the first-order context words of the near-synonyms is used. Conversely, the model with the highest value, 0.13, is the one which also uses the union-based procedure, but considers ten context words to each side of the target, without additional filtering. The second-order context features are again the union of the first-order context words of all the near-synonyms. If one had, for any reason, chosen to create only token spaces based on either of these two sets of parameter combinations, one would have ended up with very different uniformity scores giving each a biased picture of the phenomenon. By taking the mean of means, across models with different parameter settings, we counteract the bias that would result from relying too heavily on individual models.

What can we now learn from Figures 9.1 and 9.2, which plot the distributions of hierarchical (de)standardization indices for respectively Belgian Dutch and Netherlandic Dutch? For Figure 9.1, which plots the results for Belgian Dutch, the across-models mean of unweighted within-models means, visualized by the blue vertical line, is 0.01. On the other hand, the across-models mean of weighted within-models means, visualized by the red vertical line, is (rounded to two decimals) 0. The difference between the weighted and unweighted uniformity values is negligible as both hover around 0. In Figure 9.2, which plots the distribution of indices for Netherlandic Dutch, we observe a very similar pattern: –0.04 across-models mean of unweighted within-models means and –0.03 across-models mean

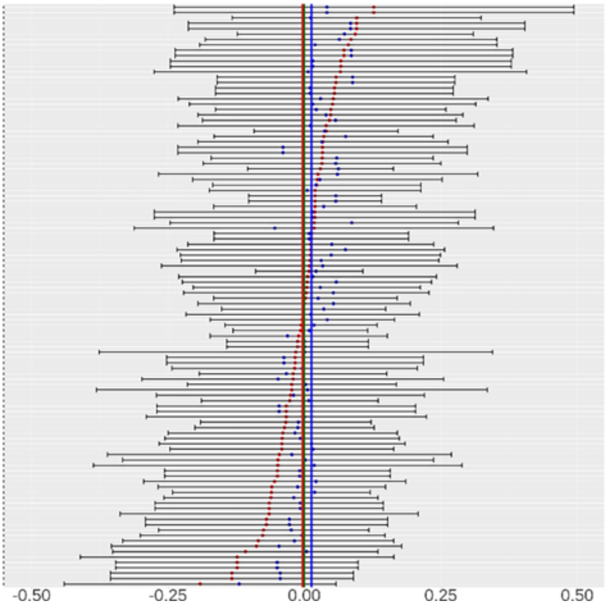


Figure 9.1 Hierarchical (de)standardization scores in Belgian Dutch (caterpillar plot with 'models' on y-axis)

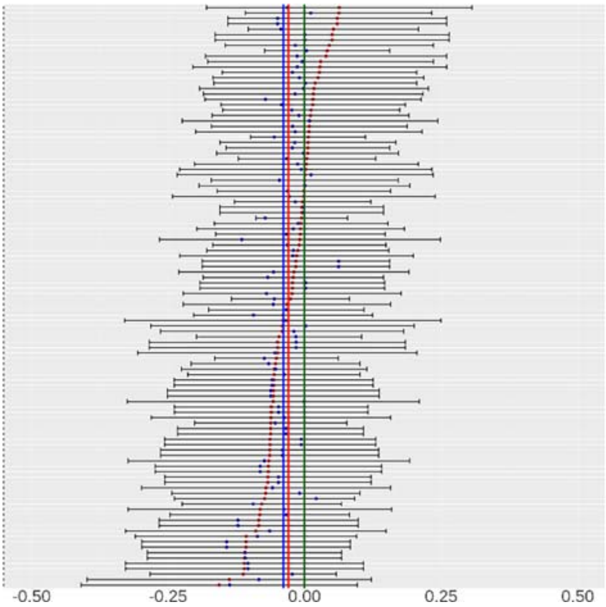


Figure 9.2 Hierarchical (de)standardization in Netherlandic Dutch (caterpillar plot with 'models' on y-axis)

of weighted within-models means. Such a negative value on this dimension signals hierarchical destandardization in Netherlandic Dutch, that is, the informal and formal strata have diverged over time, as the uniformity value between informal and formal varieties in the period 2017–2018 is smaller than the uniformity values between those same varieties in the period 1999–2004. However, both figures are very close to the reference value of 0, which indicates no change at all. It is therefore safer to assume that no real change in the stratificational distances has occurred, and that the dimension of hierarchical (de)standardization has remained stable in Netherlandic Dutch, and certainly also in Belgian Dutch. In conclusion and going back to the two goals of this chapter, these graphs have shown two things: on a descriptive level, neither in Belgium nor the Netherlands has the distance between the informal and formal strata changed much; on a methodological level, the measurement of hierarchical (de)standardization is not dependent on model parameters, and indices remain relatively stable over parameter choices.

The next pair of plots, Figures 9.3 and 9.4, relies on a similar organization as the caterpillar plots in Figures 9.1 and 9.2, and for the interpretation a similar perspective can be taken. The x-axis again represents the same reformulated hierarchical (de)standardization score, the blue dots and lines the unweighted means, the red dots and lines the weighted means, and the whiskers a distribution of these (de)standardization scores. The ordering of the distribution is again based on the

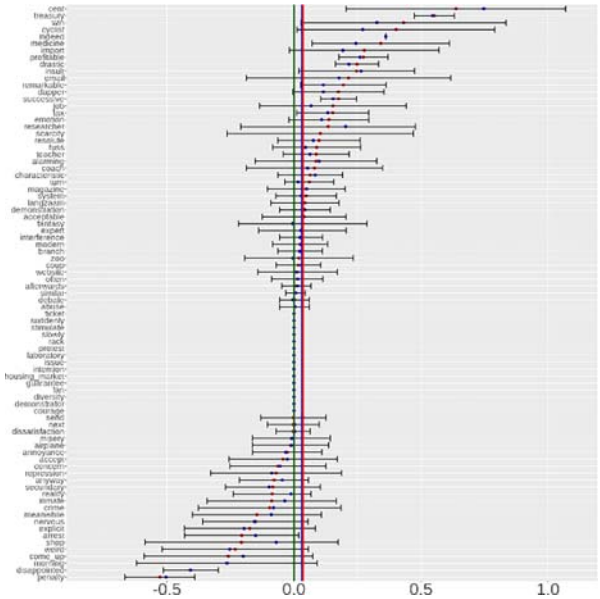


Figure 9.3 Hierarchical (de)standardization scores in Belgian Dutch (caterpillar plot with 'concepts' on y-axis)

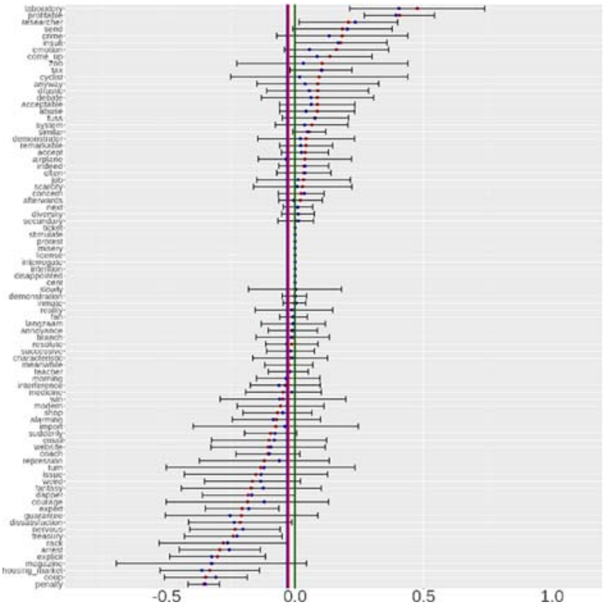


Figure 9.4 Hierarchical (de)standardization scores in Netherlandic Dutch (caterpillar plot with 'concepts' on y-axis)

weighted means (i.e. the red dots). The main difference lies on the factor and levels displayed on the y-axis, with the 85 concepts instead of the 108 models. By consequence, instead of showing the distribution of (de)standardization scores within a certain model, consisting of 85 individual scores for each concept, like in Figures 9.1 and 9.2, the relation is reversed: Figures 9.3 and 9.4 show the distribution of (de)standardization scores within a certain concept, which consists of the up to 108 individual scores for that concept in each model. The main information that we can extract from such plots has again to do with the (in)stability of the lectometric measurements, but this time we do not compare models but concepts.

What do Figures 9.3 and 9.4 tell us, for respectively Belgian Dutch and Netherlandic Dutch? First, just like in Figures 9.1 and 9.2, the ordering of the levels on the y-axis, that is, the 85 concepts, differs between the two regions. In our third level of analyses we will build on this observation by looking at semantic fields, but here it is already possible to capture differences at the level of individual concepts. Second, for some concepts the distribution of the (de)standardization scores lies fully to the right or left of the green reference line indicating 'no change'. This means that, regardless of the choice of combination of parameters, these concepts show a robust tendency towards destandardization or standardization. For Belgian Dutch, for example, CENT, TREASURY, and WIN show a standardization trend (above 0), while PENALTY and DISAPPOINTED are the only concepts having consistently destandardized. In general, more concepts are standardizing than destandardizing in Belgium, which explains the red/blue lines to the right of the green reference line. For Netherlandic Dutch the situation is the reverse: only LABORATORY, PROFITABLE, and RESEARCHER have standardized, while the group of clearly destandardized concepts is larger (for instance, PENALTY, COUP, EXPLICIT). In both countries PENALTY and PROFITABLE are clear examples of hierarchical destandardization and hierarchical standardization respectively. Third, not all concepts are equally affected by the variability generated by the different parameter combinations. For instance, in Figure 9.3 the (de)standardization score of a concept like DISAPPOINTED is located in a much narrower window (between −0.51 and −0.30) than a concept like WIN, whose scores range from a minimum of 0.02 to a maximum of 0.84 (with a weighted mean of 0.43). In other words, the standard deviation for the lectometric scores of concepts can vary widely, which directly depends on the variability caused by having different model solutions and also by how many models were actually kept after the stepwise procedure for building onomasiological profiles introduced in Section 8.3. The plots also show concepts in which apparently no distribution of values could be computed, and that only show red and blue dots that overlap with the green reference line of 'no change' (such as MISERY in Netherlandic Dutch, and PROTEST in Belgian Dutch). For these concepts the causes should be more directly sought in data scarcity. In our stepwise procedure it is possible that, at the very end, one only keeps small clusters. For these clusters the Log-likelihood Ratio test never reaches significance, the U-values are therefore set automatically to 1 (i.e. fully similar profiles) and by consequence no changes between U-values in the (de)standardization score is detected.

Having focused on the issue of (in)stability of lectometric measurements, from the perspective of the variability in parameters settings and that of concepts alike, we now turn to our third level of analysis. In order to substantiate the descriptive analysis we constructed a concept list with concepts from several word classes, and within the noun class we distinguished different semantic fields. In the next step we therefore also conducted a mixed-effects regression analysis on the subset of nominal concepts with the hierarchical (de)standardization score as dependent variable, the model parameters, semantic fields, and regiolect as independent variables, and concepts and individual token models as random effects. As expected, the individual models were not retained as a significant random effect, while the concepts were. The interaction between semantic fields and regiolects also turned out to be significant (F=33.6925; df=5; $p < 0.001$). Figure 9.5 plots the differences between the Belgian Dutch and Netherlandic Dutch scores within semantic fields. For all regression analyses reported in this chapter the same backward stepwise model selection procedure was used to filter out non-significant predictors. We started from a maximal model that included the interaction between 'region' and the concept-related predictors 'semantic field' and 'concept size' or model-related predictors like the first-order window size (see Table 9.3 for an overview

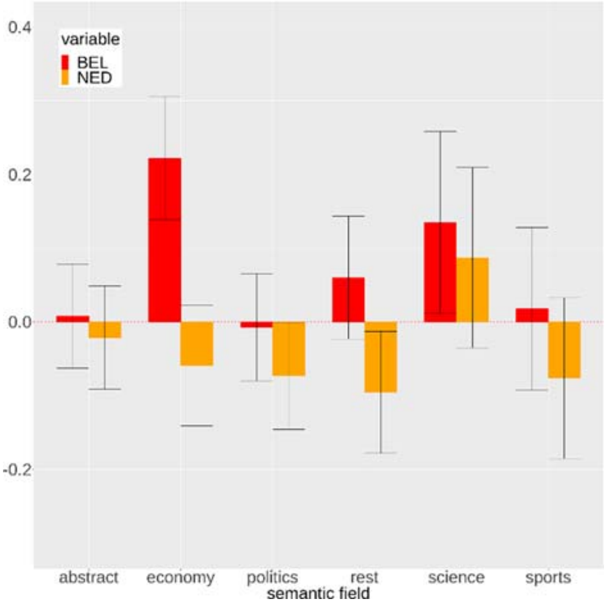


Figure 9.5 Hierarchical (de)standardization scores across semantic fields

of implemented parameters). As for this level of analysis we are interested in questions related to the lectometric distance between Belgian and Netherlandic Dutch, we will only report on the interaction between 'region' and the concept-related predictors when significant, and consider the interactions between 'region' and model-related predictors as control variables.

Overall, there is a lot of variability across semantic fields, even across regiolects. In Netherlandic Dutch most of them have undergone destandardization, with the only exception being the scientific concepts. On the other hand, in the Belgian Dutch data the situation is more variable: economy-related and science-related concepts have clearly standardized (and thus convergence between the informal and formal varieties took place), while no real change seems to have occurred in the other fields (i.e. the error bar covers the 0 reference line). It is interesting to note how in the group of scientific concepts, both Netherlandic Dutch and Belgian Dutch have seen an increase in the lexical overlap between formal and informal varieties, while for the sports and economy semantic fields the trajectories are mirrored: hierarchical destandardization in Netherlandic Dutch, but hierarchical standardization in Belgian Dutch.

9.4 Formalization and informalization

Following the pattern of the previous section, Figures 9.6 and 9.7 represent our first level of analysis, where we focus on the (in)stability over the 108 model solutions. The plots are designed in the same way as the figures for hierarchical (de)standardization: values higher than 0 denote formalization, that is, the informal strata move towards the formal strata, while values lower than 0 signal informalization, that is, the formal strata move towards the informal strata. Similar to the development regarding hierarchical (de)standardization, where apparently no large changes were detected for any regiolect, the across-models aggregate values for the (in)formalization score do not seem to differ in a significant way from a scenario of no change, in either regiolect. In Belgian Dutch we observe a slight movement towards informalization. The across-models averages are (rounded) 0 for the unweighted profiles, and −0.04 for the weighted profiles. It is hard to say whether these figures provide a significant proof of the informalization dynamics in Belgium, given that all models include the 0 reference line within their distribution. In the Netherlands, however, the absence of any (in)formalization tendency is even more outspoken: 0.02 for the across-models unweighted profiles and −0.01 for the across-models weighted profiles.

On the second level of our analysis, which focuses on the variability across and within concepts instead of tokens-based models (in Figures 9.8 and 9.9), we can observe once again that individual concepts might pull towards formalization rather than informalization, that for some concepts we can be confident this trend is robust against variation in specific model parameters (for example, MORNING and RESEARCHER in Belgian Dutch, or ZOO and COUP in Netherlandic Dutch),

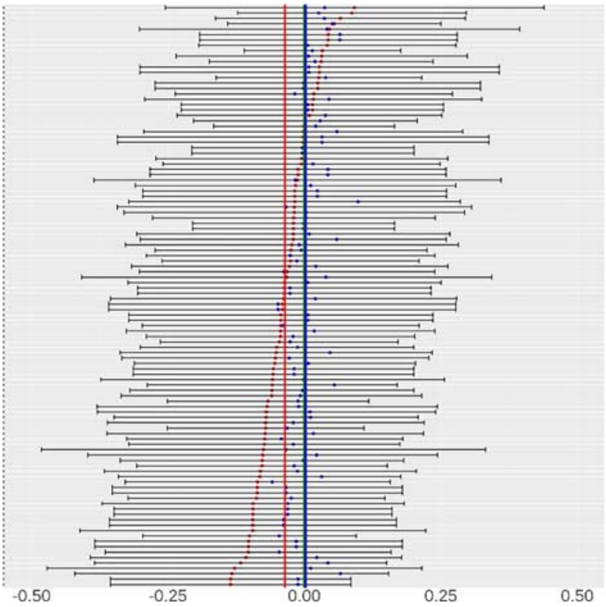


Figure 9.6 (In)formalization scores for Belgian Dutch (caterpillar plot with 'models' on y-axis)

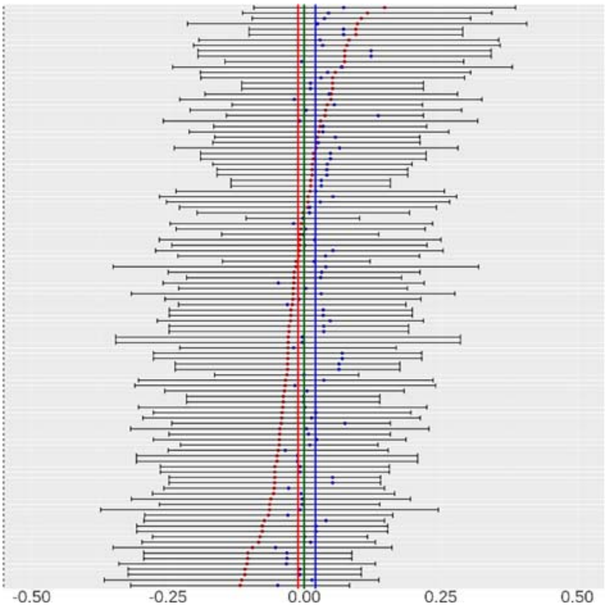


Figure 9.7 (In)formalization scores for Netherlandic Dutch (caterpillar plot with 'models' on y-axis)

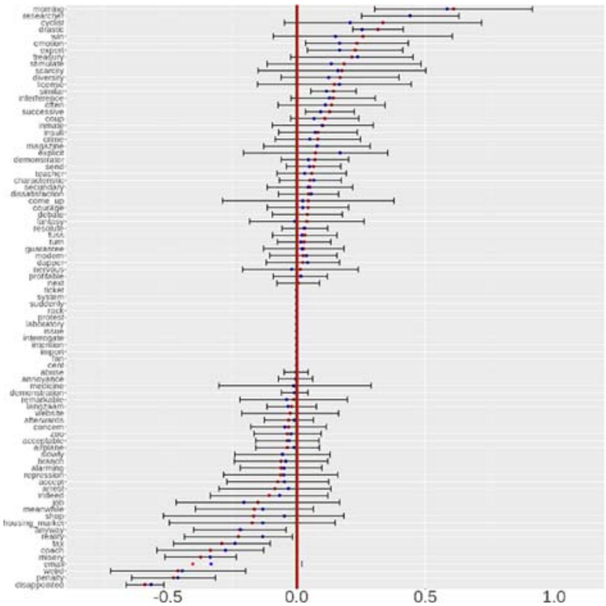


Figure 9.8 (In)formalization scores for Belgian Dutch (caterpillar plot with 'concepts' on y-axis)

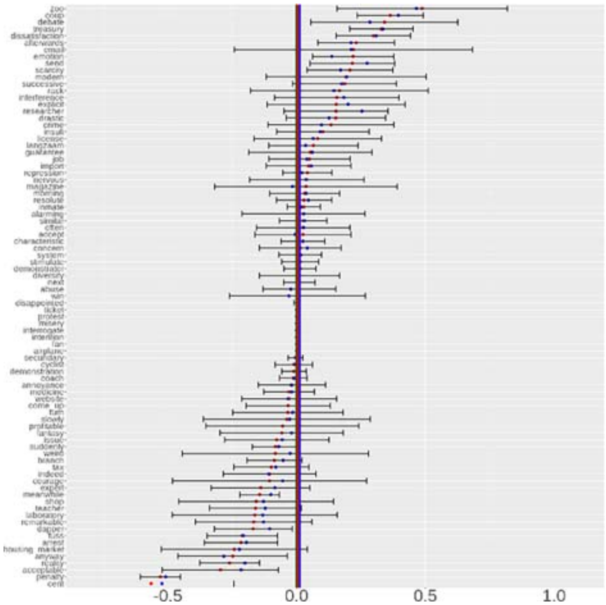


Figure 9.9 (In)formalization scores for Netherlandic Dutch (caterpillar plot with 'concepts' on y-axis)

while for other concepts the dynamics are much less outspoken, and by consequence it matters a lot by means of which parameters their token-based vectors have been built. In the four plots showing the phenomenon of (de)standardization, the overall weighted and unweighted means (i.e. the red and blue lines) matched independently of the specific aggregation perspective, that is, model-based (in Figures 9.1 and 9.2) or concept-based (in Figures 9.3 and 9.4). For the analysis of (in)formalization this is no longer the case. In fact it seems that, both for Belgian as for Netherlandic Dutch, weighted and unweighted across-models means do differ from the weighted and unweighted across-concepts means. In the latter case weighted and unweighted means fully overlap with the reference line of 'no change'.

As with the analysis for the hierarchical (de)standardization dimension, we carried out a mixed-effects regression with the same model design on the dimensions of (in)formalization. In this regression model the interaction between the region and semantic field was significant (F=32.1326; 5; p < 0.001). Judging from the caterpillar plots in Figures 9.8 and 9.9, one would not expect Netherlandic Dutch and Belgian Dutch to behave very differently on this dimension. However, when focusing only on the noun class and the variability among semantic fields (as plotted in Figure 9.10), much more disagreement is observed. The semantic field of science-related terms shows strong differences, with a clear movement towards

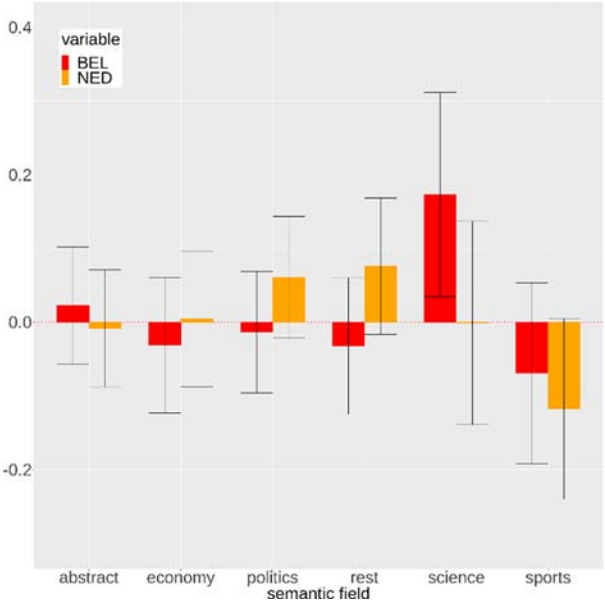


Figure 9.10 (In)formalization scores across semantic fields

formalization in Belgian Dutch (as the error bar is located fully above the 0 reference line), but an absence of shift in Netherlandic Dutch. Furthermore, the sports concepts behave rather differently compared to the other semantic fields. Here we observe a clear shift towards informalization, which is strong in both regiolects. The other semantic fields present a more mixed picture: no clear tendency for abstract and economy concepts, in both regional varieties, and rather formalization in the Netherlandic Dutch use of political concepts, as opposed to informalization for those same concepts in Belgian Dutch.

9.5 Homogenization and dehomogenization

In order to answer the question about increased or decreased dehomogenization in the two regiolects, we turn to the comparison of I-measures across time periods. Figures 9.11 and 9.12 once again plot the distribution per model, respectively for the Belgian Dutch and the Netherlandic Dutch formal variety. Both figures show the lack of any type of change, either in the direction of homogenization or that of dehomogenization, in the two regional varieties. The across-models weighted score and the across-model unweighted scores (i.e. the red and blue vertical lines) are barely distinguishable from the green reference line. On top of that, the stability of scores across models is even more evident in this dimension, as signalled by the smaller distributions per model. This is not a surprise, given that the operationalization of dehomogenization only involves the estimation of values in two profiles per regiolect, whereas the other dimensions, building on U-values instead of I-values, involve four profiles, and therefore potentially an increase in variability and noise. The across-concepts analyses (Figures 9.13 and 9.14) show a similar picture, with generally smaller ranges of scores per concepts but also fewer concepts that pull in either one of the two directions.

Zooming in on only the nominal concepts, we conducted the same type of mixed-effects regression analysis. We can see several patterns (in Figure 9.15). First, only the group of abstract concepts seems to have consistently undergone homogenization in both regional varieties (i.e. their error bars are all above the 0 reference point), which implies that for those concepts the internal uniformity has increased over time in the formal stratum. Second, the science-related and economy-related fields behave differently in the two regional lects. In the economy field, the Netherlands seems to exhibit a development towards more homogeneity, while in Belgium, these concepts tend to heterogeneity; for the scientific concepts the development is the opposite: dehomogenization in the Netherlands, but homogenization in Belgium. Last, the sports semantic field shows trends that go in the opposite direction compared to the other fields, namely, dehomogenization in both regiolects.

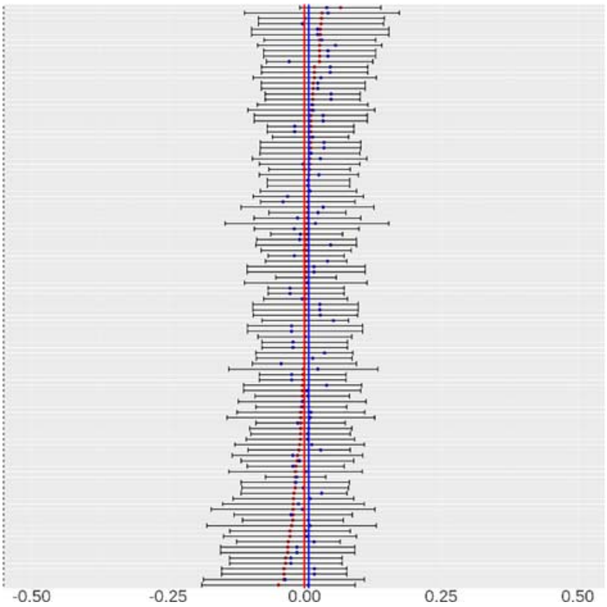


Figure 9.11 (De)homogenization scores for Belgian Dutch (caterpillar plot with 'models' on y-axis)

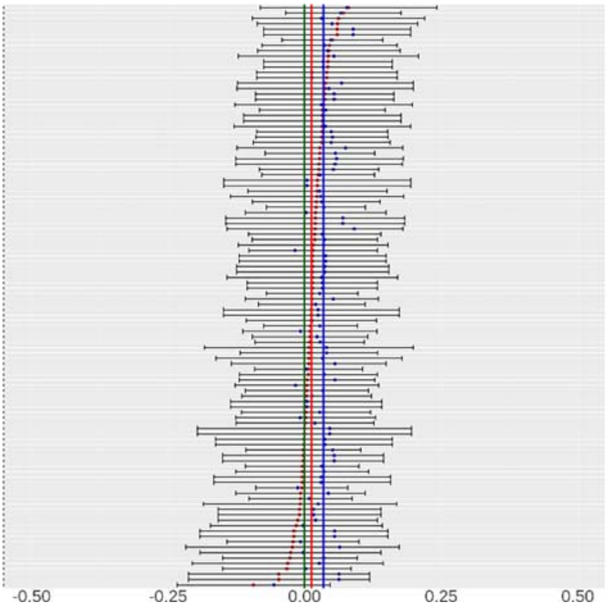


Figure 9.12 (De)homogenization scores for Netherlandic Dutch (caterpillar plots with 'models' on y-axis)

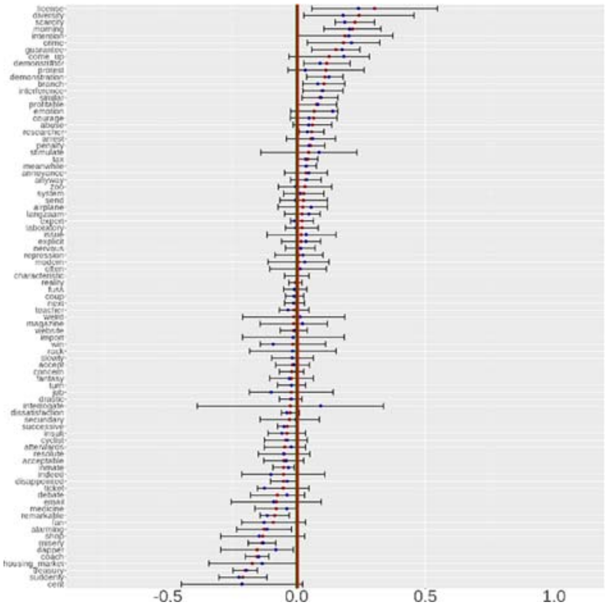


Figure 9.13 (De)homogenization scores for Belgian Dutch (caterpillar plots with 'concepts' on y-axis)

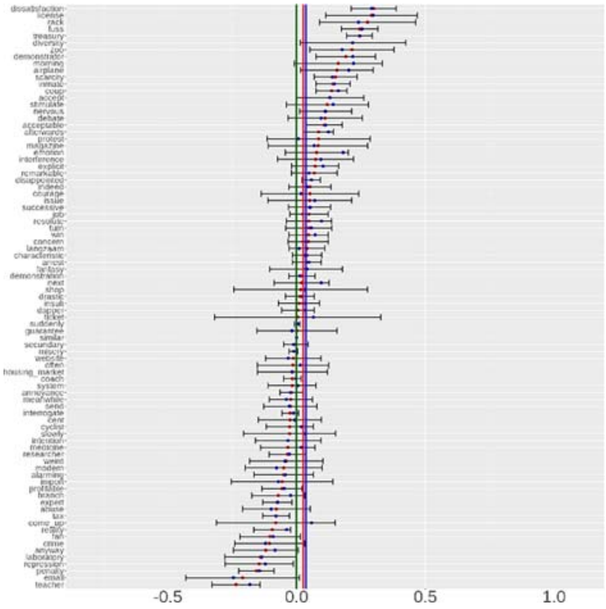


Figure 9.14 (De)homogenization scores for Netherlandic Dutch (caterpillar plot with 'concepts' on y-axis)

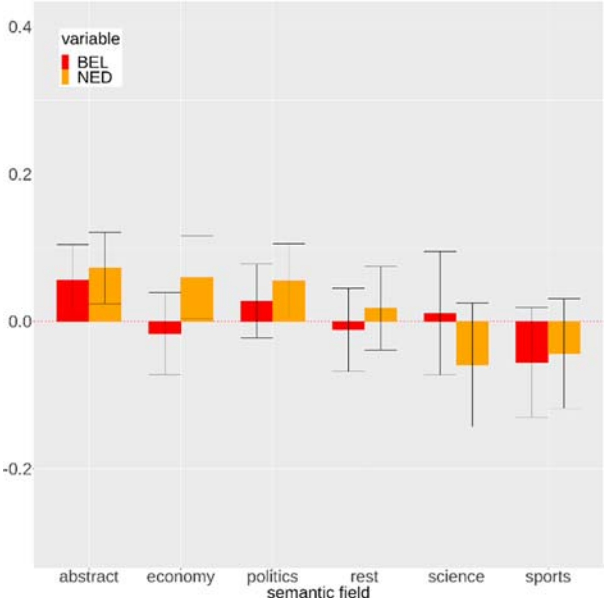


Figure 9.15 (De)homogenization scores across semantic fields

9.6 The evolution of Belgian and Netherlandic Dutch

In Section 7.2 in Chapter 7, with the introduction of the three operational definitions and formulae for the different destandardization dimensions, a small-scale study was carried out on the lexical field of 14 clothing terms. The results showed a development towards more standardization, rather than destandardization. The trends were strong in Netherlandic Dutch, in which hierarchical destandardization and formalization were observed (i.e. the advergence of the informal varieties towards the formal varieties) along with a reduction of the lexical heterogeneity in the most formal stratum. In Belgian Dutch similar dynamics were found, with the exception of informalization, that is, the advergence of the formal variety towards the informal variety. How does our large semi-automatic and token-based analysis, based on multiple lexical fields and parts-of-speech, compare to this smaller, manually conducted case study?

The results described in Chapter 7 are repeated in Table 9.5 side by side with the results obtained in the present chapter. In the table we focus on the (original) U-indices and not on the derived hierarchical (de)standardization, (in)formalization, and (de)homogenization scores of the above plots. In each pair of columns, the inequality signs >, < point in the same direction when trends in both case studies agree, within a certain dimension. If the directions are

Table 9.5 Overview of destandardization scores in Belgian Dutch and Netherlandic Dutch

BELGIUM		**THE NETHERLANDS**	
Chapter 7	***Chapter 9***	***Chapter 7***	***Chapter 9***
Hierarchical (de)standardization		*Hierarchical (de)standardization*	
U(B90, LeuKor90) = 50.47 < U(B12, LeuKor12) = 73.72	U(B00,B-Usenet) = 90.92 </≈ U(B18,B-twitter) = 91.85	U(N90, LeiMaa90) = 69.07 </≈ U(N12, LeiMaa12) = 73.62	U(N00,N-Usenet) = 93.69 >/≈ U(N18,N-twitter) = 89.40
(In)formalization		*(In)formalization*	
U(LeuKor90, B12) = 60.25 > U(B90, LeuKor12) = 53.12	U(B-Usenet, B18) = 91.92 </≈ U(B00,B-twitter) = 92.20	U(LeiMaa90, N12) = 61.57 < U(N90, LeiMaa12) = 84.93	U(N-Usenet, N18) = 89.40 </≈ U(N00,N-twitter) = 91.42
(De)homogenization		*(De)homogenization*	
I(B90) = 69.21 < I(B12) = 74.96	I(B00) = 65.17 </≈ I(B18) = 66.10	I(N90) = 68.48 </≈ I(N12) = 71.06	I(N00) = 67.99 </≈ I(N18) = 71.91

opposite, the trends disagree. What emerges from the comparison is a relatively large match between the two case studies: full agreements of trends regarding dehomogenization, for both Belgian Dutch and Netherlandic Dutch, but partial disagreement regarding the hierarchical (de)standardization dimension in the Netherlandic data and the (in)formalization dimension in Belgian Dutch: whereas Chapter 7 noted a standardization trend in the Netherlandic data, we saw a trajectory towards destandardization, and while in that case study Belgium was informalizing, in our data we observe a trend towards formalization. We have to specify here that in neither of the two studies the differences between the U-indices for hierarchical (de)standardization are larger than the five points that are customarily used to signal a true difference. On top of that we can see that, while in the Chapter 7 study the majority of uniformity index differences are usually larger than this customary five-point difference (four out of six), that is not the case for the indices calculated in the present case study, where all of the comparisons remain within the five uniformity index points difference.

Overall, our results do not seem to differ that much from the study in Chapter 7, when the directionality of the phenomena is compared. However, a striking difference concerns the absolute values of the U- and I-values in the different studies. For instance, even though both chapters signal standardization in Belgian Dutch, the baselines are very different, with U-values around 90 in our semi-automatic replication compared to 50.47 and 73.72 in the manual study. As U-indices are mainly meant to be interpreted relative to one another, judgements based on the absolute values might not be appropriate, certainly given that at present we have no knowledge about the theoretical statistical distribution of these values. Leaving aside the fact that differences between U-values for the hierarchical (de)standardization dimension in Netherlandic Dutch do not seem to be significant in both studies, the only consistent exception seems to concern the (in)formalization dimension in Belgian Dutch. Furthermore, in the regression analysis involving the semantic fields, we noticed a discrepancy between the field of sports concepts, where clear informalization had occurred, compared to all other fields, where formalization (or no change at all) seem to have taken hold.

To round off, we may now take a more general perspective and review our attempt to integrate token-based vector space modelling into the lectometric framework. What have we learned from adopting such a methodological innovation? At the start of the chapter we defined two goals, a descriptive one and a methodological one. A comparison between the results based on the manually collected data in Chapter 7 and our semi-automatic retrieval and processing workflow resulted in a mixed picture. A straightforward comparison has actually not always been evident, given the different underlying materials of the analyses. For instance, none of the calculations on our data shows a significant shift on any of the three destandardization dimensions (as represented by the approximation symbol ≈ in Table 9.5). One reason could be that enlarging the set of concepts from just a small set from one lexical field, as in the Chapter 7 study, to a very large and varied list, might have led to a situation in which the aggregation hides the presence of very different and even opposing trends, effectively neutralizing one another. Ironically, the large-scale aggregation, which is fundamental to a lectometric inquiry, might sometimes hide individual, or group-level tendencies. Therefore we also conducted analyses at the level of semantic fields. Those regressions indeed pointed towards a lot of across-field variation, which is summarized in Table 9.6.

Table 9.6 Summary of standard language change scores across semantic fields

	SEMANTIC FIELDS									
	science		*abstract*		*economy*		*politics*		*sports*	
	BE	NL	BE	NL	BE	NL	BE	NL	BE	NL
hierarchical (de)standardization	+	0	0	0	+	0	0	–	0	–
(in)formalization	+	0	0	0	0	0	0	0	0	–
(de)homogenization	0	0	+	+	0	0	0	+	0	0

Although the lack of strong shifts seems to be confirmed also at the level of individual semantic fields, we can observe some variability between the fields. In Table 9.6 these differences are encoded in the following way: when a certain score was significantly lower or higher than the reference value 0, it receives a minus sign and a plus sign respectively; 0 indicates no significant difference from a scenario of no change. Abstract concepts and sports concepts seem to show opposing trends, but in the two regiolects separately. In fact, if we were to place the semantic fields on a scale going from strong evidence for a shift towards standardization to strong evidence for a shift towards destandardization (i.e. understood as showing a significant tendency on one, two, or all of the operationalized dimensions of destandardization), some patterns become discernible. In Table 9.6 the semantic fields are organized precisely so as to reflect this continuum. A very cautious conclusion can be made: concepts appearing in more intellectual discourse of a general nature show more evidence towards standardization (i.e. the scientific and abstract concepts), while concepts appearing in less intellectual discourse, mainly involving local and localized events and referents, show more evidence towards destandardization. This is of course only a tentative explanation of the weak patterns we are observing in our data, and more focused research, putting the choice of semantic fields centre stage, will hopefully shed more light on these patterns. Daems (2022), applying our lectometric measures on manually analysed corpus data for disparate fields such as clothing, traffic, information technology, and emotion terms, provides further evidence that semantic fields differ substantially in how they contribute to standard language change in Dutch. As talking about certain topics varies by situation, speakers, media, and attitudes, it is not surprising that semantic fields show idiosyncrasies tightly linked to their domains of use, which in turn might explain heterogeneous developments.

The study presented in this chapter is the first to enrich the lectometric workflow with the large-scale application of token-based distributional semantics. How did this integration fare? As we asked in the beginning of the chapter: does the integration need small modifications, or will future research have to address more fundamental shortcomings? We believe that the general design is robust, but that a number of caveats have to be considered. First, while the graded downsampling scheme ensured that we could work with distance matrices of manageable size, we did expose ourselves to the risk of ending up with sparse onomasiological profiles. Second, further research should pay even more attention to the nature of the underlying corpora and try to avoid composite and ad-hoc compilations. The absence of a systematically stratified monitor corpus forced us to bring together corpora that were compiled and processed according to different protocols, and this in turn necessitated working with raw word forms, which may introduce a fair amount of noise in the creation of vectors (similar issues were observed in Chapter 6). Also, the ad-hoc nature of the corpus raises questions about the comparability of lects. Are Usenet and Twitter both equally good representatives of informal language, or rather, is the informal status of Usenet conversations of the same kind as that of Twitter interactions? The discursive environments in both types of data are not completely equivalent, and accordingly, not just the linguistic phenomena but the communicative situation as a whole may be said to change. Further lectometric research of a diachronic nature could take such changes in the communicative structure of the language more explicitly into account than has been the case in this chapter.

The bottom line

- Exploring a wide variety of model parameters and integrating that variability into the lectometric workflow allows for counteracting the bias of single, potentially idiosyncratic models. It ensures robustness in the aggregate results: it presents the results within a dispersion range of possible outcomes, instead of single measuring points.
- A lexical perspective on standard language change, operationalized along the three different dimensions of 'hierarchical (de)standardization', '(in)formalization', and '(de)homogenization', reveals considerable variability in the way semantic fields pull in different, sometimes opposing directions of change.
- Large-scale lectometric studies involving multiple lectal comparisons at different points in time are particularly sensitive to the degree of commensurability between registers, and in particular informal materials. The task is to ensure maximal comparability between subcorpora representing similar lects.

10
Pluricentricity from a quantitative point of view

This chapter illustrates how the steps in the distributional semantic lectometric workflow as detailed in Chapter 8 and applied to Dutch data in Chapter 9 can be applied to Spanish language data in a large-scale, exploratory lectometric study. In line with the previous chapter, the focus here will be both methodological and descriptive.

On the methodological side, two components will be added to the methodological repertoire for lectometric studies. First, as discussed in Section 8.4, the procedure that we apply to fine-tune and prune models may result in the removal of entire models. Accordingly, some concepts will be more richly modelled than others, in the sense that they retain more models. In the present chapter, we explicitly include this effect in the lectometric exploration, by comparing the results obtained by considering all concepts that pass the initial pruning procedure to results obtained from only those concepts that are modelled by a fair number of models. The cut-off point will be set at 50% of all models per concept.

In addition, yet another way of restricting the number of concepts included in the calculations will be illustrated. As mentioned in Section 7.1, by default we retain concepts with complete uniformity or without attested significant variation in the aggregate calculations: the concepts for which lects do not differ in their linguistic habits contribute to lectal distances no less than the concepts for which their lexicalization preferences differ. But omitting the concepts without demonstrable variation from the aggregate calculations allows the researcher to zoom in on the cases where significant differences do show up, and thus to see more clearly which lectal factors influence the differences.

We will apply both approaches to the reduction of the number of concepts in the calculations and examine what effect the number of models retained in the analysis has on the results. Further, we will compare the various calculations made on modelled data with calculations made on the full, unmodelled sample from which the modelled tokens were drawn.

Second, we add non-metric multidimensional scaling to the analysis (also see Chapter 3). Multidimensional scaling provides a way of mapping the (dis)similarity of data onto a 2D or 3D space, which allows us to visualize the relationships between various data points at once (as previously discussed in Section 4.1). The more similar two items are, the closer they are in the space,

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0010

while dissimilar items are further apart. Interpretations are then made upon this mapping of such (dis)similarities. Different forms of multidimensional scaling have been applied in lectometric studies (see Rosseel, Franco, and Röthlisberger 2020), and the technique has also been applied to Spanish language data, as in Asención-Delaney and Collentine (2011).

On the descriptive side, the overarching question is what the distributional semantic lectometric approach can tell us about the pluricentricity of Spanish. What can we learn about the relationship between the six national varieties examined in this study, especially regarding the current topics in the literature on Spanish pluricentricity discussed in the following section? We will be considering two perspectives in this regard: *pan-Hispanic* (including Spain in the analysis) and *pan-American* (excluding Spain from the analysis). Such a study is the first of its kind for Spanish. While Iberian and Ibero-American languages are becoming familiar faces in lectometric research (see Asención-Delaney and Collentine 2011 for L2 Spanish; Aurrekoetxea, Iglesia, Clua, Usobiaga, and Salicrú 2020 for Basque; Soares da Silva 2014 for Portuguese; and Sousa and Dubert García 2020 for Galician), a scaled-up analysis of Spanish has hitherto remained absent from the literature.

10.1 Spanish as an international language

Spanish holds a rather unique position among other commonly cited pluricentric languages, such as Dutch, Portuguese, English, French, or German. An official language in nearly two dozen countries and with approximately 400 million L1 speakers spread across five continents, an accurate and comprehensive description of the pluricentricity of the language quickly becomes a complex undertaking and there is considerable disagreement among scholars as to how it ought to be conceptualized.

At the institutional level, the language currently boasts 23 national language academies (the most for any language in the world), all of which make up the Asociación de las Academias de la Lengua Española (Association of Spanish Language Academies, or ASALE), with its oldest member being the Real Academia Española (Spanish Royal Academy), which was founded in 1713 and modelled after the already existing French and German language academies. Created at a time when Spain was approaching the height of its colonial empire, the Real Academia sought to centralize the codification of the language as a way to ensure its *unity* and *purity* (Méndez García de Paredes 2012) and maintained a staunchly conservative and euro-centric view on language use even well into the 20th century, all the way up to the formation of the Asociación de las Academias in the mid-20th century (Süselbeck 2012), and in some ways beyond, as evident by its motto of 'limpia, fija y da esplendor', or 'purifies, stabilizes and gives splendor' (translation taken from Thompson 1992). Throughout the 19th century when the former colonies began to gain independence, the fear that Spanish would go the way of Latin and fragment into mutually unintelligible pieces of the original led the institution to push the unity of the language to the forefront of its mission where it remains to this day. Nowadays, however, this emphasis on unity functions as a way to adapt to the challenges and opportunities posed by globalization and compete with the dominance of English (Lebsanft 2007).

Though now in collaboration with the Asociación de las Academias, the Real Academia Española remains to this day the central body through which many widely used dictionaries and grammars are published. While a growing number of descriptive works have been published on specific varieties of Spanish, only a relatively small number of those have actually been affiliated with the academies (Lebsanft 2007). Those that do share such an affiliation, such as the *Diccionario de mexicanismos* (Dictionary of Mexicanisms) (2010), have been criticized for continuing to maintain an Old-World perspective in their descriptions, defining American realities through the language of the peninsula (Süselbeck 2012). Even those that do not share an affiliation with the academies are susceptible to similar criticisms, such as the *Diccionario del español de Cuba: Español de Cuba–Español de España* (Dictionary of Cuban Spanish: Cuban Spanish–Peninsular Spanish) (2000) by continuing to place an American variety in contrast to the peninsula rather than to other American centres of influence, such as Mexico or Argentina (Ávila 2003). However, as an in-depth examination of the Real Academia Española and the Asociación de las Academias is beyond the scope of this chapter, see Süselbeck (2012) for a review of the institutions' evolution and development of a shared pan-Hispanic language policy.

The focus on Spanish in the study of pluricentric languages really began to gain momentum with Clyne's seminal 1992 book, *Pluricentric Languages*. In his chapter on Spanish, Thompson describes the language as undoubtedly pluricentric, but also comprised of standard varieties that remain highly united. Clyne and Thompson's work has been followed up on and expanded upon by a number of researchers (Bierbach 2000; Zimmermann 2001, 2008; Oesterreicher 2002; Ávila 2003; Lebsanft 2004; Bravo García 2008; Del Valle 2012; Greußlich 2015, among others). As a result, differing but frequently overlapping conceptualizations have emerged on the topic in recent decades. Maldonado Cárdenas (2012) provides an adequate summary of a few of the most influential perspectives, explaining that some, such as Bierbach (2000), understand the situation as a political one and see the nation-state as the arbiter of a standard, which would allow for aspects of culture and identity to be incorporated into the analysis. Others, such as Oesterreicher (2000), see nations as insufficient descriptors for empirical inquiry, opting instead for a regional perspective in which geographic areas are classified based on large urban centres of influence, such as Mexico City or Buenos Aires, what he calls *espacios comunicativos*—'communicative spaces'. Finally, there's the pan-Hispanic ideal described in Lebsanft (2004) and actively promoted by the Real Academia Española, which acknowledges the multitude of varieties within the language, but aims at also maintaining a broad, overarching standard that is supposed to belong to no one and yet is shared by everyone (which in the literature receives a number of monikers, from *international Spanish* to *neutral Spanish* to *the CNN standard* to *global Spanish*).

Of course, an international norm is not necessarily incompatible with pluricentrism and may serve as a solution to a language as geographically extensive as Spanish (Pöll 2012). This idea is promoted by many in the Spanish-speaking media as well (Gómez Font 2012), especially in countries like the United States with a diverse group of Spanish speakers from a wide range of Spanish-speaking backgrounds. This has led many to consider the United States as a kind of testing ground for an international standard of the language, particularly in mass media (hence the term *CNN standard*); see Lebsanft, Mihatsch, and Polzin-Haumann (2012). Yet, this embrace of a pan-Hispanic norm is not without its sceptics. Clyne (1992: 463) questions its practicality, noting that 'attempts at an 'international standard' are rarely successful since they tend to favour the varieties of D[ominant] countries'. Meanwhile, Del Valle (2012) questions the motives behind its promotion and posits that the Real Academia Española's pan-Hispanic discourse serves a kind of 'hispanofonía mercantil', or 'commercial hispanophonism', in which Spain is the *primus inter pares*.

In light of the differing views surrounding the pluricentricity of Spanish, we defer to the description offered by Lebsanft, Mihatsch, and Polzin-Haumann (2012), in which they describe the situation of Spanish as a mixed one, 'históricamente a medio camino entre el monocentrismo tradicional y una creciente aceptación de la diversidad de las normas emergentes o existentes, acompañada por el ideal de una norma panhispánica' ('historically halfway between traditional monocentrism and a growing acceptance of the diversity of existing or emerging standards, accompanied by the pan-Hispanic ideal standard').

Not surprisingly, the variation in global Spanish is reflected on the lexical level. There are a number of lexical differences that are quite well known among speakers of the language and are often used in a prototypical way to illustrate the differences between the so-called Old-World/New-World divide, such as *coche-carro* 'car', *ordenador-computadora* 'computer', or *zumo-jugo* 'a drink containing the liquid squeezed from a fruit, for instance orange juice', respectively, to name a few. For researchers, such an unhelpful dichotomy grossly distorts and oversimplifies the actual state of Spanish variation by reducing variants to *americanisms* or *peninsularisms* (Lipski 2012), especially as the movement of people and ideas across physical and cyber space is continuously altering the communicative environments in which speakers participate. Additionally, what is really taking place is often more interesting, as in the case of *zumo-jugo* mentioned above. Semasiologically, both lexemes display a certain amount of diatopic variation, in that *zumo* is in fact used in Mexico, but rather to refer to the small vapor or mist released from the skin of an acidic fruit when peeling or puncturing it (Ávila 2003), a use not documented in other parts of Latin America, such as, say, Colombia. Ávila and the present authors are of the opinion that Spanish lexical variation should be studied individually between regions or countries, which is precisely what the lectal comparisons in this study aim to achieve by opting for the latter. As Sorenson (2021: 3) points out, 'a country-by-country assessment [of Spanish] can be particularly advantageous in the lexical realm, as vocabulary is often more apt to be confined within the borders of a single country than phonological or morphosyntactic characteristics'.

Various researchers have proposed different ways to divide up the dialects of Spanish (Henriquez Ureña 1921; Lipski 1994; Lebsanft 2007; Hualde, Olarrea, Escobar, and Travis 2010), and in the classifications of the Americas, there are many overlapping consistencies. We will be following the description found in Lebsanft (2007), which coincides with many of the others, but treats the United States as its own area rather than only acknowledging its southwest region, which is often classified with Mexico. These are, from north to south:

- United States and Puerto Rico
- Mexico and Central America
- The Caribbean
- Venezuela and Colombia
- The Andean countries: Ecuador, Peru, and Bolivia
- Chile
- River Plate countries: Argentina, Uruguay, and Paraguay

It is also difficult to talk about lexical variation in Spanish without mentioning language contact. Not too long after the arrival of the Spanish in the Americas do koinés begin to emerge which incorporated words and expressions from Amerindian languages (Parodi 2001). The minority, indigenous, African, and other co-official languages found alongside Spanish have helped shape its varieties in all the countries where native speakers can be found. Additionally, many of the languages in contact with Spanish are also pluricentric languages themselves, such as Quechua and Guaraní in the Americas (Magadán, Rizzo, and Kleifgen 2020), and Catalán in the Iberian peninsula (Hawkey and Mooney 2021). Adding to this linguistic richness are the languages brought by immigrants, such as those who emigrated to Argentina from Spain and Italy in the mid-20th century, or more recently, the Central American immigrants who have relocated to Mexico, to name a couple examples.

Lectometry, then, particularly when approached from a distributional semantic framework, provides a wealth of possibilities for a language as diverse and widespread as Spanish, a well that has remained by and large untapped given the possibilities. What a distributional semantic lectometric approach can contribute to Spanish is especially relevant given the aforementioned debates around pluricentricity and the issue of authority over such a widespread language: providing lectal comparisons with varieties on equal footing by controlling sample sizes (i.e. downsizing subcorpora to equal sizes) and locating lexical variants in a corpus-based way (i.e. Clustering by Committee; see Chapters 8 and 9 and Section 10.2 below). This lectometric study also complements parallel computational research on Spanish, both in distributional semantics (Serigos 2017) and lexical-semantic variation (Baldissin, Schlechtweg, and Schulte im Walde 2022).

10.2 Corpus and concept selection

Due to its size and lectal diversity, the Web/Dialects corpus from the Corpus del Español (Davies 2016) is a rich source of Spanish language data and optimal for a large-scale study. The corpus contains over 2 billion words from 21 different varieties, each of which are systematically split by register: *general*, containing a variety of sources such as mass media and possibly blogs, and *blog*, which is strictly limited to that domain. However, because it is not immediately clear what *general* actually entails (without a meticulous examination of the sources beyond practicality for the current study), and to what degree informal language such as blogs comprises this section, we have opted to combine the registers together and treat the lects as national units. The corpus is also lemmatized and tagged for part-of-speech.

To limit the scope of our study to a workable size, we selected only a reduced number of varieties based upon the aforementioned categorization of Spanish dialects found in Lebsanft (2007). The final set included six countries: Argentina, Colombia, Mexico, Peru, Spain, and the United States; their frequencies are summarized in Table 10.1. Not only were these chosen because they are the largest dialects in terms of number of words in the corpus, but also because they provide samples from differing dialect zones, with the exception of Central America and the Antilles, although some, such as Lipski (2012), include parts of Colombia in both categories thanks to its unique position as point of convergence in which Central meets South America, and the fact that it borders both the Pacific Ocean and the Caribbean Sea. Colombia also borders Peru to the south, which is consistently placed with the Andean dialects. Included in these six countries are the major centres of influence proposed by Oesterreicher (2002): Buenos Aires, Mexico City, and an Andean dialect, in this case Peru. Given that the smallest subcorpus, that of *Peru-general*, consisted of around 47 million words, concepts were sampled from subcorpora rounded to 45 million tokens per register per lect. As explained in the previous chapter, by keeping the various subcorpora from which the concepts were sampled the same size, the resulting frequencies

Table 10.1 Sizes of the six lects in the Web/Dialects corpus in the Corpus del Español

LECT	GENERAL	BLOG	TOTAL
Argentina	93 195 550	89 509 348	182 704 898
Colombia	84 285 729	95 859 929	180 145 658
Mexico	132 651 925	127 946 347	260 598 272
Peru	47 120 271	68 204 165	115 324 436
Spain	208 808 667	250 504 154	459 312 821
United States	94 603 634	85 324 555	179 928 189

will be representative of differences in lects and not reflect a bias in the corpus structure itself.

In the previous chapter, concept selection was strictly a bottom-up process, known as Clustering by Committee, following the concept selection algorithm detailed in Chapter 8 and in De Pascale (2019). This study, however, incorporates both a bottom-up and a top-down approach to concept selection, closer to that carried out in Ruette, Ehret, and Szmrecsanyi (2016), in which potential variants were added to concepts as a way to capture a more complete picture of variation.

As in the previous chapter, the selection and categorization of concepts was the result of a post-hoc examination of the Clustering by Committee data. Variants were limited to nouns only and were detected by using collocate matrices from the full corpus in a bag-of-words approach with a context window of four in both directions, setting minimum frequency to 400. This resulted in a final list of nearly 10 000 committees. The committees were then ranked in order of similarity score, from highest to lowest. From there, an inspection of the first 1000 generated an initial set of committees. To be able to create and expand lexical fields, synonym sets, or *synsets*, in the Multilingual WordNet (Fernández-Montraveta, Vázquez, and Fellbaum 2008) were consulted and concepts were added if they contained corresponding committees in the full Clustering by Committee. As lexical fields emerged, further concepts were formed by variants found within separate but related committees in the Clustering by Committee, so long as they were found to function as (near) synonyms in random concordance samples from the corpus data. Finally, additional variants were added to concepts based on the authors' own judgements in order to provide a more well-rounded picture, after again cross-checking concordance samples to verify the decisions empirically. Food and clothing lexemes were explicitly excluded from the analysis due to the complexity of their variation and the difficulty of accounting for such a high number of variants. Sorenson (2021) shows how easy it is for concepts within the lexical fields of food and clothing to contain over ten variants in Spanish. And while we did include concepts with significant variation (for example, BELLY; see Table 10.2),

Table 10.2 List of all 142 concepts in their respective lexical fields

LEXICAL FIELD	CONCEPT	NEAR-SYNONYMS	FREQUENCY
JOB, POSITION (23 CONCEPTS)	APPOINTMENT TO A POSITION	*designación/nombramiento*	2835
	ATHLETE	*atleta/deportista*	4411
	BOXER	*púgil/boxeador/pugilista*	1081
	BULLFIGHTER	*matador(a)/torer(a/o)*	1063
	CREDITOR	*prestamista/fiador*	413
	EMPLOYEE	*emplead(a/o)/trabajador(a)*	**19 805**
	EMPLOYER	*empleador(a)/patrón/ patron(a/o)/jef(a/e)*	19 863
	ILLUSTRATOR	*dibujante/ilustrador(a)*	1216
	JOB	*tarea/trabajo*	**20 471**
	LANDOWNER	*hacendado/terrateniente/ latifundista*	1002
	MEDICINE MAN	*chamán(a)/curander(a/o)*	670
	OCCUPATION	*profesión/ocupación*	10 070
	OWNER	*propietari(a/o)/dueñ(a/o)*	17 909
	PAID DRIVER	*conductor(a)/chofer*	7528
	POLICE OFFICER	*agente/policía/vigilante*	**20 021**
	SECURITY ESCORT	*escolta/guardaespaldas*	732
	SERVANT	*criad(a/o)/sirvient(a/e)*	1634
	STAFF	*personal/staff*	16 157
	STUDENT	*alumn(a/o)/estudiante*	**20 033**
	TEACHER (IN A SCHOOL)	*profesor(a)/maestr(a/o)/ educador(a)/docente*	20 196
	WAITER	*camarer(a/o)/meser(a/o)*	1042
	WATCHMAN	*guarda/vigilante/guardián/ guardia/seren(a/o)/guachimán*	7742
	WRITER	*escritor(a)/autor(a)*	**19 929**
SOCIAL RELATIONSHIPS (11 CONCEPTS)	ANCESTOR	*ancestro/antepasado*	2362
	BULLYING	*bullying/acoso*	2409
	DUTY	*deber/responsabilidad /obligación*	**20 072**
	FELLOWSHIP	*compañerismo/camaradería*	548
	FOOLISH ACT	*boludez/pavada*	617
	HOMOSEXUAL	*gay/homosexual*	4290
	HUSBAND	*marido/esposo*	14 319
	LINEAGE	*linaje/abolengo*	2111
	NICKNAME	*apodo/sobrenombre/mote/ remoquete*	1417

Continued

Table 10.2 *Continued*

LEXICAL FIELD	CONCEPT	NEAR-SYNONYMS	FREQUENCY
	PERSONAL FREEDOM	*autonomía/independencia*	10 821
	RELATEDNESS (PEOPLE)	*parentesco/filiación*	1197
RELIGION, MORALITY (11 CONCEPTS)	CATHOLIC PRIEST	*cura/sacerdote*	12 020
	CUSTOM	*costumbre/tradición*	**19 954**
	DEITY	*divinidad/deidad*	2375
	JESUS CHRIST	*cristo/jesucristo/jesús*	**19 712**
	POPE	*papa/pontífice*	15 106
	PRAYER	*rezo/plegaria/oración*	12 910
	PUNISHMENT	*pena/castigo/vergüenza*	**19 891**
	SATAN	*satán/satanás*	3314
	SERMON	*sermón/predicación*	1971
	THE DEVIL	*diablo/demonio*	9156
	WORSHIP	*adoración/culto*	5866
POLITICS (11 CONCEPTS)	ATTEMPT TO INFLUENCE POLITICS	*cabildeo/lobby*	920
	BALLOT	*boleta/papeleta*	1778
	CRIME	*crimen/delito*	**20 039**
	DICTATORSHIP	*dictadura/régimen*	20 026
	FORM (DOCUMENT)	*planilla/formulario*	4218
	MUNICIPAL GOVERNMENT	*municipio/consistorio/ayuntamiento*	4412
	POLITICAL TOTALITARIANISM	*autoritarismo/totalitarismo*	1006
	PRISON	*cárcel/prisión*	12 805
	PROTEST DEMONSTRATION	*manifestación/protesta/movilización*	**19 724**
	PUBLIC DISTURBANCE	*conmoción/disturbio*	1265
	REFERENDUM	*referendo/referéndum*	1149
BUSINESS, ECONOMICS (20 CONCEPTS)	ADVERTISING (INDUSTRY)	*mercadeo/publicidad/mercadotecnia*	14 325
	BANK LOAN	*crédito/préstamo*	17 707
	CASH	*efectivo/cash/contado*	15 454
	CURRENCY	*moneda/divisa*	9906
	DISPLAY WINDOW	*escaparate/vitrina/vidriera*	1251
	ECONOMIC DEREGULATION	*liberalización/desregulación*	551

LEXICAL FIELD	CONCEPT	NEAR-SYNONYMS	FREQUENCY
	FINANCIAL COST	*coste/costo/costa*	**20 075**
	FORMAL COMPLAINT	*queja/reclamación/reclamo*	10 252
	GREED	*avaricia/codicia*	1394
	INHERITANCE (ESTATE)	*herencia/legado*	5745
	POVERTY	*pobreza/miseria*	14 771
	PRODUCT INVENTORY	*stock/mercancía/inventario*	6456
	(FINANCIAL) RETRIBUTION	*retribución/compensación*	2972
	SLOGAN	*slogan/eslogan*	951
	STAND	*puesto/stand/estand*	14 490
	SUPPLY DEPLETION	*agotamiento/desabastecimiento/escasez*	3196
	TV COMMERCIAL	*spot/anuncio/publicidad/aviso/comercial*	**19 896**
	UNEMPLOYMENT	*desempleo/desocupación/paro*	10 112
	WAGE	*salario/sueldo*	14 493
	WEALTH	*patrimonio/riqueza*	14 815
SCIENCE, ANATOMY (19 CONCEPTS)	A COLD	*catarro/resfrío/resfriado*	830
	BELLY	*vientre/barriga/panza/guata/abdomen/tripa/estómago/pancita*	8817
	BIRTH CONTROL	*anticonceptivo/pastilla/píldora*	5028
	BUTTOCKS	*culo/trasero/poto*	2631
	CHIN	*barbilla/mentón*	509
	DENTIST	*dentista/odontólog(a/o)*	1157
	DOCTOR OF MEDICINE	*médic(a/o)/doctor(a)/galen(a/o)/facultativ(a/o)*	**20 000**
	FINDING	*hallazgo/descubrimiento*	8650
	HAIR ON A HUMAN HEAD	*cabello/pelo*	15 200
	HEARTBEAT	*pulsación/latido*	1102
	INVENTION	*invención/invento*	3686
	MEDICAL CHECK-UP	*revisión/chequeo/reconocimiento*	17 708
	MEDICAL OFFICE	*consultorio/clínica/consulta*	14 432
	MEDICATION	*fármaco/medicina/medicamento*	**20 074**

Continued

Table 10.2 *Continued*

LEXICAL FIELD	CONCEPT	NEAR-SYNONYMS	FREQUENCY
	NECK	*cuello/pescuezo*	5471
	SICKNESS	*patología/enfermedad/ afección*	**20 035**
	SKIN	*piel/pellejo*	16 985
	THE FLU	*gripe/gripa/influenza*	3103
	ULTRASOUND (PROCEDURE)	*ecografía/ultrasonido*	1456
HUMAN STATES, EMOTIONS (15 CONCEPTS)	**BEHAVIOUR**	*comportamiento/conducta*	**20 104**
	BRAVERY	*coraje/valentía*	3905
	CHILDHOOD	*infancia/niñez*	8695
	DISCONTENTMENT	*descontento/inconformidad/ disconformidad*	1905
	DISCOURAGEMENT	*desánimo/desaliento*	607
	DRUNKENNESS	*ebriedad/embriaguez*	616
	ELDERLY PERSON	*anciano/viejo*	10 034
	FEAR	*miedo/temor*	20 141
	HAPPINESS	*felicidad/alegría*	20 080
	HUNGER	*apetito/hambre*	10 570
	PASSING	*fallecimiento/deceso/ defunción/muerte*	**19 943**
	PITY	*pena/lástima*	**19 902**
	RESENTMENT	*rencor/resentimiento*	2826
	SENSIBILITY	*sensatez/cordura*	1019
	TANTRUM	*berrinche/rabieta*	570
SPORTS, LEISURE (12 CONCEPTS)	A PLACE TO DRINK ALCOHOL	*pub/taberna/bar*	5222
	ARTISTIC PERFORMANCE	*performance/actuación*	11 003
	BASEBALL	*baseball/béisbol*	678
	CONCERT	*concierto/show*	12 266
	FILMING	*rodaje/filmación*	2006
	HOBBY	*hobby/pasatiempo*	1382
	MOVIE	*film/filme/película /peli*	**20 012**
	PUZZLE (HOBBY)	*puzzle/puzle/rompecabezas*	1186
	REVIEW	*review/reseña*	4700
	TV NETWORK	*network/emisora*	2398
	TV RATINGS	*rating/audiencia*	9296
	VIEWING AUDIENCE	*público/audiencia*	19 949

LEXICAL FIELD	CONCEPT	NEAR-SYNONYMS	FREQUENCY
TECHNOLOGY (9 CONCEPTS)	CLICK	*click/clic*	10 908
	COMPUTER	*ordenador/computador/computadora*	16 626
	EMAIL	*correo/email/e-mail/mail/correo-e*	**20 061**
	ENCRYPTION	*cifrado/encriptación*	467
	DIGITAL FOLDER	*folder (fólder)/carpeta*	3162
	FOLLOWER (SOCIAL MEDIA)	*seguidor/follower*	7978
	HABITUAL INTERNET USER	*cibernauta/internauta*	1212
	HYPERLINK	*link/hipervínculo/enlace*	17 029
	WEBSITE	*site/website/página/sitio*	**20 125**
MISCELLANEOUS (11 CONCEPTS)	A SHOT FROM A FIREARM	*tiro/disparo*	5886
	FORTRESS	*fortaleza/fuerte/alcázar*	7641
	GREAT GROWTH IN SUCCESS ETC.	*auge/boom*	2877
	LEGION	*hueste/legión*	956
	MILITARY STRIKE	*ataque/golpe*	**19 865**
	PERFORMANCE (RESULTS)	*performance/rendimiento*	9439
	PISTOL	*pistola/revólver*	2068
	RIFLE	*fusil /rifle*	1224
	SNACK	*snack/tentempié/refrigerio*	565
	TATTOO	*tattoo/tatuaje/tatú*	1143
	TOPIC	*tema/asunto/materia*	**20 170**

to do so across a large number of concepts would have made a large-scale study such as this rather impractical. Lastly, for the same reasons discussed in the previous chapter, such as computing power, concept size was capped at around 20 000 tokens, meaning that all variants within the concept were reduced proportionally. Table 10.2 shows all initial 142 concepts. The ones that have been reduced in size are shown in bold.

Sorting concepts into lexical fields is hardly a straightforward process (see Geeraerts, Grondelaers, and Bakema 1994: 117–153) and concepts may fit into more than one lexical field (as in the case of BOXER above, which is placed in *job/position* but could arguably be more representative of *sports/leisure,* which is itself a broad category), or a concept's inclusion in a given lexical field may be questionable. Here, the idea is not so much to have a perfect set of lexical fields, but rather locate thematic similarities between the concepts on the basis that they are in some way representative of the field to which they belong. Nevertheless, the lexical field in which a given concept is placed will have repercussions in the disambiguation of the tokens, as in the case of the lexemes *totalitarismo* and *autoritarismo,* whose meaning in this study is restricted to the lexical field of *politics*, but could also be used elsewhere to describe, say, a tyrannical boss or a domineering member of one's household.

Many nouns in Spanish that refer to people have a feminine and masculine form, such as *jefa/jefe* (BOSS) or *odontóloga/odontólogo* (DENTIST). These are tagged in the corpus as separate lemmas, but considered the same variant for the purposes of our study.

Lastly, there were five lemmas that were found in more than one concept. These included *pena* (in both PUNISHMENT and PITY), *vigilante* (POLICE OFFICER and WATCHMAN), *publicidad* (ADVERTISING [INDUSTRY] and TV COMMERCIAL), *performance* (ARTISTIC PERFORMANCE and PERFORMANCE [RESULTS]), and *audiencia* (VIEWING PUBLIC and TV RATINGS). Of the ten concepts these lemmas belong to, five were kept separate at every step of the workflow to avoid the tokens being mixed together, and every process carried out on the rest of the concepts was also done to these five in parallel. They are WATCHMAN, TV COMMERCIAL, PITY, TV RATINGS, PERFORMANCE (RESULTS). They were then incorporated back into the workflow when it came time to calculate the uniformity values.

10.3 Distributional modelling

Once the tokens were sampled, 36 models were created for each concept based on four parameters, as seen in Table 10.3. In line with the token selection process described in Chapter 8 and implemented in Chapter 9, a series of filtering mechanisms were put in place to obtain the onomasiological profiles. Tokens from each model were clustered via HDBSCAN (see Chapter 3) and those considered noise by the algorithm were discarded, along with those pertaining to monolectal clusters (for instance, a cluster made up of only tokens from Peru). As in the previous chapter, those pertaining to monolexical clusters were not removed. Because of how drastically the process for building onomasiological profiles may reduce the sizes of certain concepts, or the diversity of lects therein (due to either a large amount of noise tokens within the models or numerous monolectal clusters, or both), only concepts whose remaining tokens totalled over 10 000 were reduced by 40%. All concepts with less than 10 000 remaining tokens were sampled as is.

As explained in both Chapters 8 and 9, incorporating manual disambiguation into the distributional semantic lectometric workflow allows us to locate those clusters with a significant number of tokens representing the sense in question. For this study we opted for an ambitious disambiguation undertaking to manually annotate nearly 60 000 tokens, almost 10 000 per lect.

Table 10.3 Overview of parameters for the Spanish pluricentricity study

PARAMETER	VALUES
First-order window size	5–5; 10–10; 15–15
First-order part-of-speech filter	common nouns, adjectives, verbs, adverbs; all
Association measure	PPMI (positive PMI)
Association measure filter	ASSOCNO (no user of filtering) SELECTION (filtering context words below threshold) WEIGHT (weighting context words by association strength)
Number of second-order features	FOC (union of first-order context words) 5000 (5000 most frequent items)

The bulk of the tokens, around 46 000 (or roughly 7500 per lect), were annotated externally by an online panel of native Spanish speakers recruited via the market research company Qualtrics, while the remaining 14 000 were annotated by the authors. This specific division of labour was chosen due to the fact that a number of variants are completely or almost completely monosemic, like *hobby* or *pasatiempo*, and therefore could be annotated quickly by the authors without the nuanced, native speaker intuition necessary for the more polysemous variants, like *consulta* (in the sense of MEDICAL OFFICE, as compared to its use for, say, 'a general inquiry' or 'a medical consultation') or *audiencia* (in the sense of VIEWING AUDIENCE, as compared to, say, its use for 'an official hearing'), which were outsourced to Qualtrics. For the internal annotation, a sample size of 20 was taken per variant per lect. Of the initial 142, 29 concepts were fully annotated by the authors.

For each variant to be disambiguated externally, a maximum of 40 tokens per lect (20 per register, if available) were taken at random from the tokens resulting from the HDBSCAN and subsequent filtering process mentioned above. The 7500 tokens in each lect were then mixed and randomly grouped into blocks of mostly 57 or 58 tokens, under the assumption that a person would spend about 30 seconds on each token annotation (a generous assumption in hindsight), as we were aiming for a 25–30 minute task. The survey was generated through a custom script in Python and then uploaded to the Qualtrics platform.

The actual task presented each token in its original context followed by three options to choose from. The sense options were either adopted from the *Diccionario de la Real Academia Española* (*Dictionary of the Spanish Royal Academy*) or from the *Diccionario de americanismos* (*Dictionary of Americanisms*) and usually altered in some way (often to make them less wordy), or written completely by the authors. The in-concept option described the sense denoted by the concept that the variant belonged to in the study. If we take *crédito* from the concept BANK LOAN as an example, the in-concept option read *Dinero que se solicita a una institución financiera con la obligación de devolverlo con interés* 'money that is solicited from a financial institution with the obligation of returning it with interest'. A second option described a sense that could be perceived as closely related to the in-concept sense, but still considered out-of-concept for our purposes. In the case of *crédito*, this out-of-concept option was stated as *Tarjeta electrónica que permite el pago sin dinero en efectivo o el acceso al cajero automático* 'An electronic card that allows for cashless payment or access to a cash machine'. Lastly, a third option, *Otro* (*Other*), was available if there was not enough context or neither options described the sense represented by that occurrence of the token in question. An example can be seen below.

> *1679. Podía saber si un email había sido abierto, en qué momento y sí* (sic) *habían hecho click en algún link. ¡Y llegaron los primeros clientes! De los cuales terminó pagando con una tarjeta de <u>crédito</u> inválida. Pero no importaba, nadie les dijo que este camino sería fácil y por eso, tomaron cada error como parte del aprendizaje para ir mejorando cada vez más.*
>
> 1) *Dinero que se solicita a una institución financiera con la obligación de devolverlo con interés*
> 2) *Tarjeta electrónica que permite el pago sin dinero en efectivo o el acceso al cajero automático*
> 3) *Otro*

Given the scale and language requirements of the disambiguation task, Qualtrics was chosen as a result of the fact that they have recruiting partners in all six of the countries included in our study. A total of around 2400 native Spanish speakers, about 400 per lect, were recruited to participate in the disambiguation task, for which they were compensated according to the Qualtrics' incentive options for that given country. Participants only annotated tokens from the lect corresponding to their geographical location.

The disambiguation task was designed in such a way so that each token would be annotated by three different people and a majority-rules approach would determine the final status of the token as in-concept or out-of-concept. However, due to the set-up of the survey on the Qualtrics platform, coupled with an unexpectedly high drop-out rate, the blocks ended up receiving uneven annotation numbers, with some blocks receiving more than three and others fewer than three. Drop-out rate refers to when participants would leave the task without completing it. Due to the initial design, which was aimed at generating three annotations per token, the counter would increase when a participant would begin annotating a block. However, if a person quit in the middle of the task, the counter was not automatically reset, leading to an imbalance in the blocks. A majority-rules approach was still followed for all tokens with three or more annotations. For tokens with an even number of annotations, the final annotation was dropped when there was no majority, except when there were only two annotations, in which case the authors inspected the first disagreement (on an uncontroversial token) between annotators and selected the participant who provided the more acceptable answer, the uncontroversial token serving as a post-hoc detector.

Lastly, after an initial explanation of the task, participants were asked to provide the province, state, or region in which they were residing, except for the US panel, who were asked to describe what variety of Spanish they felt best represented the variety they spoke, the United States included.

When it comes to online panels, non-cooperation (that is, participants not answering seriously for a variety of reasons) is a persistent issue researchers must always keep in mind (see Häussler and Juzek 2015 for more on participant non-cooperation in linguistics tasks). We implemented two in-survey mechanisms to discourage non-cooperative behaviour. These included, first, a commitment request at the beginning in which the participant recognizes the value of quality answers to our study and commits to providing thoughtful responses, and second, speed checks, which pop up on the screen when the participant is going too fast and ask the participant to slow down and make sure they are reflecting on their answers.

The responses were then analysed in Python via the Qualtrics API, where we were able to filter out poor-quality responses through a set of specific criteria, mainly time, response proportionality, and spamming. Participants who completed in under six minutes were automatically discarded and those who finished in under ten minutes were flagged but not discarded outright. As for response proportionality, a very high number of *Others* or a very low number of in-concept responses, often coupled with a fast response time, indicated that responses were not serious. When at the extremes, these participants were discarded, otherwise they were merely flagged. Participants who were double-flagged for both time and proportionality were also discarded. Finally, spamming was the easiest form of non-cooperation to detect since this took the form of click-throughs, that is, clicking the same option, such as the second one, for the entire task. These participants were also discarded outright. Discarded participants were replaced by Qualtrics, except in the final round of sampling in which the number of discarded participants was low enough (<12) to be able to finalize the annotation task.

Once the in-concept tokens were obtained, the uniformity values could be calculated for the lectal comparisons, following the formulas described in Chapter 7. In order to do so, the clusters in which a majority of tokens correspond to the sense in question (as assigned by HDBSCAN) had to be located in each model by using the disambiguated token set. Similar to the cluster selection procedure in the previous chapter, if a cluster contained a minimum of five annotated tokens and at least 80% of which were in-concept, the cluster was kept and the tokens included in the lectal comparisons. If no clusters within a model met these criteria and none of the clusters were kept, the model was discarded. If none of the models for a given concept were retained, the concept was discarded. This occurred with ten concepts (see Table 10.4), leaving the final total number of concepts for analysis at 132. Lastly, monolexical clusters were included if they met the previously outlined selection criteria.

In line with Speelman, Grondelaers, and Geeraerts (2003), Ruette, Ehret, and Szmrecsanyi (2016), and the discussion in Section 7.1, a significance test was applied to each lectal comparison. If p > 0.05, the uniformity value was set to 1. We opted for the Fisher Exact test as this is ideal for smaller values and limited

Table 10.4 Concepts for which no models were retained

CONCEPT	NEAR-SYNONYMS	FREQUENCY
ARTISTIC PERFORMANCE	*performance/actuación*	11 003
ATTEMPT TO INFLUENCE POLITICS	*cabildeo/lobby*	920
BRAVERY	*coraje/valentía*	3905
FORTRESS	*fortaleza/fuerte/alcázar*	7641
LEGION	*hueste/legion*	956
MILITARY	*ataque/golpe*	19 865
POLICE OFFICER	*agente/policía/vigilante*	20 021
THE DEVIL	*demonio/diablo*	9156
TV NETWORK	*network/emisora*	2398
TV RATINGS	*rating/audiencia*	9296

Table 10.5 Completely uniform concepts resulting from the application of a significance test to the lectal comparisons

CONCEPT	NEAR-SYNONYMS	NO. MODELS	FREQUENCY
CASH	*efectivo/cash/contado*	1	15 454
CUSTOM	*costumbre/tradición*	2	19 954
ECONOMIC DEREGULATION	*liberalización/desregulación*	3	551
DIGITAL FOLDER	*folder (fólder)/carpeta*	3	3162
GREAT GROWTH IN SUCCESS ETC.	*auge/boom*	3	2877
HUNGER	*apetito/hambre*	3	10 570
MEDICAL CHECK-UP	*revisión/chequeo/ reconocimiento*	1	17 708
MEDICINE MAN	*chamán(a)/curander(a/o)*	15	670
OWNER*	*propietari(a/o)/dueñ(a/o)*	36	17 909
SERVANT	*criad(a/o)/sirvient(a/e)*	1	1634
TATTOO	*tattoo/tatuaje/tatú*	36	1143
TV COMMERCIAL	*spot/anuncio/publicidad/ aviso/comercial*	2	19 896
VIEWING AUDIENCE	*público/audiencia*	1	19 949

* This concept was uniform in in-concept tokens, but not in modelled tokens.

data and prevents extremely small lectal comparison sizes from disproportionally influencing the results. Any concepts in which all lectal comparisons resulted in 1 (total uniformity) were removed from the analysis. As noted in Chapter 7, the exclusion of 100% uniform concepts is not implemented in the other chapters, but for this study we opted to do so in order to magnify the distances in the data, given that on such a large scale, the impact of smaller differences is more easily hidden. Therefore, by excluding uniform concepts and increasing the impact of variation, we can gain a clearer glimpse into where such differences occur and what they can reveal about the underlying structure. This step led to the removal of an additional 13 concepts, seen in Table 10.5. Only 39% of the total number of concepts (55 out of 142) retained all 36 models, while 45% of the concepts (64) retained at least half.

10.4 The impact of model retention

To try and understand what effect the number of models has on the results, we chose to analyse the U-values in three ways: first, by including all non-uniform concepts, as long as any number of models were retained (henceforth *AnyMod*); second, by including only those non-uniform concepts with at least half (19+) of the models retained (henceforth *19+Mod*), and third, by including only those non-uniform concepts with at most half (≤18) of the models retained (henceforth *<18Mod*). Besides looking at *how* the concepts modelled, it is also possible to get a sense of *how well* they modelled, at least in comparison to two other sets of tokens: the disambiguated set of in-concept tokens and all the tokens sampled for each concept to create the models without applying any procedure to the data prior to calculating the uniformity values. Thus, we were able to compare three different token sets: *in-concept* (*IC*, from the disambiguation task), *modelled* (*MO*, via the distributional approach), and *all sampled* (*AS*, all sampled tokens, no modelling).

Thanks to the significant size of the *in-concept* set, by calculating uniformity values over tokens that we know represent the sense in question, we can treat this token set as a point of comparison to gauge the outcome of the modelling. However, it must be noted that the *in-concept* results should be interpreted with caution, as beyond the aforementioned filtering mechanisms for non-cooperation, the annotations themselves were not verified by the authors and likely contain errors, the degree to which is not known. Furthermore, due to the nature of the sampling procedure (taking the same number of tokens for each variant, depending, of course, on whether it was annotated externally or internally; 40 versus 20, respectively), the raw in-concept frequencies do not adequately reflect the relationships between the variants within a concept. We therefore applied the formula (9.1) in Chapter 9 in order to reconstruct these relationships.

To prepare the data for visualization, weighted averages of the uniformity values were taken across concepts for each lectal comparison. These averages then form a dissimilarity matrix by subtracting each value from 1. We then multiplied the dissimilarity matrices by a factor of 100 to achieve clearer scaling and generated 2D visualizations via multidimensional scaling in R. Six plots were generated for each of the three subsets based on the number of models retained: *AnyMod* versus *19+Mod* versus *<18Mod*. Three of the six plots examined the *in-concept*, *modelled*, and *all sampled* data from a pan-Hispanic perspective (i.e. including Spain in the analysis), and the other three plots examined the same data from a pan-American perspective (i.e. excluding Spain). An overview of the three groups can be seen in Table 10.6.

We first look at an analysis that includes all non-uniform concepts. The six *AnyMod* plots in which all non-uniform concepts were included as long as they retained a single model can be seen in Figure 10.1, with the pan-Hispanic perspective on the left and the pan-American on the right. Scales are kept consistent across all six plots. Immediately clear from the pan-Hispanic plots on the left is the separation of Argentina and Spain from the other four, and each other. Also apparent in the same three plots is the way the *in-concept* and *modelled* data transform the initial *all sampled* data. The *in-concept* distribution appears to distance the initial relationships in the *all sampled* data, while the *modelled* results appear to do the same for Argentina and Spain, yet the opposite for the other four by bringing them together into two tight clusters, one comprised of USA and Mexico, and the other of Peru and Colombia. When Spain is removed from the visualizations, we can observe a clear delineation along the x-axis across all three sets separating Argentina from the other four American lects. The *modelled* results can be viewed as forming three clusters: Argentina, USA and Mexico, Peru and Colombia, whereas this is less pronounced in the *in-concept* results, and even less so in the *all sampled* results, where Mexico appears much more central among the other three nearest lects. It is important to note that in the R implementation of multidimensional scaling, more significant variation is represented along the x-axis. It therefore makes sense that this is where the most distance is observed between Spain and the other lects in the pan-Hispanic perspective, and then Argentina and the other lects in the pan-American perspective.

Table 10.6 Overview of the concepts by model retention

MODEL GROUPING	TOTAL CONCEPTS	AVG. NUMBER OF MODELS	AVG. CONCEPT SIZE
AnyMod	119	22	8698
19+Mod	61	35	6422
<18Mod	58	8	11 091

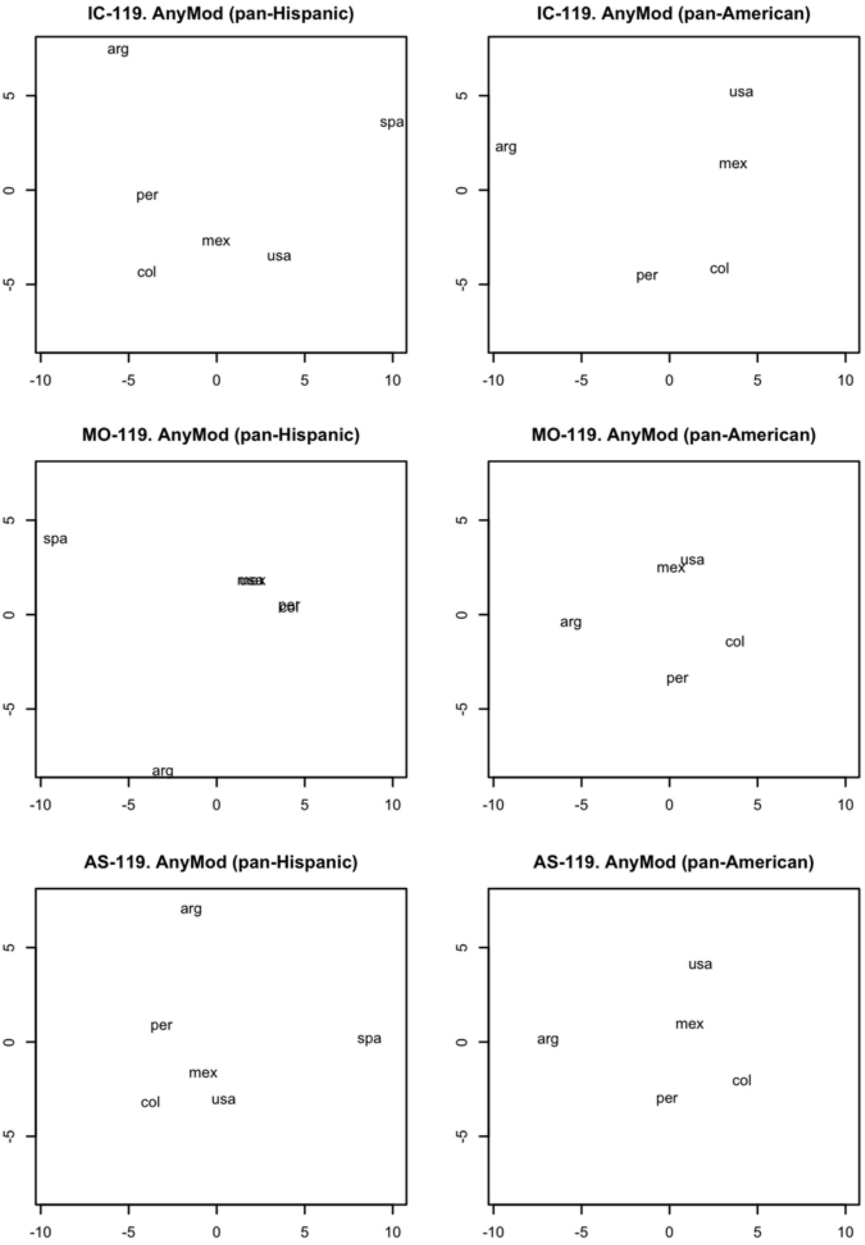


Figure 10.1 Results for 119 concepts. Left-hand panels including Spain, right-hand panels without. First row in-concept tokens, second row modelled tokens, third row all sampled tokens

Table 10.7 U-values for all three *AnyMod* token sets from both a pan-Hispanic and pan-American perspective

PAN-HISPANIC		PAN-AMERICAN	
in-concept	.900	*in-concept*	.915
modelled	.931	*modelled*	.946
all sampled	.921	*all sampled*	.931

We can also better understand the effect Spain has on the plots by looking at the uniformity values used to create the dissimilarity matrices, as shown in Table 10.7. When Spain is excluded from the data, we see a similar increase in both the average *in-concept* and *modelled* U-values, while the average *all sampled* U-value increases by a smaller margin. This also helps explain the differences between the plot pairs, in which we see greater distance between lects in the *in-concept* plots, while the *modelled* sets brings the lects closer together, particularly USA and Mexico.

Next, the results for the 61 concepts included in the *19+Mod* subset can be seen in Figure 10.2. The overall picture from the *19+Mod* plots shows little difference in the distributions between the *in-concept*, *modelled*, and *all sampled* results, respective to their pan-Hispanic or pan-American perspectives. On the pan-Hispanic side, the *modelled* results show a reduction in the distance between Peru and Argentina compared to the other two, while the *in-concept* results show Mexico in a central position relative to Peru, Colombia, and USA. At the same time, Spain is consistently situated at a significant distance from the other lects along the x-axis. On the pan-American side, the *modelled* and *all sampled* plots are practicality identical, while the *in-concept* plot echoes the distribution and relative distances of the pan-American *AS-119* results from the initial *AnyMod* concepts. Argentina continues to show significant distance from USA, Mexico, and Colombia, although less so from Peru, which sits more equidistantly between Argentina and Colombia than in the *AnyMod* plots. The U-values for both perspectives can be seen in Table 10.8. The values show relatively small variation across the three sets. The average *modelled* and *all sampled* U-values are identical, as reflected in the plots above, while the *in-concept* data shows only slightly less uniformity, though the difference is negligible.

Lastly, the six plots in Figure 10.3 show the multidimensional scaling results for those 58 non-uniform concepts that retained 18 or fewer models. Compared to the *19+Mod* results, the *<18Mod* plots show more noticeable variation between the three sets (top to bottom) and the two perspectives (left to right). On the

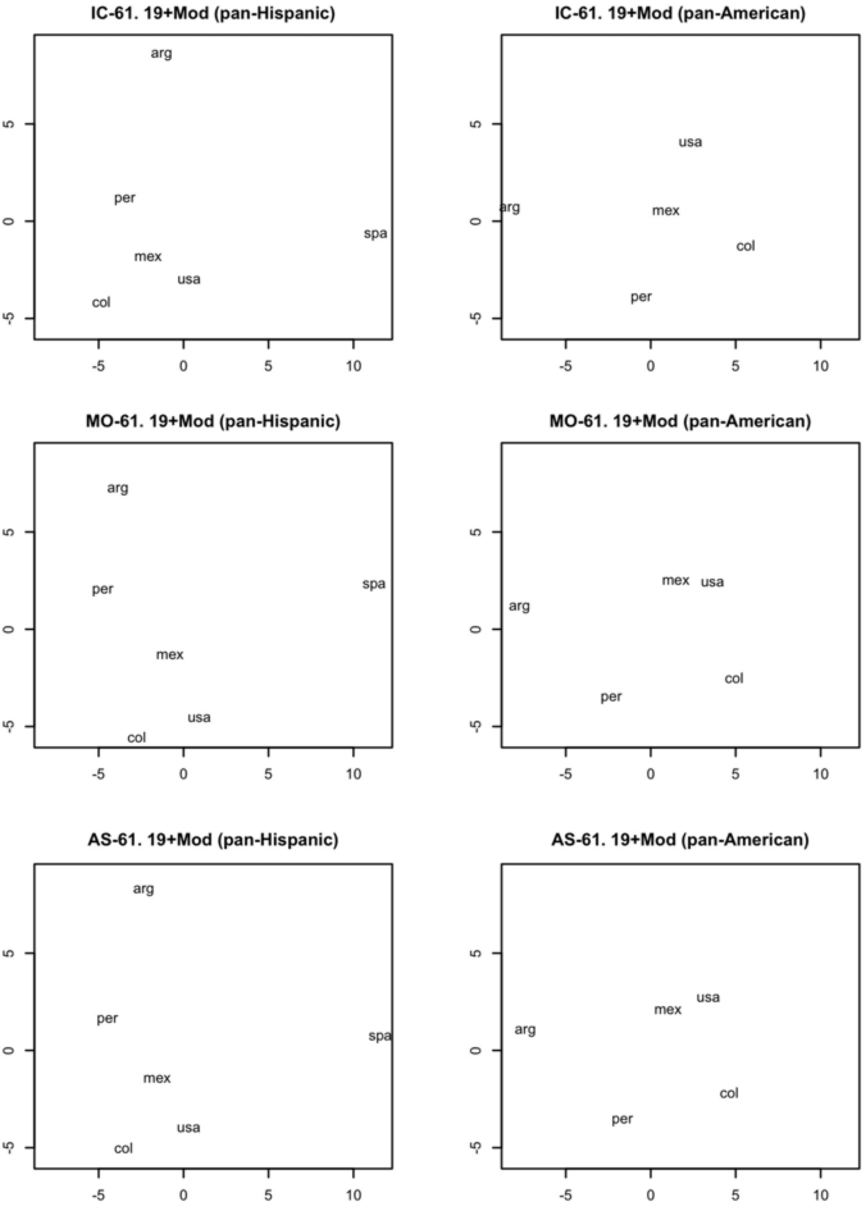


Figure 10.2 Results for 61 concepts. Left-hand panels including Spain, right-hand panels without. First row in-concept tokens, second row modelled tokens, third row all sampled tokens

Table 10.8 U-values for all three *19+Mod* token sets from both a pan-Hispanic and pan-American perspective

PAN-HISPANIC		PAN-AMERICAN	
in-concept	.904	*in-concept*	.921
modelled	.910	*modelled*	.925
all sampled	.910	*all sampled*	.925

pan-Hispanic side, the dissimilarity of Spain and Argentina is readily apparent, particularly in the *in-concept* and *all sampled* data. However, unlike the *all sampled* set, the *in-concept* set shows a differentiation between the other four lects, separating Colombia and Peru into one tight cluster that gravitates towards Argentina on the x-axis, and USA and Mexico in another tight cluster gravitating towards Spain on that same axis. This plot echoes that of the *modelled* pan-Hispanic data from the 119 *AnyMod* concepts. This picture is different here in the *<18Mod* data, as the *modelled* set shows a significantly reduced distance between Argentina and the other American lects, while the distance between the American lects and Spain is maintained. Mexico and USA continue to form one cluster while Colombia and Peru form another. On the pan-American side, both the *in-concept* and *modelled* sets closely resemble their *AnyMod* counterparts. The *in-concept* set again shows much greater distance between all the lects compared to the *modelled* set, although the same clustering of Mexico and USA, and Colombia and Peru can be observed, with Argentina separated at a distance from the others. The *all sampled* data is only informative with respect to Argentina's relationship to the other American lects, as these are all collapsed into a single cluster. Looking at the average U-values (see Table 10.9), the *all sampled* data is indeed the least affected by the removal of Spain. Furthermore, the same relationships between the three sets in the *AnyMod* data re-emerges here, as the *in-concept* set shows dissimilarity when compared to the *all sampled* set, while the *modelled* data shows greater uniformity.

Given the similarities in the plots between the *<18Mod* subset and the full *AnyMod* set of concepts, the former appears to have a larger influence over the latter than the *19+Mod* subset. A comparison of the characteristics of the *AnyMod*, *19+Mod*, and *<18Mod* groups, as seen in Table 10.10, helps to illustrate why this is. Given the notable difference in concept size and considering the fact that the concepts are weighted, it is not surprising that the *<18Mod* subset has a larger influence on the *AnyMod* distributions. Yet, the size of the lectal comparisons for

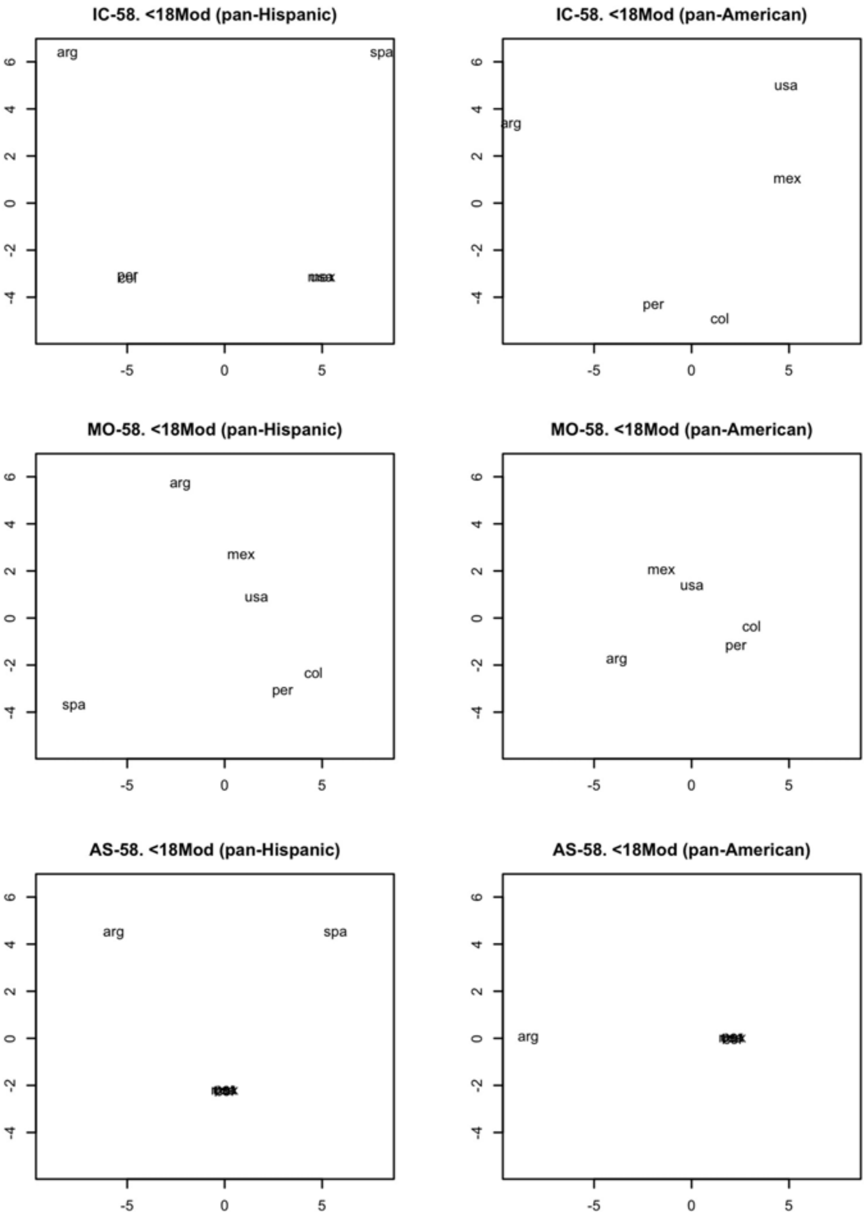


Figure 10.3 Results for 58 concepts. Left-hand panels including Spain, right-hand panels without. First row *in-concept* tokens, second row *modelled* tokens, third row *all sampled* tokens

Table 10.9 U-values for all three *<18Mod* token sets from both a pan-Hispanic and pan-American perspective

PAN-HISPANIC		PAN-AMERICAN	
in-concept	.898	*in-concept*	.912
modelled	.945	*modelled*	.959
all sampled	.927	*all sampled*	.935

Table 10.10 Characteristics of the three groups based on model retention

	AVERAGE LECTAL COMPARISON			AVERAGE	% OF IN- CONCEPT
	in-concept	modelled	all samples	CONCEPT SIZE	TOKENS
AnyMod	2141	908	2899	8698	78%
19+Mod	1880	1649	2141	6422	90%
<18Mod	2418	128	3697	11 091	65%

the *modelled* tokens in the *<18Mod* subset are drastically smaller compared to the others, both within that subset and compared to the other sets of concepts. So even though the *<18Mod* concepts are initially very large, after clustering and pruning (i.e. removing noise tokens and monolectal clusters) we are left with relatively few tokens in the *modelled* set. Finally, it is worth noting the differing proportions of annotated tokens that were labelled as *in-concept*, as the *<18Mod* subset concepts received 25% fewer of such annotations. This certainly also contributed to the significantly smaller sizes of the lectal comparisons for the *modelled* tokens in the *<18Mod* subset.

10.5 The impact of lexical fields

We saw in the previous chapter that lectometric results are not necessarily identical across lexical fields, so to get a better understanding of what is going on lexically within the different data sets, we can break them down by their lexical fields. Figure 10.4 provides a comparison of the distribution of lexical fields by model retention. The three largest lexical fields overall are *Job/Position*, *Science & Anatomy*, and *Business & Economics*. These are also the three that are more represented in the *<18Mod* subset, although the fourth largest group, *Human States & Emotions*, is over three times more frequent in the *19+Mod* subset. Above we saw the overall U-values for the *in-concept*, *modelled*, and *all sampled* sets according to the number of models retained (*AnyMod*, *19+Mod*, and *<18Mod*), but how are the actual lexical fields influencing those values?

Figures 10.5 and 10.6 show the average U-value per lexical field for each set within the different *19+Mod* and *<18Mod* subsets, respectively, and further,

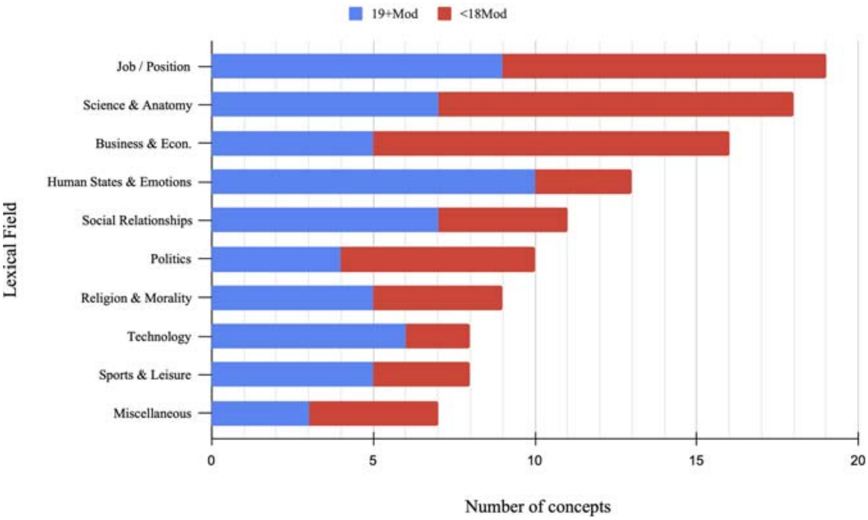


Figure 10.4 The distribution of lexical fields by subset

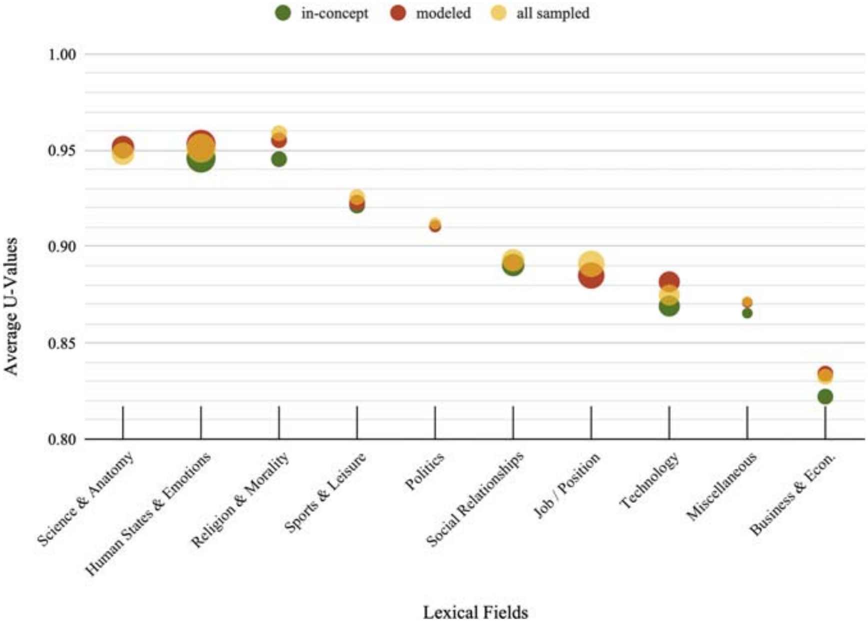


Figure 10.5 Average U-values per lexical field in the *19+Mod* subset

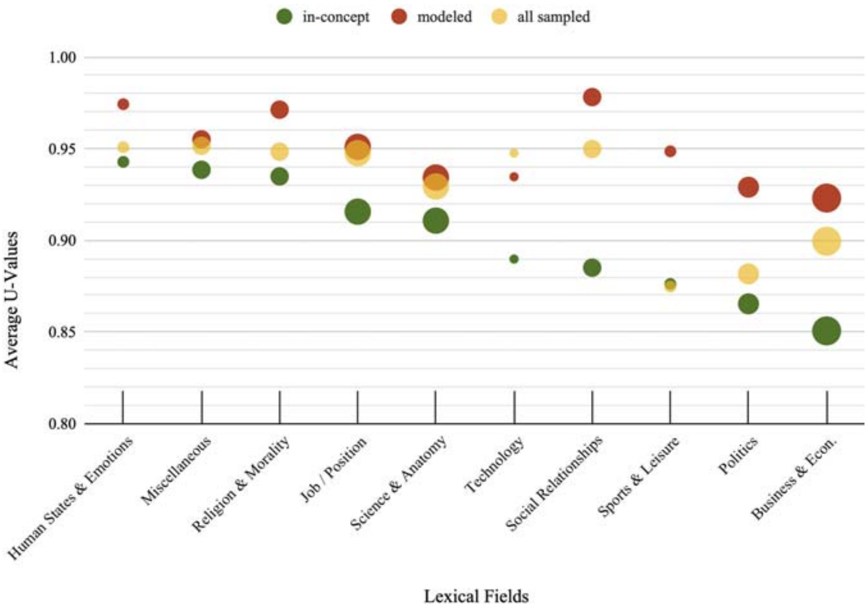


Figure 10.6 Average U-values per lexical field in the *<18Mod* subset

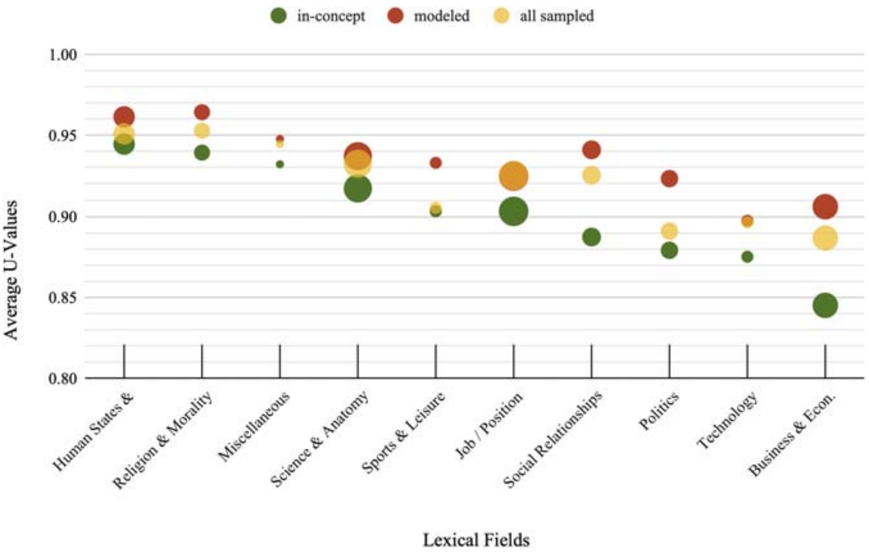


Figure 10.7 Average U-values per lexical field across all 119 concepts

Figure 10.7 shows how they combine into the *AnyMod* plot. Note that dot size is proportional to the frequency of the lexical field within the subset: larger fields are represented by larger dots. As observed in the multidimensional scaling results, the results from the *19+Mod* subset show consistency across all three sets. When breaking this subset down by lexical fields, we can find the largest discrepancies between *Religion & Morality* and *Business & Economics*. Of the three largest fields, only in one, that of *Job/Position*, do we see the *modelled* data coinciding almost perfectly with the *in-concept* data, although this is also observed in two other less represented fields, *Science & Anatomy* and *Social Relationships*.

On the other hand, the *<18Mod* subset displays a high amount of variation between the three sets, particularly once the *in-concept* U-values fall below .90, as it is in these five fields that the largest differences can be observed. In no field does the modelled data coincide with the *in-concept* data. In fact, the widest observable gap between the *modelled* and *in-concept* sets can be found in the field of *Social Relationships*, one of the fields with least amount of distance between these two sets in the *19+Mod* chart. Furthermore, in all but one field, that of *Technology* (which only includes two concepts), does the *all sampled* data come closer to the *in-concept* data than the *modelled* does. Despite the differences, certain similarities are still apparent between the two subsets. In both cases *Business & Economics* has the lowest U-value for the *in-concept* set. The *in-concept* U-values for *Technology* and *Social Relationships* are also similar in both subsets, and *Human States & Emotions* and *Religion & Morality* are among the top three.

10.6 Pluricentricity and the plurality of models

To round off, let us now come back to the two dimensions—descriptive and methodological—that we identified at the outset of the chapter. On the descriptive side, what can our results tell us about the pluricentricity of Spanish? Because one of the central aspects of a pluricentric language is understanding the distance between varieties, it is essential to uncover in what ways or across which dimensions—national, regional, pan-American, or pan-Hispanic—varieties relate to one another. In our analysis, we examined six national varieties from both a pan-Hispanic perspective and a pan-American one. If we assume the *in-concept* as the most reliable of the three sets of tokens, and the *19+Mod* as the most reliable subset of concepts, then a few observations can be made. The first is the obvious distance of Argentina and Spain from the other lects. In the case of Spain, the implications are clear if we recall the discussion at the beginning of this chapter around the criticisms concerning the Real Academia's preference for peninsular Spanish in its elaboration of dictionaries and grammars. If the objective of the institution is to provide a pan-Hispanic view of the language, then it follows that the elaboration of reference works would also reflect these realities at the textual and meta-textual level.

From a strictly pan-American perspective, if Argentina, and Buenos Aires specifically, is indeed one of the major centres of influence as Oesterreicher proposes, then the results suggest that its influence is most likely limited to the River Plate area and only minimally reaches as far up into the Andes as Peru, which shows greater similarity to Colombia and Mexico. The case for Mexico as a major centre of influence is easier to make, as the plot situates it at a relatively similar distance from Peru, Colombia, and the United States. Yet, it is equally plausible that the underlying reason for Mexico's similarity to these three is that of immigration, both to and from Mexico, meaning that the influence may just as well be bidirectional. This would be especially true for the United States, given that Mexicans make up the largest Spanish-speaking immigrant group in the country and Mexican-Americans often move regularly between the two. It is therefore no surprise that the plots show USA in closest proximity to Mexico.

Lastly, what can these results tell us about a pan-Hispanic norm or an international variety of Spanish? If we recall Lebsanft, Mihatsch, and Polzin-Haumann's (2012) notion, mentioned at the beginning of the chapter, of the United States as a kind of testing ground for *international* or *neutral Spanish* where numerous varieties converge to form a type of supranational koiné, there does not appear to be any evidence for such a phenomenon in the results of this study. If the Spanish of the United States is indeed ground zero for a new international variety of Spanish, the results tells us that this variety will display lexical similarity higher to that of Mexico than the others. Interestingly, of all the American lects, it is the one that gravitates the most towards Spain in the pan-Hispanic plot. More likely, however, is that rather than being a simple mishmash of other varieties, the United States actually boasts its own variety. If we consider the plots, the United States relates to the other lects much like Colombia and Peru do, not clinging to a single lect or a nearest neighbour. Perhaps the fact that when asked to identify what country best represents the variety of Spanish they speak, 47% of the US participants for the disambiguation task chose 'USA' indicates that the United States has indeed developed a variety in its own right.

On the methodological side, the analysis demonstrates the advantages—but also the difficulties and limits—of three specific additions to the lectometric methodology as covered in Chapters 8 and 9. First, the chapter showcases the unmistakable convenience of multidimensional scaling for visualizing the relations between a set of lectal datapoints and for comparing lectometric results that are achieved under different modelling conditions. Plots need not have the final word, but even if they are primarily an incentive for going back to the actual figures or for performing additional analyses, their heuristic function is undeniable.

Second, along similar heuristic lines, the chapter illustrates the exploratory usefulness of excluding uniform concepts from the lectometric calculations. By increasing the impact of variation, the lectal structure behind the differences becomes more outspoken—but clearly, the interpretation of the resulting picture will need to keep in mind that it is based on a deliberate exaggeration of the actual similarities and dissimilarities.

Third, building on the discussion in Section 8.4, the present chapter explores the effect of restricting the calculations to concepts that retain a high number of models after the pruning step in the methodological workflow as described in Chapter 8. If we assume that, of the three sets (*in-concept*, *modelled*, and *all sampled*), the manually disambiguated tokens provide the closest approximation to the semantic realities of the corpus data, successful models should uncover similar lectal relationships. In this sense, we can consider the subset of those concepts that retained the (often vast) majority of their models, the *19+Mod* subset, as the more successful one. This shows up from an inspection of the multidimensional scaling plots, but also from Figures 10.5 and 10.6, which reveal more deviation from the *in-concept* point of comparison in the *<18Mod* subset than in the *19+Mod* subset. However, two important caveats need to be formulated. To begin with, restricting the analysis to the *19+Mod* subset changes the distribution of concepts over lexical fields, as can be seen in Figure 10.4. And as we know from Chapter 8, and as is further confirmed by Figures 10.5 and 10.6, uniformity values are sensitive to the lexical field of the concepts in question (and probably also to other concept-related features that await further scrutiny). Restricting the number of concepts under consideration changes the distribution over lexical fields in the dataset, and the distribution over lexical field changes the lectometric results. Accordingly, going for the more secure results (the ones derived from concepts with many models) entails descriptively more limited results: this tradeoff between stability and scope needs to be observed when interpreting the effects of model retention.

As a second remark concerning the effects of model retention, there are indications that differences between the *19+Mod* subset and the *<18Mod* subset may be a side-effect of the degree of polysemy of the items. As a first step, note that of the 61 concepts in the *19+Mod* subset, 26 were fully annotated internally, while in the *<18Mod* subset no concepts were fully annotated internally. But because the choice for internal annotation was based on an overriding likelihood of monosemy, the *19+Mod* subset contains less polysemous items than the *<18Mod* subset. And monosemous items are obviously unlikely to lose models through the removal of out-of-concept clusters. As a second step, this may be related to concept size, given that high-frequency lexical items have a greater tendency towards polysemy, and a random sample will likely contain many examples of different senses. Recall the average concept sizes from Table 10.10 in which a substantial difference of 4600 tokens could be observed. It is possible that this affected the reduced amount of *in-concept* tokens in the *<18Mod* subset, for as we also saw in that same table, of all the annotated tokens, the percentage of those considered in-concept reached 90% for the *19+Mod* subset, compared to 65% from the *<18Mod* concepts. Therefore, the resulting discrepancies in the two subsets make sense when we consider how the process for model retention relied on 80% of the annotated tokens in a cluster to be *in-concept*. A high-frequency, highly polysemous variant will likely lead to a reduced number of *in-concept* tokens, which in turn will reduce the likelihood of a cluster being included in the calculations, and subsequently reduce the models that are actually retained. In short, the observation that the *19+Mod* subset may be biased towards less polysemous items constitutes a further restriction on its scope.

Next to exhibiting the relevance and the restrictions of these three methodological add-ons, the chapter also raises a methodological question of a more fundamental nature. Given that the *all sampled* token set achieved very similar (or sometimes identical) results to the *in-concept* and *modelled* sets, is distributional modelling superfluous? To answer that question, let us consider what factors could explain the fact that by merely sampling the subcorpora and calculating the uniformity values we are able to achieve similar results as a distributional approach. At least the following three factors can be mentioned as potentially contributing to an explanation. First, even though we take an approach that amplifies the differences by ignoring concepts with full uniformity, the overall level of uniformity is very high. This approximates a ceiling effect: if the level is very high, it is difficult to do better. Second, the relatively high degree of monosemy in the data mitigates the impact that disambiguation may have. Third, we cannot exclude that the lexical preferences that we register in our in-concept data also show up in the senses that we discard as out-of-concept. At least in a number of cases, the items in a lexical profile share an out-of-concept polysemy. Take, for example, the concept A SHOT FROM A FIREARM: *tiro/disparo*, derived from the Clustering by Committee algorithm. Both lexemes are highly polysemous and can overlap in their polysemy, such as their usage in sports to refer to a 'shot on goal'. Another example is that of PERSONAL FREEDOM: *autonomía/independencia*, also derived via Clustering by Committee. Both lexemes can also refer to auto-determination in a political sense. In cases such as these, the choice between *tiro* and *disparo* in the 'shot on a goal' case may exhibit the same lectal pattern as in the 'shot from a firearm' case. Similarly, the choice between *autonomía* and *independencia* may be subject to the same lectal preferences in 'political auto-determination' sense as in the 'personal freedom' sense. An *all sampled* analysis that includes the out-of-concept 'shot on a goal' and 'poltical freedom' contexts will then automatically reveal the same lectal configuration as the *in-concept* or *modelled* analysis that focuses on 'shot from a firearm' and 'personal freedom'.

In other words, the correspondence between the undisambiguated *all sampled* and the disambiguated *in-concept* or *modelled* results may hide actual differences—differences that will emerge after disambiguation. For a further example of such a confound, we may return to the results from the *<18Mod* subset. Recall from Figure 10.6 how a number of larger fields such as *Business & Economics* or *Job/Position* yielded similar uniformity values for both the *modelled* tokens and the *all sampled*. Without the distributional modelling, we would have no way of knowing which concepts or fields would be distorting the *all sampled* data, and the overall picture would be skewed by these problematic concepts. This is especially relevant when weighting the concepts, as those in the *<18Mod* set were much larger on average than those in the *19+Mod* set. And while the multidimensional scaling results for the *AnyMod* group do not show a major distortion overall between the *in-concept* and *all sampled*, examining the U-values by lexical fields shows how taking the *all sampled* data at face value still proves unreliable, especially for lexical fields with lower U-values. Taking the case of *Business & Economics*, the *all sampled* set seems unable to adequately deal with lower levels of uniformity. Furthermore, the *all sampled* multidimensional scaling results may indeed lead us to make interpretations not corroborated by the presumably more trustworthy *in-concept* plots, particularly the relationship between Mexico, Peru, and Colombia. The *all sampled* distribution shows Mexico in a much more central position relative to Peru and Colombia, whereas the *in-concept* plot shows Peru situated primarily between Argentina and Colombia with Mexico between USA and Colombia. It follows that a distributional workflow is far from superfluous: in spite of its current imperfections and perhaps intrinsic limitations, a method for semantic analysis remains crucial for lexical variation research.

The bottom line

- The lectometric workflow presented so far may be enriched with a number of methodological add-ons: a multidimensional scaling analysis to visualize patterns of lectal relations, a restriction to non-uniform concepts to bring out differences more clearly, and a focus on richly modelled concepts to increase methodological stability.
- For a pluricentric language with several national varieties like Spanish, the multidimensional scaling analysis is particularly apt to bring out lectal distinctiveness.
- Comparing models to both a disambiguated set of tokens and the full sample from which the *modelled* tokens were drawn may serve as a way to gauge the success of the distributional approach. However, the results need to be analysed cautiously in terms of the interaction between the specifics of the distributional workflow, the annotation process, and the characteristics of the concepts.
- The results for a full unmodelled sample were just as successful at approximating the distances within the *in-concept* tokens as a modelled analysis, but this success cannot be taken at face value, since hidden variation may act as a confound.
- Concepts that retained a majority of their models were the most successful at approximating the results of the annotated *in-concept* tokens, but this success comes at the cost of descriptive scope and a bias towards concepts with a low degree of polysemy.

Conclusions

In this monograph, we explored how token-based distributional semantics can contribute to the study of lexical variation and change. Accordingly, our argumentation unfolded along two interwoven strands of research.

Along the methodological dimension, Chapter 2 informally introduced the basic ideas behind our count-based, token-based distributional approach. Chapter 3 presented the framework from a more technically oriented perspective, with Chapter 4 as an introduction to the visualization tool that we developed to aid the interpretation of distributional models. Chapter 3 in particular pointed to the variability of distributional modelling. At each step of the procedure, alternatives have to be considered, and taken together, these parameter settings and algorithmic choices define a wide variety of potentially relevant models. Importantly, Chapter 5 showed that selecting an optimal solution within that space is not self-evident. Tested against a set of manually analysed lexical items, it appears that no single model setting yields the best results across the board. In response to this methodological underdetermination (which is in line with the introductory theoretical remarks made in Chapter 2) we suggested two strategies: either to select models on the basis of external evidence, or to look for the stability of descriptive results across a range of models. The latter strategy was illustrated in the lectometrical studies in Chapters 9 and 10, the former in Chapter 6, where we relied on existing studies to zoom in on specific model settings.

Along the descriptive dimension of the monograph, Chapter 1 presented a systematic and structured framework for lexical variation research. Within the lexeme-lection-lect triangle, as we called it, we distinguished between a semasiological, an onomasiological, and a lectometric perspective. The semasiological perspective was pursued in Chapter 5, the onomasiological one in Chapter 6, and the lectometric one in Chapters 7 to 10. Throughout, these descriptive chapters were closely intertwined with the methodological strand: Chapter 5 probes the semasiological accuracy of distributional modelling, and as mentioned, the ensuing case studies embody different ways of dealing with the methodological issues deriving from that test. In addition, Chapters 7 and 8 introduce methodological features—quantitative and procedural notions—that are specific to a lectometric workflow. The descriptive lexicological chapters thus mainly serve purposes of illustration, providing case studies of how specific lexical questions can be supported by a distributional analysis.

Lexical Variation and Change. Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang, Oxford University Press. © Dirk Geeraerts, Dirk Speelman, Kris Heylen, Mariana Montes, Stefano De Pascale, Karlien Franco, and Michael Lang (2024). DOI: 10.1093/oso/9780198890676.003.0011

Overall, then, we have tried to define a double-sided research programme for lexical research, by theoretically mapping out its domain of enquiry, and by methodologically specifying and illustrating a number of distributional workflows geared towards specific aspects of the domain. The research programme is a programme indeed: not a finished edifice, but a plan, a prospect, a perspective, a structured set of topics, questions, and methods that need to be developed further. The way we have introduced the programme then also suggests how we believe it should primarily be elaborated.

On the descriptive side, the scope of lexicological research can be expanded well beyond studies straightforwardly modelled on the examples presented in this book. Large-scale diachronic studies as in Chapter 6, or broadly conceived sociolectometric studies as in Chapters 9 and 10 could take lexicology a major step forward, tapping into fundamental aspects of lexical structure: the cognitive principles that underly the conceptual organization and reorganization of the lexicon, and the cultural and social forces that shape the evolution of the vocabulary. But there is more to be done. In particular, we have paid only limited attention to the interaction between the various perspectives that can be identified within the lexeme-lection-lect triangle. We have shown in Chapter 6 how semasiological and onomasiological changes interact, but such combined perspectives can be taken much further. To name just a few, at the interface of semasiological and lectal variation one could ask whether some lects have a more or less polysemous structure than others, while at the interface of lectal and onomasiological variation, the question could be whether on average certain lects have more synonyms for a given concept than others. And do frequent senses (a semasiological feature) on average have more or less synonyms (an onomasiological feature) than less frequent ones, and if there is an effect, is it the same in all language varieties (a lectal feature)? Such questions that arise in the force field of semasiological, onomasiological, and lectal variation are not unknown to existing lexicological research but answering them with a sufficiently large empirical grounding requires a systematic, large-scale investigation of the kind that distributional corpus semantics might offer.

The hedge in the previous sentence is not accidental or rhetorical, though. We have illustrated the potential of a count-based, token-based distributional approach, but we have also pointed at restrictions and limitations. Crucially, the semantic phenomena captured by distributional models are semasiologically speaking diverse (they are not just senses of the kind we think of in a lexicographical context), and there is no model setting that produces optimal results across a wide range of lexical items. It follows that on the methodological side of the research programme, more work needs to be done to scrutinize, refine, and expand the distributional method. We can distinguish three layers in that process, each step broadening the investigation beyond the approach presented in these pages. First, staying within the confines of the methods illustrated in the previous chapters, the effect of parameter settings and workflow variants should be probed more systematically than we have been able to do here. Which settings yield which kind of semantic information with which kinds of words? If a methodical study of the question reveals relevant regularities, we will be able to apply count-based distributional workflows in ways that are tailored to the specific research topics and lexical materials at hand. Second, the count-based framework of the present monograph should be confronted with a deep learning implementation of the Distributional Hypothesis. Given the success and the current popularity of transformer models, such a comparison is an immediate priority. An initial test that we were able to carry out on the dataset of Chapter 5 does not reveal an obvious superiority of transformer models for the classification of the target senses (Sevenants 2022). But this is obviously only scratching the surface; again, a more exhaustive and thoroughgoing comparison is called for. And third, as we specified in Chapter 2 already, distributional corpus semantics should be triangulated with referential and psycho-experimental perspectives on meaning.

Given the open questions with regard to the methodology of meaning research, we may round off with a more distant and theoretical look at the research programme that emerges from the above considerations. What is it that we do when we describe semantic variation? Given that each utterance is different to begin with, what we are doing when we look for different meanings at the level of utterances is to identify equivalence classes, i.e. sets of utterances that are identical or near-identical from the point of view of meaning. But the equivalence classes that we find may be influenced by the method we use and the specific parameters we include in the application of that method. This is not essentially different from the situation in any empirical enquiry, which is necessarily constrained by the particular observational tools and analytic instruments at its disposal, but it does imply that we will want to see clear in the nature and the extent of those constraints. If we don't hypostatize meaning, then, the research programme will have to address a number of questions. First, to what extent do the various methodological approaches correlate with each other? And second, what external phenomena do the resulting classifications of meaning correlate with? If a given method yields a specific set of equivalence classes among utterances, for which other aspects of linguistic behaviour is that specific classification relevant? For instance, it could be that a distributional analysis yields a classification of semantic verb classes that plays a significant role in the choice of auxiliaries with those verbs, whereas a classification resulting from experimental association data has explanatory value in a multimodal analysis of the spontaneous gestures accompanying language. These are imaginary examples, but the point will be evident: in the context of a research programme that doesn't simply take meaning or semantic method for granted, a large-scale exploration of such correspondences should be pursued.

The underlying epistemological question here is whether meaning is a unitary phenomenon. If there is no one-to-one correspondence between the results of the methodological perspectives, we could say that there are aspects of meaning that are identified by method A and others that we measure with method B, but we will have to leave open the possibility that the phenomena under scrutiny will eventually be recognized as different entities altogether, rather than as different aspects of the same phenomenon. To get a better grip on what is at stake here, we may refer to well-known examples from the exact sciences. On the one hand, meaning could be like light, which must be conceived in terms of particles or in terms of waves depending on the kind of experiment with which its properties are investigated. In physical theory light is still, ontologically speaking, thought of as one thing, but under the perspective of different methods, different properties are foregrounded. How those apparently contradictory properties can be reconciled into a single theoretical model of light is a highly technical (and not entirely settled) matter, but that difficulty does not detract from the fact that light is considered a unitary phenomenon. On the other hand, meaning could be like the notion of a vital force, which in large parts of pre-20th century biology was believed to be a unitary principle of life underlying the full spectrum of biological phenomena. Within a reductionist biochemical framework, however, that spectrum is resolved in different, ontologically distinct systems, like metabolism and evolutionary selection, each with its own appropriate methods of investigation. The current situation in lexical semantics could then be described as undecided between these two models: is meaning a unitary phenomenon appearing in different guises according to the perspective we take, or should it be broken down into a complex of distinct phenomena? The question is open for investigation but is far beyond the scope of what we could offer in the previous chapters. To come back to the image that we used at the end of Chapter 2: the fog of meaning has not yet lifted...

Software resources

A number of tools have been developed within the project that have facilitated the various case studies presented in this volume. On the one hand, some of them involve Python 3 and R code used to execute the steps described in Chapter 3: extract information from corpora, create distributional models and apply clustering and other statistical analyses. On the other, visualization tools have been developed within the context of the semasiological workflow for the qualitative examination of token-level models; these are described in Chapter 4. In this section we offer a brief overview of the Python and R tools used for the case studies, with links to the source code, documentation, and tutorials of each tool.

The main Python package is **nephosem** (QLVL 2021), written by Tao Chen with input and testing by the rest of the team members and further refinements by Stefano De Pascale and Mariana Montes. This package provides an array of functionalities covering all the steps described up to Section 3.3 in Chapter 3: from parsing the corpus and generating frequency lists to generating both bag-of-words and dependency-based token-level models and computing distances between the tokens. As such, it lies at the base of all the case studies discussed in the present volume.

The studies in Chapter 5 and Section 6.3 have been further supported by the Python package semasioFlow (Montes 2022) and the R package semcloud (Montes 2021c), both developed by Mariana Montes. These packages are geared towards the generation of multiple models based on combinations of parameter settings and the preparation of the data for the NephoVis visualization tool illustrated in Chapter 4 (see below). On the one hand, **semasioFlow** offers wrappers for efficiently looping across different combinations of parameter settings with nephosem code, covering the first part of the workflow: from parsing the corpus to generating multiple token-level distance matrices. It also already registers the parameter space as well as the context words captured for each token by each model. On the other, the **semcloud** package is meant to process the output run with semasioFlow and generate the data needed by the visualization tool: coordinates from dimensionality reduction, distances between models, HDBSCAN clustering, and counts of context words per cluster.

Schematically, the source code for the packages can be found in the following locations:

nephosem

Language: Python
GitHub repository: qlvl/nephosem
Documentation: https://qlvl.github.io/nephosem/
Tutorial: https://qlvl.github.io/nephosem/tutorials/all-in-one.html

semasioFlow

Language: Python
GitHub repository: qlvl/semasioFlow
Documentation: https://qlvl.github.io/semasioFlow/
Tutorial: https://qlvl.github.io/nephosem/tutorials/createClouds.html

semcloud

Language: R
GitHub repository: qlvl/semcloud
Documentation: https://qlvl.github.io/semcloud/
Tutorial: https://qlvl.github.io/semcloud/articles/processClouds.html

Further, our suite of tools includes:

- the **NephoVis** tool (Sevenants, Montes, and Wielfaert 2022) introduced in Chapter 4, which allows an interactive exploration of semasiological and onomasiological models at different levels. It is demonstrated at https://qlvl.github.io/NephoVis and the source code can be found in https://github.com/qlvl/NephoVis.
- a ShinyApp, also illustrated in Chapter 4, that can be used to explore individual models in detail. It is demonstrated at https://marianamontes.shinyapps.io/Level3/ and the source code can be found in https://github.com/montesmariana/Level3.
- **nephoNeural**, a Python package developed by Anthe Sevenants that parses output from transformer models and prepares it for NephoVis: https://github.com/AntheSevenants/NephoNeural

For the onomasiological and lectometric studies in Chapter 6, 9, and 10, both **Python scripts and R scripts** were developed by Karlien Franco, Stefano De Pascale, and Michael Lang. They can be found at https://github.com/qlvl/NephoSem-BookMaterials. These scripts cover, among other things, the modelling choices for the token-based models, the lectometric calculations, the processing of the Qualtrics questionnaire data used in Chapter 10, and the compilation of the corpora (where applicable). Specific **tutorials** for conducting lectometric studies, along with the other tutorials for the various components of our approach, can be found at https://github.com/qlvl/tutorials.

References

Allwood, Jens (2003). 'Meaning potentials and context: Some consequences for the analysis of variation in meaning', in Hubert Cuyckens, René Dirven, and John Taylor (eds), *Cognitive Linguistic Approaches to Lexical Semantics*. Berlin: Mouton de Gruyter, 29–66.

Ameel, Eef, Barbara Malt, Gert Storms, and Fons Van Assche (2009). 'Semantic convergence in the bilingual lexicon', *Journal of Memory and Language* 60: 270–90.

Amrami, Asaf, and Yoav Goldberg (2019). 'Towards better substitution-based word sense induction', *arXiv:1905.12598 [cs.CL]*.

Anishchanka, Alena, Dirk Speelman, and Dirk Geeraerts (2014). 'Referential meaning in basic and non-basic color terms', in Wendy Anderson, Carole P. Biggam, Carole Hough, and Christian Kay (eds), *Colour Studies: A Broad Spectrum*. Amsterdam: John Benjamins, 323–38.

Anishchanka, Alena, Dirk Speelman, and Dirk Geeraerts (2015a). 'Measuring the diversity of colour naming in advertising', in Victoria Bogushevskaya and Elisabetta Colla (eds), *Thinking Colours: Perception, Translation and Representation*. Edinburgh: Cambridge Scholars Publishing, 45–72.

Anishchanka, Alena, Dirk Speelman, and Dirk Geeraerts (2015b). 'Usage-related variation in the referential range of blue in marketing context', *Functions of Language* 22: 20–43.

Asención-Delaney, Yuly, and Joseph Collentine (2011). 'A multidimensional analysis of a written L2 Spanish corpus', *Applied Linguistics* 32 (3): 299–322.

Auer, Peter (2005). 'Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations', in Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts (eds), *Perspectives on Variation. Sociolinguistic, Historical, Comparative*. Berlin: Mouton de Gruyter, 7–42.

Auer, Peter (2011). 'Dialect vs. standard: a typology of scenarios in Europe', in Bernd Kortmann and Johan van der Auwera (eds), *The Languages and Linguistics of Europe. A comprehensive guide*. Berlin: De Gruyter Mouton, 485–500.

Aurrekoetxea, Gotzon, Aitor Iglesias, Esteve Clua, Iker Usobiaga, and Miquel Salicrú (2020). 'Analysis of transitional areas in dialectology: Approach with fuzzy logic', *Journal of Quantitative Linguistics* 28 (4): 337–55.

Ávila, Raúl (2003). 'La lengua española y sus variantes en los medios de comunicación masiva', in Ávila Raúl, José Antonio Samper, and Hiroto Ueda (eds), *Pautas y pistas en el análisis del léxico hispano(americano)*. Madrid: Iberoamericana Vervuert, 11–26.

Baldinger, Kurt. 1980. *Semantic Theory*. Oxford: Blackwell.

Baldissin, Gioia, Dominik Schlechtweg, and Sabine Schulte im Walde (2022). 'DiaWUG: A dataset for diatopic lexical semantic variation in Spanish', in Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2601–9.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). 'Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors', in Kristina Toutanova and Hua Wu (eds), *Proceedings of the 52nd Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 238–47.

Beal, Joan, and Lourdes Burbano-Elizondo (2012). ‘“All the lads and lasses”: lexical variation in Tyne and Wear’, *English Today* 28: 10–22.

Berlin, Brent (1978). ‘Ethnobiological classification’, in Eleanor Rosch and Barbara B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates, 9–26.

Berlin, Brent (1992). *Ethnobiological Classification. Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton: Princeton University Press.

Bertels, Ann, and Dirk Speelman (2014). ‘Clustering for semantic purposes. Exploration of semantic similarity in a technical corpus’, *Terminology* 20: 279–303.

Bierbach, Mechtild (2000). ‘Spanisch—eine plurizentrische Sprache? Zum Problem von norma culta und Varietät in der hispanophonen Welt’, *Vox Romanica* 59: 143–70.

Biria, Reza, and Ali Bahadorian-Baghbaderani (2016). ‘Cross-cultural analysis of prototypicality norms used by male and female Persian and American speakers’, *Journal of Psycholinguistic Research* 45: 1301–14.

Bostock, Michael, Vadim Ogievetsky, and Jeffrey Heer (2011). ‘D Data-Driven Documents’, *IEEE Transactions on Visualization and Computer Graphics* 17 (12): 2301–9.

Bravo García, Eva (2008). *El Español Internacional—Conceptos, Contextos y Aplicaciones*. Madrid: Arco Libros.

Brugman, Claudia (1988). *The Story of ‘Over’. Polysemy, Semantics and the Structure of the Lexicon*. New York: Garland.

Bullinaria, John A., and Joseph P. Levy (2007). ‘Extracting semantic representations from word co-occurrence statistics: A computational study’, *Behavior Research Methods* 39 (3): 510–26.

Campello, Ricardo J. G. B., Davoud Moulavi, and Joerg Sander (2013). ‘Density-based clustering based on hierarchical density estimates’, in Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (eds), *Advances in Knowledge Discovery and Data Mining* (Lecture Notes in Computer Science). Berlin, Heidelberg: Springer, 160–72.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman (1999). *Readings in Information Visualization: Using Vision to Think*. San Francisco, California: Morgan Kaufmann Publishers.

Casas, Bernardino, Antoni Hernández-Fernández, Neus Català, Ramon Ferrer-i-cancho, and Jaume Baixeries (2019). ‘Polysemy and brevity versus frequency in language’, *Computer Speech and Language* 58: 19–50.

Catlin, Jane-Carol, and Jack Catlin (1972). ‘Intentionality: A source of ambiguity in English?’, *Linguistic Inquiry* 3: 504–8.

Chang, Winston, Joe Cheng, Joseph J. Allaire, Carson Sievert, Barret Schloerke, Yihui Xie, Jeff Allen, Jonathan McPherson, Alan Dipert, and Barbara Borges (2022). *Shiny: Web application framework for R*. Software, available at https://shiny.posit.co/.

Church, Kenneth Ward, and Patrick Hanks (1989). ‘Word association norms, mutual information, and lexicography’, in Julia Hirschberg (ed), *Proceedings of the 27th annual meeting on Association for Computational Linguistic*. Vancouver, British Columbia, Canada: Association for Computational Linguistics, 76–83.

Clyne, Michael G. (ed) (1992). *Pluricentric Languages: Differing Norms in Different Nations*. Berlin: Mouton de Gruyter.

Corpus Hedendaags Nederlands—CHN (2021). Online service, available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-s8. Consulted in July 2022.

Coseriu, Eugenio (1981). 'Los conceptos de "dialect", "nivel" y "estilo de lengua" y el sentido propio de la dialectologia', *Lingüística española actual* 3: 1–32.

Coupland, Nikolas, and Tore Kristiansen (2011). 'SLICE: Critical perspectives on language (de)standardisation', in Tore Kristiansen and Nikolas Coupland (eds), *Standard Languages and Language Standards in a Changing Europe*. Oslo: Novus, 11–35.

Cox, Michael A. A., and Trevor F. Cox (2008). 'Multidimensional scaling', in Chunhouh Chen, Wolfgang Härdle, and Antony Unwin (eds), *Handbook of Data Visualization* (Springer Handbooks of Computational Statistics). Berlin, Heidelberg: Springer, 315–48.

Cruse, D. Alan (1982). 'On lexical ambiguity', *Nottingham Linguistic Circular* 11: 65–80.

Dąbrowska, Ewa, and Dagmar Divjak (eds) (2015). *Handbook of Cognitive Linguistics*. Berlin: De Gruyter Mouton.

Daems, Jocelyne, Kris Heylen, and Dirk Geeraerts (2015). 'Wat dragen we vandaag: een hemd met blazer of een shirt met jasje?', *Taal en Tongval* 67: 307–42.

Daems, Jocelyne, Eline Zenner, and Dirk Geeraerts (2016). 'Lexicale homogeniteit en lexicale voorkeur in de Nederlandse woordenschat van emoties', *Tijdschrift voor Nederlandse Taal- en Letterkunde* 132: 276–319.

Daems, Jocelyne (2022). *Profile-based measures of lexical variation. Four case studies on variation in word choice between Belgian and Netherlandic Dutch*. PhD thesis, University of Leuven.

Davies, Mark (2016). *Corpus del Español: Two billion words, 21 countries (Web/Dialects)*. Provo, Utah: Brigham Young University.

Davies, Mark, and Robert Fuchs (2015). 'Expanding horizons in the study of World Englishes with the 1.9 billion word Global Web-based English Corpus (GloWbE)', *English World-Wide*, *36* (1): 1–28.

De Deyne, Simon, Danielle J. Navarro, Andrew Perfors, Marc Brysbaert, and Gert Storms (2019). 'The "Small World of Words" English word association norms for over 12,000 cue words', *Behavior Research Methods* 51: 987–1006.

Delarue, Steven, and Anne-Sophie Ghyselen (2016). 'Setting the standard. Are teachers better speakers of Standard Dutch?', *Dutch Journal of Applied Linguistics* 5(1): 34–64.

Delobelle, Pieter, Thomas Winters, and Bettina Berendt (2020). 'RobBERT: a Dutch RoBERTa-based language model', in Trevor Cohn, Yulan He, and Yang Liu (eds), *Findings of the Association for Computational Linguistics: EMNLP 2020*. Stroudsburg, PA, USA: Association for Computational Linguistics, 3255–65.

Del Valle, José (2012). 'Panhispanismo e hispanofonía: breve historia de dos ideologías siamesas', *Sociolinguistic Studies* 5 (3): 465–84.

De Pascale, Stefano (2019). *Token-based vector space models as semantic control in lexical lectometry*. PhD thesis, University of Leuven.

Depuydt, Katrien, and Hennie Brugman (2019). 'Turning digitised material into a diachronic corpus: Metadata challenges in the Nederlab Project', in Stefan Pletschacher and Isabel Martínez (eds), *DATeCH2019: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*. New York: Association for Computing Machinery, 169–73.

De Saussure, Ferdinand (1916). *Cours de Linguistique Générale*. Paris: Payot.

De Sutter, Gert (ed) (2017). *De vele gezichten van het Nederlands in Vlaanderen*. Leuven: Acco.

De Sutter, Gert, Dirk Speelman, and Dirk Geeraerts (2005). 'Regionale en stilistische effecten op de woordvolgorde in werkwoordelijke eindgroepen', *Nederlandse taalkunde* 10: 97–128.

De Sutter, Gert, Dirk Speelman, and Dirk Geeraerts (2008). 'Prosodic and syntactic-pragmatic mechanisms of grammatical variation: the impact of a postverbal constituent on the word order in Dutch clause final verb clusters', *International Journal of Corpus Linguistics* 13: 194–224.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). 'BERT: Pre-training of deep bidirectional transformers for language understanding', in Jill Burstein, Christy Doran, and Thamar Solorio (eds), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1)*. Minneapolis, Minnesota: Association for Computational Linguistics, 4171–86.

Divjak, Dagmar (2006). 'Ways of intending: Delineating and structuring near-synonyms', in Stefan Th. Gries and Anatol Stefanowitsch (eds), *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter, 19–56.

Divjak, Dagmar (2010). *Structuring the Lexicon. A Clustered Model for Near-Synonymy*. Berlin: De Gruyter Mouton.

Dollinger, Stefan (2017). '*Take up* #9 as a semantic isogloss on the Canada-US border', *World Englishes* 36: 80–103.

Dunning, Ted (1993). 'Accurate methods for the statistics of surprise and coincidence', *Computational Linguistics* 19 (1): 61–74.

Durkin, Philip (2012). 'Variation in the lexicon: the "Cinderella" of sociolinguistics?', *English Today* 112: 3–9.

Erk, Katrin, and Sebastian Padó (2008). 'A structured vector space model for word meaning in context', in Mirella Lapata and Hwee Tou Ng (eds), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, Hawaii: Association for Computational Linguistics, 897–906.

Escoriza Morera, Luis (2015). 'The influence of the degree of formality on lexical variation in Spanish', *Spanish in Context* 12: 199–220

Evans, Vyvyan (2009). *How Words Mean. Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press.

Evert, Stefan (2004). *The statistics of word cooccurrences. Word pairs and collocations*. PhD thesis, Universität Stuttgart.

Evert, Stefan (2009). 'Corpora and collocations', in Anke Lüdeling and Merja Kytö (eds), *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 1212–48.

Fernández-Montraveta, Ana, Gloria Vázquez, and Christine Fellbaum (2008). 'The Spanish version of WordNet 3.0', in Angelika Storrer, Alexander Geyken, Alexander Siebert, and Kay-Michael Würzner (eds), *Text Resources and Lexical Knowledge*. Berlin: Mouton de Gruyter, 175–82.

Firth, John R. (1957). 'A synopsis of linguistic theory 1930–1955', in John R. Firth (ed), *Studies in Linguistic Analysis*. Oxford: Philological Society, 1–32.

Franco, Karlien, and Dirk Geeraerts (2019). 'Botany meets lexicology: the relationship between experiential salience and lexical diversity', in Janice Fon (ed), *Dimensions of Diffusion and Diversity*. Berlin: De Gruyter Mouton, 115–48.

Franco, Karlien, Dirk Geeraerts, Dirk Speelman, and Roeland Van Hout (2019a). 'Concept characteristics and variation in lexical diversity in two Dutch dialect areas', *Cognitive Linguistics* 30: 205–42.

Franco, Karlien, Dirk Geeraerts, Dirk Speelman, and Roeland Van Hout (2019b). 'Maps, meanings and loanwords: The interaction of geography and semantics in lexical borrowing', *Journal of Linguistic Geography* 7: 1–19.

Franco, Karlien, Mariana Montes, and Kris Heylen (2022). 'Deconstructing destruction: A Cognitive Linguistics perspective on a computational analysis of diachronic change', in Nina Tahmasebi, Syrielle Montariol, Andrey Kutuzov, Simon Hengchen, Haim Dubossarsky, and Lars Borin (eds), *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*. Dublin, Ireland: Association for Computational Linguistics, 23–32.

Franco, Karlien, and Sali A. Tagliamonte (2021). 'Interesting fellow or tough old bird? 3rd person male referents in Ontario', *American Speech* 96: 192–216.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery (2017). 'Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence', *Language Learning* 67: 155–79.

Gabrielatos, Costas, and Paul Baker (2008). 'Fleeing, sneaking, flooding. A corpus analysis of discursive constructions of refugees and asylum seekers in the UK press', *Journal of English Linguistics* 36: 5–38.

Galván Torres, Adriana R. (2021). '*Macho*. The singularity of a mock Spanish item', *Borealis* 10: 63–85.

Geeraerts, Dirk (1985). 'Preponderantieverschillen bij bijna-synoniemen', *De Nieuwe Taalgids* 78: 18–27.

Geeraerts, Dirk (1988). 'Where does prototypicality come from?', in Brygida Rudzka-Ostyn (ed), *Topics in Cognitive Linguistics*. Amsterdam: John Benjamins, 207–29.

Geeraerts, Dirk (1989). 'Prospects and problems of prototype theory', *Linguistics* 27: 587–612.

Geeraerts, Dirk (1993). 'Vagueness's puzzles, polysemy's vagaries', *Cognitive Linguistics* 4: 223–72.

Geeraerts, Dirk (1995). 'Representational formats in Cognitive Semantics', *Folia Linguistica* 39: 21–41.

Geeraerts, Dirk (1997). *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*. Oxford: Clarendon Press.

Geeraerts, Dirk (2003). 'The interaction of metaphor and metonymy in composite expressions', in René Dirven and Ralf Pörings (eds), *Metaphor and metonymy in comparison and contrast*. Berlin; New York, NY: Mouton de Gruyter, 435–66.

Geeraerts, Dirk (2005). 'Lectal variation and empirical data in Cognitive Linguistics', in Francesco Ruiz de Mendoza Ibáñez and Sandra Peña Cervel (eds), *Cognitive Linguistics. Internal Dynamics and Interdisciplinary Interactions*. Berlin: Mouton de Gruyter, 163–89.

Geeraerts, Dirk (2010a). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Geeraerts, Dirk (2010b). 'Schmidt redux: How systematic is the linguistic system if variation is rampant?', in Kasper Boye and Elisabeth Engberg-Pedersen (eds), *Language Usage and Language Structure*. Berlin: De Gruyter Mouton, 237–62.

Geeraerts, Dirk (2015). 'How words and vocabularies change', in John Taylor (ed), *The Oxford Handbook of the Word*. Oxford: Oxford University Press, 416–30.

Geeraerts, Dirk (2016a). 'Entrenchment as onomasiological salience', in Hans-Jörg Schmid (ed), *Entrenchment and the Psychology of Language Learning. How We Reorganize and Adapt Linguistic Knowledge*. Berlin: De Gruyter Mouton. 153–74.

Geeraerts, Dirk (2016b). 'The sociosemiotic commitment', *Cognitive Linguistics* 27: 527–42.

Geeraerts, Dirk (2016c). 'Sense individuation', in Nick Riemer (ed), *The Routledge Handbook of Semantics*. London: Routledge, 233–47.

Geeraerts, Dirk (2017). 'Distributionalism, old and new', in Anastasia Makarova, Stephen M. Dickey, and Dagmar Divjak (eds), *Each Venture a New Beginning. Studies in Honor of Laura A. Janda*. Bloomington, Indiana: Slavica, 29–38.

Geeraerts, Dirk (2018a). *Ten Lectures on Cognitive Sociolinguistics*. Leiden: Brill.

Geeraerts, Dirk (2018b). 'A lectometric definition of lexical destandardization', in Stefan Engelberg, Henning Lobin, Kathrin Steyer, and Sascha Wolfer (eds), *Wortschätze. Dynamik, Muster, Komplexität* (Institut für Deutsche Sprache Jahrbuch 2017). Berlin: De Gruyter Mouton, 233–44.

Geeraerts, Dirk, and Hubert Cuyckens (eds) (2007). *The Oxford Handbook of Cognitive Linguistics*. New York: Oxford University Press.

Geeraerts, Dirk, and Caroline Gevaert (2008). 'Hearts and (angry) minds in Old English', in Farzad Sharifian, René Dirven, Ning Yu, and Susanne Niemeier (eds), *Culture, Body, and Language. Conceptualizations of Internal Body Organs across Cultures and Languages*. Berlin: Mouton de Gruyter, 319–47.

Geeraerts, Dirk, Caroline Gevaert, and Dirk Speelman (2012). 'How "anger" rose. Hypothesis testing in diachronic semantics', in Kathryn Allan and Justyna Robinson (eds), *Current Methods in Historical Semantics*. Berlin: De Gruyter Mouton, 109–32.

Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema (1994). *The Structure of Lexical Variation. Meaning, Naming, and Context*. Berlin: Mouton de Gruyter.

Geeraerts, Dirk, Stefan Grondelaers, and Dirk Speelman (1999). *Convergentie en Divergentie in de Nederlandse Woordenschat: Een Onderzoek naar Kleding- en Voetbaltermen*. Amsterdam: Meertens Instituut.

Geeraerts, Dirk, Gitte Kristiansen, and Yves Peirsman (eds) (2010). *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton.

Geeraerts, Dirk, and Dirk Speelman (2010). 'Heterodox concept features and onomasiological heterogeneity in dialects', in Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman (eds), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton, 23–40.

Geeraerts, Dirk, and Hans Van de Velde (2013). 'Supra-regional characteristics of colloquial Dutch', in Frans Hinskens and Johan Taeldeman (eds), *Language and Space 3. Dutch*. Berlin: De Gruyter Mouton, 532–56.

Gillmann, Melitta (2018). 'Causal inference or conventionalized meaning? A corpus study of the German connector *nachdem* "after" in regional standard varieties', *Folia Linguistica* 52: 483–522.

Glynn, Dylan (2008). 'Lexical fields, grammatical constructions and synonymy: A study in usage-based Cognitive Semantics', in Hans-Jörg Schmid and Sandra Handl (eds), *Cognitive foundations of linguistic usage-patterns*: *Empirical studies*. Berlin: Mouton de Gruyter, 89–118.

Glynn, Dylan (2009). 'Polysemy, syntax, and variation: A usage-based method for Cognitive Semantics', in Vyvyan Evans and Stéphanie Pourcel (eds), *New directions in Cognitive Linguistics*. Amsterdam: John Benjamins, 77–106.

Glynn, Dylan (2014). 'The many uses of *run*: Corpus methods and Socio-Cognitive Semantics', in Dylan Glynn and Justyna Robinson (eds), *Corpus methods for semantics: Quantitative studies in polysemy and synonymy*. Amsterdam: John Benjamins, 117–44.

Glynn, Dylan (2016). 'Quantifying polysemy: Corpus methodology for prototype theory', *Folia Linguistica* 50: 413–47.

Goebl, Hans (2011). 'Dialectometry and quantitative mapping', in Alfred Lameli, Roland Kehrein, and Stefan Rabanus (eds), *Language and Space 2. Language Mapping*. Berlin: De Gruyter Mouton, 433–64.

Gómez Font, Alberto (2012). 'El español global en la prensa del siglo XXI', in Franz Lebsanft, Wiltrud Mihatsch, and Claudia Polzin-Haumann (eds), *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert, 19–26.

Greußlich, Sebastian (2015). 'El pluricentrismo de la cultura lingüística hispánica: política lingüística, los estándares regionales y la cuestión de su codificación', *Lexis* 39 (1): 57–99.

Gries, Stefan Th. (2003). *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement*. London: Continuum Press.

Gries, Stefan Th. (2006). 'Corpus-based methods and cognitive semantics: the many senses of 'to run'', in Stefan Th. Gries and Anatol Stefanowitsch (eds), *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis*. Berlin: Mouton de Gruyter, 57–99.

Gries, Stefan Th. (2010). 'Behavioral profiles: A fine-grained and quantitative approach in corpus-based lexical semantics', *The Mental Lexicon* 5: 323–46.

Gries, Stefan Th. (2013). '50-something years of work on collocations: What is or should be next…', *International Journal of Corpus Linguistics 18*(1): 137–66.

Grieve, Jack (2016). *Regional Variation in Written American English*. Cambridge: Cambridge University Press.

Grieve, Jack, Costanza Asnaghi, and Tom Ruette (2013). 'Site-restricted web searches for data collection in regional dialectology', *American speech* 88: 413–40.

Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo (2019). 'Mapping Lexical Dialect Variation in British English Using Twitter', *Frontiers in Artificial Intelligence* 2: 11.

Grieve, Jack, Andrea Nini, and Diansheng Guo (2018). 'Mapping lexical innovation on American social media', *Journal of English Linguistics* 46: 293–319.

Grön, Leonie, and Ann Bertels (2018). 'Clinical sublanguages: Vocabulary structure and its impact on term weighting', *Terminology* 24: 41–65.

Grondelaers, Stefan, Dirk Speelman, and Dirk Geeraerts (2002). 'Regressing on "er". Statistical analysis of texts and language variation', in Anne Morin and Pascale Sébillot (eds), *6ièmes Journées internationales d'Analyse statistique des Données Textuelles—6th International Conference on Textual Data Statistical Analysis*. Rennes: Institut National de Recherche en Informatique et en Automatique, 335–46.

Grondelaers, Stefan, Dirk Speelman, and Dirk Geeraerts (2008). 'National variation in the use of er "there". Regional and diachronic constraints on cognitive explanations', in Gitte Kristiansen and René Dirven (eds), *Cognitive Sociolinguistics. Language Variation, Cultural Models, Social Systems*. Berlin: Mouton de Gruyter, 153–203.

Grondelaers, Stefan, Katrien Deygers, Hilde Van Aken, Vicky Van den Heede, and Dirk Speelman (2000). 'DigiTaal: het CONDIV-corpus geschreven Nederlands', *Nederlandse taalkunde* 5: 356–63.

Hahsler, Michael, Matthew Piekenbrock, and Derek Doran (2019). 'dbscan: Fast density-based clustering with R', *Journal of Statistical Software* 91 (1): 1–30.

Hahsler, Michael, and Matthew Piekenbrock (2021). *dbscan: Density based clustering of applications with noise (DBSCAN) and related algorithms*. Software, available at https://CRAN.R-project.org/package=dbscan.

Hampton, James A. (2016). 'Categories, prototypes and exemplars', in Nick Riemer (ed), *The Routledge Handbook of Semantics*. London: Routledge, 125–41.

Hanks, Patrick W. (2013). *Lexical Analysis. Norms and Exploitations*. Cambridge, MA: MIT Press.

Harris, Zellig (1946). 'From morpheme to utterance', *Language* 22: 161–83.

Harris, Zellig (1951). *Methods in Structural Linguistics*. Chicago: The University of Chicago Press.

Harris, Zellig (1954). 'Distributional structure', *Word* 10: 146–62.

Häussler, Jana, and Thomas Juzek (2015). 'Detecting and discouraging noncooperative behavior in online experiments using an acceptability judgment task', in Hanna Christ, Daniel Klenovsak, Lukas Sonning, and Valentin Werner (eds), *A blend of MaLT: Selected contributions from the methods and linguistic theories symposium 2015*. Bamberg: University of Bamberg Press, 73–100.

Hawkey, James, and Damien Mooney (2021). 'The ideological construction of legitimacy for pluricentric standards: Occitan and Catalan in France', *Journal of Multilingual and Multicultural Development* 42 (9): 854–68.

Heylen, Kris (2005). 'A quantitative corpus study of German word order variation', in Stephan Kepser and Marga Reis (eds), *Linguistic Evidence: Empirical, Theoretical and Computational Perspectives*. Berlin: Mouton de Gruyter, 241–64.

Heylen, Kris, and Ann Bertels (2016). 'Sémantique distributionnelle en linguistique de corpus', *Langages* 201: 51–64.

Heylen, Kris, and Dirk De Hertog (2015). 'Automatic term extraction', in Hendrik J. Kockaert and Frieda Steurs (eds), *Handbook of Terminology 1*. Amsterdam: John Benjamins, 199–219.

Heylen, Kris, Yves Peirsman, and Dirk Geeraerts (2008). 'Automatic synonymy extraction', in Suzan Verberne, Hans van Halteren, and Peter-Arno Coppen (eds), *Computational linguistics in the Netherlands 2*007. Amsterdam: Rodopi, 101–16.

Heylen, Kris, Yves Peirsman, Dirk Geeraerts, and Dirk Speelman (2008). 'Modelling word similarity: An evaluation of automatic synonymy extraction algorithms', in *Proceedings of the Sixth International Language Resources and Evaluation Conference*. Marrakech: European Language Resources Association, 3243–3249.

Heylen, Kris, Dirk Speelman, and Dirk Geeraerts (2012). 'Looking at word meaning. An interactive visualization of Semantic Vector Spaces for Dutch synsets', in Hannes Wettig, Kirill Reshetnikov, and Roman Yangarber (eds), *Proceedings of the EACL 2012 Joint Workshop of LINGVIS and UNCLH*. Avignon: Association for Computational Linguistics, 16–24.

Heylen, Kris, Thomas Wielfaert, Dirk Speelman, and Dirk Geeraerts (2015). 'Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis', *Lingua* 157: 153–72.

Hilpert, Martin, and David Correia Saavedra (2017). 'Using token-based semantic vector spaces for corpus-linguistic analyses: From practical applications to tests of theoretical claims', *Corpus Linguistics and Linguistic Theory* 16 (2): 393–424.

Hilpert, Martin, and Susanne Flach (2020). 'Disentangling modal meanings with distributional semantics', *Digital Scholarship in the Humanities* 36 (2): 307–21.

Hilte, Lisa, Walter Daelemans, and Reinhild Vandekerckhove (2020). 'Lexical patterns in adolescents' online writing: the impact of age, gender, and education', *Written Communication* 37(3): 365–400.

Hoey, Michael (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

Hoey, Michael (2005). *Lexical Priming. A New Theory of Words and Language*. London: Routledge.

Hothorn, Torsten, Kurt Hornik and Achim Zeileis (2006). 'Unbiased recursive partitioning: A conditional inference framework', *Journal of Computational and Graphical Statistics* 15(3): 651–74.

Hothorn, Torsten and Achim Zeileis (2015). 'partykit: A modular toolkit for recursive partytioning in R', *Journal of Machine Learning Research* 16: 3905–09.

Hothorn, Torsten and Achim Zeileis (2021). *Partykit: A toolkit for recursive partytioning* Software, available at http://partykit.r-forge.r-project.org/partykit/.

Hualde, José Ignacio, Antxon Olarrea, Anna María Escobar, and Catherine E. Travis (2010). *Introducción a la Lingüística Hispánica* (2nd edition). Cambridge: Cambridge University Press.

Impe, Leen, Dirk Geeraerts, and Dirk Speelman (2009). 'Mutual intelligibility of standard and regional Dutch language varieties', *International Journal of Humanities and Arts Computing* 2: 101–17.

Jansegers, Marlies, and Stefan Th. Gries (2017). 'Towards a dynamic behavioral profile: A diachronic study of polysemous *sentir* in Spanish', *Corpus Linguistics and Linguistic Theory* 16 (1): 145–87.

Jaspers, Jürgen, and Sarah Van Hoof (2015). 'Ceci n'est pas une Tussentaal: Evoking standard and vernacular language through mixed Dutch in Flemish telecinematic discourse', *Journal of Germanic Linguistics* 27(1): 1–44.

Jenset, Gard B., and Barbara McGillivray (2017). *Quantitative Historical Linguistics. A Corpus Framework*. Oxford: Oxford University Press.

Johnson, Kathy E. (2001). 'Impact of varying levels of expertise on decisions of category typicality', *Memory and Cognition* 29: 1036–50.

Jurafsky, Daniel, and James H. Martin (2023). *Speech and Language Processing* (3rd edition draft). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf.

Kaufman, Leonard, and Peter J. Rousseeuw (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons.

Kempton, Willet (1981). *The Folk Classification of Ceramics: A Study of Cognitive Prototypes*. New York: Academic Press.

Kiela, Douwe, and Stephen Clark (2014). 'A systematic study of semantic vector space model parameters', in Alexandre Allauzen, Raffaella Bernardi, Edward Grefenstette, Hugo Larochelle, Christopher Manning, and Scott Wen-tau Yih (eds), *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality*. Gothenburg, Sweden: Association for Computational Linguistics, 21–30.

Kiesling, Scott F. (2004). 'Dude', *American Speech* 79: 281–305.

Kilgarriff, Adam (1997). 'I don't believe in word senses', *Computers and the Humanities* 31: 91–113.

Kilgarriff, Adam, and Michael Rundell (2002). 'Lexical profiling software and its lexicographic applications: Case study', in Anna Braasch and Claus Povlsen (eds), *Proceedings of the 10th EURALEX International Congress*. Copenhagen: Center for Sprogteknologi, 807–19.

Kleiber, Georges (1990). *La Sémantique du Prototype. Catégories et Sens Lexical*. Paris: Presses Universitaires de France.

Konopka, Tomasz (2022). *Umap: Uniform manifold approximation and projection*. Software, available at https://github.com/tkonopka/umap.

Koptjevskaja-Tamm, Maria, and Magnus Sahlgren (2014). 'Temperature in the word space: Sense exploration of temperature expressions using word-space modelling', in Benedikt Szmrecsanyi and Bernhard Wälchli (eds), *Aggregating Dialectology, Typology, and Register Analysis*. Berlin: De Gruyter Mouton, 231–67.

Krawczak, Karolina (2014). 'Corpus evidence for the cross-cultural structure of social emotions: Shame, embarrassment, and guilt in English and Polish', *Poznań Studies in Contemporary Linguistics* 50: 441–75.

Krawczak, Karolina, and Dylan Glynn (2015). 'Operationalising mirativity: A usage-based quantitative study on constructional construal in English', *Review of Cognitive Linguistics* 13: 253–82.

Krijthe, Jesse H. (2018). *Rtsne: T-distributed Stochastic Neighbor Embedding using Barnes-Hut implementation*. Software, available at https://github.com/jkrijthe/Rtsne.

Kristiansen, Gitte, and René Dirven (eds) (2008). *Cognitive Sociolinguistics: Language Variation, Cultural Models, Social Systems*. Berlin: Mouton de Gruyter.

Kristiansen, Gitte, Karlien Franco, Stefano De Pascale, Laura Rosseel, and Weiwei Zhang (eds) (2021). *Cognitive Sociolinguistics Revisited*. Berlin: De Gruyter Mouton.

Kristiansen, Tore (2016). 'Contemporary standard language change. Weakening or strengthening?', *Taal en Tongval* 68: 93–118.

Kruskal, J. B. (1964). 'Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis', *Psychometrika* 29 (1): 1–27.

Labov, William (1966). *The Social Stratification of English in New York City*. Washington: Center for Applied Linguistics.

Labov, William (1972). *Sociolinguistic Patterns*. Philadelphia: University of Pennsylvania Press.

Labov, William (1973). 'The boundaries of words and their meanings', in Charles-James Bailey and Roger W. Shuy (eds), *New Ways of Analysing Variation in English*. Washington DC: Georgetown University Press, 340–71.

Lakoff, George (1970). 'A note on vagueness and ambiguity', *Linguistic Inquiry* 1: 357–359.

Langacker, Ronald W. (1991). 'A usage-based model', in Ronald W. Langacker (ed), *Concept, Image, and Symbol. The Cognitive Basis of Grammar*. Berlin: Mouton de Gruyter, 261–88.

Lapesa, Gabriella, and Stefan Evert (2014). 'A large scale evaluation of distributional semantic models: Parameters, interactions and model selection', *Transactions of the Association for Computational Linguistics* 2: 531–46.

Lavandera, Beatriz (1978). 'Where does the sociolinguistic variable stop?', *Language in Society* 7: 171–83.

Lebsanft, Franz (2004): 'Plurizentrische Sprachkultur in der spanischsprachigen Welt', in Alberto Gil, Dietmar Osthus, and Claudia Polzin-Haumann (eds), *Romanische Sprachwissenschaft. Zeugnisse für Vielfalt und Profil eines Faches. Festschrift für Christian Schmitt zum 60. Geburtstag*. Frankfurt am Main: Peter Lang, 205–20.

Lebsanft, Franz (2007). 'Norma pluricéntrica del español y Academias de la Lengua', in Christopher F. Laferl and Bernhard Pöll (eds), *Amerika und die Norm. Literatursprache als Modell?* Tübingen: Niemeyer, 227–46.

Lebsanft, Franz, Wiltrud Mihatsch, and Claudia Polzin-Haumann (2012). *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert.

Leemann, Adrian, Marie-José Kolly, and David Britain (2018). 'The English Dialects App: The creation of a crowdsourced dialect corpus', *Ampersand* 5: 1–17.

Lenci, Alessandro (2018). 'Distributional models of word meaning', *Annual Review of Linguistics* 4: 151–71.

Levshina, Natalia, Dirk Geeraerts, and Dirk Speelman (2013a). 'Mapping constructional spaces: A contrastive analysis of English and Dutch analytic causatives', *Linguistics* 51: 825–54.

Levshina, Natalia, Dirk Geeraerts, and Dirk Speelman (2013b). 'Towards a 3D-Grammar: Interaction of linguistic and extralinguistic factors in the use of Dutch causative constructions', *Journal of Pragmatics* 52: 34–48.

Levshina, Natalia, and Kris Heylen (2014). 'A radically data-driven Construction Grammar: Experiments with Dutch causative constructions', in Ronny Boogaart, Timothy Colleman, and Gijsbert Rutten (eds), *Extending the Scope of Construction Grammar*. Berlin: De Gruyter Mouton, 17–46.

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). 'Improving distributional similarity with lessons learned from word embeddings', *Transactions of the Association for Computational Linguistics* 3: 211–25.

Lin, Dekang, and Patrick Pantel (2002), 'Concept discovery from text', in *COLING 2002: The 19th International Conference on Computational Linguistics*. Taipei, Taiwan: Association for Computational Linguistics, 1–7. https://aclanthology.org/C02–1144.

Lipski, John (1994). *Latin American Spanish*. London: Longman.

Lipski, John (2012). 'Geographical and social varieties of Spanish: An overview', in José Ignacio Hualde, Antxon Olarrea, and Erin O'Rourke (eds), *The Handbook of Hispanic Linguistics*. Malden, MA and Oxford: Wiley-Blackwell, 1–26.

Littlemore, Jeannette, and John R. Taylor (eds) (2014). *The Bloomsbury Companion to Cognitive Linguistics*. London: Bloomsbury.

Loper, Edward, and Steven Bird (2002). 'NLTK: The Natural Language Toolkit', in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*. Philadelphia: Association for Computational Linguistics, 63–70.

Lötscher, Andreas (2017). *Areale Diversität und Sprachwandel im Dialektwortschatz* (Zeitschrift für Dialektologie und Linguistik—Beiheft 169). Stuttgart: Franz Steiner.

Maechler, Martin, Peter Rousseeuw, Anja Struyf, and Mia Hubert (2022). *Cluster: 'Finding groups in data'*. Software, available at https://svn.r-project.org/R-packages/trunk/cluster/.

Magadán, Cecilia, Florencia Rizzo, and Jo Anne Kleifgen (2020). 'Language and territory—Part I', *WORD* 66 (4): 239–54.

Maldonado Cárdenas, Mireya (2012). 'Español como lengua pluricéntrica: Algunas formas ejemplares del español peninsular y del español en América', in Franz Lebsanft, Wiltrud Mihatsch, and Claudia Polzin-Haumann (eds), *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert, 95–122.

Malt, Barbara, and Edward E. Smith (1982). 'The role of familiarity in determining typicality', *Memory and Cognition* 10: 69–75.

Manandhar, Suresh, Ioannis P. Klapaftis, Dmitriy Dligach, and Sameer S. Pradhan (2010). 'SemEval-2010 task 14: Word sense induction and disambiguation', in Katrin Erk and Carlo Strapparava (eds), *Proceedings of the 5th International Workshop on Semantic Evaluation*. Uppsala: Association for Computational Linguistics, 63–68.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mattheier, Klaus J. (1997). 'Über Destandardisierung, Umstandardisierung und Standardisierung in modernen europäischen Standardsprachen', in Klaus J. Mattheier and Edgar Radtke (eds), *Standardisierung und Destandardisierung europäischer Nationalsprachen*. Frankfurt: Lang, 1–9.

McColl Millar, Robert, William Barras, and Lisa Marie Bonnici (2014). *Lexical Variation and Attrition in the Scottish Fishing Communities*. Edinburgh: Edinburgh University Press.

McEnery, Tony and Andrew Hardie (2012). *Corpus Linguistics: Method, Theory and Practice*. Cambridge; New York: Cambridge University Press.

McEnery, Tony, Richard Xiao, and Yukio Tono (2010). *Corpus-based Language Studies: An Advanced Resource Book* (Reprinted). London: Routledge.

McInnes, Leland, John Healy, and James Melville (2020). 'UMAP: Uniform Manifold Approximation and Projection for dimension reduction', *arXiv:1802.03426 [stat.ML]*.

McInnes, Leland, John Healy, and Steve Astels (2016). 'How HDBSCAN Works — hdbscan 0.8.1 documentation'. https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html.

Meillet, Antoine (1906). 'Comment les mots changent de sens', *Année Sociologique* 9: 1–38.

Méndez García de Paredes, Elena (2012). 'Los retos de la codificación normativa del español: cómo conciliar los conceptos de español pluricéntrico y español panhispánico', in Franz Lebsanft, Wiltrud Mihatsch, and Claudia Polzin-Haumann (eds), *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert, 281–312.

Michel, Jean-Baptiste, and Erez Lieberman Aiden (2010). 'Quantitative analysis of culture using millions of digitized books', *Science* 331: 176–82.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013). 'Distributed representations of words and phrases and their compositionality', *Advances in Neural Information Processing Systems* 26: 3111–19.

Mitchell, Jeff, and Mirella Lapata (2008). 'Vector-based models of semantic composition', in Johanna D. Moore, Simone Teufel, James Allan, and Sadaoki Furui (eds), *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics, 236–44.

Montes, Mariana (2021a). *Cloudspotting: visual analytics for distributional semantics*. PhD thesis, University of Leuven.

Montes, Mariana (2021b). 'Modelling meaning granularity of nouns with vector space models. *Papers of the Linguistics Society of Belgium* 15.

Montes, Mariana (2021c). *semcloud 0.1.0*. Software, available at https://doi.org/10.5281/zenodo.5596374.

Montes, Mariana (2022). *semasioFlow 0.1.0*. Software, available at https://doi.org/10.5281/zenodo.5900998

Montes, Mariana, Karlien Franco, and Kris Heylen (2021). 'Indestructible insights. A case study in distributional prototype semantics', in Gitte Kristiansen, Karlien Franco, Stefano De Pascale, Laura Rosseel, and Weiwei Zhang (eds), *Cognitive Sociolinguistics Revisited*. Berlin: De Gruyter Mouton, 251–64.

Nasiruddin, Mohammad (2013). 'A state of the art of word sense induction: A way towards word sense disambiguation for under-resourced languages', in Florian Boudin and Loïc Barrault (eds), *Proceedings of RECITAL 2013*. Les Sables d'Olonne: ATALA, 192–205.

Navigli, Roberto (2012). 'A quick tour of word sense disambiguation, induction and related approaches', in Mária Bieliková, Gerhard Friedrich, Georg Gottlob, Stefan Katzenbeisser, and György Turán (eds), *SOFSEM 2012: Theory and Practice of Computer Science* (Lecture Notes in Computer Science). Berlin; Heidelberg: Springer, 115–29.

Navigli, Roberto, and Daniele Vannella (2013). 'SemEval-2013 task 11: Word sense induction and disambiguation within an end-user application', in Suresh Manandhar and Deniz Yuret (eds), *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013*). Atlanta, Georgia: Association for Computational Linguistics, 193–201.

Nerlich, Brigitte, and Nelya Koteyko (2009). 'Carbon reduction activism in the UK: Lexical creativity and lexical framing in the context of climate change', *Environmental Communication* 3: 206–23.

Norrick, Neal R. (1981). *Semiotic Principles in Semantic Theory*. Amsterdam: John Benjamins.

Nunberg, Geoffrey (1979). 'The non-uniqueness of semantic solutions: polysemy', *Linguistics and Philosophy* 2: 143–84.

Oesterreicher, Wulf (2000). 'Plurizentrische Sprachkultur—der Varietätenraum des Spanischen', *Romanistisches Jahrbuch* 51 (1): 287–318.

Oesterreicher, Wulf (2002). 'El español, lengua pluricéntrica: perspectivas y límites de una autoafirmación lingüística nacional en Hispanoamérica. El caso mexicano', *Lexis* 26 (2): 275–304.

Okabe, Masataka, and Kei Ito (2002). 'Color Universal Design (CUD). How to make figures and presentations that are friendly to Colorblind people', *J* Fly Data Depository for Drosophila researchers*. https://jfly.uni-koeln.de/color/.

Oksanen, Jari, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R.B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazi Evangelista, Rich FitzJohn, Michael Friendly, Brendan Furneaux, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Dan McGlinn, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J.F. Ter Braak, and James Weedon (2022). *Vegan: Community ecology package*. Software, available at https://github.com/vegandevs/vegan.

Ordelman, Roeland J.F., Franciska M.G. de Jong, Adrianus J. van Hessen, and G.H.W. Hondorp (2007). 'TwNC: A multifaceted dutch news corpus', *ELRA Newsletter* 12, no. 3–4.

Oskolkov, Nikolay (2021). 'How exactly UMAP works'. (Online documentation) https://towardsdatascience.com/how-exactly-umap-works-13e3040e1668 (consulted on 7 May, 2021)

Padó, Sebastian, and Mirella Lapata (2007). 'Dependency-based construction of semantic space models', *Computational Linguistics* 33 (2): 161–99.

Pantel, Patrick (2003). *Clustering-By-Committee*. PhD thesis, University of Alberta.

Pantel, Patrick, and Dekang Lin (2002a). 'Discovering word senses from text', in Osmar R. Zaïane, Randy Goebel, David Hand, Daniel Keim, and Raymond Ng (eds), *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '02)*. New York: Association for Computing Machinery, 613–19.

Pantel, Patrick, and Dekang Lin (2002b). 'Efficiently clustering documents with committees', in Mitsuru Ishizuka and Abdul Sattar (eds), *PRICAI 2002: Trends in Artificial Intelligence (7th Pacific Rim International Conference on Artificial Intelligence, Tokyo, Japan, August 18–22, 2002. Proceedings)*. Berlin, Heidelberg: Springer, 424–33.

Parodi, Claudia (2001). 'Contacto de dialectos y lenguas en el Nuevo Mundo: La vernacularización del español en América', *International Journal of the Sociology of Language*, 33–53.

Partington, Alan (1998). *Patterns and Meanings. Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

Paul, Hermann (1920). *Prinzipien der Sprachgeschichte* (5th edition). Halle: Max Niemeyer Verlag.

Peirsman, Yves, Simon De Deyne, Kris Heylen, and Dirk Geeraerts (2008). 'The construction and evaluation of word space models', in Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias (eds), *Proceedings of the Sixth International Language Resources and Evaluation Conference*. Marrakech: European Language Resources Association, 3082–88.

Peirsman, Yves, and Dirk Geeraerts (2009). 'Predicting strong associations on the basis of corpus data', in Joakim Nivre and Claire Gardent (eds), *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Athens: Association for Computational Linguistics, 648–56.

Peirsman, Yves, Dirk Geeraerts, and Dirk Speelman (2010). 'The automatic identification of lexical variation between language varieties', *Natural Language Engineering* 16: 469–91.

Peirsman, Yves, Dirk Geeraerts, and Dirk Speelman (2015). 'The corpus-based identification of cross-lectal synonyms in pluricentric languages', *International Journal of Corpus Linguistics* 20: 54–80.

Peirsman, Yves, Kris Heylen, and Dirk Geeraerts (2008). 'Size matters: Tight and loose context definitions in English word space models', in Marco Baroni, Stefan Evert, and Alessandro Lenci (eds), *Proceedings of the ESSLLI Workshop on Distributional Lexical Semantics. Bridging the gap between semantic theory and computational simulations*. Hamburg: ESSLLI, 34–41.

Peirsman, Yves, Kris Heylen, and Dirk Geeraerts (2010). 'Applying word space models to sociolinguistics. Religion names before and after 9/11', in Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman (eds), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton, 111–37.

Perek, Florent (2016). 'Using distributional semantics to study syntactic productivity in diachrony: A case study', *Linguistics* 54 (1): 149–88.

Perek, Florent (2018). 'Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis', *Corpus Linguistics and Linguistic Theory* 14 (1): 65–97.

Pettersson-Traba, Daniela (2022). *The Development of the Concept of 'Smell' in American English. A Usage-based View of Near-Synonymy*. Berlin: De Gruyter Mouton.

Pickl, Simon (2013). 'Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry', *Literary and Linguistic Computing* 28: 63–81.

Plank, Barbara (2022). 'The "problem" of human label variation: On ground truth in data, modeling and evaluation', *arXiv.2211.02570 [cs.CL]*.

Plevoets, Koen, Dirk Speelman, and Dirk Geeraerts (2007). 'A corpus-based study of modern colloquial Flemish', in Stephan Elspaß, Nils Langer, Joachim Scharloth, and Wim Vandenbussche (eds), *Germanic Language Histories 'from Below'*. Berlin: Mouton de Gruyter, 179–88.

Pöll, Bernhard (2012). 'Situaciones pluricéntricas en comparación: el español frente a otras lenguas pluricéntricas', in Franz Lebsanft, Wiltrud Mihatsch, and Claudia Polzin-Haumann (eds), *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert, 29–45.

Pustejovsky, James (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.

QLVL (2021). *Nephosem 0.1.0*. Software, available at https://doi.org/10.5281/zenodo.5710426.

Quine, Willard V.O. (1960). *Word and Object*. Cambridge, MA: MIT Press.

R Core Team (2022). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Software, available at https://www.R-project.org/.

Robinson, Justyna A. (2010). '*Awesome* insights into semantic variation', in Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman (eds), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton, 85–110.

Robinson, Justyna A. (2012). 'A gay paper: Why should sociolinguistics bother with semantics?', *English Today* 28: 38–54.

Rosch, Eleanor, and Carolyn B. Mervis (1975). 'Family resemblances: Studies in the internal structure of categories', *Cognitive Psychology* 7: 573–605.

Rosch, Eleanor (1975). 'Cognitive reference points', *Cognitive Psychology* 7: 532–47.

Rosch, Eleanor (1978). 'Principles of categorization', in Eleanor Rosch and Barbara B. Lloyd (eds), *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum Associates, 27–48.

Rosenberg, Andrew, and Julia Hirschberg (2007). 'V-measure: A conditional entropy-based external cluster evaluation measure', in Jason Eisner (ed), *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, Prague: Association for Computational Linguistics, 410–20.

Rosseel, Laura, Karlien Franco, and Melanie Röthlisberger (2020). 'Extending the scope of lectometry', *Zeitschrift für Dialektologie und Linguistik* 87 (2): 131–43.

Rosseel, Laura, Dirk Speelman, and Dirk Geeraerts (2018). 'Measuring language attitudes using the Personalized Implicit Association Test: A case study on regional varieties of Dutch in Belgium', *Journal of Linguistic Geography* 6: 20–39.

Rosseel, Laura, Dirk Speelman, and Dirk Geeraerts (2019a). 'Measuring language attitudes in context: Exploring the potential of the Personalized Implicit Association Test', *Language in Society* 48: 1–33.

Rosseel, Laura, Dirk Speelman, and Dirk Geeraerts (2019b). 'The relational responding task (RRT): A novel approach to measuring social meaning of language variation', *Linguistics Vanguard* 5: 1–13.

Ruette, Tom (2012). *Aggregating lexical variation: Towards large-scale lexical lectometry*. PhD thesis, University of Leuven.

Ruette, Tom, Katharina Ehret, and Benedikt Szmrecsanyi (2016). 'A lectometric analysis of aggregated lexical variation in written Standard English with Semantic Vector Space models', *International Journal of Corpus Linguistics* 21 (1): 48–79.

Ruette, Tom, Dirk Geeraerts, Yves Peirsman, and Dirk Speelman (2014). 'Semantic weighting mechanisms in scalable lexical sociolectometry', in Benedikt Szmrecsanyi and Bernhard Wälchli (eds), *Aggregating Dialectology, Typology, and Register. Analysis Linguistic Variation in Text and Speech*. Berlin: De Gruyter Mouton, 205–30.

Ruette, Tom, and Dirk Speelman (2014). 'Transparent aggregation of variables with individual differences scaling', *Literary and Linguistic Computing* 29: 89–106.

Ruette, Tom, Dirk Speelman, and Dirk Geeraerts (2011). 'Measuring the lexical distance between registers in national variaties of Dutch', in Augusto Soares Da Silva, Amadeu Torres, and Miguel Gonçalves (eds), *Pluricentric Languages: Linguistic Variation and Sociocognitive Dimensions*. Braga: Aletheia, Publicações da Faculdade de Filosofia da Universidade Católica Portuguesa, 541–54.

Ruette, Tom, Dirk Speelman, and Dirk Geeraerts (2014). 'Lexical variation in aggregate perspective' in Augusto Soares da Silva (ed), *Pluricentricity. Language Variation and Sociocognitive Dimensions*. Berlin: De Gruyter Mouton, 103–26.

Ruhl, Charles (1989). *On Monosemy. A Study in Linguistic Semantics*. Albany: State University of New York Press.

Sahlgren, Magnus (2008). 'The distributional hypothesis', *Italian Journal of Linguistics* 20 (1): 33–53.

Schlechtweg, Dominik, Nina Tahmasebi, Simon Hengchen, Haim Dubossarsky, and Barbara McGillivray (2021). 'DWUG: A large resource of diachronic word usage graphs in four languages', in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana: Association for Computational Linguistics, 7079–91.

Schmid, Hans-Jörg (2010). 'Does frequency in text instantiate entrenchment in the cognitive system?' in Dylan Glynn and Kerstin Fischer (eds), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin: De Gruyter Mouton, 101–33.

Schütze, Hinrich (1998). 'Automatic word sense discrimination', *Computational Linguistics* 24 (1): 97–123.

Serigos, Jacqueline (2017). 'Using distributional semantics in loanword research: A concept-based approach to quantifying semantic specificity of anglicisms in Spanish', *International Journal of Bilingualism* 21 (5): 521–40.

Sevenants, Anthe (2022). *Into the vector space with BERT* (QLVL internship report). MA thesis, University of Leuven.

Sevenants, Anthe, Mariana Montes, and Thomas Wielfaert. (2022). *NephoVis 1.1.0*. Software, available at https://doi.org/10.5281/zenodo.6629350.

Shneiderman, Ben (1996). 'The eyes have it: a task by data type taxonomy for information visualization', in Pieter van Zee, Margaret Burnett, and Maureen Chesire (eds), *Proceedings 1996 IEEE Symposium on Visual Languages*. Boulder, CO: IEEE Computer Society, 336–43.

Sievert, Carson, Chris Parmer, Toby Hocking, Scott Chamberlain, Karthik Ram, Marianne Corvellec, and Pedro Despouy (2021). *Plotly: Create interactive web graphics via plotly.js*. Software, available at https://CRAN.R-project.org/package=plotly.

Sinclair, John M. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John M. (1996). 'The search for units of meaning', *Textus* 9: 75–106.

Sinclair, John M. (1998). 'The lexical item', in Edda Weigand (ed), *Contrastive lexical semantics*. Amsterdam: Benjamins, 1–24.

Sinclair, John M. and Patrick Hanks (1987). *Collins Cobuild English Language Dictionary*. London: Collins.

Smilkov, Daniel, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg (2016). 'Embedding projector: Interactive visualization and interpretation of embeddings', *arXiv:1611.05469 [stat,ML]*.

Soares da Silva, Augusto (2010). 'Measuring and parameterizing lexical convergence and divergence between European and Brazilian Portuguese', in Dirk Geeraerts, Gitte Kristiansen, and Yves Peirsman (eds), *Advances in Cognitive Sociolinguistics*. Berlin: De Gruyter Mouton, 41–83.

Soares da Silva, Augusto (2014). 'The pluricentricity of Portuguese: A sociolectometrical approach to divergence between European and Brazilian Portuguese', in Augusto Soares da Silva (ed), *Pluricentricity. Language variation and sociocognitive dimensions*. Berlin: De Gruyter Mouton, 143–88.

Sorenson, Travis D. (2021). *The Dialects of Spanish: A Lexical Introduction*. Cambridge: Cambridge University Press.

Sousa, Xulio, and Francisco Dubert García (2020). 'Measuring language contact in geographical space: Spanish loanwords in Galician', *Zeitschrift Für Dialektologie Und Linguistik* 87 (2): 285–306.

Speelman, Dirk (2021). 'Profiles visiting Procrustes', in Gitte Kristiansen, Karlien Franco, Stefano De Pascale, Laura Rosseel, and Weiwei Zhang (eds), *Cognitive Sociolinguistics Revisited*. Berlin: De Gruyter Mouton, 139–52.

Speelman, Dirk, and Dirk Geeraerts (2009a). 'Causes for causatives: the case of Dutch "doen" and "laten", in Ted Sanders and Eve Sweetser (eds), *Causal Categories in Discourse and Cognition*. Berlin: Mouton de Gruyter, 173–204.

Speelman, Dirk, and Dirk Geeraerts (2009b). 'The role of concept characteristics in lexical dialectometry', *International Journal of Humanities and Arts Computing* 2: 221–42.

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts (2003). 'Profile-based linguistic uniformity as a generic method for comparing language varieties', *Computers and the Humanities* 37: 317–37.

Speelman, Dirk, Stefan Grondelaers, and Dirk Geeraerts (2006). 'A profile-based calculation of region and register variation: the synchronic and diachronic status of the two main national varieties of Dutch', in Andrew Wilson, Dawn Archer, and Paul Rayson (eds), *Corpus Linguistics around the World*. Amsterdam: Rodopi, 195–202.

Speelman, Dirk and Kris Heylen (2017). 'From dialectometry to semantics', in Martijn Wieling, Martin Kroon, Gertjan van Noord, and Gosse Bouma (eds), *From Semantics to Dialectometry. Festschrift in Honor of John Nerbonne*. Groningen: College Publications, 325–34.

Speelman, Dirk, Leen Impe, and Dirk Geeraerts (2014). 'Phonetic distance and intelligibility in Dutch', in Augusto Soares da Silva (ed), *Pluricentricity. Language Variation and Sociocognitive Dimensions*. Berlin: De Gruyter Mouton, 227–41.

Speelman, Dirk, Adriaan Spruyt, Leen Impe, and Dirk Geeraerts (2013). 'Language attitudes revisited: auditory affective priming', *Journal of Pragmatics* 52: 83–92.

Stefanowitsch, Anatol (2010). 'Empirical cognitive semantics: Some thoughts', in Dylan Glynn and Kerstin Fischer (eds), *Quantitative Methods in Cognitive Semantics: Corpus-Driven Approaches*. Berlin: De Gruyter Mouton, 355–80.

Stubbs, Michael (2002). *Words and Phrases. Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, Michael (2009). 'Memorial article: John Sinclair (1933)', *Applied Linguistics 30*(1): 115–37.

Stukken, Loes, Steven Verheyen, and Gert Storms (2013). 'Representation and criterion differences between men and women in semantic categorization', in Markus Knauff, Michael Pauen, Natalie Sebanz, and Ipke Wachsmuth (eds), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 3474–79.

Süselbeck, Kirsten (2012). 'Las relaciones institucionales entre las Academias de la Lengua Española y su colaboración en la elaboración de la norma lingüística de 1950 hasta hoy', in Franz Lebsanft, Wiltrud Mihatsch, and Claudia Polzin-Haumann (eds), *El español, ¿desde las variedades a la lengua pluricéntrica?* Madrid: Iberoamericana Vervuert, 257–80.

Swanenberg, Jos (2001). 'Brabantse etnozoologische termen in semantisch perspectief', *Taal en Tongval* 53: 63–82.

Taylor, John R. (1989). *Linguistic Categorization. Prototypes in Linguistic Theory*. Oxford: Clarendon Press.

Taylor, John R. (1992). 'How many meanings does a word have?', *Stellenbosch Papers in Linguistics* 25: 133–68.

Tahmasebi, Nina, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen (eds) (2021). *Computational Approaches to Semantic Change*. Berlin: Language Science Press.

Thompson, Robert W. (1992). 'Spanish as a Pluricentric Language', in Michael Clyne (ed), *Pluricentric Languages. Differing Norms in Different Nations*. Berlin: Mouton de Gruyter, 45–70.

Tuggy, David (1993). 'Ambiguity, polysemy, and vagueness', *Cognitive Linguistics* 4: 273–90.

Tummers, José, Dirk Speelman, and Dirk Geeraerts (2004). 'Quantifying semantic effects. The impact of lexical collocations on the inflectional variation of Dutch attributive adjectives', in Gérald Purnelle, Cédrick Fairon, and Anne Dister (eds), *Le poids des mots. Actes des 7es Journées internationales d'Analyse statistique des Données Textuelles*. Louvain-la-Neuve: Presses Universitaires de Louvain, 1079–88.

Tummers, José, Dirk Speelman, and Dirk Geeraerts (2005). 'Inflectional variation in Belgian and Netherlandic Dutch: A usage-based account of the adjectival inflection', in Nicole Delbecque, Johan van der Auwera, and Dirk Geeraerts (eds), *Perspectives on Variation. Sociolinguistic, Historical, Comparative*. Berlin: Mouton de Gruyter, 93–110.

Tummers, José, Dirk Speelman, Kris Heylen, and Dirk Geeraerts (2015). 'Lectal constraining of lexical collocations. How a word's company is influenced by the usage settings', *Constructions and Frames* 7: 1–46.

Turney, Peter D., and Patrick Pantel (2010). 'From frequency to meaning: Vector space models of semantics', *Journal of Artificial Intelligence Research* 37: 141–88.

Ureña, P. Henríquez (1921). 'Observaciones sobre el español en América', *Revista de filología española* 8: 357–90.

Van de Cruys, Tim (2021). 'Classifying Northern and Southern Dutch', lecture presented at the *European Language Grid workshop on Resources for Luxemburgish and Flemish* (8 July 2021).

Van de Cruys, Tim, and Marianna Apidianaki (2011). 'Latent semantic word sense induction and disambiguation' in Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds), *ACL HLT 2011–49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, OR: Association for Computational Linguistics, 1476–85.

Van de Cruys, Tim, Thierry Poibeau, and Anna Korhonen (2013). 'A tensor-based factorization model of semantic compositionality', in Colin Cherry and Matt Post (eds), *Proceedings of NAACL 2013*. Atlanta, GA: Association for Computational Linguistics, 1142–51.

van der Maaten, Laurens J.P. (2014). 'Accelerating t-SNE using tree-based algorithms', *Journal of Machine Learning Research* 15: 3221–45.

van der Maaten, Laurens J.P., and G.E. Hinton (2008). 'Visualizing high-dimensional data using t-SNE', *Journal of Machine Learning Research* 9: 2579–605.

Van Gijsel, Sofie, Dirk Speelman, and Dirk Geeraerts (2008). 'Style shifting in commercials', *Journal of Pragmatics* 40: 205–26.

Vankrunkelsven, Hendrik, Lara Vankelecom, Gert Storms, Simon De Deyne, and Wouter Voorspoels (2021). 'Guessing words. A comparison of tekst corpus and word association models as the basis for the mental lexicon', in Gitte Kristiansen, Karlien Franco, Stefano De Pascale, Laura Rosseel, and Weiwei Zhang (eds), *Cognitive Sociolinguistics Revisited*. Berlin: De Gruyter Mouton, 572–83.

Vankrunkelsven, Hendrik, Steven Verheyen, Gert Storms, and Simon De Deyne (2018). 'Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models', *Journal of Cognition* 1 (1): 45, 1–14.

van Noord, Gertjan (2006). 'At Last Parsing Is Now Operational', in Piet Mertens, Cédrick Fairon, Anne Dister, and Patrick Watrin (eds), *Actes de la 13ème conférence sur le Traitement Automatique des Langues Naturelles*. Leuven: ATALA, 20–42.

Verheyen, Steven, Eef Ameel, and Gert Storms (2011). 'Overextensions that extend into adolescence: Insights from a threshold model of categorization' in Laura Carlson, Christoph Hölscher, and Thomas F. Shipley (eds), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society, 2000–5.

Wattenberg, Martin, Fernanda Viégas, and Ian Johnson (2016). 'How to Use t-SNE Effectively', *Distill* 1: 10. http://doi.org/10.23915/distill.00002

Weeds, Julie, David Weir, and Diana McCarthy (2004). 'Characterising measures of lexical distributional similarity', in *COLING 2004: Proceedings of the 20th international conference on Computational Linguistics*. Geneva: COLING, 1015–21.

White, Anne, Gert Storms, Barbara Malt, and Steven Verheyen (2018). 'Mind the generation gap: Differences between young and old in everyday lexical categories', *Journal of Memory and Language* 98: 12–25.

Wiechmann, Daniel (2008). 'On the computation of collostruction strength', *Corpus Linguistics and Linguistic Theory* 4 (2): 253–90.

Wielfaert, Thomas, Kris Heylen, Dirk Speelman, and Dirk Geeraerts (2019). 'Visual analytics for parameter tuning of Semantic Vector Space models', in Miriam Butt, Annette Hautli-Janisz and Verena Lyding (eds), *LingVis: Visual analytics for linguistics*. Stanford, CA: CSLI Publications, 215–45.

Wieling, Martijn, Simonetta Montemagni, John Nerbonne, and Harald Baayen (2014). 'Lexical differences between Tuscan dialects and Standard Italian: Accounting for geographic and sociodemographic variation using Generalized Additive Mixed Modeling', *Language* 90: 669–92.

Wieling, Martijn, and John Nerbonne (2015). 'Advances in dialectometry', *Annual Review of Linguistics* 1: 243–64.

Wierzbicka, Anna (1985). *Lexicography and Conceptual Analysis*. Ann Arbor, MI: Karoma.

Wright, Laura, and Christopher Langmuir (2019). 'Interpreting Charles Lamb's 'Neat-Bound Books'', *Studia Anglica Posnaniensia* 54: 157–77.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik (2008). 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics* 17(2): 492–514.

Zenner, Eline, Dirk Speelman, and Dirk Geeraerts (2012). 'Cognitive Sociolinguistics meets loanword research: Measuring variation in the success of anglicisms in Dutch', *Cognitive Linguistics* 23: 749–92.

Zenner, Eline, Dirk Speelman, and Dirk Geeraerts (2013). 'Macro and micro perspectives on the distribution of English in Dutch. A quantitative usage-based analysis of job ads', *Linguistics* 51: 1019–64.

Zenner, Eline, Dirk Speelman, and Dirk Geeraerts (2014). 'Core vocabulary, borrowability, and entrenchment: A usage-based onomasiological approach', *Diachronica* 31: 74–105.

Zenner, Eline, Dirk Speelman, and Dirk Geeraerts (2015). 'A sociolinguistic analysis of borrowing in weak contact situations: English loanwords and phrases in expressive utterances in a Dutch reality TV show', *International Journal of Bilingualism* 19: 333–46.

Zenner, Eline, and Dorien Van De Mieroop (2021). 'The (near) absence of English in Flemish dinner table conversations', *Applied Linguistics Review* 12(2): 299–330.

Zhang, Weiwei, Dirk Geeraerts, and Dirk Speelman (2015). 'Visualizing onomasiological change: Diachronic variation in metonymic patterns for WOMAN in Chinese', *Cognitive Linguistics* 26: 289–330.

Zhang, Weiwei (2016). *Variation in Metonymy. Cross-linguistic, Historical and Lectal Perspectives*. Berlin: De Gruyter Mouton.

Zimmermann, Klaus (2001). 'Interculturalidad y contacto de lenguas: condiciones de la influencia mutua de las lenguas amerindias con el español', in Klaus Zimmermann, and Thomas Stolz (eds), *Lo propio y lo ajeno en las lenguas austronésicas amerindias: procesos interculturales en el contacto de lenguas indígenas con el español en el Pacífico e Hispanoamérica*. Madrid: Iberoamericano Vervuert, 17–34.

Zimmermann, Klaus (2008). 'La invención de la norma y del estándar para limitar la variación lingüística y su cuestionamiento actual en términos de pluricentrismo (mundo hispánico)', in Jürgen Erfurt, and Gabriele Budach (eds), *Standardisation et déstandardisation: le francais et l'espagnol au XXe siècle./Estandarización y desestandarización: el francés y el español en el siglo XX*. Frankfurt am Main: Peter Lang, 187–207.

Zipf, George Kingsley (1945). 'The meaning-frequency relationship of words', *The Journal of General Psychology* 33 (2): 251–56.

Zwicky, Arnold and Jerry Sadock (1975). 'Ambiguity tests and how to fail them', in John Kimball (ed), *Syntax and Semantics 4*. New York: Academic Press, 1–36.

Index