HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE, VOLUME 2

Data Science, Statistical Modelling, and Machine Learning Methods

Edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu and Lars Lyberg

First published 2022

ISBN: 978-0-367-45780-8 (hbk)
ISBN: 978-1-032-07770-3 (pbk)
ISBN: 978-1-003-02524-5 (ebk)

21

FROM FREQUENCY COUNTS TO CONTEXTUALIZED WORD EMBEDDINGS

The Saussurean turn in automatic content analysis

Gregor Wiedemann and Cornelia Fedtke

DOI: 10.4324/9781003025245-25

The funder for this chapter is Johannes Kepler University Linz, Institute of Sociology

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

21

FROM FREQUENCY COUNTS TO CONTEXTUALIZED WORD EMBEDDINGS

The Saussurean turn in automatic content analysis

Gregor Wiedemann and Cornelia Fedtke

1 Introduction

Text, the written representation of human thought and communication in natural language, has been a major source of data for social science research since its early beginnings. While quantitative approaches seek to make certain contents measurable, for example through word counts or reliable categorization (coding) of longer text sequences, qualitative social researchers put more emphasis on systematic ways to generate a deep understanding of social phenomena from text. For the latter, several qualitative research methods such as qualitative content analysis (Mayring, 2010), grounded theory methodology (Glaser & Strauss, 2005), and (critical) discourse analysis (Foucault, 1982) have been developed. Although their methodological foundations differ widely, both currents of empirical research need to rely to some extent on the interpretation of text data against the background of its context. At the latest with the global expansion of the internet in the digital era and the emergence of social networks, the huge mass of text data poses a significant problem to empirical research relying on human interpretation. For their studies, social scientists have access to newspaper texts representing public media discourse, web documents from companies, parties, or NGO websites, political documents from legislative processes such as parliamentary protocols, bills and corresponding press releases, and for some years now micro-posts and user comments from social media. Computational support is inevitable even to process samples of such document volumes that could easily comprise millions of documents.

Although automatic methods of content analysis and text mining already have been employed in social and political science for decades, for a long time their application was restricted to rather simplistic methods of quantifying word usage (Wiedemann, 2013). Especially qualitatively oriented researchers remained skeptical about computational methods since they do not seem to contribute to a deeper understanding of expressed contents (ibid.).

Fortunately, natural language processing (NLP), a subfield of computer science and artificial intelligence (AI) research, made remarkable progress in recent years improving computational capabilities to represent semantics. The rapid evolvement in neural networks for so-called *deep learning* with big data currently revolutionizes the way natural language is processed in computer science. Of particular interest is the representation of meaning in the form of high-dimensional,

 DOI: 10.4324/9781003025245-25

dense vectors, so-called *embeddings*, which we will discuss in more detail in this article. For applied sciences like empirical social research, this new development bears a huge potential to address concerns about the suitability of previously existing automatic content analysis approaches. However, the earlier development also has shown that the transfer and adaptation of state-of-the-art models from computer science into social science applications can take quite a long time. For instance, it took about 15 years for the machine learning approach of topic modeling from the publication of a seminal paper (Blei et al., 2003) to enter the standard toolbox of the social sciences (Maier et al., 2018). Other NLP methods with an even longer history such as supervised text classification with logistic regression or support vector machines are still rarely used (Wiedemann, 2019).

Grimmer and Stewart (2013) argue in a widely cited review article about automatic content analysis, "all quantitative models of language are wrong, but some are useful" (p. 269f). Their usefulness should be evaluated regarding their ability to perform social science tasks such as document categorization, the discovery of new useful categorization schemes and the measuring of theoretically relevant quantities from large collections. Regarding this claim, it is interesting to reflect on the evolution of computational models for text throughout the last decades, especially on their capability to represent certain semantics. It is still true for each human-written text that its "data generation process [. . . remains] a mystery – even to linguists" (ibid.) and, hence, we cannot know whether some newer computational model is more correct than its predecessor. However, we can evaluate the fit between computational models and basic linguistic theories to assess their suitability to perform social science tasks. The hypothesis underlying this chapter is that computational models of semantics eventually converge towards basic linguistic models of semantics and, thus, comply much better with basic theoretical assumptions of many of today's text-based empirical research methodologies. Especially the compliance of the new embedding-based models with the methodological assumptions underlying social research may pave the way for exciting new research opportunities. In this light, we strive to answer two questions: (1) how do models of computational semantics relate to linguistic structuralism as the provenance of many modern text research methods, and (2) how may the recent advancements impact automatic content analysis in the social sciences?

We will start our argumentation in section 2 with a brief discussion of theoretical foundations of linguistic structuralism by Ferdinand de Saussure along with an illustrative example about the semantic change of the term "Goldstück" (English: gold piece). This example was selected because it vividly demonstrates the rapid evolvement of its meaning from a positively connoted term to a hate speech expression in German social media communication. The fourth section then introduces a recent turning point in natural language processing: models of latent space semantics and embedding semantics that, as we argue, to some extent replicate the linguistic turn of the humanities and social sciences on a technical level. This argument, again, is illustrated with the empirical example of the term "Goldstücke" to reveal the potentials and limitations of the individual computational methods. Eventually, as we conclude in the final section, computational semantics 'after the Saussurean turn' enables semantic text search and context-sensitive automatic coding, which offers exciting opportunities for the integration of qualitative and quantitative empirical social research.

2 Linguistic Semantics

2.1 Linguistic structuralism

Due to his book *Cours de linguistique générale* (1916), Ferdinand de Saussure is considered the founder of linguistic structuralism. With his works, Saussure paved the way for a new, systematic

analysis of language. Decades later, his theory became the decisive preparatory groundwork for the *linguistic turn* in the social sciences and humanities. The linguistic turn describes a paradigmatic shift towards methodological research "approaches that assume that knowledge is structured by language and other sign systems" (Wrana, 2014, p. 247, own translation). This basic assumption significantly influenced qualitatively oriented empirical social research, and especially (poststructuralist) discourse research in the second half of the 20th century. Saussure's structural model presents language as an interrelation between individual elements that are organized by differences. In the social sciences, these interrelationships became a subject of research as a proxy to study social reality (cp. Posselt & Flatscher, 2018, p. 211; Angermüller, 2007, p. 161). Besides the central assumption of language as a system of differences, two further conceptual differentiations are introduced that are essential for our considerations – the differentiation of language into *langue* and *parole*, and the linguistic sign as a composition of the *signifier* and the *signified*.

The differentiation of language into langue and parole is the starting point of Saussure's theory and essential to structuralist language analysis. Langue describes the system of language and its rules as a social institution, while parole refers to the act of speaking, that is, the concrete use of language. Those two mutually depend on each other because the language system only exists when language is actually used, and conversely, language use can only lead to successful communication if it follows the rules of a commonly shared language system. Saussure, thus, specifically states that langue is either social or does not exist (cf. Saussure, 2001, p. 20). However, according to Saussure, only langue is systematic and therefore suitable as a subject of research. Pierre Bourdieu sharply criticizes this privilege of langue over parole, pointing to Saussure's neglect of not just the usage of language but also of its historicity, materiality, and social embedding (Bourdieu, 2003).

The conceptualization of language as a system of linguistic signs, in which each element does not a priori have a fixed and objective meaning but rather receives this meaning only through the difference to other elements, represents a radical break with essentialist traditions (cf. Posselt & Flatscher, 2018, p. 207). The rejection of signs as 'positive' elements leads to the core idea of structuralism, which states that the position of a sign in the structured language system is relevant to determine its meaning. Referring to the game of chess, Saussure illustrates this idea: "The value of the individual figures depends on their respective position on the chessboard, just as in language each element has its value through its positional relationship to other elements" (Saussure, 2001, p. 105; own translation).

As in chess, there are also rules in language when it comes to determining the meaning of signs. Hereby, the guiding principle is the distinction between syntagmatic and paradigmatic relationships. Syntagmatic relationships exist between successive elements of a text sequence, for example the sentence "He paid five gold pieces". Due to their sequential order and specific relative positions in a sentence, the individual words in their role as linguistic signs are given a certain syntactic function, such as subject, predicate, and object, as well as a certain semantic function, such as the fact that gold pieces can be used by someone to pay. A paradigmatic relationship exists, in contrast, between elements of different sentences that are functionally interchangeable while retaining the same context. The sentence "He paid five gold coins", for example, together with the previous example, establishes a paradigmatic relationship between gold pieces and gold coins as a means of payment (cf. Saussure, 2001, p. 147f.).

But what are signs anyway? For Saussure, a sign is composed of two sides and arises in the link of a concept with a sound-image (Saussure, 2001, p. 78). The sound-image, also called signifier, is primarily the psychological representation of a word (cf. Saussure, 2001, p. 77). The signified, on the other hand, is the hearer's impression of that sound, also called the concept. In

the course of the linguistic turn, this definition has been extended to "describe the signifier as the form that the sign takes and the signified as the concept to which it refers" (Chandler, 2007, p. 18). This interpretation includes the acoustic representation of the spoken word as well as its written representation as a sequence of characters of an alphabet into the notion of the signifier, which was important especially for the methodological foundations of empirical social research with text. An essential aspect is that precisely this connection between signifier and signified is arbitrary, though neither random nor necessary (cf. Saussure, 2001, p. 79). Instead, the connection is socially constructed and, thus, not fixed. Consequently, signs can change their meaning, with the result that language cannot transport meaning unambiguously. The assignment of meaning therefore always arises through the act of interpretation against the background of a language system as well as the context of an utterance.

Following empirical research methodologies rooted in structuralism, one needs to take these assumptions into account when social reality is to be examined based on language data, whether as a mental concept, spoken language, or written text. In this respect, automatic content analysis faces a severe problem because it can only observe the surface of language, that is, the signifier level. For a computer, it is impossible to interpret what meaning is actually to be conveyed in a given text. However, a human interpretation of the linguistic surface form can be encoded in a semantic model by a process of operationalization of the signified, that is, the concept level as the actual study objective. Therefore, it is crucial for the support of empirical social research to determine to what extent computer algorithms are able to represent ambiguous surface structures together with their linguistic contexts to disambiguate their (interpreted) meaning.

2.2 Levels of semantics

Given that signifier and signified are only arbitrarily linked in the language system (langue), the context of their use (parole) is decisive in order to be able to deduce the meaning of a sign. Especially for qualitative text analyses, it is important to look at statements in their respective context to validly interpret them. Thereby, context can refer to very different things. In general, this can be a historical linguistic context, which corresponds roughly to what Saussure describes

Table 21.1 Three levels of semantics for "Goldstücke" (Eng. translation in square brackets)

Semantic level	*Example*
Word (1)	***Goldstücke***
	[gold pieces]
Sentence (2)	*Sind doch alles* ***Goldstücke****, diese Menschen, die wir geschenkt bekommen haben.*
	[They are all **pieces of gold**, these people that we have been given as gifts.]
Discourse (3)	*Das darf doch gar nicht sein. Sind doch alles* ***Goldstücke****, diese Menschen, die wir geschenkt bekommen haben. Eines steht immer mehr fest – Wenn Europa und insbesondere Deutschland bereits vor zwei Jahren eine effektive Grenzsicherung betrieben hätte, wäre unsere Gesellschaft nicht noch um 50 Milliarden € reicher, sondern auch viel sicherer.*
	[This is impossible. They are all **pieces of gold**, these people that we have been given as gifts. One thing is becoming increasingly clear – if Europe, and Germany in particular, had already implemented effective border control two years ago, our society would not only be €50 billion richer but also much safer.]
	(Anonymized Facebook user, 28 April 2017, commenting on a German news article from DER SPIEGEL about a suspected Islamist terrorist)

with langue as the actual subject of synchronic linguistics. More specifically, context can also refer to a surrounding text unit or a contemporary social debate.

To systematize the origin of meaning, linguists generally distinguish between word and sentence meaning while discourse theory further introduces the notion of discourse meaning (Pêcheux et al., 1995, p. 102). The meaning of a sign can change, therefore, depending on the context level from which it is looked at. This can be illustrated vividly with the example of the German term "Goldstücke". In Table 21.1, an example is given for each of the three levels. On the level of the word (1) we only see the string "Goldstücke". Due to the current German language system, most readers will associate this string with the concept of shiny metallic coins, which were used as means of payment in earlier times. Another meaning would be a metaphorical compliment to another person. If the individual word is considered in a concrete sentence, not only the sentence itself carries a certain meaning. The sentence context also modifies the meaning of the individual word. In example (2), one can deduce that the term is intended to designate a group of people. According to the conventional interpretation of the language system, persons with this designation are attributed to a positive esteem. However, this interpretation changes drastically when a broader context is considered. At the level of discourse meaning (3), what has been said before must be included in the interpretation. This can refer to actually present text, previous statements in a dialogical exchange of texts, or even a broad social debate that has been conducted across many participants and texts. In the example, the context of discourse makes it clear that the term "gold pieces" is used in a derogatory way to refer to refugees who came to Germany in large numbers in 2015 in an attempt to seek asylum.

The repeated use of the word "gold pieces" (parole) in this particular sense would be capable of altering its meaning in the linguistic system (langue) by adding a third meaning to the inventory of its aforementioned senses. In general, this points to the phenomena of polysemy and synonymy in language. The same word (represented in written language by the same character string) can take on several meanings or be linked to different mental concepts due to the characteristic of polysemy. On the other hand, synonymy describes the phenomenon that a mental concept can be expressed by different linguistic signifiers. These characteristics of language result in the heterogeneity and diversity with which similar thoughts can be expressed, which as a whole shape social reality, but also make the study of this reality through text such a complex endeavor.

3 In search for gold: an illustrative example

Due to its ambiguity and context-dependency, the interpretation of language use for empirical research is a demanding task for analysts. Computational models are capable of encoding the semantics of natural language with varying degrees of complexity to make it accessible to human interpretation. To illustrate these differences, we present the capabilities of different algorithmic models along with a selected example from German social media communication.

In June 2016, the chairman of the German Social Democrats (SPD) and president of the European Parliament Martin Schulz made an influential public statement about the increase of people seeking asylum in Germany at that time: "What the refugees bring to us is more valuable than gold. It is the unwavering belief in the dream of Europe. A dream that was lost to us at some point" (Riemer, 2016, own translation). In German social media communication, especially in debates of right-wing Facebook groups and commentators, this quotation was quickly altered into the sarcastic aggravation that refugees themselves are "more valuable than gold". Eventually, this culminated in a substantial semantic change of the term "Goldstücke". Originally, the German reference dictionary Duden describes two regular meanings: a gold

coin deemed to be legal tender, or metaphorically a very appreciated person or object (Duden, 2020). In social media, it rapidly became a derogatory term to mock migrants that frequently occurred in hate speech contexts (Fedtke & Wiedemann, 2020).

To reveal the semantic capabilities of different computational models, we investigate the semantic shift of the term "gold pieces" in social media compared to a reference corpus of standard language use. Our social media corpus contains roughly 360,000 user comments downloaded from German Facebook pages operated by major German mass media outlets such as Spiegel Online or Tagesschau.[1] Comments and replies to comments were crawled via the Facebook API from about 3,000 discussion threads during a time range of five months in 2017. Threads were selected only if the corresponding official post text of the mass media account contained at least one term of a list of filter keywords.[2] This key term filter was applied to retrieve a corpus of (potentially) controversial topics around refugees, migration, and discrimination. For our analysis, we split the user comments into roughly two million sentences to investigate the usage of terms in their sentence context.

In a previous study, this corpus was examined to analyze the dynamics of hate speech and counter-speech in German social media (Fedtke & Wiedemann, 2020). As one outcome, the new way of using the word "gold pieces" was particularly striking. Accordingly, the corpus serves us in this chapter as a basis for a closer examination of this shift in meaning. We contrast the new use of the word "gold pieces" in social media with a thematically nonspecific corpus of Wikipedia articles. For this reference corpus, we combined two standardized corpora from the Leipzig Corpora Collection, each containing one million sentences randomly sampled from Wikipedia in 2014 and in 2016 (Quasthoff et al., 2014).

4 Three challenges in automatic content analysis

In light of the linguistic principles discussed earlier, three challenges for automatic content analysis can be derived. *First*, to comply with the principle of syntagmatic relations between words, a computational text model should be able to take the sequentiality of text for adequate meaning representation into account. *Second*, the principle of paradigmatic relation requires a model to capture linguistic phenomena such as polysemy and synonymy concerning a specific usage context. *Third*, content analysis requires a representation of meaning for different semantic levels.

To meet these challenges, in our view it is essential that a model is able to separate a linguistic surface form, for example a fixed character string as a representative for a sign, from its corresponding meaning representation. However, since the first attempts to use computers for content analysis in social and communication studies, researchers relied mostly on word retrieval and frequency counts that are strictly confined to the surface form. The French philosopher and linguist Michel Pêcheux pointed to a tautological problem for the endeavor to infer understanding of texts based on such a technology: automatic searches for word surfaces "presuppose a knowledge of the very result we are trying to obtain" (Pêcheux et al., 1995, p. 121). Put another way, one must already know which meaning is communicated with a word to validly quantify its usage in texts. From quantification alone, however, one cannot learn about any meaning or change of it. This problem explains a lot of the discomfort qualitatively oriented social scientists expressed for decades when being confronted with research narratives based on computational analysis (for an early critique see Kracauer, 1952).

With Saussure's distinction of signifier and signified, Pêcheux argues that discourse has to be studied by observing language within its contexts of production and its use with as few presumptions as possible. Approaches that just count character sequences assigned to categories suffer

from the underlying false assumption of a bi-unique relation between signifier and signified. They are, thus, to be considered as "pre-Saussurean" (Pêcheux et al., 1995, p. 65). Along with our illustrative example, we will discuss opportunities and limitations of two "pre-Saussurean" semantic representations widely used in automatic content analysis: string patterns and the vector space model.

4.1 String pattern matching

The most basic form of computational semantic representation is also the most widely used in social science contexts: sequences of alphabet characters, referred to as a string, serve as query input for a matching procedure on some target text encoded as a string as well. Strings can be of arbitrary lengths and, thus, represent single words, multi-word units, phrases, sentences, or entire documents. For instance, a target document can be split at punctuation marks into sentences, and sentences can be split into isolated words at whitespace characters. The pattern matching algorithm can then check how often a query string, for example a keyword, occurs in a given target text. More complex extensions to this approach combine queries for individual words to word lists, so-called dictionaries, in which each term of a dictionary category is a representative for a more abstract concept (e.g. positive or negative sentiment). Dictionary terms can further be combined with certain rules (e.g. by AND, OR, NOT conditions) or regular expressions to match a desired observation in target texts. Among other things, regular expressions allow using wildcard characters and character classes (e.g. numbers or word characters) to look for. Such dictionaries are the basis for most automatic content analyses from the early document categorizations in media studies (Stone et al., 1966) to nowadays widely conducted sentiment analyses (e.g. Young & Soroka, 2012).

With regard to our example, Table 21.2 shows potential information that can be obtained automatically within this paradigm. We see that after splitting our corpora into sentences and words, Wikipedia contains more terms (word tokens) and a richer vocabulary (word types). For both corpora, we can measure how often we observe the occurrence of the exact character string "Goldstück" (singular), its plural form "Goldstücke", or its diminutive ("Goldstückchen"). Since the matching is for exact character strings only, query variations spelled with an initial lowercase letter produce drastically different results. The last query pattern, the regular expression "(?i)Goldstück.*|Goldmünze.*" matches a more complex pattern of words

Table 21.2 Frequencies of textual entities, character strings, and regular expression patterns in the two example corpora

Pattern	*Absolute*		*Relative*	
	Wikipedia	*Facebook*	*Wikipedia*	*Facebook*
Sentences	2,000,050	2,056,795	-	-
Tokens	39,387,827	27,863,340	-	-
Types	1,560,839	492,660	-	-
Goldstück	18	47	0.46	1.69
Goldstücke	31	276	0.79	9.91
Goldstückchen	0	80	0.00	2.87
goldstück	0	0	0.00	0.00
goldstücke	0	13	0.00	0.47
(?i)Goldstück.*|Goldmünze.*	146	494	3.71	17.73

beginning with Goldstück or its synonym term Goldmünze (gold coin) followed by any potential suffix while ignoring letter casing for any match. In total, we can observe quite large differences in absolute frequencies between the two sources. Also, nonstandard spelling variations occur only in Facebook. Yet, absolute numbers are hard to interpret regarding the different corpus sizes. Computing a relative measure, such as frequency per million words, shows that there is significant overuse of "Goldstück" and especially of the plural form "Goldstücke" in social media.

From our theoretical perspective, the important characteristic is the underlying assumption that each matching of a query in a target text is identical with one predefined meaningful event. In other terms, the Saussurean distinction of signifier and signified is bluntly ignored, as soon as meaning simply collapses with its linguistic surface form. It is impossible to learn about how these quantitative differences relate to actual semantics just from frequencies alone. Although string query patterns retain the sequential order of text, they are mostly restricted to the word level and do not contain any abstract representation of meaning to capture polysemy

Table 21.3 Keyword in context display for the target term "Goldstücke"

Corpus	*Left context*	*Keyword*	*Right context*
Wikipedia	samt dem Großgrundbesitz für 275 500	*Goldstücke*	, wobei ein Teil des Kaufpreises
	[along with the large land estate worth 275 500 of]	*[gold pieces]*	[where as part of the price]
	tausend "Goldmännchen", kleine	*Goldstücke*	mit eingravierten menschlichen Figuren aus der
	[thousand "little men of gold", small]	*[gold pieces]*	[with engraved human figures from the]
	Die letzten behandelt er bedächtlich wie	*Goldstücke*	; keine verfehlt das Ziel.
	[the last he treated cautiously like]	*[gold pieces]*	[; none misses the target.]
	Straßburg gegen eine Zahlung von 130.000	*Goldstücke*	und eine lebenslange Rente von 9.000
	[Straßbourg at a payment of 130,000]	*[gold pieces]*	[and a life annuity of 9,000]
	, die unter dem Namen „	*Goldstücke*	" vertrieben werden.
	[, those by the name of "]	*[gold pieces]*	[" being distributed.]
Facebook	und Massenschlägereien meinen da sind die	*Goldstücke*	sehr aktiv aber leider nicht wirklich
	[mean . . . and mass brawls, there are]	*[gold pieces]*	[very active but unfortunately not really]
	hat, allerdings gibt es wirklich	*Goldstücke*	unter den Leuten, die hier
	[have, however there are truly]	*[gold pieces]*	[among those people that here]
	Kannst ja paar	*Goldstücke*	haben vielleicht denkst dann anders:
	[You can [have] a couple of]	*[gold pieces]*	[maybe then you change your mind]
	Kohls	*Goldstücke*	und ihr Neid auf Alles und
	[Kohls']	*[gold pieces]*	[and their jealousy of everything and]
	Für Merkel und den Altparteien ihre	*Goldstücke*	25 Mrd.
	[For Merkel's and the old-parties']	*[gold pieces]*	[25 billion.]

or synonymy. To learn about actual differences in usage contexts and corresponding semantic change, we need to read closely each matching example of the search query. In Table 21.3, a so-called keyword in context (KWIC) list is given for five matches from each corpus that displays six tokens to the left and the right of each occurrence of an examined key term. The sample reveals that in Wikipedia "gold pieces" is utilized in the context of a means of payment, while in Facebook the term presumably designates a group of persons. To further validate this hypothesis, the analyst is set back to the method of qualitative reading of many more samples.

4.2 Vector space model

More elaborated methods of text mining and automatic content analysis are based on the idea to convert texts into some numeric representation ready for statistical analysis (Grimmer & Stewart, 2013, p. 272). For this, the complexity of natural language must be severely reduced. The two most severe simplification assumptions are that meaning can be observed through frequency counting of isolated word surface forms (statistical semantics hypothesis), and that word order in sentences or documents is not relevant to encode their meaning (bag of words hypothesis) (cp. Turney & Pantel, 2010). In the vector space model (VSM), each context unit (e.g. a sentence or document) of a given text collection is converted into a vector containing a count value for each word of a defined vocabulary. The vocabulary may comprise all distinct word types occurring in the text collection. This way, the entire corpus can be represented as a so-called document-term matrix (DTM) with n rows, the number of documents in the collection, and m columns, the number of word types in the vocabulary (e.g. 2,000,050 sentences and 1,560,839 distinct words for the Wikipedia corpus, cp. Table 21.2). A variant of the VSM is the term-term matrix (TTM). In $m \times m$ dimensions, the TTM encodes how often (or how significantly) two words co-occur in the same context window such as left/right neighborhood, sentence, or document throughout the entire collection.

Computationally such large data objects are hard to process and suffer from information sparsity.[3] Thus, the complexity of the original texts needs to be reduced even further. Typical measures are to restrict the vocabulary (e.g. ignoring less frequent terms) or to aggregate different word surface forms into one single form, for example counting each occurrence of an inflected form or the plural "Goldstücke" as an observation of the corresponding lemma form (this preprocessing step is called lemmatization). Of course, substantial information is lost from the text during this process of abstraction. However, proponents of automatic content analysis argue that enough information for certain analysis types such as thematic categorization, document similarity, or the study of word semantics is preserved. For the latter, co-occurrence analysis (also referred to as collocation analysis if conducted for direct neighboring contexts) strives to extract typical context words that co-occur with a given target word based on a statistical test of the TTM frequencies. From this analysis, co-occurrence graphs of typical usage contexts of words can be drawn for visual inspection.

Figure 21.1 contrasts co-occurrence graphs for our example target word "Goldstücke" in the two corpora based on a sentence context window and a log-likelihood test for significance. For complexity reduction, all words have been preprocessed to their lemma form with lower casing. By this step, we lost the information that the plural form is much more likely to occur in the Facebook corpus than the corresponding singular form and, hence, may be associated with a particularly interesting usage context. Yet, the graphs reveal very distinct contexts for the two sources. In Wikipedia, we see terms from a bakery context, as well as a range of other terms that first seem somewhat unrelated but actually stem from sentences of board game instructions. In Facebook, we also see a context of bakery terms and a set of terms related to politics and

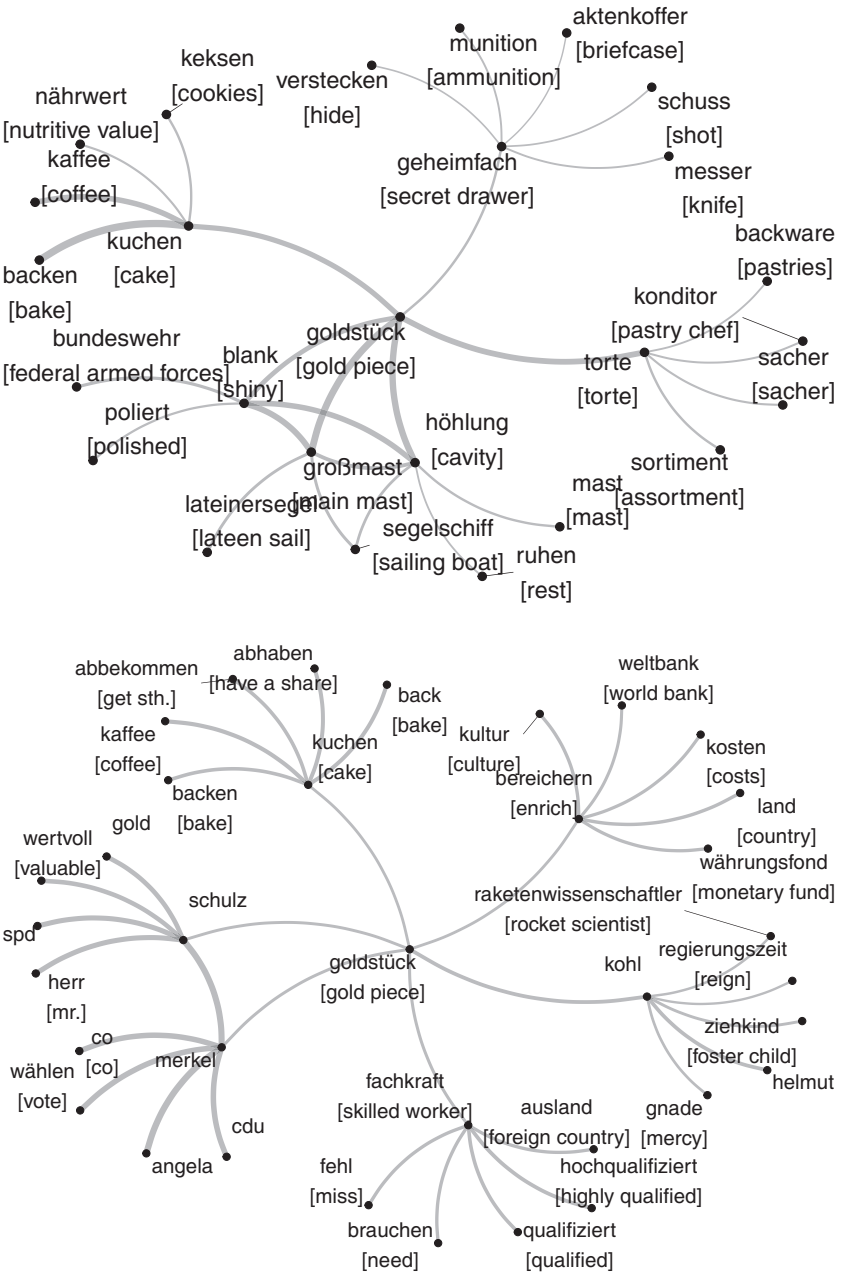

Figure 21.1 Co-occurrence graphs based on Wikipedia (above) and Facebook (below): starting with the term "goldstück", its top five co-occurrence terms and again their co-occurrences are displayed. Edge thickness represents its significance strength

the migration discourse in Germany. Words such as "Fachkraft" (skilled worker) and "hochqualifiziert" (highly qualified), for instance, point to debates about skill levels of the migrant workforce. However, to be able to validly interpret the actual meaning behind these word pairs, again we need to closely read (a sample of) sentence structures containing them.

In contrast to basic string matching, the VSM is much more flexible to encode meaning for content analysis. Due to its severe reduction of complexity of natural language during the conversion of words or text sequences into DTMs or TTMs, a representation of different semantic levels for complex analysis setups can be achieved easily. This flexibility comes at the expense of the loss of all information from sequentiality, that is, word order. However, the numeric 'bag-of-words' encoding of text allows for the use of advanced statistics to reveal meaningful structures. In terms of Saussurean structuralism, a statistically significant co-occurrence of two words in the same context window can be interpreted as a syntagmatic relation between those words (e.g. "Kohls"[4] and "Goldstücke" frequently occur next to each other). Similar patterns of syntagmatic relations, that is, small distances in the vector space of the TTM, can be interpreted as a paradigmatic relation between those words (e.g. "Goldstück" and "Fachkraft" seem to be used often in a similar sentence context). However, the major drawback of string matching also remains for representations of words in the VSM. Still, word meaning is collapsed with its surface form. Thus, specific usage contexts signaling distinct word senses are not separated in lexicometric statistics. Even statistically significant patterns of surface forms that appear quite similar at the first glance, such as the semantic relation of "Goldstück" and "Kuchen" [cake] in both of our example data sets, may actually relate to drastically different contents.

5 The 'Saussurean turn' in automatic content analysis

The earlier examples demonstrated that Pêcheux's criticism applies to a range of computational text analysis approaches widely used to date. The shortcoming of meaning representations collapsed into surface forms not only affects empirical social science but also sets an upper bound to the performance of many inference tasks in NLP. For this reason, language models that allow a separation of the meaning from its surface form while incorporating sequential and contextual information were studied in computer science. In the following, we sketch two major model types from this development which we, in analogy to the earlier methodological turn in the social sciences, interpret as the 'Saussurean turn' in automatic content analysis.

5.1 Latent space models

The idea that meaning should not be identical with observable variables in the data paved the way for the success of latent variable models in NLP. A first attempt by Deerwester et al. (1990), called 'latent semantic analysis', utilized the idea of singular value decomposition (SVD), a matrix factorization approach that allows for compression of the information contained in the high-dimensional, sparse DTM. LSA results in a lower-dimensional matrix representation of the original text that keeps the largest possible variance of the data. This way, the information encoded in document-wise counts of the several thousands of vocabulary words can be reduced to a vector of typically $K = 50$, 100, or 300 dimensions. In this latent vector space, semantically similar documents are placed near to each other due to correlation patterns of the vocabulary they contain. As a result, documents can be retrieved as similar, not necessarily because they contain a large overlap in their observable vocabulary, but because they share a certain latent meaning that stems from semantically related terms. Although LSA was successfully applied for document retrieval, it did not gain much attention in social science applications because the latent dimensions resulting from SVD are hardly interpretable in a meaningful way. This changed with "pLSA", the probabilistic reformulation of the approach by Hofmann (1999), and "latent Dirichlet allocation" by Blei et al. (2003), who combined the pLSA idea with a Bayesian statistical framework. Also referred to as the first 'topic model', LDA especially attracts a lot of

attention in computational social science nowadays. The model contains two latent variables, a document-topic distribution θ and a topic-term distribution ϕ to model the thematic composition of a text collection. Most probable terms of each of the *K* term distributions in ϕ (usually) form semantically coherent groups that can be interpreted as topics. Topic distributions in θ contain information about how large the proportion of a certain topic in every single document of the collection is. With this information, contents of very large collections can be analyzed thematically as well as with respect to collection metadata such as topic distribution over time, across different authors, or publications (Blei, 2012). The adaptation of topic models in the social sciences has led to research on best practices for their use as well as on solutions to the reliability problem of the seminal model (Maier et al., 2018).

Regarding our illustrative example, we used LDA to compute one model with $K = 200$ topics from a new corpus combining both the Wikipedia and the Facebook source. To keep the thematic usage of the term gold pieces distinguishable by source, we replaced each occurrence of the term with "goldstück_wiki" in the first source, and "goldstück_fb" in the second source. Further, lemmatization, lowercase reduction, and stopword removal were applied as preprocessing steps.

Table 21.4 displays the top 20 terms of the three topics with the highest probability of our target term in the two sources. Since "goldstück" itself is not overly frequent in both corpora, it is not among the top 20 terms shown in the table. However, the list of topic terms reveals six distinguishable thematic contexts, three for each source, which further enlightens the semantic shift in social media compared to the standardized text source. While the finance context dominates the use in Wikipedia, Facebook users tend to use the term when they express their opinion about migration and the welfare system.

Regarding the three challenges to automatic content analysis, latent topic models still are based on the DTM representation suffering from a complete loss of word order. Since the goal is to infer unobservable latent variables as an explanation for thematic coherence of entire texts, topic models are also restricted to operate on the document level of semantics only. However, compared to earlier more simple approaches, for the first time, they introduce a separation of meaning representation encoded in latent variables from the linguistic surface form of words used in a text. Since the model assigns a probability to each term of the vocabulary for each topic, any single term such as gold pieces can have a greater or lesser influence on each topic. This implicitly allows capturing the polysemic nature of terms to some extent.

5.2 Static word embeddings

In the adaptation of Saussure's notion of paradigmatic word relations for statistical semantics, Harris (1954) stated his famous distributional hypothesis that words that occur in similar contexts tend to have a similar meaning. Based on this assumption, language models that strive to encode word meaning by observing statistical word co-occurrence patterns in empirical language data were developed in computational linguistics. A very successful approach to distributional semantics is so-called word embedding vectors computed by artificial neural networks. Latent semantic models reduce the dimensionality of the vocabulary to represent the meaning of entire documents as a vector. Word embedding models, in contrast, learn meaningful low-dimensional vectors for each word of the vocabulary by observing their neighboring context words in a large text collection. Popular models such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) result in vectors for which the proximity of words in their vector space indicates their semantic relatedness, for example similarity. As in Saussure's chess analogy, an embedding model tries to find a positioning of all the words to each other to optimally

Table 21.4 Top 20 topic terms of the three topics with the highest probability of "Goldstücke" in Wikipedia and Facebook (Eng. translation in brackets). The word lists reveal thematic contexts around the term. Topic labels originate from manual interpretation.

Source	*Wikipedia*			*Facebook*		
Label	*Money*	*Routine*	*Order*	*Migration*	*Opinion*	*Welfare*
1	million [million]	tage [days]	grund [reason]	papier [paper]	passen [fit]	arbeiten [work]
2	euro [euro]	nächst [next]	einverstanden [agreed]	bundeswehr [federal armed forces]	schön [nice]	bekommen [get]
3	pro [per]	schön [nice]	ha [hectare]	wertvoll [valuable]	weltbild [world view]	geld [money]
4	rund [about]	woche [week]	heißen [called]	ausweis [ID]	echt [real]	arbeit [work]
5	mio [million]	abend [evening]	erlauben [allow]	bringen [bring]	gut [good]	job
6	betragen [amount]	stunde [hour]	strafrecht [criminal law]	gold	voll [full]	verdienen [earn]
7	milliarde [billion]	wochenende [weekend]	überdenken [review]	handy [cell phone]	klappen [work out]	rente [pension]
8	zahlen [pay]	essen [food]	vortäuschen [pretend]	schnell [fast]	raus [out]	wohnung [apartment]
9	jährlich [annual]	tag [day]	hinrichtung [execution]	identität [identity]	glauben [believe]	hartz [welfare]
10	us-dollar	nacht [night]	weide [meadow]	paß [passport]	halt [just]	rentner [retiree]
11	kosten [cost]	paar [pair]	angemessen [adequate]	verlieren [lose]	hirn [brain]	arbeitslos [unemployed]
12	insgesamt [total]	täglich [daily]	acker [acre]	holen [fetch]	lustig [funny]	leisten [afford]
13	dollar	weisheit [wisdom]	unsicher [insecure]	kontrollieren [control]	richtig [right]	lebensunterhalt [subsistence]
14	verkaufen [sell]	traum [dream]	gelten [apply]	bw	denken [think]	empfänger [receiver]
15	dm	brot [bread]	zschäpe	gefälscht [faked]	bitten [ask]	arbeitsplatz [job]
16	mrd [billion]	löffel [spoon]	aufhalten [impede]	überprüfen [check]	bisschen [a little]	bezahlen [pay]
17	höhe [amount]	lecker [delicious]	asiat [asian]	rein [clean]	ok	mindestlohn [minimum wage]
18	monat [month]	ruhen [rest]	abwertung [devaluation]	behörde [authority]	sorry	verdient [earn]
19	schätzen [estimate]	nah [near]	geltung [application]	prüfen [check]	passend [fitting]	vermieter [landlord]
20	knapp [scarce]	schaukel [swing]	schlumpf [smurf]	smartphone	maul [yap]	arbeiter [worker]

describe the language use it is presented with during its training phase. But, instead of the two-dimensional chessboard, it typically uses between 50 and 300 dimensions to position its elements and encode different aspects of meaning along with these dimensions. This way, it is not only the relative proximity of words that implies their meaning. To some extent, vector arithmetic can also be applied to reveal more complex semantic properties than solely word similarity. For instance, the vector operation 'king' – 'man' + 'woman' yields 'queen' as the closest vector indicating that some notion of gender is encoded in the model. Kozlowski et al. (2019) use this property of word2vec embeddings to their approach of the "geometry of culture". The main idea is to investigate cultural patterns simply by studying distances of words in an embedding model trained on a large corpus of texts representative for some base population (e.g. millions of US-American newspaper articles from one decade). The embedding vectors for music genres, sports, or professions can be projected along with gender, class, or ideological dimensions through the relative word distance to opposite terms of an imagined continuum (e.g. male-female, rich-poor, or liberal-conservative). The approach can reveal interesting cultural connotations especially in a diachronic perspective, for example the feminization of the occupation "journalist" in the second half of the 20th century (ibid.). In natural language processing, the use of word embeddings pretrained on very large generic corpora drastically improved the state of the art for almost any inference task due to the circumstance that they allow for some form of knowledge transfer in machine learning. For instance, a sentiment classifier presented a sentence containing the attribute 'good' during training already knows something about other sentences containing closely related terms such as 'great' or 'awesome' (Rudkowsky et al., 2018).

Embedding models can be trained on very large, generic data sets to encode general linguistic knowledge or with a domain-specific data set to capture particular aspects of a subset of the language system. For our gold piece example, we show the results for two embedding models. Since word embedding models require a lot of text to learn vectors from, we computed a word2vec model based on the entire German Wikipedia, and a GloVe model based on our entire Facebook corpus instead of our previously used samples.[5]

Table 21.5 shows a list of the nearest neighbors of "Goldstück" ranked by cosine similarity in the two embedding models. We can see easily that the Wikipedia list is dominated by semantically equivalent terms from the context of means of payment or treasure. The closest 15 terms can be regarded as more or less synonymous with gold pieces. In Facebook, however, we see terms such as "Facharbeiter", "Asylant" [asylum seeker], or "Neubürger" [new citizen]. In fact, these terms are used by right-wing commentators in Facebook to belittle refugees sarcastically. In this sense, they become synonymous with gold pieces which can automatically be revealed by the model. This way, word embeddings solve the synonymy problem of natural language to some extent by effectively separating word surface forms from their meaning.

5.3 Contextualized word embeddings

Saussure's dualism of the linguistic sign implies an arbitrary connection between the mental concept and a corresponding linguistic pattern, which results in the aforementioned synonymy problem. Conversely, linguistic patterns such as words can be assigned to distinct mental concepts. Since static word embeddings still mingle all different senses of a word surface form into one vector, this polysemy problem of natural language requires an extension of embedding semantics. Our gold piece example has already shown that the meaning of the term is quite different in our two sources. To represent the sense as a means of payment vs. a sarcastic reference to refugees, actually, two vectors would be required. For empirical analysis, it is impractical to assume a fixed sense inventory, since we do want to find out about different (new) usage

Table 21.5 The 25 terms with highest cosine similarity to "Goldstück" (descending) in Wikipedia and Facebook (Eng. translations in brackets)

Rank	*Wikipedia*		*Facebook*	
1	*Goldstück [gold piece]*	1	*goldstück [gold piece]*	1
2	Münze [coin]	0.63	fachkraft [skilled worker]	0.42
3	Goldbarren [gold bar]	0.62	facharbeiter [skilled worker]	0.37
4	Silbermünze [silver coin]	0.59	mutti [mommy]	0.36
5	Goldmünze [gold coin]	0.59	asylant [asylum seeker]	0.34
6	Schmuckstück [jewelry]	0.54	merkel	0.34
7	Geldstück [coin]	0.53	neubürger [new citizen]	0.32
8	Geldschein [banknote]	0.52	traumatisiert [traumatized]	0.32
9	Silberstück [silver piece]	0.51	unsere [our]	0.32
10	Silberbarren [silver bar]	0.50	volksvertreter [representative]	0.31
11	Juwel [jewel]	0.50	einzelfall [single case]	0.31
12	Taler [thaler]	0.50	kerl [guy]	0.30
13	Banknote [banknote]	0.50	kanzlerin [chancellor]	0.30
14	Dukaten [ducat]	0.50	schulz	0.30
15	Schatz [treasure]	0.50	migrant	0.29
16	Gold [gold]	0.50	ronny	0.28
17	Goldschatz [treasure of gold]	0.49	flüchtling [refugee]	0.28
18	Edelstein [gem]	0.48	wir [we]	0.27
19	Schatulle [casket]	0.48	unserer [our]	0.27
20	Kostbarkeit [preciousness]	0.48	gast [guest]	0.27
21	Goldklumpen [nugget]	0.48	unser [our]	0.27
22	Truhe [chest]	0.47	billig [cheap]	0.27
23	Schatztruhe [treasure chest]	0.47	wertvoll [valuable]	0.26
24	Papiergeld [paper money]	0.47	wirtschaftsflüchtling [economic migrant]	0.26
25	Silberschatz [treasure of silver]	0.47	goldjungs [golden boys]	0.26

contexts in the first place. To deal with this context-dependency of meaning, contextualized word embeddings have been developed as the recent major milestone in NLP. Contextualized word embedding models such as ELMO (Peters et al., 2018) or flair (Akbik et al., 2018) combine static word embeddings with an embedding representation of the surrounding context for each word in a text sequence, which allows an embedding of the same word surface form to vary slightly depending on its left and right neighboring terms. The most recent innovation are language models based on the transformer neural network architecture such as BERT (Devlin et al., 2019). Transformer networks create contextualized word embeddings without relying on precomputed static embeddings as an initial layer of word meaning representation. Instead, the entire model is pretrained on very large text collections to learn about language regularities. In its deep architecture, it produces embeddings on each network layer that encode complex syntactic and semantic information. For the BERT model, Wiedemann et al. (2019) show that its last embedding layer places the same word into separable regions of the vector space along with their different sense contexts. This suggests that some internal mechanism of word sense disambiguation is performed by the model automatically. The incorporation of long context sequences for word embeddings is so successful that newer models based on BERT even outperform the average human performance for many natural language inference tasks such as reading comprehension, question answering, and text classification (Liu et al., 2019).

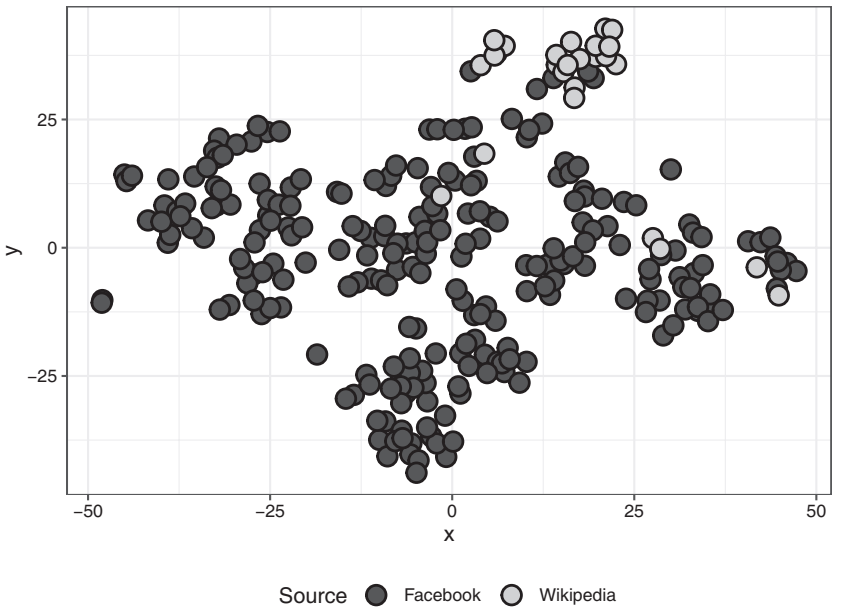

Figure 21.2 Two-dimensional t-SNE projection (Maaten & Hinton, 2008) of contextualized word embeddings in Wikipedia and Facebook. Each point represents one occurrence of "Goldstücke" in a specific sentence. The closer the points, the more similar is their contextual meaning

For our example case, we use a BERT model pretrained on German newspaper data to create contextualized word embeddings for all occurrences of "Goldstücke" in a combined corpus of our Wikipedia and Facebook source. Figure 21.2 displays a two-dimensional projection of the 768-dimensional embedding vectors. In this plot, the proximity of points indicates the semantic similarity of their usage contexts. Embeddings from Wikipedia are mainly placed in a separate region in the upper part of the projection grouping the contexts around the act of payment. Other regions of the projection reveal distinct derogatory contexts of "gold pieces", for instance accusations of sexual violence in the bottom center.

The clustering or projection of contextualized embeddings allows for a semantic grouping and quick visual inspection of patterns of similar usage contexts. Moreover, information about polysemy and synonymy of terms captured with contextual embeddings can drastically improve classifiers for the automatic coding of text. To automatically code the context of accusations of sexual violence as one specific form of hate speech against refugees in social media would require only a few training examples for a BERT-based text classifier. As an integral part of the transformer neural network architecture, the way of encoding semantics by contextual embeddings also retains the information from the sequential order of the text. Finally, yet importantly, the aggregation of word embeddings allows, depending on the network architecture, to compute sentence or document embeddings, thus, a flexible encoding of different semantic levels.

6 Conclusion

In this chapter, we derived three challenges for automatic content analysis based on structuralist linguistic principles that were highly influential for the development of text-based empirical

Table 21.6 Summary of linguistic challenges addressed by computational text models

Challenge	*String matching*	*Vector space model*	*Latent semantic model*	*(Contextualized) Embeddings*
Sequentiality	yes	no	no	yes
Semantic levels	no	yes	no	yes
Sign dualism	no	no	(yes)	yes

social research. Along with an illustrative example, we showed that widely used computational methods actually struggle to meet these challenges.

Table 21.6 summarizes how these challenges are addressed by computational models of increasing complexity. For the less complex models, information from the sequentiality of text is only retained when retrieving words as consecutive strings directly. Word order information is lost in the 'bag-of-words' approach of the VSM and the latent semantic model based on it. Second, a flexible encoding of semantics from varying levels of granularity is only possible with the VSM in which the vectors can easily represent different context units such as sentences, documents, or entire collections. The latent semantic model, in contrast, is only meaningfully defined for the document level, since documents are considered as a mixture of topics. The Saussurean dualism of the linguistic sign can be operationalized in computational models by separating a meaning representation from its linguistic surface form, that is, the character string representing a word or a word sequence. In part, this is realized by latent semantic models such as the LDA topic model, which models the same word contributing with varying share to each topic. This way, a high proportion of a word in two or more thematically distinct topics is a hint for its polysemic character. The automatic, unsupervised detection of such meaningful thematic coherence between words that abstracts from the isolated, ambiguous word surface form constitutes the success of topic modeling in the social sciences (Maier et al., 2018).

Regarding our posed challenges, the latest paradigm shift in NLP towards contextualized word embeddings in combination with deep neural networks (DNN) can be seen as a major step forward for computational content analysis. This way of modeling natural language allows taking word order into account. It also can represent semantics on higher levels than words through various methods of aggregating word embeddings. Last but not least, embeddings consequently realize Saussure's sign duality. A semantic concept can be interpreted as some specific region in the embedding vector space that may contain several, distinct terms referring synonymously to it. The contextualization of embeddings further allows for the same linguistic sign to refer to distinct regions in the vector space, thus modeling the polysemy of words.

To further elaborate the analogy between linguistic structuralism and embedding models from NLP, we can again look at the distinction between langue and parole. Language models such as word2vec and BERT need to be trained on very large text collections to encode their knowledge about semantics and language structure in an embedding vector space. Hereby, training data fits the notion of parole, the actual language use. Based on this empirical data, a language model is learned that can be viewed as an approximation of langue, the language system that originally generated the training data.

What's in it for computational social sciences? In contrast to earlier automatic approaches, the new DNN models are particularly appealing for qualitatively oriented social research. In empirical research with text, we do not look for signifiers as an end in itself but as references to underlying ideas and concepts. Contextualized embeddings, eventually, will enable researchers to conduct semantic searches, that is, to search for meaning instead of word surface forms. Like our illustrative example revealed, with the help of embedding semantics we can discover the

shift of meaning of the term "Goldstücke" from a compliment, or means of payment, to a set of sarcastic, derogatory accusations against refugees in social media. In combination with active learning (Wiedemann, 2019) where the machine and the content analyst fruitfully interact for automatic text coding, the new NLP-based language models will enable us to study social discourse with sophisticated depth in qualitative terms while retaining the advantages of quantitative analysis at large scale. Michel Pêcheux, to our knowledge the earliest advocate for a research program towards automatic discourse analysis (1969), would probably have been very excited about this 'Saussurean turn' in automatic content analysis.

Notes

1 See https://spiegel.de, and https://tagesschau.de
2 The keyterm list contains the following German word stems (translation, respective explanations are given in square brackets): afd [AfD, a German right-wing radical party], afrika [Africa], anschlag [attack], asyl [asylum], ausländ* [foreign], flucht [escape], flücht* [refugee], frau [woman], humanit* [humanitarian], islam [Islam], kopftuch [headscarf], missbrauch [abuse], muslim [Muslim], nazi [Nazi], npd [NPD, a German right-wing extremist party], rassis* [racism], schleier [scarf], sexuell [sexual], sudan [Sudan], syr* [Syria], terror [terror], vergewalt* [rape].
3 The absolute majority of cells in each row of such a matrix contains 0 values since only a few terms of the *m*-length vocabulary occur in each single document.
4 A reference to the former German chancellor Helmut Kohl.
5 We decided for the GloVe model for our Facebook corpus because it performs significantly better than word2vec for smaller data sets.

References

Akbik, A., Blythe, D., & Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638–1649). Association for Computational Linguistics. http://aclweb.org/anthology/C18-1139

Angermüller, J. (2007). Was fordert die Hegemonietheorie? Zu den Möglichkeiten und Grenzen ihrer methodischen Umsetzung. In M. Nonhoff, E. Laclau, & C. Mouffe (Eds.), *Diskurs – radikale Demokratie – Hegemonie. Zum politischen Denken von Ernesto Laclau und Chantal Mouffe* (pp. 159–172). Transcript.

Blei, D. M. (2012). Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. www.cs.princeton.edu/~blei/papers/BleiNgJordan2003.pdf

Bourdieu, P. (2003). The production and reproduction of legitimate language. In P. Bourdieu, J. B. Thompson, & G. Raymond (Eds.), *Language and symbolic power* (pp. 43–65). Harvard University Press.

Chandler, D. (2007). *Semiotics: The basics*. Routledge.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies (long and short papers)* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Duden. (2020). *Goldstück*. www.duden.de/rechtschreibung/Goldstueck

Fedtke, C., & Wiedemann, G. (2020). Hass- und Gegenrede in der Kommentierung massenmedialer Berichterstattung auf Facebook: Eine computergestützte kritische Diskursanalyse. In P. Klimczak, C. Petersen, & S. Breidenbach (Eds.), *Soziale Medien? Interdisziplinäre Zugänge zur Onlinekommunikation* (pp. 91–120). Springer. https://doi.org/10.1007/978-3-658–30702–8_5

Foucault, M. (1982). *The archaeology of knowledge*. Pantheon Books.

Glaser, B. G., & Strauss, A. L. (2005). *Grounded theory: Strategien qualitativer Forschung* (2nd ed.). Huber.

Grimmer, J., & Stewart, B. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Harris, Z. (1954). Distributional structure. *Word*, *10*(23), 146–162.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *SIGIR'99, proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 50–57). ACM. https://doi.org/10.1145/312624.312649

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905–949. https://doi.org/10.1177/0003122419877135

Kracauer, S. (1952). The challenge of qualitative content analysis. *Public Opinion Quarterly*, *16*(4), 631–642. www.jstor.org/stable/2746123

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, *abs/1907.11692*.

Maaten, L. van der, & Hinton, G. (2008, November). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., . . . Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Mayring, P. (2010). Qualitative content analysis. In U. Flick, E. V. Kardorff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 266–269). Sage.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio & Y. LeCun (Chairs.), *1st International conference on learning representations (ICLR)*.

Pêcheux, M. (1969). *Analyse automatique du discours*. Dunod.

Pêcheux, M., Hak, T., & Helsloot, N. (Eds.). (1995). *Automatic discourse analysis*. Rodopi.

Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe. Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)*. www.aclweb.org/anthology/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics (NAACL)* (pp. 2227–2237). ACL. https://doi.org/10.18653/v1/N18-1202

Posselt, G., & Flatscher, M. (2018). *Sprachphilosophie: Eine Einführung* (2nd ed.). UTB: Vol. 4126. Facultas.

Quasthoff, U., Goldhahn, D., & Eckart, T. (2014). Building large resources for text mining: The Leipzig corpora collection. In C. Biemann & A. Mehler (Eds.), *Theory and applications of natural language processing. Text mining* (pp. 3–24). Springer. https://doi.org/10.1007/978-3-30819–12655–5_1

Riemer, S. (2016, June 11). "Was die Flüchtlinge uns bringen, ist wertvoller als Gold". *Rhein-Neckar-Zeitung*. www.rnz.de/nachrichten/heidelberg_artikel,-Heidelberg-Was-die-Fluechtlinge-uns-bringen-ist-wertvoller-als-Gold-_arid,198565.html

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Saussure, F. de. (1916). *Cours de linguistique générale*. Payot.

Saussure, F. de. (2001). *Grundfragen der allgemeinen Sprachwissenschaft* (C. Bally, & A. Sechehaye, Eds., 3rd ed.). De Gruyter.

Stäheli, U. (2000). *Poststrukturalistische Soziologien. Einsichten. Themen der Soziologie*. Transcript.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis*. MIT Press.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188. www.jair.org/media/2934/live-2934-4846-jair.pdf

Wiedemann, G. (2013). Opening up to big data: Computer-assisted analysis of textual data in social sciences. *Historical Social Research*, *38*(4), 332–357.

Wiedemann, G. (2019). Proportional classification revisited: Automatic content analysis of political manifestos using active learning. *Social Science Computer Review*, *37*, 135–159. https://doi.org/10.1177/0894439318758389

Wiedemann, G., Remus, S., Chawla, A., & Biemann, C. (2019). Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. In *Proceedings of the 15th conference on natural language processing (KONVENS)* (pp. 161–170). German Society for Computational Linguistics & Language Technology.

Wrana, D. (2014). Linguistic turn. In D. Wrana, A. Ziem, M. Reisigl, M. Nonhoff, & J. Angermüller (Eds.), *Suhrkamp-Taschenbuch Wissenschaft: Vol. 2097. DiskursNetz: Wörterbuch der interdisziplinären Diskursforschung* (1st ed., pp. 247–248). Suhrkamp.

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234