Helene Schmolz

Anaphora Resolution and Text Retrieval

Empirische Linguistik / Empirical Linguistics

Herausgegeben von
Wolfgang Imo und Constanze Spieß

Band 3

Helene Schmolz

Anaphora Resolution and Text Retrieval

A Linguistic Analysis of Hypertexts

DE GRUYTER

Veröffentlicht mit finanzieller Unterstützung der Universität Passau.

ISBN 978-3-11-041674-9
e-ISBN (PDF) 978-3-11-041675-6
e-ISBN (EPUB) 978-3-11-041681-7
ISSN 2198-8676

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Druck und Bindung: CPI books GmbH, Leck
♾ Gedruckt auf säurefreiem Papier
Printed in Germany

www.degruyter.com

Foreword

Examining a topic from different perspectives has coined my career since the very first academic research. I have experienced that such an approach often leads to results which would not have been possible if looking just at one field. The book follows this tradition and takes into consideration the linguistic and information technological disciplines.

However, this book would not have come into existence if it had not been for many supporters. As it is based on my doctoral thesis and is a revised version of it, I would like to thank my supervisors first of all: my main supervisor Prof. Dr. Daniela Wawra, for her never-ending encouragement, for providing ample suggestions for improvement and for her student-friendly policy. I was always welcome to ask questions and her answers came back quickly. My second supervisor Prof. Dr. Harald Kosch encouraged me to write a thesis with such an interdisciplinary approach. He supported me patiently in all computer-related issues and paved my way to delve into the information technological part of this book. Special thanks also to his (former) employees, Prof. Dr. Mario Döller and Dr. David Coquil, who both helped me in discussing the research procedure with me in plenty of meetings. Furthermore, I would like to thank Prof. Dr. Ursula Reutner, who returned her report as third supervisor not only quickly but also offered helpful comments. Prof. Dr. Rudolf Emons first brought into my mind the idea of writing a doctoral thesis. I would also like to thank him for sharing his profound expertise in linguistics, which was indispensable for my book. A note of thanks also goes to my colleagues. With them I had the opportunity to develop initial thoughts in linguistics. At this point I also thank all those that have contributed to this book in different ways. Finally, I would like to thank my parents, who always supported me emotionally over the years and without whom I would not have been able to invest that much effort and time in the book.

Helene Schmolz

Contents

1 Introduction

Searching for information on the World Wide Web is a method of acquiring information useful to many people in varied guises. As the Internet continues to grow rapidly, with evermore resources such as hypertexts, documents or multimedia files being added every second, the challenge of providing users with the specific contents they need becomes more and more important. Current search engines are far from perfect as they hardly ever return completely satisfying results. One reason for this is that search engines, or more specifically text retrieval systems, usually do not consider the semantics of a text but rather just conduct a statistical analysis. Search engines, for example, will tend to rank a text in which *university* occurs ten times more highly than a text containing four occurrences of that item. This approach, however, cannot adequately represent the text's content, nor is a simple "bag-of-words" approach, where each text is merely seen as containing "an unordered set of words" (Baeza-Yates & Castillo 2006: 527), sufficient (cf. Jurafsky & Martin 2009: 801). Accordingly, the results returned by Web retrieval systems are commonly not subjected to any closer examination of the text's topics, let alone a linguistic analysis. Thus, search systems tend to use purely quantitative, rather than qualitative linguistic methods.

One approach to analyse a text linguistically is to investigate its cohesion, and here, more specifically, to pay attention to anaphors. The aim of this book is to outline anaphors of English and to examine to what extent they are worth being considered in text retrieval systems. Although anaphors and their resolution is a highly debated issue in present research, there are few studies that explore anaphors in the context of text retrieval. Even research in the field of anaphora resolution that is not intended for text retrieval shows a number of deficiencies.

To start with, a comprehensive classification of anaphor types based on linguistic description and also with regard to text retrieval systems is missing. Text retrieval systems would profit from a thorough examination because more precise rules for resolving anaphors could be formulated. The standard work for computational anaphora resolution is Mitkov's book *Anaphora Resolution* (2002). However, Mitkov's classification is not satisfying from a linguistic point of view because it does not take into account the many different types of anaphors and their features.

A further weakness is that in the discussion of anaphor types, no approach pays attention to *non-finite clause* anaphors. Not only are non-finite clauses disregarded as one type of anaphor in existing text retrieval systems, but they

DOI 10.1515/9783110416756

are even frequently ignored as one type of anaphor in linguistics, e.g. in Stirling & Huddleston (2010). Although Quirk et al. (2012: 910) mention non-finite clauses as a special type of ellipsis, they do not discuss whether non-finite clauses are a type of anaphor or not.

An additional shortcoming lies in the scarcity of annotated corpora. The few corpora that are annotated are mostly not freely accessible. Furthermore, current annotations of corpora only contain a few anaphor types, which is why these corpora are not adequate for an examination in this book.

All in all, there are many desiderata in the field of anaphora resolution. This book will contribute to their investigation from a linguistic and computational point of view. It draws particularly on syntactic, text linguistic and corpus linguistic methods as well as on methods from text retrieval and natural language processing. This book will first examine the question of what anaphor types the English language shows. For that purpose, a linguistic definition of anaphors is needed (cf. chapter 2), before a classification of anaphor types is presented (cf. chapter 3). Second, the frequency of each type of anaphor in hypertexts will be analysed (cf. chapter 4). From these insights, research questions for computational anaphora resolution can then be formulated (cf. chapter 4.5).

In more detail, the book is structured as follows. The second chapter will define anaphors and discuss related concepts. It will conclude with six conditions or characteristics of anaphors, all of which have to apply to items in order to be regarded as anaphors. In the third chapter, the twelve types of anaphors will be described in detail. The grammatical features of each anaphor type will be explained in depth, which is subsequently also of importance for computational anaphora resolution. The fourth chapter will examine the frequency of anaphors in hypertexts. Here, a corpus including different types of hypertexts will be introduced and statistically investigated with regard to anaphor types. A further chapter will describe text retrieval systems in general and for retrieving hypertexts from the Internet specifically, and the types of natural language processing methods these systems use. The sixth chapter will then present computational anaphora resolution, i.e. current approaches and applications, and the structure and evaluation of anaphora resolution systems. In the last chapter, *non-finite clause* anaphors will be analysed with respect to computational anaphora resolution, applying the insights of chapter four about the frequency of anaphors. Rules for identifying anaphors as well as for assigning antecedents will be established. Finally, the results will be discussed and perspectives for future research will be offered.

2 Linguistic fundamentals of anaphors and anaphora

2.1 Basic definitions

The word *anaphora* originates from Greek *ana-* (“back”) and *pherein* (“to bear”) and entered English via Latin transmission (cf. “Anaphora” 2010). In English, it is documented for the first time in 1589 (cf. Simpson & Weiner 1989: 436-437):

> *Anaphora*, or the Figure of Report. Repetition in the firft degree we call the figure of *Report* according to the Greeke originall, and is when we make one word begin, and as they are wont to fay, lead the daunce to many verfes in fute, as thus.
> *To thinke on death it is a miferie,*
> *To think on life it is a vanitie:*
> *To thinke on the world verily it is,*
> *To thinke that heare man hath no perfit bliffe.*
> (Puttenham 1589: 165)

“Anaphora” here denotes the rhetoric figure of repetition. The first written evidence of a use in grammar is not found until 1933, when the term appeared in Bloomfield’s work *Language*:

> [W]hen we say *Ask that policeman, and he will tell you*, the substitute *he* means, among other things, that the singular male substantive expression which is replaced by *he*, has been recently uttered. A substitute which implies this, is an *anaphoric* or *dependent* substitute, and the recently-uttered replaced form is the *antecedent*. (Bloomfield 1984: 249)

Later he gives another example:

> The word *one* [...] replaces *a* with anaphora of the noun [...] when no other modifier is present (*Here are some apples; take one*); [...] it is the anaphoric substitute for nouns after an adjective, and in this use forms a plural, *ones* (*the big box and the small one, these boxes and the ones in the kitchen* [...]). (ibid.: 265-266)

As for derivations, the adjective *anaphoric* and the adverb *anaphorically* are first mentioned in 1914 (cf. Bloomfield 1984: 249-251; Simpson & Weiner 1989: 436-437). According to the *Oxford Dictionary of English* (2005, 2nd rev. ed.), the noun *anaphor* has its origin in a backformation of *anaphora*, which dates back to the 1970s (cf. Soanes & Stevenson 2005: 55).

When consulting current dictionaries, the word *anaphora* often divides up into different senses, depending on its use in various contexts. First, the term denotes a part of the mass in liturgics. Second, “anaphora” describes the “repetition of a word or phrase at the beginning of successive clauses, lines of verse, etc." (Agnes et al. 2007: 51) in rhetoric (cf. Wilpert 2001: 27). So it is still used in the sense it was for the first time (cf. Puttenham 1589, above). Third, "anaphora" is used in music for the repetition of a voice, usually the bass (cf. Bartel 2007: 90-95).

The fourth – grammatical – definition is of importance here: "anaphora" is "[t]he use of a word which refers to, or is a substitute for, a preceding word or group of words" (Simpson & Weiner 1989: 436). According to Valentin (1996: 179), this meaning has developed from the use of anaphora in rhetoric. The following example illustrates what an anaphor is in the grammatical sense of the word:

(1) **Susan** plays the piano. She likes music.[1]

In example (1), the word *she* is an anaphor[2] and refers back to a preceding expression, in this case *Susan*. As can be seen in this example, an anaphor is an item that commonly points backwards.[3] Anaphors derive their interpretation from the expressions they refer to because their own meaning is often rather general (cf. Finch 2005: 199-200; Trask & Stockwell 2007: 16-17; Huddleston 2010a: 68; Quirk et al. 2012: 335, 862). This becomes obvious if the second sentence in example (1), *She likes music*, appears on its own. In such circumstances, it is not possible to find out the person meant by *she*. We can only state that it is most likely a female person.[4] But if both sentences are present, *she* is undoubtedly used in place of *Susan*.

The linguistic element or elements to which an anaphor refers is called "an antecedent". The antecedent in the preceding example is the expression *Susan*. The relationship between anaphor and antecedent is termed "anaphora" (cf. Huddleston 2010a: 68-69). "Anaphora resolution" or "anaphor resolution" is

[1] Throughout this work, anaphors are underlined, antecedents are set in boldface.

[2] In his binding theory, Chomsky uses the word *anaphor* differently. For him, an anaphor encompasses only reflexive (e.g. *himself*) and reciprocal pronouns (e.g. *each other*). All other anaphors, such as personal pronouns, are called "pronominals". In this book, though, Chomsky's definition is not relevant. Subsequently, *she* is regarded as anaphor in example (1) (cf. Chomsky 1993: 188, 211-212; Garnham 2001: 53; Bußmann 2008: 40-41; Jurafsky & Martin 2009: 736).

[3] An anaphor can also point forwards. This, however, is not very frequent (see chapter 2.3.1).

[4] If a person has a close connection with an animal, it is also possible to refer to these animals with *she* or *he*. This especially happens with domesticated animals, such as pets (cf. Quirk et al. 2012: 316-317).

the process of finding the correct antecedent of an anaphor (cf. Kübler n.d.: 5; Mitkov 2004a: 269; Crystal 2009: 25). In addition, so-called "anaphoric chains" can arise, if anaphors are themselves antecedents. In example (2), the anaphor *she* refers to the antecedent *Ann*, and *she* is also the antecedent of *herself* (cf. Halliday & Hasan 2008: 15, 52; Stirling & Huddleston 2010: 1457).

(2) **Ann** knew that <u>**she**</u> had written the letter <u>herself</u>.

Another central aspect of anaphors is that they can vary with regard to the importance of the antecedent for determining reference. Anaphoric noun phrases with a definite article are a case in point. They, for instance, can have antecedents that are not needed for determining the referent of the anaphor, as is shown here:

(3) I went to **an amusing show** recently where I met two friends. ... As they were sitting next to me during <u>**the show**</u> [1] I was able to ask them about the presenter. However, they could not tell me anything about <u>the show</u> [2].

In this example, the second anaphor *the show* [2] has the antecedent *the show* [1]. At first sight, the second anaphor [2] does not seem to gain new information through this relation to the antecedent [1]. But as the antecedent [1] itself is an anaphor and refers to *an amusing show*, the second anaphor [2] also gains information through these links. In consequence, it makes sense that the second anaphor [2] is interpreted in relation to its identical antecedent [1] (cf. Quirk et al. 2012: 1464-1465).

Recognising anaphors whose antecedents are literally identical with themselves is also important for computational anaphora resolution systems because anaphoric chains can be established through that process. Additionally, when detecting anaphoric chains, the distance between anaphor and antecedent does not become unnaturally large. Stirling & Huddleston (2010) argue:

> There can be a very large distance between the first antecedent in a chain and the final anaphor, greater than would typically be permitted for a direct link: it is the intermediate links that keep the referent salient in the context of discourse so that reference to it can be made by means of a personal pronoun or other anaphor with little intrinsic content. (ibid.: 1457)

With regard to the form of anaphors and antecedents, different realisations are possible. Anaphors can be whole noun or verb phrases, nouns, pronouns, adjectives, adverbs, infinitive markers; even "gaps", i.e. ellipses (see chapter 3), can work as anaphors. An anaphor, therefore, can constitute a gap, a single word or a phrase. The antecedent can be a word, a phrase – especially a noun phrase – a clause or even one or more sentences. There are antecedents that are coordinated, for example with an *and*-conjunction (example (4)):

(4) Yesterday **Kate and Joshua** went to the new restaurant. They enjoyed the food immensely.

Antecedents, namely so-called “split antecedents”, can also be made up from two or more separate parts (cf. Kübler n.d.: 20; Stirling & Huddleston 2010: 1458-1460), as in the following example (5), where the anaphor *they* refers to *Sue* and *Peter*:

(5) **Sue** got married to **Peter**. They were very much in love.

Finally, anaphors can vary whether their antecedent is in the same sentence (intrasentential anaphors), as in example (2), or in another sentence (intersentential anaphors), as in example (4) and (5) (cf. Botley & McEnery 2000: 4).

2.2 The relationship between anaphor and antecedent

2.2.1 Coreference

As the examples so far have shown, an anaphor is often related coreferentially to its antecedent, for instance, in (4), in which *they* and *Kate and Joshua* are coreferential. In fact, coreference represents the prototypical and simplest anaphoric relation. Fillmore (1975), for instance, formulated: “by the anaphoric use of an expression I mean that use which can be correctly interpreted by knowing what other portion of the same discourse the expression is coreferential with”[5] (ibid.: 40). Lewandowski (1994) describes anaphors similarly:

> Es handelt sich um Ausdrücke, die innerhalb eines Satzes oder Textes auf vorausgehende Ausdrücke Bezug nehmen, indem sie Identität herstellen. Die Bedeutung dieses Bezugs oder dieser Identitätsbeziehung läßt sich als Referenzidentität oder Koreferenz bezeichnen [...]. (ibid.: 64)

[5] Underlining of *anaphoric* and *coreferential* removed.

In the same way, Glück (2010: 40-41) and Bußmann (2008: 40-41) take coreference as criterion in their definition of anaphors or anaphoric relations. Quirk et al. (2012) provide a similar definition of coreference: "Coreference, as the name implies, means the bond of 'cross-reference' between two items or expressions which refer to the same thing or set of things" (ibid.: 863-864). Hence, an anaphor is associated to its antecedent on the level of meaning (cf. Lewandowski 1994: 611; Matthews 2007: 83; Crystal 2009: 116-117).

It should be noted in this context that anaphora resolution, as is the issue here, is not the same as coreference resolution. The focus of coreference resolution is to establish chains of coreferential expressions. Consequently, coreference resolution does not consider anaphors that are not coreferential (cf. Kübler n.d.: 9; Mitkov 2002: 7).

2.2.2 Substitution

Another relation, namely substitution, is also possible with anaphora (cf. Halliday & Hasan 2008: 3; Stirling & Huddleston 2010: 1458), although many researchers do not mention it with anaphora (cf. chapter 2.2.1):

(6) Tom bought a blue **shirt**. Simon bought a green <u>one</u>.

Here, the anaphor *one* has the antecedent *shirt*. The element *one* substitutes the word *shirt* and therefore shows a substitutional relation. The anaphor shows no coreferential relation to its antecedent because the green shirt is not identical with the blue shirt.

Substitution occurs where an anaphor replaces the antecedent, without necessarily being coreferential. This replacement must not lead to a change of meaning. Yet substitution is not restricted to instances where the antecedent can be inserted in lieu of the anaphor in exactly the same form. Morphological differences between anaphor and antecedent are no excluding criterion, i.e. if morphological changes are necessary, this instance belongs to the category of substitution nevertheless (cf. Quirk et al. 2012: 863-864). To demonstrate this, consider example (7), in which *one* has to be replaced by *tomato*, rather than *tomatoes*:

(7) Mr Tailor only likes red **tomatoes**. He has never tried a yellow <u>one</u>.

2.2.3 Comparison of coreference and substitution

Halliday & Hasan (2008: 88-90) argue that the difference between coreference, or “reference”[6] in their terminology, and substitution mainly lies in the type of linguistic level involved. They explain:

> In terms of the linguistic system, reference is a relation on the semantic level, whereas substitution is a relation on the lexicogrammatical level, the level of grammar and vocabulary, or linguistic ‘form’. (ibid.: 89)

According to Halliday & Hasan, “reference” means that two expressions are related in meaning, namely by being coreferential.[7] Substitution, however, relies on identities on the level of grammar and vocabulary. In most cases of substitution, both anaphor and antecedent have the same function in a phrase or clause. In example (6) the anaphor *one* and the antecedent *shirt* are heads of a noun phrase in object position. Taking the same function is not necessarily the case with coreferential anaphors: in example (8), *the parents* is a noun phrase in subject position, *themselves* is a noun phrase in object position.

(8) **The parents** blamed <u>themselves</u> for the accident.

Related to the functional similarity is the extent of the replaceability of an anaphor. As defined with substitution, a substitutional anaphor must be able to be replaced by the antecedent (see example (6)). With coreference, on the contrary, the possibility of replacements is no defining criterion, although replace-

[6] Halliday & Hasan (2008: 88-90) draw a similar distinction as coreference and substitution, but use the terms “reference” and “substitution”. As to the difference of “coreference” and “reference”, the term “reference” comprises coreference, so if expressions are coreferential, they are usually also referential (cf. Esser 2009: 35).

[7] In general, “reference” is used, especially in semantics and speech act theory, to describe the relation between a linguistic expression and an object in the extralinguistic world this expression designates (cf. Linsky 1974: 76-77; Lyons 1977: 177-178, 660; Lyons 2005: 295; Vater 2005: 11; Esser 2009: 35). Lyons (1977), for example, argues: “we will not say that a pronoun refers to its antecedent but rather that it refers to the referent of the antecedent expression with which it is correlated” (ibid.: 660). However, Halliday & Hasan (2008) use “reference” not only to describe relations between a text and the world, but also to denote relations within a text (cf. ibid.: 31-32). They state that reference occurs if “the interpretation of the reference item DEPENDS IN SOME WAY on that of the presupposed” (ibid.: 314). As the focus of this book is on text-immanent approaches, the term “reference” is only applied to the relation of anaphor and antecedent. Thus, an anaphor “refers” to the antecedent, where both anaphor and antecedent are items that occur in the text.

ments are sometimes possible, as in (4). Example (8) provides an instance where replacement does not work with coreference. The resulting sentence *The parents blamed the parents for the accident* would imply that we speak of two different groups of parents (cf. Quirk et al. 2012: 863-864).[8]

Ellipsis is a special case and can either be seen as separate category or counted as part of substitution. Here is an example of an ellipsis first:

(9) I will **invite Simon to the party** if you want me to ___.

The writer or speaker does not utter the expression *invite Simon to the party* at the end of example (9), he or she only implies it. Finch (2005: 199-200), for instance, considers ellipsis as a third type of anaphoric relationship. Halliday & Hasan (2008: 142-146), however, see ellipsis as a special form of substitution in which an expression is replaced by zero. This last definition is also adopted here. As a result, unspoken expressions represent substitutions by gap.

It should also be pointed out that to draw an absolute boundary between coreference and substitution is not possible because not all items within one anaphor type can usually be classified as belonging clearly either to coreference or substitution (cf. Halliday & Hasan 2008: 88-90; Quirk et al. 2012: 863, 865-867). Items in some contexts are even neither coreferential nor substitutional (see e.g. chapter 3.1.2.5). As such items have an explicit antecedent in the text, they are not excluded. In fact, they can be valuable in the same degree for anaphora resolution systems as items showing coreference or substitution. These items are then counted as part of a separate, miscellaneous category. Consequently, three categories are distinguished for the relation between anaphor and antecedent: the category of coreference; the category of substitution; and the miscellaneous category comprising items not belonging clearly to the other two. What types of anaphors belong to each of these categories is discussed in chapter 3.

[8] Moreover, “identity of reference” and “identity of sense” are found for denoting relations between anaphor and antecedent. Mitkov (2002: 16-17) and Huang (2000: 2-3), for instance, use these expressions. “Identity of reference” is synonymous with “coreference”, but “identity of sense” is not quite the same as “substitution”. “Identity of sense” is a part of substitution, but not the other way round (cf. Kübler n.d.: 12-13, 19-20; Huang 2000: 2-3; Garnham 2001: 46-47; Halliday & Hasan 2008: 3).

2.2.4 Sense relations

Apart from relationships of coreference and substitution, anaphors and antecedents can contain words that additionally show meaning relationships. Generally, three basic sense relations can occur between anaphor and antecedent: synonymy, hyponymy/hypernymy and meronymy[9]. Synonymy occurs if two words show "a relationship of 'sameness of meaning'" (Jackson & Amvela 2007: 107). "'Strict' synonymy" means that two words

> would have to be interchangeable in all their possible contexts of use: a free choice would exist for a speaker or writer of either one or the other word in any given context. The choice would have no effect on the meaning, style or connotation of what was being said or written. Linguists argue that such strict synonymy does not exist [...]. (Jackson & Amvela 2007: 108)

Therefore, we define synonymy as "'loose' synonymy". This means that two words are also considered synonyms even if they are not interchangeable in all contexts (cf. ibid.: 107-109). For instance, in (10) *satchel* and *schoolbag* are synonyms:

(10) I have got **a new satchel**. ... <u>This schoolbag</u> is bigger than the last one.

Next, hyponymy and meronymy are both hierarchical relationships. Hyponymy or "'kind of' relation" occurs if "the meaning of one form is included in the meaning of another" (Yule 2010: 118). Here, a superordinate term (hypernym) contains subordinate terms (hyponyms). Example (11) contains the hyponym *dogs*, which is a type of *animal* or, formulated differently, has as hypernym *animal*. Finally, meronymy is a "'part of' relationship" where one term is a part (meronym) of a superordinate term (holonym). To give a concrete example, *tyres* are a part of a *car*, as in (12) (cf. Jackson & Amvela 2007: 107-109, 117-120; Matthews 2007: 180; Jurafsky & Martin 2009: 646-651; Yule 2010: 117, 119; Schubert 2012: 48-51).

(11) A lot of people like **dogs**. Researchers say that <u>these animals</u> contribute to physical and psychological well-being.
(12) If Barbara drives to work by **car** she first has to change <u>the tyres</u>.

[9] Relations of meronymy are treated under the headings of indirect, bridging or associative anaphora in anaphora resolution (cf. Kübler n.d.: 18; Markert & Nissim 2005: 371).

Apart from that, homography and polysemy might be of interest with anaphora resolution, in order to find out if two items are anaphoric or not. Homography means that two items have the same spelling, but different meanings, e.g. *bank* (financial institution) – *bank* (of a river). With polysemy, two orthographically identical items are related in meaning, e.g. *foot* (part of body) – *foot* (measure) (cf. Herbst, Stoll & Westermayr 1991: 42; Yule 2010: 120).

2.3 Related concepts: cataphora, endophora, exophora and deixis

This section first considers cataphors and classifies them as a special form of anaphors. Cataphors will not be investigated in more detail due to their low frequency. Anaphora and cataphora together form endophora. Consequently, endophora will be discussed, as well as exophora and deixis, which can be contrasted with endophora.

2.3.1 Cataphora

Other terms that are found instead of "cataphora" include "anticipatory anaphora" and "backwards anaphora". These terms point out that cataphors are rather a special type of anaphor than an opposed concept (cf. Hoffmann 2000: 303; Stirling & Huddleston 2010: 1455-1456). It also implies that an item that can be used cataphorically, such as the personal pronoun *she*, can always also be used anaphorically, but not the other way round. Cataphors are defined as pointing forward (cf. Herbst, Stoll & Westermayr 1991: 182). The following sentence demonstrates such a case:

(13) After she had come home, **Susan** did her homework.

Here, the cataphor *she* refers to *Susan*. It is quite common to use the term "antecedent" also for an expression to which a cataphor refers (cf. Stirling & Huddleston 2010: 1455; Quirk et al. 2012: 862). The term "postcedent" instead of "antecedent" is also found (cf. Baicchi 2004: 30; Sladovníková 2010: 71). The relationship between cataphor and antecedent is named cataphora. Cataphors tend not to be very frequent in texts. They occur only in specific circumstances and constructions, mostly as rhetoric devices in fiction and in journalism. There is one effect for which cataphors are particularly used: they can create suspense (cf.

Jackson & Moulinier 2002: 179; Finch 2005: 201-202; Carter & McCarthy 2006: 245; Biber et al. 2007: 331).

2.3.2 Endophora and exophora

Halliday & Hasan (2008: 31-37) subsume anaphors and cataphors under the category endophora, which they in turn distinguish from exophora. This categorisation of reference is visualised in Figure 1.

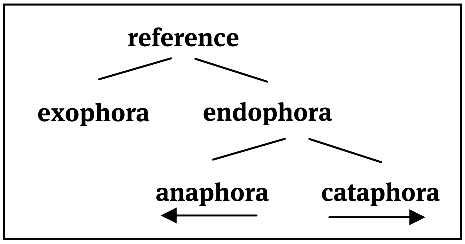


Fig. 1: Reference

Although both endophora and exophora constitute referential processes, they are different in one aspect fundamentally important for this work: whether the antecedents are found in the text or context. By definition, endophoric expressions have their point of reference in the text. Exophoric items, however, are references to the situation, so the referred item is retrievable from the verbal or nonverbal situation. It depends on the context, if an item has its point of reference in the surrounding text or in the situation, i.e. if this item is endophoric or exophoric. By reading or hearing only the sentence *She likes music* an outside observer does not know who *she* is. Yet, if a pointing gesture to the person who is meant by *she* accompanies this sentence in discourse, the reference is identifiable, and so it is exophoric. If the sentence *Susan plays the piano* precedes (see example (1)), the reference is endophoric. This also means that it is not an expression itself that is endophoric or exophoric, but the specific interpretation depends on the individual context. Some items, however, are nearly always endophoric, such as *herself*, or exophoric, such as *here*. Other items can usually be both endophoric or exophoric, depending on the context, such as *this* (cf. Matthews 2007: 131; Halliday & Hasan 2008: 31-37).

2.3.3 Deixis

A good definition of deixis[10] comes from Stirling & Huddleston (2010), for whom deixis is

> [t]he reference of certain kinds of expression[s] [that] is determined in relation to features of the utterance-act: the time, the place, and the participants, i.e. those with the role of speaker or addressee. (ibid.: 1451)

This implies that the point of reference to which deictic expressions refer varies from utterance to utterance. More specifically, their meanings relate to the utterance act in a particular way. Deictic *this* and *that*, for instance, differ in respect to the distance from the speaker, i.e. *this* is nearer, *that* is farther away. This means that deictic expressions are interpreted relative to the speaker. As stated in the quotation, deixis divides up into time deixis (e.g. *now*, *then*), place deixis (e.g. *here*, *there*, *this*, *that*) and person deixis (e.g. *I*, *you*). Moreover, deictic expressions are often used together with indexing acts, e.g. gestures or eye movements, in order to make clear what the point of reference is (cf. Finch 2005: 210-211; Bublitz 2009: 243-256; Stirling & Huddleston 2010: 1451-1453).

Anaphora and deixis share some features that emphasise their close relation (cf. Stirling & Huddleston 2010: 1454-1455): first, some items are anaphoric and deictic at the same time under certain circumstances. In example (14), the form *we* refers to *Bill* anaphorically and to implied *I* for the speaker deictically. Second, some expressions are either anaphoric or deictic, depending on the context. The item *that* in example (15) is deictic, as it refers to something that is evident in the situation and which is not close to the speaker. By contrast, *that* in example (16) is anaphoric, referring to *20 euros*.

(14) **Bill** came home earlier than usual. <u>We</u> will go to the cinema.
(15) *That* is not enough.[11]
(16) Jim gave me **20 euros**, but <u>that</u> is not enough.

It should also be mentioned here that there is a special form of deixis: “discourse deixis”, also called “textual deixis”. Stirling & Huddleston (2010) characterise it as follows: “the referent is not physically present in the situation of

[10] Bühler (1982), who discusses the connection of anaphora and deixis from a psychological point of view, defines “deixis” differently: For him, anaphora constitutes one type of deixis (cf. Bühler 1982: 105-106, 119-124; Bühler 1990: 120-122, 135-140; Lenz 1997: 23-32).

[11] Throughout this work, non-anaphoric uses are set in italics.

utterance but is located in the discourse itself" (ibid.: 1460). An example of discourse deixis is given in (17), in which *that* refers to the word, rather than to the concept bilingualism.

(17) She is currently writing about **bilingualism**. Shall I spell that out for you?

Discourse deixis is in a way similar to anaphora because explicit antecedents can be present, though antecedents are not obligatory for discourse deixis (cf. Cornish 2006: 632; Bublitz 2009: 256-257; Stirling & Huddleston 2010: 1460-1461). With regard to the difference of discourse deixis and anaphora, Lenz (1997) remarks:

> Die Tatsache, daß sowohl die Diskursdeixis als auch die Anaphora auf Diskursstellen verweisen, hat einige Verwirrung hinsichtlich ihrer Unterscheidung gestiftet. In der Literatur zu den beiden Phänomenen taucht immer wieder die Auseinandersetzung mit Grenzfällen auf, in denen die Interpretation als Diskursdeixis und Anaphora schwer entscheidbar sei. (ibid.: 75)[12]

It is also not entirely clear where the boundary between anaphora and deixis in general lies (cf. Stirling & Huddleston 2010: 1461). According to Cornish (2006), discourse deixis "provide[s] a transition between the notions of deixis and anaphora" (ibid.: 632). He argues that discourse deixis "consist[s] in using the deictic procedure to point to part of a pre- or postexisting textual or memory representation, but which is not necessarily highly activated" (ibid.: 632). Likewise, Matthews (2007) explains that discourse deixis includes "[a]ll forms of *anaphora and *exophora in discourse: i.e. of relations in fact distinguished from *deixis proper" (ibid.: 108). Expressions that are attributed discourse deictic function will also be considered in this book if the antecedent is present explicitly in the text. This condition applies in general: an item is only considered anaphoric if the antecedent is present in the same text (cf. Trask & Stockwell 2007: 16-17). As the aim is to carry out anaphora resolution with computational programmes, this condition is necessary otherwise systems would face great difficulties in resolving anaphors. The condition, however, also leads to the exclusion of some items of the so-called "discourse anaphors".[13] Those items are excluded that do not have an explicit textual antecedent (cf. Cornish

[12] A detailed discussion of possible distinctions between anaphora and discourse deixis is, for example, provided by Consten (2004: 29-35) and Lenz (1997: 68-70).

[13] "Discourse anaphors" must not be confused with "discourse deixis".

2006: 631). The reason is that the inferred antecedent might be very vague (cf. example (18)), which would make it difficult for systems to resolve such items correctly. Furthermore, discourse anaphors occur mostly in informal language (cf. Stirling & Huddleston 2010: 1471).

Cornish (2006) remarks on discourse anaphors: “not all referents will have been introduced via an explicit textual antecedent; it is also possible for them to be evoked ‘obliquely’ in terms of an association or a (stereotypical) inference of some kind” (ibid.: 631). For instance, the antecedent of the discourse anaphor *it* has to be inferred in (18). It could be *the cake* or a similar expression. Other terms for such cases include “quasi-anaphoric uses” (Stirling & Huddleston 2010: 1470-1471; see also chapter 3.1.1.3), “associative anaphora” (Meyer & Dale 2002), and “indirect anaphora” (Mitkov 2002: 15).

(18) Could you call me when you have found out? I’d like to prepare *it* for the birthday party.

As to the relationship between deixis, exophora and endophora, there are different views: Halliday & Hasan (2008: 33) use the term “exophora” instead of the term “deixis” (cf. Consten 2004: 80), whereas Cornish (1999: 112-115 and 1996: 20), for example, argues that exophora belongs to anaphora and not deixis: “‘EXOPHORA’ falls under the heading of anaphora proper and not (proto-typical) deixis” (Cornish 1996: 20). Finch (2005), in contrast, states that deixis is both exophoric and endophoric. He explains that “[one] form of deixis is EXOPHORIC in character in that it is situationally, or contextually, bound. A secondary form of deixis is ENDOPHORIC and serves to locate items textually” (ibid.: 210-211). The latter occurs in examples such as (19), in which *this* refers to the following sentence and so is cataphoric while being deictic in Finch’s point of view. In this book, *this* in (19) is a case of a cataphoric interpretation.

(19) Keep <u>this</u> in mind: **you must not smoke**.

All in all, the tendency is to see deixis as exophoric and anaphora as endophoric (cf. Green 2006: 417). This could go back to the fact that the words *exophora* and *endophora* by Halliday & Hasan (2008) did not come into extensive use but are instead frequently replaced by *deixis* and *anaphora* (cf. Vater 2005: 18). For instance, Huddleston & Pullum’s *The Cambridge Grammar of the English Language* (2010) includes a chapter called “deixis and anaphora”, rather than “exophora and endophora”.

2.4 Anaphors as cohesive devices in texts

Anaphors linguistically belong to the concept of cohesion. Anaphors are cohesive devices because they establish relations between linguistic elements in texts and therefore contribute to the cohesion of a text. Therefore, this chapter investigates cohesion and other basic concepts of text linguistics. Finally, the role anaphors play in reducing texts will be discussed.

2.4.1 Texts and their features

As this book analyses texts, the question of what a text is will now be examined. To start with, there is no agreed definition of the term “text” among linguists. In a strict sense, “text” only refers to written forms of language, which is also consistent with the way “text” is used in everyday language. Yet, it can also be used for spoken forms of language (cf. Matthews 2007: 405-406). Instead of *text*, the word *discourse* is often used synonymously (cf. Esser 2009: 9; Malmkjær 2010: 538; Van Dijk 2010: 116; Hoffmann 2012: 5; Żebrowska 2013: 54). In some cases, the two terms are also distinguished. A good example is provided by Brinker (2010: 18-19). He argues that to define a “text”, a differentiation between monologue and dialogue is essential. A monologue is a product of one person, whereas a dialogue is an interactive process. As a result, the term “text” is mainly used for monologues, whereas “discourse” is more applied to dialogues, irrespective of their written or spoken form (cf. Gansel & Jürgens 2009: 16-17).

De Beaugrande & Dressler (1981: 3) suggest a communication-oriented definition of “text”, which is often cited. According to De Beaugrande & Dressler (1981: 3), there are seven criteria that define textuality, and, therefore, what a text is: cohesion, coherence, intentionality, acceptability, informativity, situationality and intertextuality. De Beaugrande & Dressler classify cohesion and coherence as referring to text-internal relations.[14] Text-external criteria are intentionality, acceptability, informativity, situationality and intertextuality. These five aspects focus on the user and therefore are based on the communication situation (cf. De Beaugrande & Dressler 1981: 7-11; Stede 2007: 26-28; Żebrowska 2013: 56). If the seven features of textuality are met, they argue, the

[14] To see coherence as text-internal is not undisputed (cf. Vater 2001: 54). For instance, Bublitz (1994: 218-220) argues that the recipient has to establish coherence: “Coherence is solely hearer-based. It is not a text-inherent property at all, but arises only through the process of interpretation and ascription of those who try to understand.” (ibid.: 220).

text is communicative and therefore a text proper. But these criteria underlie criticism, for example from Schubert (2012: 23). He argues that not each criterion is necessary for textuality. For Vater (2001), coherence seems to be dominating the other features (cf. ibid.: 28-36, 52-54).

Halliday & Hasan (2008: 1, 13) only use cohesion, which also includes coherence in their terminology, to define a text:

> [T]he concept of cohesion accounts for the essential semantic relations whereby any passage of speech or writing is enabled to function as text. We can systematize this concept by classifying it into a small number of distinct categories – reference, substitution, ellipsis, conjunction, and lexical cohesion; categories which have a theoretical basis as distinct TYPES of cohesive relation, but which also provide a practical means for describing and analysing texts. [...] There are, of course, other types of semantic relation associated with a text which are not embodied in this concept; but the one that it does embody is in some ways the most important, since it is common to text of every kind and is, in fact, what makes a text a text. (Halliday & Hasan 2008: 13)

In sum, it is difficult to say what distinguishes a text from a non-text and what features are obligatory for texts. A text is thus often seen as a “prototype concept” (Esser 2009: 12): some texts fulfil more, others fewer criteria. Consequently, it is a definition on a scale (cf. Vater 2001: 17, 20-21; Wawra 2008: 116, 119-120).

Text linguistics[15] was initially concerned with the systematic analysis of language and examined syntactic-semantic relations between linguistic elements, especially across sentences. Later on, text linguistics was viewed from a pragmatic perspective, and explored the communicative function of texts (cf. Brinker 2010: 12-20; Schubert 2012: 27). Although both directions of text linguistics are at best combined when dealing with texts, they together are still insufficient for defining texts in all their aspects (cf. Vater 2001: 20-21).

When discussing texts, the concept of multimodality has also to be mentioned. Especially on the Web, texts do not only use written language, but pictures, videos, animations, in short, any signs for communication. All these different signs in sum contribute to the meaning (cf. Żebrowska 2013: 69, 91). This means that a definition of “text” has also to encompass such multimodal aspects (cf. Schütte 2004: 92).

[15] In the Anglo-American context, text linguistics is part of discourse analysis (Schubert 2012: 14-15).

2.4.2 Cohesion

One feature of textuality – cohesion – will here be discussed in more detail because it is a central aspect for anaphora. In general, cohesion deals with how expressions are connected, within and especially across sentences. Cohesion makes obvious the texture resulting from cohesive ties, so only endophoric, i.e. anaphoric and cataphoric, reference is cohesive (cf. Halliday & Hasan 2008: 36-37). Moreover, both coreference and substitution are means of cohesion (cf. Quirk et al. 2012: 864). Coreference establishes cohesion in that it shows "continuity of referential meaning" (Halliday & Hasan 2008: 323), substitution through the "continuity of lexicogrammatical meaning" (ibid.: 322). The above-mentioned anaphoric chains also play an important role in establishing cohesion, "since it creates a kind of network of lines of reference, each occurrence being linked to all its predecessors up to and including the initial reference" (Halliday & Hasan 2008: 52; cf. also Biber et al. 2007: 234-235).

Additionally, grammatical cohesion, which encompasses function words, such as pronouns, can be distinguished from lexical cohesion, which concerns content words, i.e. lexical items.[16] The pronoun *she* in example (20), for instance, shows grammatical cohesion in that a link to *the teacher* is established. Example (21) illustrates lexical cohesion because of the expression *the actor*. *Actor* is not a pronoun or a different function word but a content word, and it refers synonymously to *Brad Pitt*. In the examples, *she* and *the actor* constitute anaphors. Hence, items of both types of cohesion can work anaphorically, as will also become clear in chapter 3. Anaphors as defined here are consequently not restricted to grammatical cohesion (cf. Quirk et al. 2012: 267-268, 351; Schubert 2012: 31, 46-48).

(20) **The teacher** entered the room. <u>She</u> carried a lot of books.
(21) **Brad Pitt** did not attend the show. <u>The actor</u> is currently filming in London.

In contrast to cohesion, there is coherence, which looks at the semantic connection of a text. In cases in which semantic relations are not explicitly found in the text, the reader has to infer them. This is why, to establish coherence, the reader or hearer needs implicit information from the text and world

[16] Further information on the categorisation of grammatical and lexical cohesion is given, for example, in Schubert (2012: 31-55). See also Quirk et al. (2012: 67-68, 72) on the difference between function and content words, i.e. closed and open word classes.

knowledge (cf. De Beaugrande & Dressler 1981: 3-4; Stede 2007: 20-21; Esser 2009: 13-15, 42; Malmkjær 2010: 540-541; Schubert 2012: 31-32, 65-66). Hoffmann (2012) explains that "whereas cohesion is fixed in textual form, coherence is a mental process or product" (ibid.: 10). Coherence is not very important here because only elements explicitly found in the text are investigated.

Another differentiation between cohesion and coherence comes from De Beaugrande & Dressler (1981). They argue that cohesion "concerns the ways in which the components of the SURFACE TEXT, i.e. the actual words we hear or see, [...] are *mutually connected within a sequence*" (ibid.: 3). Cohesion is distinguished from coherence because the latter "concerns the ways in which the components of the TEXTUAL WORLD, i.e. the configuration of CONCEPTS and RELATIONS which *underlie* the surface text, are *mutually accessible* and *relevant*" (ibid.: 4).

Halliday & Hasan (2008) do not distinguish between cohesion and coherence but use the term "cohesion" to refer to both. They identify five categories that belong to grammatical or lexical cohesion, or fall in between. First, grammatical cohesion consists of two types: reference and substitution. Second, lexical cohesion also splits up into two types, according to Halliday & Hasan: reiteration and collocation. Reiteration is achieved by the use of repetition or sense relations, such as synonymy. Example (3) contains *the show* in the second sentence, which is repeated in the third sentence, and thereby establishes cohesion. Collocation, according to Halliday & Hasan, includes items that share some lexicosemantic relation, such as *north* and *south*, which denote the opposite direction. They show "complementarity" and so are "related by a particular type of oppositeness" (ibid.: 285). Moreover, there is conjunction. Conjunction lies in between grammatical and lexical cohesion but belongs more to grammatical cohesion. Conjunction does not refer to a specific expression but establishes logical relations between the preceding and the following sentence. It includes expressions such as *but*, *therefore* and *then* (cf. Halliday & Hasan 2008: 5-9, 226-227, 242-243, 278-279, 288, 303-304; Vater 2001: 30; Malmkjær 2010: 540). Example (22) shows this type of cohesion:

(22) We filled in the form. *Then* we talked to the person in charge.

Halliday & Hasan (2008) state:

> Conjunctive elements are cohesive not in themselves but indirectly, by virtue of their specific meanings; they are not primarily devices for reaching out into the preceding (or following) text, but they express certain meanings which presuppose the presence of other components in the discourse. (ibid.: 226)

As conjunctions are not “phoric” (Halliday & Hasan 2008: 321), they are not relevant for anaphora (cf. Christiansen 2011: 160).

Halliday & Hasan’s approach remains the exception because cohesion and coherence are usually treated separately, e.g. in De Beaugrande & Dressler (1981). Nevertheless, both concepts are in a way related to each other. For instance, cohesion contributes to the creation of coherence (cf. Esser 2009: 15). As a result, Stede (2007: 25-26) views cohesion as the linguistic reflection of coherence. Cohesion makes obvious the implicit ties in a text:

> Wir betrachten daher die Kohäsion (an der Textoberfläche sichtbare Verknüpfung) als die linguistische Reflexion von Kohärenz (unter der Textoberfläche liegende, vom Rezipienten zu rekonstruierende, inhaltliche Verknüpfung). (ibid.: 25)

Likewise, Storrer (2003: 275-276) looks upon cohesion as a special case of coherence, which makes explicit the relation of expressions through grammatical means. Even though cohesion is not an absolutely necessary criterion of textuality, most texts also show cohesion. For instance, when people express ideas, this has to happen within the grammar of a language and by using the correct pronouns. Not paying attention to grammar might lead to misunderstandings or make it hard to understand the text (cf. Brinker 2010: 40). Bublitz (1998: 5-7) formulates in this context:

> Cohesive means are cues which ‘signal’ or indicate the preferred line of coherence interpretation. A lack of cohesive means may disturb the hearer’s/reader’s interpretation of coherence. (ibid.: 5-6)

All in all, paying attention to cohesion plays a fundamental role in finding out what the text is about.

2.4.3 Cohesive devices as a form of reduction

The use of anaphors as cohesive devices in texts can be a form of reduction. Reduction means that shorter forms replace full expressions; the text gets more compressed and redundancies are avoided. Syntax plays an important role and decides what forms of reduction are possible. Quirk et al. (2012) stress:

> Although reduction may in general be regarded in semantic or pragmatic terms as a means of avoiding redundancy of expression, what kinds of reduction are permitted is largely a matter of syntax. (ibid.: 859)

Generally, the shortest and so the most economical form of reduction is chosen in a text. These reductive forms must have an explicit antecedent, i.e. they have to be recoverable from the text. Consider example (23), which shows various degrees of reduction – a) being the shortest form. Furthermore, reduction, and especially the most economical form of it, can contribute to clarity because the hearer or reader can concentrate on new information. If, however, ambiguity or misunderstandings arise, reduction is avoided (cf. Biber et al. 2007: 327; Quirk et al. 2012: 858-862).

(23) a) I will **invite Simon to the party** if you want me to ___.
b) I will **invite Simon to the party** if you want me to do so.
c) I will invite **Simon** to the party if you want me to invite him.
d) I will invite **Simon** to the party – IF you want me to invite him to the party.

To what extent anaphors reduce a text, depends on the type of anaphor. Ellipses are the most radical form of reduction. Other types of anaphors cannot reduce much but rather serve to avoid excessive repetition.[17] Poitou (1996) argues that pro-forms[18] are more a device for establishing coherence than for simply reducing a text: “Sicher ist die Pro-Form meistens kürzer als das von ihr vertretene Segment, aber nicht immer” (ibid.: 124). Using alternative forms is a sign of the competence of an adult speaker or writer, whereas children in their early stages of language development rather use the same expression (cf. Clark 2009: 199-200; Brinker 2010: 33-34; Christiansen 2011: 326).[19] Subsequently, compressed and alternative expressions as anaphors also serve to connect linguistic elements in a text and play an important role as cohesive devices.

[17] The types of anaphors and their degree of reduction and/or repetition are listed in chapter 3.

[18] “Pro-forms” are “ITEMS in a SENTENCE which substitute for other items or constructions” (Crystal 2009: 390). Sladovníková (2010) argues: “Sie bilden in der Sprache Minimalformen, da sie sprachlich inhaltsärmer sind” (ibid.: 67). Personal pronouns, indefinite pronouns, and *so*, for instance, belong to pro-forms (cf. Quirk et al. 2012: 865; see also Crystal 1994: 315 and Lewandowski 1994: 836-837).

[19] Repetition can be used on purpose to cause rhetorical effects, for example, to create emphasis. Furthermore, legal language uses more repetition than the language we use for everyday communication, in order to avoid misunderstandings (cf. Quirk et al. 2012: 860, 1441).

2.5 Anaphors in the present book

To sum up, the following characteristics or conditions have to apply to anaphors in this book:

- *Anaphors can refer backwards as well as forwards, i.e. the antecedent either precedes or follows the anaphor.*

  This means that cataphoric interpretations are also examined, with those anaphor items with which they can occur.
- *Anaphors must have an explicit antecedent, i.e. an antecedent that occurs in the same text.*

  As a result, only endophoric elements are seen as anaphors. Such a definition, on the one hand also includes those discourse deictic items with an explicit antecedent. On the other hand this means that discourse anaphors are excluded.
- *The referent of anaphors is determined in relation to their antecedents.*

  In how far the antecedent is absolutely necessary for determining the referent of anaphors depends on the type of anaphor. Nevertheless, each anaphor relies – more or less – on its antecedent for interpretation.
- *The relation between anaphor and antecedent is coreferential, substitutional or neither.*

  Not only anaphors that show coreference will be investigated, but also those that show substitution. In addition, some anaphors in specific contexts that show neither a clear coreferential nor substitutional relationship will be considered. Subsequently, an anaphor belongs to one of these three categories: category of coreference, category of substitution, miscellaneous category.
- *The use of an anaphor mostly leads to a reduction of a text and/or usually avoids excessive repetition.*

  Some types of anaphors reduce a text and introduce alternative expressions. Other types are more responsible for avoiding repetition and restricting repetition to an acceptable amount.
- *Anaphors contribute to the cohesion of a text, and thus disclose a text's content.*

  Anaphors are cohesive devices. Cohesion is not absolutely necessary for texts but can often be regarded as visualised semantic relations that reflect the content of texts.

3 Types of anaphors

Moving from the definition and characteristics of anaphors to the types of anaphors, this chapter will detail the nomenclature of anaphor types established for this book. In general, anaphors can be categorised according to: their form; the type of relationship to their antecedent; the form of their antecedents; the position of anaphors and antecedents, i.e. intrasentential or intersentential; and other features (cf. Mitkov 2002: 8-17). The procedure adopted here is to categorise anaphors according to their form. It should be stressed that the types distinguished in this book are not universal categories, so the proposed classification is not the only possible solution. For instance, personal, possessive and reflexive pronouns can be seen as three types or as one type. With the latter, the three pronoun classes are subsumed under the term “central pronouns”, as it is adopted here.

Linguistic classifications of anaphors can be found in two established grammar books, namely in Quirk et al.’s *A Comprehensive Grammar of the English Language* (2012: 865) and in *The Cambridge Grammar of the English Language* (Stirling & Huddleston 2010: 1449-1564). Quirk et al. include a chapter of proforms and here distinguish between coreference and substitution. However, they do not take anaphors as their starting point of categorisation. Additionally, Stirling & Huddleston do not consider anaphors on their own but together with deixis. As a result, anaphoric noun phrases with a definite article, for example, are not included in both categorisations. Furthermore, Schubert (2012: 31-55) presents a text-linguistic view, of which anaphors are part, but his classification is similarly unsuitable because it does not focus on the anaphoric items specifically. For instance, it is doubtful if extended reference, i.e. *it* and *this*/*that* referring to a clause, belongs (as he details it) to his category of “comparative reference”, or to “personal reference” and “demonstrative reference” because these are personal/demonstrative pronouns (cf. ibid.: 35).

In addition, other classifications, for instance, from Huang (2000: 2-5) could be considered. He divides anaphora up into two syntactic categories: noun phrase- (including noun-) anaphora and verb phrase-anaphora. However, these classes are too broad and unspecific for computational tasks. Mitkov (2002: 8-15) proposes a further classification giving more weight to the computational aspect of anaphora resolution. He distinguishes between pronominal anaphora, lexical noun phrase anaphora, noun anaphora, verb anaphora, adverb anaphora and zero anaphora. Such categories are too vague from a linguistic point of view. For instance, lexical noun phrase anaphora is realised, per definition, as definite noun phrase or proper name and its antecedent is a full noun phrase.

DOI 10.1515/9783110416756

Noun anaphora, such as *one* in example (6), however, does not have a full noun phrase but only a noun as antecedent. One problem of such a classification is, for instance, whether *this* belongs more to lexical noun phrase anaphora because it is a definite noun phrase, or to noun anaphora because it often takes only a noun as antecedent (see chapter 3.3). Additionally, clauses or sentences as antecedents are not considered in any of Mitkov's categories (cf. also Mitkov 2004a: 268-269). It is obvious that all these categorisations are not entirely adequate for anaphors.[1]

The criteria that are taken into account in establishing the classification in this book rely on both linguistic viewpoints and practicability for computational tasks, with the linguistic aspect in the foreground. This means that the categorisation of anaphors predominantly follows linguistic criteria. Computational features are particularly taken into consideration in contexts where items classified as anaphors have to be distinguished from their non-anaphoric uses. Anaphors are here divided into 12 categories, which are: central pronouns; reciprocal pronouns; demonstrative pronouns; relative pronouns; adverbs; noun phrases with a definite article; proper names; indefinite pronouns; other forms of coreference and substitution; verb phrases with *do* and combinations with *so*, *this*, *that*, *it* and *the same (thing)*; ellipses; and non-finite clauses. These anaphor types and their items as well as a detailed description of their features – which will be important for anaphora resolution – are discussed in chapters 3.1 to 3.12.

3.1 Central pronouns

The expression "central pronouns" is an umbrella term covering personal, possessive and reflexive pronouns. According to Quirk et al. (2012: 345-346), these three types of pronouns form one category because they belong to each other more than do the remainder of pronouns. Personal, possessive and reflexive pronouns all differentiate between person, number and gender. More importantly, the characteristics of person, number and gender do not only unite central pronouns but also serve a fundamental role in finding the antecedent because anaphors and their antecedents usually have to show concord in these three features. Consequently, person, number and gender are of great impor-

[1] It should be mentioned that there are further classifications in the context of computational anaphora resolution. These classifications are not generally accepted but rather have been devised by individual researchers/authors. A selection of such classifications can be found with anaphora resolution systems in chapter 6.3.

tance for anaphora resolution. Furthermore, Quirk et al. (2012: 335-336, 346) state that central pronouns are by far the most important of all pronouns, especially personal pronouns, because of their frequency and grammatical features.

3.1.1 Personal pronouns

3.1.1.1 Subjective and objective forms

Personal pronouns divide up into subjective and objective forms, depending on the case that is required.[2] The subjective forms are *I*, *he*, *she*, *we* and *they*, and the corresponding objective forms are *me*, *him*, *her*, *us* and *them* respectively. *You* and *it* occur in both subjective and objective positions with one and the same form. The distinction between subjective and objective case forms goes back to the function a pronoun takes in a clause (see example (24) and (25); cf. Aarts & Aarts 1986: 48-49; Quirk et al. 2012: 335-339).[3]

(24) *He* was at home.
(25) I met *him*.

3.1.1.2 Person, number and gender

As mentioned in the introduction, the forms of personal pronouns distinguish between person, number and gender. As to person, personal pronouns fall into the categories of first person (*I*/*me*, *we*/*us*), second person (*you*) and third person (*he*/*him*, *she*/*her*, *it*, *they*/*them*). The first person is typically used for the speaker/writer (addresser) or a group including the speaker/writer. The second person typically denotes one or more hearers/readers (addressees) or a group of which the addressee is part. The third person is characteristically used for third parties that do not include addresser or addressee(s) (cf. Stirling & Huddleston 2010: 1463).

Furthermore, number distinguishes between singular and plural forms. The singular forms are *I*/*me*, *he*/*him*, *she*/*her* and *it*. The plural forms include *we*/*us*

[2] There is also a genitive form of pronouns – possessive pronouns – which constitutes a separate chapter (cf. chapter 3.1.2).

[3] There are five functions that constituents can fulfil in a clause: subject, verb, (direct or indirect) object, (subject or object) complement, adverbial. These functions are then realised by clauses or phrases. There are five types of phrases: noun phrase, verb phrase, adjective phrase, adverb phrase, prepositional phrase. For an overview see, for example, Quirk et al. (2012), pp. 49-59 for functions, pp. 1047-1048 for clauses, and pp. 60-67 for phrases.

and *they*/*them*. The form *you* is used for both singular and plural. This classification does not mean that a plural form always refers to plural entities because it is possible for plural forms to refer to expressions with singular meaning in some situations. For example, *they* in (26) is interpreted as referring to an expression in the singular (cf. Carter & McCarthy 2006: 376-382; Quirk et al. 2012: 343-345). Moreover, if a personal pronoun refers to a collective noun[4] such as *government*, singular and plural forms can be used (example (27) a) and b)). According to Quirk et al. (2012), the decision whether singular or plural pronouns are used indicates "a difference in point of view: the singular stresses the nonpersonal collectivity of the group, and the plural stresses the personal individuality within the group" (ibid.: 316). This is also reflected in the verb, which is singular or plural if it is in present tense. Third person -*s* occurs if the subject is understood as a unit (example (28) b)); the base form of the verb is used if the individuals are stressed (example (28) a)).

(26) **Someone** who has never been skiing will not know what equipment they will need.
(27) a) **The team** wins in every competition. It seems unbeatable.
b) **The team** have decided that they will not take part in the next competition.
(28) a) The team seem highly motivated.
b) The team seems highly motivated.

Finally, masculine, feminine and neuter gender is distinguished. Gender is principally not so important in English as in other languages because in English, gender depends on the sex of the person (cf. Biber et al. 2007: 311). Therefore, only third person singular has different forms of gender. These are *he* for masculine, *she* for feminine and *it* for neuter. Masculine and feminine forms are subsumed under "personal" gender, which is contrasted to the "nonpersonal" neuter form. Personal gender forms are not only used for human beings, but for all living beings that are regarded as belonging to the human race. Because of this, personal gender forms can also refer to supernatural beings, for example, gods and angels, or to higher animals such as dogs (example (29)) (cf. Halliday & Hasan 2008: 47; Quirk et al. 2012: 341).

It is used in cases where *he* or *she* is not acceptable, i.e. *it* can refer to things, abstractions or even to a clause, or one or more sentences. In sum, *it* can refer to "any identifiable portion of text" (Halliday & Hasan 2008: 52), which

[4] A "collective noun" is a "noun which denotes a group of entities" (Crystal 1994: 70).

Halliday & Hasan (2008: 52) call “text reference” and Quirk et al. (2012: 1461-1462) term “discourse reference”. A good example is (30) where *it* refers to the preceding sentence. Replacing the anaphor with the antecedent leads to *That David won the ski race was a great surprise*. Another term in this context is “extended reference”[5], as shown in example (31), that Schubert (2012: 36) and also Halliday & Hasan (2008: 52-53) use if the antecedent “is more than just a person or object, it is a process or sequence of processes (grammatically, a clause or string of clauses, not just a single nominal)” (Halliday & Hasan 2008: 52).

Apart from that, some rarer uses are found: for instance, *it* can also refer to children, particularly in scientific reports with an emotional distance to the human being (cf. Quirk et al. 2012: 316-317). Additionally, personal gender forms, normally used for people, serve to personify objects. For example, *she* can refer to ships, countries or cars, although many people object to such a use. Personification is common in informal language and is notably a means in fiction and poetry, where everything can, in fact, be personified (example (32)) (cf. Biber et al. 2007: 317-318).

As for the choice of masculine or feminine gender forms, this decision relies on the sex of the person or animal referred to. As may be known, discussions about gender neutrality and sexual bias in language began in the second half of the 20th century within the feminist movement in the USA (cf. Wawra 2004: 2). As a result, new forms and practices have found their way into English. In order to avoid mentioning only masculine gender forms with personal pronouns, the expressions *s/he* or *(s)he* have developed in writing. However, these items are not possible in speech. Other forms such as *he/she*, *he or she* or singular *they* are common in both speech and writing. As *he/she*, *he or she* is very formal, singular *they* is commonly used, particularly if the reference is to expressions such as *person*, *someone* or *anyone* (example (26)) (cf. Carter & McCarthy 2006: 376-380; Biber et al. 2007: 316-317; Payne & Huddleston 2010: 426, 492-494; Quirk et al. 2012: 341-343, 347-348).

(29) **My dog “Snoopy”** is very lazy. He always sleeps in the afternoon.
(30) **David won the ski race**. It was a great surprise.
(31) **Peel the potatoes!** At least think about it.

[5] Halliday & Hasan (2008: 52-53) distinguish between “extended reference” and “text reference”. However, Consten (2004) points out: “Halliday/Hasan (1976) prägen das Begriffspaar „extended reference“, deren Unterscheidung von „textual reference“ undurchsichtig bleibt.” (ibid.: 33).

(32) I cannot start **the computer**. **He** always refuses to work when I need him the most.

The classification of personal pronouns regarding person, number and gender is visualised in Table 1. With respect to case, the subjective case form is given first for those personal pronouns that have different forms for subjective and objective case.

Table 1: Personal pronouns

Number Person	Singular			Plural
1st	*I / me*			*we / us*
2nd	*you*			*you*
3rd	Masc.	*he / him*	*he/she / him/her,* *he or she / him or her,* *s/he, s(he), they / them*	*they / them*
	Fem.	*she / her*		
	Neuter	*it*		

3.1.1.3 Anaphoric and non-anaphoric use

In general, personal pronouns have definite meaning as they refer to entities that are identifiable without needing further information (cf. "Personal pronoun" n.d.; Stirling & Huddleston 2010: 1468; Quirk et al. 2012: 335).[6] This does not mean that all personal pronouns are anaphoric. In more detail, personal pronouns of first and second person refer to entities present in the specific situation, so they are typically used deictically. However, *we/us* can be used anaphorically if it refers to a group including the addresser, namely if this group or person apart from the addresser is mentioned explicitly (example (33)). Apart from *we* in this use, personal pronouns of first and second person could be attributed anaphoric use in dialogues if the person concerned is mentioned. For example, in *She said: 'I do not know him.'* the item *I* is related indirectly to *she* (cf. Halliday & Hasan 2008: 48-50). Such a use is not considered to be representing an anaphor here because the relation is indirect and therefore does not show real explicitness: *I* does not refer directly to *she*. *I* rather refers to the

[6] An example of an indefinite use of personal pronouns is *it* in *It was a nice evening* (cf. Swan n.d.).

speaker and the speaker herself is introduced by *she*. However, there is no need to mention the speaker and in fact he or she is often left unstated.

With regard to third person personal pronouns, "the characteristic use of the 3rd person personal pronouns ***he***, ***she***, ***it***, and ***they*** is anaphoric" (Stirling & Huddleston 2010: 1468) because the antecedent is usually present linguistically. The relationship between third person personal pronouns and their antecedents is commonly coreferential, since both refer to the same person or thing (cf. Siddiqui & Tiwary 2008: 185; Stirling & Huddleston 2010: 1465, 1468; Quirk et al. 2012: 865).

However, the third person forms *he/him*, *she/her*, *it* and *they/them* do not always take anaphoric interpretations in all contexts. First, these personal pronouns can refer to entities not present linguistically but identifiable from the context (cf. Stirling & Huddleston 2010: 1469-1470). Example (34) is a case in point. Here, the situation might show a man driving into a parking space and hitting a parked car. In this context, *him* and *he* refer to the man in this car.

Second, pronouns fall in between proper anaphoric and non-anaphoric uses in certain contexts. Stirling & Huddleston (2010) call these "quasi-anaphoric" uses (ibid.: 1470). The referent is not mentioned explicitly in such cases but interpretable from a related expression. For example, *they* refers to *Peter and his girlfriend* in (35), although only *Peter* occurs in the preceding sentence. It has to be inferred from the context that *and his girlfriend* is understood. Thus, *they* relates in some way to Peter but *Peter* is not itself the antecedent. For that reason, such expressions will not be considered here as they contradict the conditions for anaphors in this book because the antecedent is not present explicitly in the text. Moreover, these "quasi-anaphoric" expressions are common in informal language (cf. Stirling & Huddleston 2010: 1470-1471). Stirling & Huddleston (2010) remark that "in more carefully monitored speech or writing one would be more likely to use more explicit expressions" (ibid.: 1471).

(33) **Luke and I** know that Ms. Thomson is <u>our</u> neighbour, but <u>we</u> are not sure if she is married.
(34) Look at *him*! *He* is going to crash into that car.
(35) Peter called me yesterday. *They* are going to marry next week.

Third, personal pronouns of third person are non-anaphoric in generic use. Expressions show generic use if these "[refer] to an entire class of individuals, events, etc., rather than to specific members" (Matthews 2007: 156). To give an example, personal pronouns that are part of proverbs and colloquial idioms show generic use. Numbers (36) and (37) are instructive examples (cf. Speake

2008: 26, 302). Non-anaphoric instances in generic use are also found in sentences beginning with *He who…*, which occurs in proverbs (see example (38)) and is familiar from the Bible. Apart from these contexts, a generic use of personal pronouns is rare. *He* in generic use stands for any person or “people in general” (Quirk et al. 2012: 353), and is mostly used even sex-neutrally. In constructions where *he* refers to a male person, a corresponding expression for a female person, i.e. *She who…*, is possible as well.[7] Furthermore, *it* in generic use refers to life in general, as demonstrated by the idiom in (39). Finally, generic *they* can be used for “people in general” (example (40)). *They* is also used to refer to an authority or institution that is not mentioned explicitly in the text, such as the government or the media (example (41)) (cf. Halliday & Hasan 2008: 53; Stirling & Huddleston 2010: 1468-1472; Quirk et al. 2012: 347-354, 1467).

An additional non-anaphoric use must not be forgotten. Consider, for example, number (42) (cf. Speake 2008: 346). Here, the first *he* is not anaphoric, and the second, third and fourth occurrence of *he* refer to the preceding *he* respectively. If such personal pronouns refer back to items that have been examined regarding their anaphoric or non-anaphoric status, but were identified as being non-anaphoric, all pronouns referring to such non-anaphoric items will not be considered anaphors. Such a use might occur in proverbs (example (42)) but also in other contexts. This procedure is adopted because establishing such relations is not relevant for computational systems. In more detail, knowing that *he* refers to the preceding non-anaphoric *he* does not help to find out about the textual content.

(36) The bigger *they* are, the harder *they* fall.
(37) The more you stir *it* the worse *it* stinks.
(38) *He* who dares wins.
(39) How’s *it* going?
(40) *They* say the German team has the best chances of winning.
(41) *They* have increased taxes for petrol again.
(42) *He* that will not when *he* may, when *he* will *he* shall have nay.

[7] The *Comprehensive Grammar of the English Language* considers *he* and *she* in *He who…* and *She who…* cataphors. Thus, *he* in (38), for example, would refer to the postmodifier *who dares* (cf. Quirk et al. 2012: 352-353). However, such instances are not cohesive. Halliday & Hasan (2008) state in that context: “The reference is within the sentence, and is determined by the structure of the sentence” (ibid.: 56). As a result, they are treated as non-anaphoric here.

A non-anaphoric use of *it* is termed "pleonastic *it*" (Lappin & Leass 1994: 538-539; Mitkov 2002: 9), "prop *it*", "empty *it*", "expletive *it*" (Quirk et al. 2012: 348-349, 749) and occurs in two more instances, apart from generic use in proverbs and idioms. These concern only *it* and not any other personal pronouns. First, *it* is especially used together with verbs or predicative adjectives[8], denoting weather (example (43)), time (44) or place (45). Here, *it* only has the syntactic function of filling the subject position. Such clauses can often be reformulated and then result in clauses that do not include prop *it*. This reformulation is possible if a temporal clause contains both a subject complement that denotes a temporal state, and an adverbial as a noun phrase. Therefore, example (46) could be paraphrased as *Next week will be February 1*. Such reformulated clauses carry a similar meaning as prop-*it* clauses (cf. Quirk et al. 2012: 748-749). Apart from atmospheric, temporal and local conditions, *it* can also be used in utterances in which *this* can substitute *it* (example (47)).

(43) *It*'s sunny today.
(44) *It*'s half past one.
(45) *It*'s only a few hundred metres to the city centre.
(46) *It* will be February 1 next week.
(47) *It* was a good film.

Second, non-anaphoric *it* occurs in extraposition and cleft sentences.[9] In extraposition with *it*, the subject is postponed, and *it* fills this subject position. As a result, the sentence has two subjects, the notional subject found at the end of the sentence and *it* as the grammatical subject (cf. Hasselgård, Lysvåg & Johansson 2012; Quirk et al. 2012: 1403). A good example of extraposition is (48). The notional subject there is *that Linda won*. Such sentences with extraposition can be reformulated so that they do not contain *it*. Number (48) would then read *That Linda won surprised me* with the notional subject, and at the same time the grammatical subject, at the beginning of the sentence. Cleft sentences have the form of *it* plus *be*, which are followed by the expression on which the focus lies and a clause (see example (49)). The non-cleft version of (49) is *I started studying English last week*. Consequently, cleft sentences always place the stressed elements at the beginning, after *it* and *be* (cf. Aarts & Aarts 1986: 97-98; Stirling & Huddleston 2010: 1481-1483; Quirk et al. 2012: 348-349, 1384-1392).

[8] These are adjectives functioning as subject or object complement (cf. Quirk et al. 2012: 403).

[9] Similarly, such constructions could be regarded as cataphors (cf. Quirk et al. 2012: 349). This position is not adopted here.

(48) *It* surprised me that Linda won.
(49) *It* was last week that I started studying English.

3.1.1.4 Cataphoric use

Apart from anaphoric interpretation, cataphoric use is also possible with personal pronouns. In general, cataphors take either integrated or non-integrated antecedents. The antecedent of the integrated form is a constituent of a clause (example (50) a)), whereas the antecedent in the non-integrated type rather forms a separate clause or sentence (example (51) a)). The cataphor and the antecedent can often be reversed in the integrated form so that the cataphor turns into a "usual" anaphor. Such a change of positions in example (50) a) is shown in (50) b). An inversion, however, is not possible in all cases (see example (52)). In contrast, cataphor and antecedent in the non-integrated form cannot change their positions (example (51) a)). If the two expressions are changed here, this does not result in an anaphoric interpretation. Instead, the anaphor is not needed any more. For instance, the reversed order in example (51) a) does not need *it* (example (51) b)). Therefore, the possibility of changing positions is the decisive criterion to differentiate the non-integrated from the integrated type (cf. Stirling & Huddleston 2010: 1456).

Furthermore, Stirling & Huddleston (2010: 1476-1477) distinguish between first-mention and repeat-mention cataphors. With regard to first-mention cataphors, the cataphor itself introduces the entity into the text. Example (50) a), for instance, is a case of a first-mention cataphor. Number (53) demonstrates a repeat-mention cataphor. As will be noticed, repeat-mention cataphors do not mention an entity for the first time, since one or more expressions introduced that entity before. Such repeat-mention cataphors are not seen as cataphors proper here because expressions following the cataphor as well as one or more previous expressions can be viewed as antecedents.

(50) a) Although he is a fan of Arnold Schwarzenegger, **Frank** is not sure whether or not he should vote for him.
b) Although **Frank** is a fan of Arnold Schwarzenegger, he is not sure whether or not he should vote for him.
(51) a) It is now clear: **The dog has eaten the sausages**.
b) That the dog has eaten the sausages is now clear.
(52) Not only do I work with her, **Cindy** is also my best friend.
(53) Yesterday **Sue** had an appointment with her GP. As she was coming directly from her workplace, **Sue** forgot to bring her insurance card.

Cataphors with personal pronouns mainly occur in three constructions: within subordinate clauses, in a subordinate position within noun phrases, and within prepositional phrases at the beginning of clauses. If the pronoun is part of a subordinate clause, the antecedent is found in the rest of the main/superordinate clause, i.e. the matrix clause (cf. Quirk et al. 2012: 991).[10] It is also possible that the antecedent is located in another subordinate clause of this main/superordinate clause. Example (50) a) shows a cataphor in the subordinate clause (the part before the comma); the second clause, which is the matrix clause, contains the antecedent. Similarly, a cataphor can take a subordinate role within noun phrases. Example (54) is a case in point, in which *the constant gossip about him* is a noun phrase. Here, *about him* postmodifies the head *gossip*, and so takes a subordinate position within the noun phrase.

Finally, cataphors are found in prepositional phrases if these are preposed, i.e. occur in the front position of a sentence (example (55) a)). Instances such as (55) a) illustrate cases where a cataphor is even necessary. Thus, if the items are reversed with the pronoun following the prepositional phrase (example (55) b)), the meaning changes. As a result, example (55) b) implies that the spider is above another person; *she* is no anaphor. Constructions such as (55) a) can be reformulated in another way, with the prepositional phrase taking the position at the end of the sentence, as in (55) c). Such a change turns the item *her* into an anaphor (cf. Stirling & Huddleston 2010: 1477-1478, 1490).

(54) The constant gossip about him made **Geoffrey** nervous.
(55) a) Above her, **Tina** saw a spider.
b) Above Tina, *she* saw a spider.
c) **Tina** saw a spider above her.

In addition, there are three special cases in which a cataphor can be found. With these, the restrictions about the subordinate position or position in a prepositional phrase at the beginning of a clause mentioned above do not apply. In one case, cataphors are used for rhetorical effect. Stirling & Huddleston (2010) explain their role as follows:

> It is a quite common feature of journalism and novels to use anticipatory anaphora [i.e. cataphors] as a device to catch the listener's or reader's attention: pronouns are used to tempt the curious reader or listener into continuing to pay attention – so that they can

[10] According to Quirk et al. (2012: 991), the subordinate clause is part of the main/superordinate clause. The term "matrix clause" is used to refer to the part of the main/superordinate clause without the subordinate clause.

> find out who or what the pronoun refers to. In these cases reference by pronoun may continue across a number of sentences before a full NP [i.e. noun phrase] provides the required identification. (ibid.: 1480)

The second and third cases are instances of the non-integrated form. The antecedent can be the whole following main clause in one case (example (51) a)), or an expression attached to the clause, usually a noun phrase, in the other case (example (56)). The latter construction is termed “right dislocation”, i.e. an expression is added to a clause at its end, which is found more frequently in informal speech (cf. Stirling & Huddleston 2010: 1411-1412, 1480-1481; Quirk et al. 2012: 352, 1310).

(56) Do you know <u>them</u>, **Harry’s parents**?

3.1.1.5 Relationship between anaphor and antecedent

Personal pronouns are mostly coreferential with their antecedents, although not in all cases. If anaphors refer to expressions with quantifiers, no coreference arises (cf. Mitkov 2002: 6-7). A “quantifier” is “[a]ny word or expression which gives a relative or indefinite indication of quantity. [...] [It is] distinguished as such from a *numeral, which gives a precise and absolute indication of quantity” (Matthews 2007: 329). For example, the quantifier *few* in *few people* as contrasted with the numeral *three* in *three people* (cf. Aarts & Aarts 1986: 58 and Quirk et al. 2012: 376-380 for a list of items that are quantifiers). Example (57) illustrates an antecedent with the quantifier *every*. *She* relates in some way to *every woman* but cannot be replaced by the antecedent. If it were replaced, the meaning of the sentence would change: *Every woman knew that every woman had to give her best* is not the same as example (57). The situation can be described better by reformulating the sentence, namely *What every woman knew was that: I have to give my best* (cf. Stirling & Huddleston 2010: 1458, 1472-1475). Stirling & Huddleston (2010) explain:

> Because the variable expressed by *she* is within the scope of a quantifier, it is said to be **bound** by that quantifier: the pronoun here therefore expresses a **bound variable**. (ibid.: 1473)

(57) **Every woman** knew that <u>she</u> had to give her best.

Items referring to interrogative pronouns as antecedents would also establish no coreferential relationship between anaphor and antecedent. But as interrogative pronouns are sorted out in the process of identifying non-anaphoric relative pronouns (see chapter 3.4), all expressions referring to interrogative pronouns are not regarded as anaphors, for example, in *Who thinks they know the answer?* (cf. Stirling & Huddleston 2010: 1473-1474).

Additionally, no coreferential relationship arises if personal pronouns refer to a clause, or one or more sentences, which is only possible with *it* (example (30)). Furthermore, items are not coreferential if a cataphoric *it* refers to a following unit (example (51) a)). Esser (2009) states: "It must be pointed out that cataphoric reference is often not used to establish an overt referential relation in a text world (i.e. coreference) but rather to inform the reader of what comes next" (ibid.: 50).

As a result, these cases – whether anaphoric or cataphoric references to clauses, and references to expressions with quantifiers – fall neither into the category of coreference nor in that of substitution, and so are counted here to the miscellaneous category introduced in chapter 2.2.3 (cf. Stirling & Huddleston 2010: 1475; Quirk et al. 2012: 864, 868, 1461-1462).

3.1.1.6 Summary

Third personal pronouns take subjective and objective forms. Both forms are generally anaphoric, because "in writing an explicit referent will normally be required" (Halliday & Hasan 2008: 51). But from these, a variety of non-anaphoric uses of personal pronouns have to be marked off. Cataphoric uses are possible but restricted to certain constructions. Furthermore, the antecedent of personal pronouns is not restricted with regard to clause functions and so can take any function such as subject or object. Finally, personal pronouns are a form of reduction and mostly coreferential.

3.1.2 Possessive pronouns

3.1.2.1 Determinative and independent possessive pronouns

Possessive pronouns constitute the genitive form of central pronouns. Possessive pronouns fall into two classes, those with determinative function and those with independent function.[11] Determinative possessive pronouns encompass the

[11] Traditionally, the term "possessive pronoun" is applied to items of both classes, i.e. determinative uses are described with pronominal uses in one and the same chapter (see e.g. Esser 2009: 37-38; Quirk et al. 2012: 361-362). This also goes for reciprocal pronouns, demonstrative

forms *my*, *your*, *his*, *her*, *its*, *our*, *your* and *their*. Independent possessive pronouns take the forms *mine*, *yours*, *his*, *hers*, *its*, *ours*, *yours* and *theirs*. Obviously, *his* and *its* take the same form in both classes of possessive pronouns (example (58) and (59)). But yet, *its* in independent function is extremely rare. Additionally, *her* is a determinative possessive pronoun but has the same form in the objective position of personal pronouns (cf. Quirk et al. 2012: 70-71, 361).

The difference between the two classes of possessive pronouns is pointed out here: as the name suggests, a determinative possessive pronoun occurs within a noun phrase and here takes the function of a determinative[12] (example (60)). In contrast, an independent possessive pronoun is head of a noun phrase. An example of an independent possessive pronoun as object is (61), in which the anaphor refers to both *Sally* and *a working calculator* (cf. ibid.: 330-331, 336, 361-363).

Furthermore, it could be argued that independent possessive pronouns in subject and object position show ellipsis, in the way that only a possessive pronoun is used and the noun elided. An instructive example is (59), which contains the independent possessive pronoun *his*. *His* refers to both *Sam* and *the fountain pen* at the same time. It could be argued, though, that *his* elides the noun *fountain pen*. As a result, Quirk et al. (2012: 891) speak of "virtual ellipsis" in the case of *his* and *its*, and of "quasi-ellipsis" in instances such as (61). However, such instances are not treated with the category of ellipsis but with possessive pronouns (cf. ibid.: 361-363). Even Quirk et al. (2012) acknowledge: "Whether quasi-ellipsis or virtual ellipsis are to be treated as cases of ellipsis or as cases of substitution is a matter of definition" (ibid.: 891).

(58) We visited **John** and saw <u>his</u> new flat.
(59) **Sam** always leaves things behind. **The fountain pen** is <u>his</u>.
(60) Did **Linda** leave <u>her</u> documents at home?
(61) **Sally** has **a working calculator**. You can borrow <u>hers</u>.

pronouns, relative pronouns and indefinite pronouns. However, Stirling & Huddleston (2010: 1499, 1504) distinguish between determinative and independent uses. When describing items that take functions both as determinatives and as pronouns, they speak of "demonstratives" instead of "demonstrative pronouns", for instance. This book follows traditional usage.

[12] A noun phrase distinguishes the functions determinative, premodification, head and postmodification. For example, the noun phrase *his early arrival in London* contains the determinative *his*, the premodification *early*, the head *arrival* and the postmodification *in London*. See Quirk et al. (2012: 60-62, 253-257) for more information about the functions of noun phrase elements and about determinatives.

3.1.2.2 Person, number and gender

Possessive pronouns distinguish between person, number and gender in the way personal pronouns do. With regard to person, first person pronouns are *my*/*our* and *mine*/*ours*. Second person has the forms *your* and *yours*. Finally, possessive pronouns of third person comprise determinative *his*, *her*, *its*, *their* and independent *his*, *hers*, *its*, *theirs* (cf. Quirk et al. 2012: 339-340).

Moving to number, the singular forms are *my*, *your*, *his*, *her*, *its* and *mine*, *yours*, *his*, *hers*, *its*. They are distinguished from the plural forms of *our*, *your*, *their* and *ours*, *yours*, *theirs*. The plural form *their*, for instance, can refer to singular forms, such as *everybody* in example (62), much the same as personal pronouns do. In addition, we find cases with collective nouns, i.e. expressions in the singular form such as *government*. If they are understood as a group consisting of individuals (example (63) a)) rather than as abstract units (example (63) b)), the pronoun is in the plural rather than in the singular. This distinction is also represented in the verb. In the case of *be*, for instance, the third person singular form *is* and the third person plural form *are* are used in the present tense (see example (63)). Furthermore, the third person singular form *was* and the third person plural form *were* can be differentiated in the past tense (cf. ibid.: 339-340, 1467).

Finally, third person singular distinguishes between forms of gender. Personal gender encompasses *his* as masculine, *her* and *hers* as feminine forms. *Its* is the only nonpersonal, i.e. neuter, form. Personal gender forms show close similarities to personal pronouns: possessive pronouns can also be found in place of expressions apart from human beings. Furthermore, personification is possible as well. Gender neutrality with possessive pronouns arises from using the form *their*, which is often preferred to the cumbersome formulation *his or her* (cf. Carter & McCarthy 2006: 382-383; Biber et al. 2007: 331-332; Payne & Huddleston 2010: 493-495; Quirk et al. 2012: 336-343, 770-771).

(62) **Everybody** should do <u>their</u> own training.
(63) a) **The government** are improving <u>their</u> programme to help people in need.
b) **The government** is improving <u>its</u> programme to help people in need.

The differentiation regarding person, number and gender is summarised in Table 2. The first item in each cell is the determinative form of the possessive pronouns; the element after the slash constitutes the independent form.

Table 2: Possessive pronouns

Number Person	Singular			Plural
1st	*my / mine*			*our / ours*
2nd	*your / yours*			*your / yours*
3rd	Masc.	*his / his*	*his/her / his/hers,* *his or her / his or hers,* *their / theirs*	*their / theirs*
	Fem.	*her / hers*		
	Neuter	*its / its*		

3.1.2.3 Anaphoric/cataphoric and non-anaphoric use

In the same way as personal pronouns are anaphoric, possessive pronouns show anaphoricity. This means that possessive pronouns of third person typically carry anaphoric reference, both in determinative and independent function. With regard to determinative possessive pronouns, it could be argued that first and second person are also anaphoric if referring to *I*, *we* or *you*. Such cases are not explicitly discussed in the literature, but in this book, such instances are not seen as anaphors. *I*, *we* and *you* usually represent speaker/writer (addresser) and hearer/reader (addressee). Yet, *our* is seen as anaphor, analogous to *we*, if it refers explicitly to a third person unit apart from the addresser (example (33)) (cf. Stirling & Huddleston 2010: 1463-1466, 1468-1469).

As regards independent possessive pronouns, not only third person pronouns can be attributed anaphoric reference, though. The anaphoricity of first and second person is not found in Quirk et al. (2012) and Stirling & Huddleston (2010). It seems, however, justified to see first and second person as anaphors as well, at least in a certain way. It is undeniable that independent possessive pronouns have two elements as their antecedents. Even Halliday & Hasan (2008) argue for third person possessive pronouns that they are "doubly anaphoric […] (i) by reference, to the possessor and (ii) by ellipsis, to the thing possessed" (ibid.: 55). Although *mine*, *ours* and *yours* refer to the addresser, addressee or a group in which these are part in one aspect, they refer to some other element in the second aspect. It is the second aspect that is the reason for treating such instances as anaphors. It will be readily apparent that *yours* in example (64) refers to the addressee in one part and to *car* in the second part. Only *car* is the antecedent that is considered computationally relevant. As for *ours*, it is possible that the part that refers to a person or group apart from the addresser is mentioned explicitly. In that case, *ours* is considered in both parts.

As to third person possessive pronouns, these also show non-anaphoric use. The contexts do not deviate much from personal pronouns. They involve sentences in which the antecedent is not present linguistically (example (65)) and quasi-anaphoric uses (example (66)). Moreover, proverbs but also sentences in other contexts can contain possessive pronouns that refer to non-anaphoric personal pronouns (example (67)). Finally, *their* and *theirs* can be non-anaphoric if these refer to an authority, an institution or people in general, i.e. anybody/anything that has not been mentioned explicitly in the text (example (68)) (cf. Speake 2008: 319; Stirling & Huddleston 2010: 1468-1472).

(64) Since my **car** did not work, I borrowed <u>yours</u>.
(65) Did you take *his* pen?
(66) Andy sent us pictures from *their* wedding.
(67) He that will thrive must first ask *his* wife.
(68) We do not like *their* programme.

When examining the cataphoric use of possessive pronouns, they can be compared to the description given with personal pronouns. The three central constructions identified above need no further adaption or extension. For the non-integrated forms, a few aspects need discussion, though. As only the personal pronoun *it* refers to the following sentence, such use is not possible with possessive pronouns. Furthermore, right dislocation seems to occur with personal pronouns only. That leaves cases in which cataphors occur for rhetorical effect. Example (69) shows a cataphor in a preposed prepositional phrase exemplarily (cf. Stirling & Huddleston 2010: 1477-1478, 1480-1481).

(69) For <u>her</u> spare time, **Linda** enjoys reading books by Sir Walter Scott.

3.1.2.4 Relationship between anaphor and antecedent

Anaphoric possessive pronouns in determinative function and their antecedents show a coreferential relationship in most instances. Some exceptions that have been mentioned with personal pronouns apply to possessive pronouns as well. Consequently, antecedents with a quantifier (example (70)) and items referring back to interrogative pronouns are not coreferential. In this case they are included in the miscellaneous category.

The relationship is different with independent possessive pronouns: they show substitution and coreference. For example, *hers* in (71) is a substitutional form of the noun phrase *her car*, which, furthermore, includes *her* as coreferential relationship. Consequently, they are classified into the miscellaneous category because they do neither fall clearly into the category of coreference nor into the category of substitution. In the case of first and second person, the relationship is also substitutional and coreferential. However, as the part with coreference is usually not relevant for these pronouns, they are classified referring to the substitutional part only (example (64)) (cf. Halliday & Hasan 2008: 55; Esser 2009: 37).

(70) **Each person** knows <u>their</u> name.

(71) Martin and **Lucy** wore similar T-shirts yesterday. His **T-shirt** was brown, <u>hers</u> was yellow.

3.1.2.5 Summary

Possessive pronouns divide up into determinative and independent uses, among which third person pronouns are usually anaphoric. Independent possessive pronouns of first and second person are anaphors as well, at least in a particular way. Cataphoric and non-anaphoric uses are in major parts similar to personal pronouns. Possessive pronouns are a form of reduction, as are personal pronouns. The relationship between anaphor and antecedent is coreferential in most instances of determinative possessive pronouns. Some determinative possessive pronouns belong to the miscellaneous category; the cases are analogous to personal pronouns. First and second person independent possessive pronouns show a substitutional relationship and third person independent possessive pronouns belong to the miscellaneous category because they are both coreferential and substitutional.

3.1.3 Reflexive pronouns

The third subtype of central pronouns is reflexive pronouns. According to their name, reflexive pronouns "'reflect' another nominal element of the clause or sentence, usually the subject, with which it is in a coreferential relation [...]" (Quirk et al. 2012: 356).[13] Reflexive pronouns are formed from the first and second determinative possessive pronouns and the objective forms of personal pronouns. This results in the forms *myself*, *yourself*, *himself*, *herself*, *itself*, *our-*

[13] "Nominal" means "[p]ertaining to nouns or to projections of nouns" (Trask 1993: 183).

selves, *yourselves* and *themselves*. Moreover, the generic form *oneself* for people in general could be mentioned (cf. ibid.: 356, 865).

3.1.3.1 Basic and emphatic use

Reflexive pronouns can be used in two ways, in basic (example (72)) or in emphatic use. The position of emphatic reflexive pronouns is variable. As a result, alternatives of example (73) a) are (73) b) and (73) c) (cf. Stirling & Huddleston 2010: 1488-1493; Quirk et al. 2012: 355-361).

(72) **Andy** blamed himself.
(73) a) **Betty** herself can do the homework.
b) **Betty** can do the homework herself.
c) **Betty** can herself do the homework.

The antecedent takes the following functions in a clause: in basic use, the antecedent is usually the subject of the clause, though there are other possibilities. Thus, the antecedent can be object (example (74)), or can even be found in a different clause as is the case with cleft sentences (example (75)). With regard to emphatic use, the antecedent is the element to which it has an appositional relation (cf. Biber et al. 2007: 343; Stirling & Huddleston 2010: 1486-1493, 1496; Quirk et al. 2012: 355-361, 387). The term "apposition" describes the "syntactic relation in which an element is juxtaposed to another element of the same kind" (Matthews 2007: 24). It might happen that the subject, and therefore the antecedent, is implied in some clauses, e.g. in *-ing*-participle clauses. In example (76), the participle clause has an implied subject, which is *Toby*. It is consequently important that such *-ing*-participle clauses are considered anaphors so that such references can be resolved (see chapter 3.12.2).

Another noteworthy fact is that reflexive pronouns contrast with personal pronouns of the objective form when looking for antecedents to anaphors. Accordingly, in example (77) the antecedent of *herself* is *Mary*. However, the antecedent of *her* cannot be *Mary* but has to be a different female person, which is not mentioned in this example (cf. Carter & McCarthy 2006: 385; Stirling & Huddleston 2010: 1484, 1489, 1492; Quirk et al. 2012: 356-357).

(74) The children asked **Mary** about herself at the age of five.
(75) It was for himself that **Tim** bought the chocolate.
(76) In cooking the meal himself, **Toby** succeeded in surprising the family.
(77) **Mary** talked to $\left\{\begin{array}{l}\textit{her.}\\ \text{herself.}\end{array}\right.$

3.1.3.2 Person, number and gender

As with personal and possessive pronouns, reflexive pronouns also distinguish between person, number and gender. First, person differentiation falls into first person with *myself* and *ourselves*, second person with *yourself* and *yourselves*, and third person encompasses *himself*, *herself*, *itself* and *themselves*. Second, number comprises the singular forms *myself*, *yourself*, *himself*, *herself* and *itself* and the plural forms *ourselves*[14], *yourselves* and *themselves*. There is also the form *themself* for singular entities (cf. Payne & Huddleston 2010: 426, 494). As a result, *themself* and *themselves* can refer to singular *they* (cf. chapter 3.1.1.2). As can readily be seen, singular reflexive pronouns end in *-self* and plural ones in *-selves*. In addition, singular and plural forms can be used with collective nouns such as *government*, depending on whether the focus lies on the group as an abstract entity (example (78) a)) or more on the people as individuals (example (78) b)). Such a use follows the rules as outlined above for possessive pronouns (cf. Quirk et al. 2012: 316-317, 771).

(78) a) **The government** committed itself to the proposed austerity measures.
b) **The government** committed themselves to the proposed austerity measures.

Third, gender distinctions encompass personal gender with *himself* for masculine and *herself* for feminine use, and nonpersonal gender, which is *itself*. As with personal and possessive pronouns, reflexive pronouns can be used for entities apart from human beings. Furthermore, reflexive pronouns can occur in personification. Gender neutral formulations stem from the use of *himself or herself*, *himself/herself*, *themself* or *themselves* (cf. Payne & Huddleston 2010: 493-494; Quirk et al. 2012: 339-345). The categories of person, number and gender with reflexive pronouns are listed in Table 3.

Table 3: Reflexive pronouns

Number	Singular	Plural
Person		
1st	*myself*	*ourselves*
2nd	*yourself*	*yourselves*

[14] There is also a singular form of *ourselves* which is *ourself*. The form *ourself* refers to royal *we*, but is very rarely used (cf. Quirk et al. 2012: 344, 356).

3rd	Masc.	*himself*	*himself/herself,* *himself or herself,* *themself, themselves*	*themselves*
	Fem.	*herself*		
	Neuter	*itself*		

3.1.3.3 Anaphoric/cataphoric and non-anaphoric use

With regard to anaphoric use, reflexive pronouns are similar to personal pronouns in that only third person pronouns typically carry anaphoric function. But yet, first and second person of reflexive pronouns could also be seen as showing some type of anaphoric reference. Stirling and Huddleston (2010: 1485), for example, argue that these pronouns have both deictic and anaphoric function. According to them, reflexive pronouns are deictic on the one hand because first and second person refer to addresser or addressee. On the other hand, reflexive pronouns are simultaneously anaphoric because they are linked to the antecedent, which is *I*, *we* or *you*. Example (79) shows first person, example (80) an instance of second person.

However, reflexive pronouns in first and second person are usually not seen as anaphors here. This goes back to the fact that first and second person reflexive pronouns only refer to addresser and addressee and not to other content words, as independent possessive pronouns do. As will be known, independent possessive pronouns consist of two parts: They refer to the addresser or addressee and the entity concerned. Only in their reference to the entity are they considered anaphors. There is one exception with reflexive pronouns: if *ourselves* refers to a third person mentioned explicitly, apart from the addresser, it is considered anaphoric, which is in analogy with *we*, *us* and *our* (cf. Stirling & Huddleston 2010: 1477-1478, 1490-1496).

(79) I carried the bags *myself.*
(80) You cannot carry the bag *yourself*!

Furthermore, some third person reflexive pronouns are treated as non-anaphoric. This includes sentences where the reflexive pronoun refers to a third person personal pronoun that is itself no anaphor. In doing so, these third person reflexive pronouns are regarded in the same way as first and second person ones. The contexts where this occurs are similar to personal pronouns, e.g. quasi-anaphoric uses or uses in proverbs (cf. ibid.: 1477-1478, 1490-1496).

In addition, reflexive pronouns can be used cataphorically. This involves preposed prepositional phrases, as is the case with personal and possessive pronouns. However, a reflexive pronoun can also be preposed even if it is not a prepositional complement (example (81) a)), although such a use is rare. The non-preposed construction, which then contains an anaphoric reference, is given in (81) b). Constructions where the anaphor is in a subordinate clause or in a subordinate position within a noun phrase seem not to be possible with reflexive pronouns because reflexive pronouns are more tied to the antecedent. Nevertheless, cataphor and antecedent can be located in different clauses. Number (82) demonstrates a construction where the antecedent is located in a subordinate clause (cf. Stirling & Huddleston 2010: 1477-1478, 1490-1496; Quirk et al. 2012: 361).

(81) a) (For) herself, **Chloe** bought some ice cream.
b) **Chloe** bought herself some ice cream.
(82) It was for themselves that **the friends** organised the party.

3.1.3.4 Summary

Reflexive pronouns show basic and emphatic use. As for anaphoric, cataphoric and non-anaphoric use, reflexive pronouns are similar to personal and possessive pronouns. For instance, third person reflexive pronouns are not anaphoric if they refer to a non-anaphoric personal pronoun. The antecedent can be subject, object or it is found in a different clause in basic use. The antecedent of anaphors in emphatic use is the item to which the anaphor has an appositional relation. Finally, reflexive pronouns are reductive and show a coreferential relationship.

3.1.4 Summary of personal, possessive and reflexive pronouns

From all central pronouns, third person pronouns are typically anaphoric. Additionally, items of first and second person with independent possessive pronouns are attributed a special anaphoric function. The items *we*, *us*, *our* and *ourselves* are anaphoric if they refer to an explicit third person expression. A cataphoric use is limited to certain constructions. Third person central pronouns also work non-anaphorically in specific situations.

Furthermore, central pronoun anaphors show reduction and contribute to the grammatical cohesion of a text. The relationship between anaphor and antecedent is mostly coreferential. Third person independent possessive pronouns belong to the miscellaneous category; independent possessive pronouns of second and third person show substitution. If the antecedent is a clause or sentence or if it includes a quantifier, they belong to the miscellaneous category because they are neither coreferential nor substitutional. The categories *person*, *number* and *gender* as well as identifying the functions of a clause can help in finding the correct antecedent.

An overview of important aspects with central pronouns, as described in this chapter, is given in Tables 4, 5 and 6. They lay out whether specific features apply to an item (marked by "×") or not (marked by "-"). The tables particularly summarise information that is relevant for the distinction between anaphoric and non-anaphoric use and the identification of the correct antecedent of each anaphor.

References from Table 4:

[1] Only relevant if person/group apart from addresser is given

[2] Being regarded only in their nonpersonal part of reference

[3] For collective nouns; to avoid gender bias

[4] May occur in the nonpersonal part of reference

[5] Only in science

[6] Only in personification

[7] Only in cleft sentences

Table 4: Anaphoric use of central pronouns

	Number		**Person/gender**					**Antecedent**			
	Singu-lar	Plural	Human		Non-human		Material & non-material	Subject	Object	Clause/ sentence	Located in another clause
			Female	Male	Beings close to human race	Other beings					
Personal pronouns											
he	×	-	-	×	×	×[6]	×[6]	×	×	-	×
him	×	-	-	×	×	×[6]	×[6]	×	×	-	×
she	×	-	×	-	×	×[6]	×[6]	×	×	-	×
her	×	-	×	-	×	×[6]	×[6]	×	×	-	×
it (subjective form)	×	-	×[5]	×[5]	×	×	×	×	×	×	×
it (objective form)	×	-	×[5]	×[5]	×	×	×	×	×	-	×
we[1]	×	×	×	×	×	×[6]	×[6]	×	×	-	×
us[1]	×	×	×	×	×	×[6]	×[6]	×	×	-	×
they	×[3]	×	×	×	×	×	×	×	×	-	×
them	×[3]	×	×	×	×	×	×	×	×	-	×
he/she, *he or she*, *s/he*, *s(he)*	×	-	×	×	-	-	-	×	×	-	×
him/her, *him or her*	×	-	×	×	-	-	-	×	×	-	×
Determinative possessive pronouns											
his	×	-	-	×	×	×[6]	×[6]	×	×	-	×
her	×	-	×	-	×	×[6]	×[6]	×	×	-	×
its	×	-	×[5]	×[5]	×	×	×	×	×	-	×
our[1]	×	×	×	×	×	×[6]	×[6]	×	×	-	×
their	×[3]	×	×	×	×	×	×	×	×	-	×
his/her, *his or her*	×	-	×	×	-	-	-	×	×	-	×

Independent possessive pronouns											
mine[2]	×	×	×	×	×	×	×	×	×	-	×
yours[2]	×	×	×	×	×	×	×	×	×	-	×
his	×	×[4]	-	×	×	×[6]	×[6]	×	×	-	×
hers	×	×[4]	×	-	×	×[6]	×[6]	×	×	-	×
its	×	×[4]	×[5]	×[5]	×	×	×	×	×	-	×
ours[(2)]	×	×	×	×	×	×	×	×	×	-	×
theirs	×[3, 4]	×	×	×	×	×	×	×	×	-	×
his/hers, his or hers	×	-	×	×	-	-	-	×	×	-	×
Reflexive pronouns: basic use											
himself	×	-	-	×	×	×[6]	×[6]	×	×	-	×[7]
herself	×	-	×	-	×	×[6]	×[6]	×	×	-	×[7]
itself	×	-	×[5]	×[5]	×	×	×	×	×	-	×[7]
ourselves[1]	×	×	×	×	×	×[6]	×[6]	×	×	-	×[7]
themself	×	-	×	×	×	×	×	×	×	-	×[7]
themselves	×[3]	×	×	×	×	×	×	×	×	-	×[7]
himself/herself, himself or herself	×	-	×	×	-	-	-	×	×	-	×[7]
Reflexive pronouns: emphatic use											
himself	×	-	-	×	×	×[6]	×[6]	×	×	-	-
herself	×	-	×	-	×	×[6]	×[6]	×	×	-	-
itself	×	-	×[5]	×[5]	×	×	×	×	×	-	-
ourselves[1]	×	×	×	×	×	×[6]	×[6]	×	×	-	-
themself	×	-	×	×	×	×	×	×	×	-	-
themselves	×[3]	×	×	×	×	×	×	×	×	-	-
himself/herself, himself or herself	×	-	×	×	-	-	-	×	×	-	-

Table 5: Non-anaphoric use of central pronouns

	Antecedent not present in text	**Quasi-anaphoric**	**Generic**			**Reference to items identified as non-anaphoric**	**Atmospheric, temporal, local condition & cases where replacing with *this* is possible**	**Cleft sentences & extra-position**
			Proverbs & idioms	People in general	Authority & institution			
Personal pronouns								
he	×	-	×	×	-	×	-	-
him	×	-	×	×	-	×	-	-
she	×	-	×	×	-	×	-	-
her	×	-	×	×	-	×	-	-
it (subjective form)	×	-	×	×	-	×	×	×
it (objective form)	×	-	×	×	-	×	-	-
we	×	×	×	×	-	×	-	-
us	×	×	×	×	-	×	-	-
they	×	×	×	×	×	×	-	-
them	×	×	×	×	×	×	-	-
he/she, *he or she*, *s/he*, *s(he)*	×	-	×	×	-	×	-	-
him/her, *him or her*	×	-	×	×	-	×	-	-
Determinative possessive pronouns								
his	×	-	×	-	-	×	-	-
her	×	-	×	-	-	×	-	-
its	×	-	×	-	-	×	-	-
our	×	×	×	×	-	×	-	-
their	×	×	×	×	×	×	-	-
his/her, *his or her*	×	-	×	-	-	×	-	-

Independent possessive pronouns								
mine	×	-	×	-	-	×	-	-
yours	×	×	×	-	-	×	-	-
ours	×	×	×	×	-	×	-	-
his	×	-	×	-	-	×	-	-
hers	×	-	×	-	-	×	-	-
its	×	-	×	-	-	×	-	-
theirs	×	×	×	×	×	×	-	-
his/hers, his or hers	×	-	×	-	-	×	-	-
Reflexive pronouns (basic and emphatic use)								
himself	-	-	×	-	-	×	-	-
herself	-	-	×	-	-	×	-	-
itself	-	-	×	-	-	×	-	-
ourselves	-	×	×	-	-	×	-	-
themself	-	-	×	-	-	×	-	-
themselves	-	×	×	-	-	×	-	-
himself/herself, himself or herself	-	-	×	-	-	×	-	-

Table 6: Cataphoric use of central pronouns

	Integrated					**Non-integrated**	
	In a subordinate clause	In subordinate position within a noun phrase	At the beginning of a sentence		For rhetorical effect	The following clause	Right dislocation
			In prepositional phrases	As head of noun phrases (on their own)			
Personal pronouns							
he	×	×	×	-	×	-	×
him	×	×	×	-	×	-	×
she	×	×	×	-	×	-	×
her	×	×	×	-	×	-	×
it (subjective form)	×	×	×	-	×	×	×
it (objective form)	×	×	×	-	×	-	×
we	×	×	×	-	×	-	×
us	×	×	×	-	×	-	×
they	×	×	×	-	×	-	×
them	×	×	×	-	×	-	×
he/she, *he or she*, *s/he*, *s(he)*	×	×	×	-	(×)	-	×
him/her, *him or her*	×	×	×	-	(×)	-	×
Determinative possessive pronouns							
his	×	×	×	-	×	-	-
her	×	×	×	-	×	-	-
its	×	×	×	-	×	-	-
our	×	×	×	-	×	-	-
their	×	×	×	-	×	-	-
his/her, *his or her*	×	×	×	-	(×)	-	-

Independent possessive pronouns							
mine	×	×	×	-	×	-	-
yours	×	×	×	-	×	-	-
ours	×	×	×	-	×	-	-
his	×	×	×	-	×	-	-
hers	×	×	×	-	×	-	-
its	×	×	×	-	×	-	-
theirs	×	×	×	-	×	-	-
his/hers, his or hers	×	×	×	-	(×)	-	-
Reflexive pronouns (basic and emphatic use)							
himself	-	-	×	×	-	-	-
herself	-	-	×	×	-	-	-
itself	-	-	×	×	-	-	-
ourselves	-	-	×	×	-	-	-
themself	-	-	×	×	-	-	-
themselves	-	-	×	×	-	-	-
himself/herself, *himself or herself*	-	-	×	×	-	-	-

Traditional grammar differentiates, apart from the three pronoun categories previously mentioned, between reciprocal, demonstrative, relative, interrogative and indefinite pronouns (cf. Sasse 1993: 669). Reciprocal pronouns will be discussed in chapter 3.2. Demonstrative and relative pronouns will be considered in chapters 3.3 and 3.4, and indefinite pronouns in chapter 3.8. They all show characteristics that they do not have in common with central pronouns. Interrogative pronouns are not discussed here because they cannot work anaphorically and cohesively, and so do not relate to an antecedent (cf. Halliday & Hasan 2008: 309; Huddleston, Pullum & Peterson 2010: 1037; Stirling & Huddleston 2010: 1462).

3.2 Reciprocal pronouns

This is a category with minor importance because its items do not occur frequently (cf. Quirk et al. 2012: 365). There are only two forms, *each other* and *one another*, which are both coreferential with their antecedent. Reciprocal pronouns resemble reflexive pronouns and, as Quirk et al. (2012) put it, "[they] are related to the reflexive pronouns in that they can be said to express a 'two-way reflexive relationship'" (ibid.: 364). But yet differences between reflexive and reciprocal pronouns can be found as well, which can best be seen if the two pronoun types are contrasted. Example (83) a) contains a reflexive pronoun, example (83) b) a reciprocal one. The difference in meaning becomes evident if these sentences are paraphrased. As a result, example (83) a) can be reformulated as *Tom saw himself, and Sue saw herself*. But example (83) b) has to be paraphrased by *Tom saw Sue, and Sue saw Tom*. This means that reciprocal pronouns imply mutual relationship (cf. Sasse 1993: 675; Stirling & Huddleston 2010: 1502; Quirk et al. 2012: 345-346).

In addition, reciprocal, but not reflexive pronouns, have genitive forms, which are *each other's* and *one another's* (example (84)). Finally, reflexive pronouns and reciprocal pronouns take a different status within pronouns: "the reflexives [i.e. reflexive pronouns] are inflectional forms of the personal pronouns, while the reciprocals [i.e. reciprocal pronouns] are independent pronouns" (Stirling & Huddleston 2010: 1502).

(83) a) **Tom and Sue** looked at <u>themselves</u> in the mirror.
b) **Tom and Sue** looked at <u>each other</u>.
(84) **The twins** used <u>each other's</u> car.

3.2.1 Compound and split construction

Both *each other* and *one another* show two uses. The elements of each form can occur as a compound together (example (85)) or they can be separated and result in a split construction (example (86)). As compounds, reciprocal pronouns function as pronouns. In split construction, *each* and *one* function as determinatives of a noun phrase, *(an)other* is a noun and functions as head of a different noun phrase. The split construction distinguishes between items of two (example (86)) and three or more (example (87)), in which *other* occurs in the singular or plural (cf. Stirling & Huddleston 2010: 1499).

The two reciprocal pronouns are semantically the same as a compound, but *each other* is more frequent and commonly used in informal contexts, whereas *one another* is more typical in formal situations (cf. Quirk et al. 2012: 364). *Each other* and *another one*, however, behave differently in the split construction concerning syntax and semantics. Syntactically, they differ in which position they occur in a sentence. Semantically, *one* plus *(an)other* can be used for more reciprocal relations than *each* plus *other* (cf. Stirling & Huddleston 2010: 1499-1501).

(85) **The children** told each other a story.
(86) Each **child** told the other a story.
(87) Each of the **pupils** knows the others.

3.2.2 Anaphoric/cataphoric and non-anaphoric use

Reciprocal pronouns are usually anaphoric in compound and split construction. But if they are split, this leads to one construction in which the elements are not reciprocal, though anaphoric. This is if *one* functions as determinative of a noun phrase in subject position (example (88)). Such examples are regarded as instances of indefinite pronouns (see chapter 3.8) because only *other* refers back to the antecedent (cf. Stirling & Huddleston 2010: 1501). Moreover, if reciprocal pronouns refer to items that are themselves non-anaphoric, the expression is not seen as an anaphor, which is in analogy with central pronouns.

With regard to the antecedent, the following features can be observed. The antecedent of reciprocal pronouns is usually a noun phrase with a plural noun as head (example (85)) or it is an expression of noun phrases coordinated with *and* (example (89)). This is unlike reflexive pronouns, which can refer to a singular antecedent. The antecedent of reciprocal pronouns can have a singular form in some cases but at the same time this item has to imply a group of people. Hence, *everyone* in example (90) is meant to include more than one person. Other singular instances are collective nouns, as in example (91). Moreover, antecedents may also be implicit, as with reflexive pronouns. To give an example, *each other* in (92) refers to the implied subject in the *-ing*-participle clause, which is *the group members* (cf. Stirling & Huddleston 2010: 1501-1504; Quirk et al. 2012: 365).

(88) One **event** follows the other.
(89) **The girls and their boyfriends** laughed at each other.
(90) **Everyone** greeted each other.
(91) **The couple** likes each other.
(92) Looking at each other, **the group members** finally agreed on the plan.

In addition, reciprocal pronouns can have a cataphoric interpretation. This is restricted to constructions in which the reciprocal pronoun and the antecedent are attached to the same noun. It is possible with prepositions, as example (93) shows. Both anaphor and antecedent are prepositional phrases introduced by *of* (cf. Stirling & Huddleston 2010: 1503).

(93) The film is about the betrayal of each other of **the friends**.

3.2.3 Summary

Reciprocal pronouns can be used in compound and split constructions. With regard to compound reciprocal pronouns, it can be said that they resemble reflexive pronouns. As to split reciprocal pronouns, it is important to mention that *one* is not a reciprocal within subject position if occurring with *(an)other*, but falls into the category of indefinite pronouns. The antecedent of reciprocal pronouns is usually plural, but can also be singular in specific circumstances. Furthermore, the antecedent can be implied, or it is found after the anaphor and therefore establishes cataphoric reference. As reflexive pronouns, reciprocal pronouns are coreferential with their antecedent and reductive devices. A visual summary is given in Table 7.

Table 7: Characteristics of reciprocal pronouns

	Anaphoric				**Cataphoric**	**Non-anaphoric**	
	Antecedent is singular		Antecedent is plural		Cataphor and antecedent attached to the same noun	In subject position	Reference to non-anaphoric items
	Collective nouns	Indefinite pronouns e.g. *everyone*	Plural noun phrases	Singular noun phrases coordinated with *and*			
Compound construction							
each other	×	×	×	×	×	-	×
one (an)other	×	×	×	×	×	-	×
Split construction							
each … other	×	×	×	×	×	-	×
one … other	×	×	×	×	×	×[1]	×

[1] Is anaphoric, but as it belongs to indefinite pronouns it is regarded as non-anaphoric for reciprocal pronouns

3.3 Demonstrative pronouns

3.3.1 Dependent and independent demonstrative pronouns

This category includes *this*, *that*, *these* and *those*, which are either used dependently or independently. With regard to dependent function, they work as determinative or modifier. Thus, *this* in example (94) determines *instrument*. The whole expression *this instrument* is an anaphor with the antecedent *the guitar*. Example (95) illustrates a demonstrative pronoun as modifier, which is only possible for *this* and *that*. The independent function is demonstrated in example (96). Here, *these* works as a pronoun that refers to the antecedent *the red ones* (cf. Stirling & Huddleston 2010: 1504-1507, 1510-1511).

(94) Many people play **the guitar**. <u>This instrument</u> is probably the most popular one.

(95) This building is **100 metres** high. I am actually surprised it is <u>this tall</u>.

(96) There are some green and red apples in the kitchen. Look at **the red ones**! <u>These</u> are especially sweet.

3.3.2 Grammatical and referential aspects

With regard to grammatical characteristics, demonstrative pronouns differentiate between forms for number. The singular forms *this* and *that* for count nouns in the singular and for mass nouns, and the plural forms *these* and *those* for count nouns in the plural are distinguished. It is, however, also possible that singular demonstrative pronouns occur with reference to plural forms (example (97)) and, the other way round, plural items are used for singular forms (cf. Halliday & Hasan 2008: 62). Halliday & Hasan (2008) point out for such demonstrative pronouns: "they refer to the meanings and not to the forms that have gone before" (ibid.: 62). As for gender, all demonstrative pronouns can be used for either personal or nonpersonal reference. The nonpersonal function, however, is especially common in independent function. Personal function with independent demonstrative pronouns is only possible if they are used as subjects of a clause (example (98)).

Apart from number, the distance between speaker/writer and the referred entity is distinguished, which constitutes the referential aspect. *This* and *these* are used if referring to "near" objects or events, or even emotions (example (99)). *That* and *those*, by contrast, imply "distant" reference (example (98)). For example, the decision between *this* and *that* in (98) and (99) relies on the spatial or temporal distance between the speaker and the person/object referred to. See Table 8 for a visualisation of these characteristics (cf. Carter & McCarthy 2006: 389; Biber et al. 2007: 347; Payne & Huddleston 2010: 373; Quirk et al. 2012: 372-373).

(97) Dad bought **sausages, carrots, cucumbers, potatoes, three pounds of beef and four packets of biscuits**. – What is he going to do with all that food?

(98) Can you see **the man in the street**? That is Mr Miller, my neighbour.

(99) Noah was playing **golf**. This new hobby of his has been taking up all of his time.

Table 8: Demonstrative pronouns

	Number	Singular	Plural
Reference			
Near reference		*this*	*these*
Distant reference		*that*	*those*

3.3.3 Anaphoric/cataphoric and non-anaphoric use

The antecedent of anaphoric demonstrative pronouns can take different forms. It can be a noun phrase (numbers (98) and (99)), an adjective phrase (example (100)) or adverb phrase (example (101)). With regard to independent demonstrative pronouns, the antecedent can also be a nominal only (see examples (102) and (103)). Even clauses or sentences can serve as antecedents of demonstrative pronouns (example (104)), but only for independent *this* and *that* (cf. Halliday & Hasan 2008: 53, 66; Quirk et al. 2012: 1461-1463; Schubert 2012: 36). Moreover, the antecedent with independent *this* can be a title or sub-title (example (105)). Here, *it* cannot occur (cf. Carter & McCarthy 2006: 246; Stirling & Huddleston 2010: 1506-1509; Quirk et al. 2012: 868).

(100) The crayon you gave me was **blue**, but I did not want that particular colour.
(101) Sue performed **incredibly well** last night. With this talent, she will certainly win tomorrow.
(102) This **watch** is more expensive than that.
(103) The **mountains** in Austria are higher than those in Germany.
(104) **Toby will travel to Australia in August**. At least, that is what I understood.
(105) **Syntax**
This is one field of linguistics.

Comparing independent demonstrative pronouns with personal pronouns, they can both be used with noun phrases or clauses as antecedents. Personal pronouns, however, are more common if the antecedent is a noun phrase. Demonstrative pronouns, by contrast, are more likely with clauses as antecedents. In such use, they show no coreferential relationship (cf. Halliday & Hasan 2008: 53, 66; Stirling & Huddleston 2010: 1506-1509; Quirk et al. 2012: 375, 868, 1461-1463).

Demonstrative pronouns are not always used anaphorically but frequently have deictic function. Quirk et al. (2012) even argue that "[t]he anaphoric and cataphoric uses of the demonstratives are extensions of their situational use" (ibid.: 375). With regard to deictic function, demonstrative pronouns relate to the spatial, temporal or emotional proximity. Example (106) a), on the one hand, could be used if the person was thinking of the specific book and thus is near in time. On the other hand, example (106) b) is likely if the person thought about the book some time ago, being distant in time.

Moreover, demonstrative pronouns can also refer to an entity which is not found in the situation or in the preceding text (example (107)). Instead, the demonstrative pronoun has to be interpreted from what is known or was experienced. Stirling & Huddleston (2010: 1510) term this case “recognitional use”, which is, however, informal. Dependent *that* or *those* are usually found with these constructions, but dependent *this* and *these* could also occur (cf. Quirk et al. 2012: 374-376).

(106) a) I have found *this book*.
b) I have found *that book*.
(107) I never saw *that* ring he gave you.

A distinction between anaphoric and deictic uses of demonstrative pronouns cannot always be made, but it is also possible that demonstrative pronouns are anaphorically and deictically at the same time. For example, *that* is anaphoric in (108) as it refers back to the antecedent *the chair next to the drawer*. At the same time, it shows deictic meaning if *that* is used for a distant entity present in the specific situation (cf. Stirling & Huddleston 2010: 1506-1509; Quirk et al. 2012: 375).

(108) Look at **the chair next to the drawer**. <u>That</u> is the one I bought.

In addition, the demonstrative pronouns *that* and *those* can even occur without having any anaphoric or deictic function. This is the case if independent demonstrative pronouns are postmodified by finite clauses (example (109)). The postmodification is enough to identify what is meant by the demonstrative pronoun here. For instance, *who do not keep their promise* in example (109) makes clear the non-anaphoric and non-deictic reference of *those*. But yet, independent demonstrative pronouns in such constructions are not always non-anaphoric, especially in formal contexts. Example (110) shows a case where independent *those* is anaphoric. There is, however, a difference between *that* and *those* in such a use. *Those* can be used if the antecedent denotes a person, animal or thing, *that* only for antecedents that are things. If it concerns dependent demonstrative pronouns, a head has to occur between demonstrative and postmodification (example (107)) (cf. Stirling & Huddleston 2010: 1510-1511).

Furthermore, two other non-anaphoric instances should not be forgotten. First, independent and dependent *those* plus head can be postmodified by a partitive (example (111)).[15] Such uses should not be confused with anaphoric cases postmodified by *of*-phrases (example (112)), which occur in formal, academic contexts (cf. Trask 1997: 163; Carter & McCarthy 2006: 251-252; Payne & Huddleston 2010: 413; Stirling & Huddleston 2010: 1510-1511). Second, non-anaphoric *this* and *that* are possible as a degree modifier (example (113)). *That* means “particularly” or “so” in this example. However, *that* can also be anaphoric if premodifying an adjective, at least in some contexts, such as example (95) (cf. Biber et al. 2007: 350; Payne & Huddleston 2010: 373; Stirling & Huddleston 2010: 1510-1511; Quirk et al. 2012: 866-867, 1466).

(109) He does not belong to *those* who do not keep their promise.

(110) The **people** in Siberia have a hard life. However, those who know how to make the best of their situation can enjoy it, too.

(111) *Those* of you who know the answer should raise their hands.

(112) This **theory** seems more plausible than that of Chomsky.

(113) He was not *that* fast in the race.

Cataphoric instances of demonstrative pronouns are possible with non-integrated antecedents (example (114)). Such uses show similarities with *it*, which also occurs in non-integrated constructions. Demonstrative pronouns that take such cataphoric interpretations are restricted to the items *this* and *these* and to the modifier *that*, though (cf. Quirk et al. 2012: 375, 1461-1463). Stirling and Huddleston (2010) point out: “It would seem that there are no cases of distal ***that*** [i.e. *that* and also *those* in their terminology] that can properly be regarded as involving anticipatory anaphora” (ibid.: 1509).

(114) This is the best news I have heard so far today: **The TV set is working again.**

3.3.4 Relationship between anaphor and antecedent

Demonstrative pronouns in independent function are a form of reduction; those in dependent function are rather a means to avoid repetition. Furthermore, demonstrative pronouns mainly show a coreferential relationship. If *this* and *that* refer anaphorically to a clause, they do not show coreference but belong to

[15] “Partitives” are “constructions denoting a part of a whole” (Quirk et al. 2012: 249), which have the form of a prepositional phrase beginning with *of* (cf. Quirk et al. 2012: 249-251).

the miscellaneous category. Furthermore, a substitutional relationship to the antecedent is possible with *that* and *those*. Here, it depends on the individual use, whether a coreferential or substitutional relationship is shown. Substitution is illustrated in the examples (102) and (103), in which the demonstrative pronouns and the antecedents do not denote the same entities. In sum, coreference is the usual case with dependent uses of demonstrative pronouns, even if the nouns of the demonstrative pronoun and the antecedent are not the same. Here, pronoun and antecedent can be related in some form of synonymy, or hyponymy/hypernymy (example (94)) (cf. Carter & McCarthy 2006: 251-252; Halliday & Hasan 2008: 63; Quirk et al. 2012: 863-865, 872-873).

3.3.5 Summary

Demonstrative pronouns show anaphoric and cataphoric use. It does not matter if the demonstrative pronoun shows a deictic use simultaneously. Such instances will only be considered here regarding their anaphoric function. Furthermore, the relationship between anaphor and antecedent is coreferential in most cases. The exceptions are independent *that* and *those* that can be substitutional, especially if they have nominals as antecedents. Independent *this* and *that* can refer to a clause, in which case they belong to neither coreference nor substitution, but to the miscellaneous category. Cataphors are only possible with *this* and *these* and *that* as modifier.

With regard to non-anaphoricity, all demonstrative pronouns can occur in non-anaphoric use. *That* and *those* in particular are often not anaphoric together with specific forms of postmodifications. Finally, *this* and *that* as modifier can be used non-anaphorically, or in some situations, anaphorically. A summary of these features is given in Tables 9 and 10. They present characteristics that anaphors and antecedents of demonstrative pronouns have to share, and a summary of non-anaphoric features.

[16] *NP* stands for “noun phrase”, *AdjP* for “adjective phrase” and *AdvP* for “adverb phrase”.

Table 9: Anaphoric use of demonstrative pronouns

	Anaphor						**Form of antecedent**[16]						**Cataphor**
	Number			Gender		Postmodification with finite clause	NP	AdjP	AdvP	Nominal	Clause	(Sub)title	Non-integrated
	Singular	Plural	Non-count	Personal	Non-personal								
Independent function													
this	×	(×)	×	×[2]	×	-	×	×	×	×	×	×	×
that	×	(×)	×	×[2]	×	×[3]	×	×	×	×	×	-	-
these	(×)	×	×[1]	×[2]	×	-	×	×	×	×	-	-	×
those	(×)	×	×[1]	×[2]	×	×[3]	×	×	×	×	-	-	-
Dependent function													
this	×	(×)	×	×	×	-	×	×	×	-	-	-	×
that	×	(×)	×	×	×	×[3]	×	×	×	-	-	-	-
these	(×)	×	-	×	×	-	×	×	×	-	-	-	×
those	(×)	×	-	×	×	×[3]	×	×	×	-	-	-	-
this as modifier	×	×	×	×	×	-	×	×	×	-	-	-	×
that as modifier	×	×	×	×	×	-	×	×	×	-	-	-	-

[1] Only if two noun phrases are coordinated
[2] Only if the demonstrative pronoun functions as subject
[3] In formal use

Table 10: Non-anaphoric use of demonstrative pronouns

	Deictic	**Recognitional use**	**Postmodification**		**Degree modifier**
			Finite clause	Partitive	
Independent function					
this	×	-	-	-	-
that	×	-	×	-	-
these	×	-	-	-	-
those	×	-	×	×	-
Dependent function					
this	×	(×)	-	-	-
that	×	×	×	-	-
these	×	(×)	-	-	-
those	×	×	×	×	-
this as modifier	×	(×)	-	-	×
that as modifier	×	×	-	-	×

3.4 Relative pronouns

3.4.1 Form and function

The forms of relative pronouns are *who*, *whom*, *which*, *whose*, *that* and zero *that*.[17] Apart from their anaphoric function relative pronouns are also part of a relative clause. To give an example, *that* in (115) refers to the antecedent *the reason*, and at the same time it serves as the object of the relative clause (cf. Quirk et al. 2012: 1247-1253, 1257-1260).

(115) **The reason** that you gave is not very convincing.

Comparing relative pronouns to personal pronouns, relative pronouns differ from personal pronouns in a number of ways. To begin with, a relative pronoun is always at the beginning of a clause, irrespective of its function. Regarding the antecedent, a relative pronoun mostly refers to the preceding noun phrase that the relative pronoun postmodifies. For instance, *that you gave* in (115) is the relative clause that postmodifies *reason*. Antecedents taking other forms such as adjective, adverb and verb phrases are possible as well, though infrequent (cf. Huddleston, Pullum & Peterson 2010: 1035, 1052, 1060). Relative

[17] Huddleston, Pullum & Peterson (2010: 1034) speak of “bare relatives” in the case of zero *that*.

and personal pronouns also have some features in common: they are both usually coreferential with their antecedents. Apart from that, relative pronouns are a form of reduction, as are personal pronouns (cf. Carter & McCarthy 2006: 387; Quirk et al. 2012: 365, 368).

3.4.2 Types of clauses and their anaphoric and non-anaphoric use

The items of relative pronouns have to be distinguished from related expressions. To start with, it is useful to differentiate between relative and appositive clauses because the form *that* occurs in appositive clauses as well as in relative clauses. Appositive clauses are not relevant here as they are not anaphoric. In appositive clauses, *that* is a conjunction (example (116)), and so can be distinguished from relative clauses in which *that* is a relative pronoun (example (117)).

Relative clauses then distinguish between adnominal, nominal and sentential relative clauses. Adnominal relative clauses (see example (117)) are the most important type of relative clauses and fall into two categories: restrictive and nonrestrictive. The restrictive category is the more frequent one. In general, these types represent how closely the adnominal relative clause and the antecedent to which the relative pronoun refers are semantically connected with each other. A restrictive clause represents a close connection, a nonrestrictive clause implies a more distant relation. Nonrestrictive clauses are usually embedded in between commas in writing. Restrictive clauses, on the one hand, delimit the semantic range of the antecedent, as in example (117) where the statement says that not all houses are enjoyed but only those on hills. On the other hand, nonrestrictive clauses further describe the antecedent and can be seen as comments inserted into a sentence (example (118)) (cf. Huddleston, Pullum & Peterson 2010: 1034-1035, 1058-1059; Quirk et al. 2012: 365-366, 1247-1250, 1257-1261).

The distinction between restrictive and nonrestrictive clauses is important because it influences which forms of relative pronouns can be used in the individual situation, as is shown in Table 11 below. Determinative *which* only occurs in nonrestrictive clauses. Furthermore, only nonrestrictive clauses take an antecedent that is a proper name or a whole clause (cf. Huddleston, Pullum & Peterson 2010: 1048, 1060-1061).

(116) It proved to be the right decision *that* we chose the more expensive machine.

(117) I like **houses** <u>that</u> are built on hills.

(118) Susan called out to **her friend Tom**, <u>who</u> was just crossing the street.

In addition, a sentential relative clause refers to an antecedent that is a clause (example (119)). Of the items mentioned above, only *which* is anaphoric and used in sentential relative clauses (Quirk et al. 2012: 1119-1120).

Finally, nominal relative clauses use the forms *which*, *whom* and *who*, where *which* and *whom* occur only with certain verbs such as *like* or *wish*. These relative clauses "are unique among relative clauses in that they 'contain' their antecedents" (ibid.: 1244). For instance, example (120) can be paraphrased as *You are not the person I was looking for*. Relative pronouns in nominal relative clauses are therefore non-anaphoric. Moreover, other forms such as *what*, *whoever*, *whichever* and *whatever* occur in nominal relative clauses. Quirk et al. (2012) describe these items as follows: "[the] *wh*-element is merged with its antecedent (the phrase to which the *wh*-element refers)" (ibid.: 1056). They could be paraphrased with *that which* in the case of *what* and *that who/which/what* in the other cases respectively (example (121)). The second element of the paraphrase refers to the first, namely *that*, which is non-anaphoric. Thus, such constructions are not relevant here. Moreover, the items *whoever*, *whichever* and *whatever* are particularly found in speech. *Whoever* is avoided in formal contexts and is instead paraphrased with *he who*. These items are again not relevant for anaphoricity (cf. Quirk et al. 2012: 1056-1059, 1244-1245, 1260-1262).

(119) **The earthquake caused the shed to collapse**, <u>which</u> means we need to clean it up now.
(120) You are not *who* I was looking for.
(121) Do *what* he tells you!

3.4.3 Further non-anaphoric uses

Apart from the distinctions above, other non-anaphoric situations should not be forgotten. For example, *wh*-relative pronouns have to be differentiated from interrogative pronouns. An interrogative pronoun is found in (122). In addition, determinative *which* usually does not work as anaphor (example (123)). Anaphoricity is more debateable in example (124), although such cases are also considered being non-anaphoric here. The reason is that the antecedent is too implicit and an acceptable antecedent such as *if it is cheap* would need a certain amount of paraphrasing. Such paraphrasing is then dependent on the context i.e. different instances of determinative *which* would need different types of paraphrasing (cf. Carter & McCarthy 2006: 392; Huddleston, Pullum & Peterson 2010: 421-422; Quirk et al. 2012: 365, 368). Another non-anaphoric case with *which* is when it is postmodified by an *of*-phrase, e.g. *which of them* (cf. Payne & Huddleston 2010: 413).

(122) *Who* helped us repair the heater?
(123) He does not know *which* song to choose from.
(124) It could be cheap, in *which* case you should buy it.

Finally, there are two more uses that are not considered anaphoric here. First, *that* also occurs as independent demonstrative pronoun, which means that relative *that* (example (117)) has to be distinguished from demonstrative use (e.g. example (108)). *That* as independent demonstrative pronoun does not belong to the relative pronoun category. Second, relative pronouns are not regarded as anaphors if they refer to an antecedent that is non-anaphoric. This is the case in example (38), which is *He who dares wins*. It contains *who* as relative pronoun referring to non-anaphoric *he*.

3.4.4 Gender and case

As to their grammatical features, all relative pronouns except *that* distinguish between forms for gender, but not for number or person. Gender involves the distinction between forms for personal use and forms for nonpersonal use. *Who* and *whom* are personal forms, *which* is nonpersonal. The distinction between personal and nonpersonal use can also be found with collective nouns, such as *government*. If these are understood as a group of individuals, *who* is used, otherwise *which* (cf. Quirk et al. 2012: 316-317, 771, 1260). *Whose* shows personal and nonpersonal use. Although *whose* is found with nonpersonal reference, the personal use is more common. Moreover, if an antecedent is coordinated with personal and nonpersonal nouns, the sequence determines whether *who* or *which* is chosen: it then depends on the gender of the last noun (example (125)) (cf. Huddleston, Pullum & Peterson 2010: 1048-1049; Quirk et al. 2012: 1245).

Additionally, there are other circumstances that define gender. For instance, *who* and *whom* can refer to antecedents that are animals, especially pets, or supernatural beings on the one hand. Such a use then implies more emotional involvement.[18] On the other hand *which* can be used for children, and implies emotional distance, which is similar to personal pronouns. Furthermore, *which* is found in two special constructions where it also refers to antecedents that are human beings. One such construction is illustrated in (126), in which the antecedent functions as subject complement with the verb *be*. The other is found with *have (got)*, as in example (127), where the antecedent is the direct object of *have*.[19] In such cases, relative pronouns are not coreferential but substitutional (cf. Huddleston, Pullum & Peterson 2010: 1048-1049; Quirk et al. 2012: 1245).

(125) He mentioned **his family and the cats** of <u>which</u> he was fond.
(126) They say that he was **the best worker**, <u>which</u> he surely was.
(127) They have **domestic servants**, <u>which</u> we do not have.

Apart from gender, relative pronouns distinguish between forms of case. The forms in subjective case are *who* and *which*, in objective case *whom* and *which* and in genitive case *whose*. The item *whose* occurs as possessive determiner of noun phrases. Moreover, *whom* is restricted to formal use. *Who* occurs in place of *whom* in informal contexts (cf. Carter & McCarthy 2006: 387; Quirk et al. 2012: 366-368, 1249-1250, 1252).

There are further forms in restrictive clauses, but not in nonrestrictive ones. For instance, *that* can be used instead of the *wh*-forms in subjective and objective cases, but it is more informal. An example is (117), in which the objective relative pronoun *that* can be substituted by *which*. Furthermore, *that* can work as neutral form that does not distinguish between personal and nonpersonal entities. In addition, the relative pronoun at the beginning of relative clauses can be left out in objective cases, as in example (128). More informal contexts generally show a preference for leaving out the relative pronoun here. The division into restrictive and non-restrictive and their forms that distinguish between case and gender are shown in Table 11 (cf. Carter & McCarthy 2006: 387; Quirk et al. 2012: 366-368, 1249-1250, 1252).

(128) **The meeting** <u>Ø</u> I attended yesterday ended at 4 p.m.

[18] In contexts where the personal pronoun *she* refers to ship, for instance, the relative pronoun *which* has to be used for that antecedent. *Who* is not found in such contexts (cf. Quirk et al. 2012: 1245).

[19] Huddleston, Pullum & Peterson (2010: 1049) do not speak of “direct object”, but “complement” in the sense of the distinction between “complement” versus “adjunct” (cf. Huddleston 2010b: 219).

Table 11: Relative pronouns

		Restrictive		Nonrestrictive	
	Gender	**Personal**	**Nonpersonal**	**Personal**	**Nonpersonal**
Case					
Subjective		*who / that*	*which / that*	*who*	*which*
Objective		*whom / that / Ø*	*which / that / Ø*	*whom*	
Genitive		*whose*		*whose*	

3.4.5 Summary

Relative pronoun forms distinguish between gender and case. Their form also depends on whether relative pronouns occur in restrictive or nonrestrictive clauses. Relative pronouns are usually anaphoric, and as Huddleston, Pullum & Peterson (2010) state "the anaphoric relation is an essential feature of the construction" (ibid.: 1036). Relative pronouns frequently refer to the preceding noun or noun phrase, or, in the case of *which*, also to a preceding clause or proper name. However, some non-anaphoric uses of relative pronoun items are possible. If items occur in appositive clauses or work as interrogative pronouns, they are not regarded as anaphors. Furthermore, relative pronouns are ignored if they refer to an antecedent that is non-anaphoric or if *that* is a demonstrative pronoun.

Finally, relative pronouns are mostly coreferential and show reduction. If referring to a clause, relative pronouns are not coreferential but belong to the miscellaneous category. Relative pronouns are substitutional if they are subject complements of the verb *be* or direct objects of the verb *have (got)*. They cannot be used cataphorically. An overview is given in Tables 12 and 13.

Table 12: Anaphoric use of relative pronouns

	Clause type			**Anaphor**		**Antecedent**	
	Adnominal relative		Sentential relative	Determiner	Pronoun as such	Personal	Nonpersonal
	Restrictive	Nonrestrictive					
who	×	×	-	-	×	×	×[2]
whom	×	×	-	-	×	×	×[2]
which	×	×	×	(×)	×	×[1]	×
whose	×	×	-	×	-	×	×
that	×	-	-	-	×	×	×
zero *that*	×	-	-	-	×	×	×

[1] Only with emotional distance, e.g. in science; as subject complement of *be* and direct object of *have (got)*; in coordination with a personal entity
[2] Only with higher animals and supernatural beings; in coordination with a nonpersonal entity

Table 13: Non-anaphoric use of relative pronouns

	Clause type		**Form**		**Postmodification by an *of*-phrase**	**Reference to non-anaphoric items**
	Nominal relative	Appositive clause (→ conjunction)	Interrogative pronoun	Independent demonstrative		
who	×	-	×	-	-	×
whom	×[1]	-	×	-	-	×
which	×[1]	-	×	-	×	×
whose	-	-	×	-	-	×
that	(×)[2]	×	-	×	-	×
zero *that*	-	-	-	-	-	×

[1] Only with verbs such as *like*, *choose*, *please*, *want*, *wish*, e.g. *You can marry whom you want.*
[2] Only in the constructions *that which/who/what*

3.5 Adverbs

The items that belong to this category are: *here*, *there*, *then*, *where*, *when*, *while* and *why*.[20] *Here* and *there* usually denote spatial, *then* temporal orientation. The four *wh*-items are relative adverbs with *where* for local, *when* and *while* for temporal, and *why* for causal relations.[21] As to the relationship between anaphor and antecedent, the adverbs listed here usually show coreference (Quirk et al. 2012: 864-867). Each of these adverbs will now be considered in more detail.

3.5.1 *Here* and *there*

In comparison to demonstrative pronouns, Stirling & Huddleston (2010) characterise *here* and *there* as follows: "*here* and *there* are distinguished as proximal

[20] Similar to *whoever* and other items in 3.4, some items of adverbs combine with *-ever*, i.e. *wherever* and *whenever* (cf. Huddleston, Pullum & Peterson 2010: 1074).

[21] *The Cambridge Grammar of the English Language* (Huddleston, Pullum & Peterson 2010: 1050-1052) regards only *why* as adverb, the other *wh*-items are classified as prepositions. Here, however, traditional grammar (e.g. Quirk et al. 2012: 865, 1253-1254) is given precedence.

and distal, like the demonstratives ***this*** and ***that*** respectively" (ibid.: 1549). *Here* and *there* mostly denote spatial location and so differ in whether the speaker perceives a place as near or distant, but they can also be used temporally with a near-distant contrast. In the first case, they are anaphoric, e.g. in (129). In the latter case, they are non-anaphoric, e.g. as *there* in (130). Here, *there* is interpreted as being distant in a temporal sense, meaning "at that point (of our discussion)".

In general, *here* is mainly deictic and occurs anaphorically only in some situations (e.g. in example (131)). Furthermore, *here* as well as *there* often show both anaphoric and deictic reference at the same time (cf. Stirling & Huddleston 2010: 1549-1550). Stirling & Huddleston (2010) explain why examples such as (129) are not purely anaphoric: "*There* [...], though primarily anaphoric, retains a distal deictic component of meaning, for it still indicates a place relatively removed from where I am now" (ibid.: 1550).

Apart from deictic occurrence, *here* and *there* show further non-anaphoric uses. The adverb *there* has to be differentiated from existential *there*, which does not refer to an entity but is required grammatically as a pronoun. In this use, *there* has no further semantic content but postpones information to a non-initial position in the sentence, similarly to *it*. Hence, example (132) a) could be paraphrased with (132) b) (cf. Carter & McCarthy 2006: 392, 789; Quirk et al. 2012: 89). Furthermore, *here* can be a noun, and both *here* and *there* can be interjections, for instance in (133) (cf. "Here" n.d.; Summers 2006: 656, 1440-1441).

As to the antecedents of *here* and *there*, the following observations are noteworthy. If the expression to which the anaphor refers contains a preposition, the antecedent usually includes this preposition (see example (134)). If the anaphor is preceded by a preposition, the antecedent constitutes the expression without the preposition (example (129)). In case neither the anaphor nor the antecedent incorporates a preposition, the preposition is understood as being of a non-special kind, such as *at* or *in*. For instance, (131) implies the preposition *at* (cf. Stirling & Huddleston 2010: 1550).

Finally, *here* can be used cataphorically. It is the only item in this category of adverbs where a cataphoric interpretation is common (example (135)). As the antecedent is a clause, no coreferential relationship is established but this item then belongs to the miscellaneous category (cf. Halliday & Hasan 2008: 68, 75).

(129) Dad was in **London**. He came back from <u>there</u> yesterday.
(130) He stopped *there* and said he would continue next time.

(131) **The new open-air swimming-pool** opened two days ago. Here, many people will go swimming during the summer months.

(132) a) *There* is still plenty of time left.
b) Plenty of time is still left.

(133) *There*! Was that what you wanted?

(134) She spent her holidays **in Cornwall**. She got to know plenty of nice people there.

(135) Here is the plan: **we will rent a car and drive to Birmingham**.

3.5.2 *Now* and *then*

Generally, *now* and *then* behave similarly to *here* and *there*. To quote Stirling & Huddleston (2010):

> Proximal *now* and distal *then* are the temporal counterparts of spatial *here* and *there* respectively. *Now* is predominantly deictic while *then*, in its temporal sense, is usually anaphoric. (ibid.: 1558)

The distinction between near and more distant time is expressed through the contrast of *now* and *then*. Example (136) is a fine illustration of deictic *now*. Stirling & Huddleston (2010: 1558-1559) give no example of anaphoric *now*, and also Quirk et al. (2012: 865) mention only *then* but not *now* in their list of coreferential and substitutional items. Moreover, Halliday & Hasan (2008) concede that "*now* is very rarely cohesive" (ibid.: 74). Consequently, *now* is not considered as anaphor here. The item *then*, if used anaphorically, refers to a time and typically takes a prepositional phrase as antecedent (example (137)).

Then also shows some non-anaphoric uses. To begin with, *then* is non-anaphoric if it points to a preceding clause, to the time this clause expresses, or to the time that comes shortly after what was mentioned. Example (138) contains a non-anaphoric *then*, which refers to the time shortly after he baked a cake. Yet, Stirling & Huddleston (2010: 1559) regard such uses as anaphors. As *then* serves to denote a temporal sequence, similar to items such as *next*, *subsequently* and *after that*, and as the antecedent lacks explicitness, which has been defined to be one essential criterion for anaphors here, these expressions are not considered anaphors (cf. Halliday & Hasan 2008: 261). Moreover, *then* is also non-anaphoric if it can be paraphrased with *also* or *besides* (example (139)) (cf. Summers 2006: 1440). Halliday & Hasan (2008: 74) speak of "conjunction *then*" in such instances. Finally, *then* can be used deictically if it refers to the time that is evident from the situation. Example (140) makes sense if, for instance, looking at photos of the childhood (cf. Stirling & Huddleston 2010: 1559).

A further non-anaphoric case is if adjective *then* occurs as modifier (example (141)) (cf. Summers 2006: 1440). Stirling & Huddleston (2010: 1559) again classify such instances as anaphoric. Here, however, such cases are not seen as anaphors because the anaphoric link of *then* is too unspecific, i.e. not explicit enough to be relevant. Stirling & Huddleston (2010) argue in these contexts that "[t]he reference [...] is to the time of the situation expressed in the clause containing *then*" (ibid.: 1559). For instance, *then* in (141) points to the verb *congratulated*, which is in the past tense. As a result, *then* would refer to the past in this example without establishing references to explicit antecedents (cf. ibid.: 1558-1559).

(136) The tea is *now* ready.
(137) Susan married **in 1919**. She was 30 years old back <u>then</u>.
(138) He baked a cake and *then* he tried it.
(139) *Then*, what did you do?
(140) We were not used to staying up late back *then*.
(141) The staff congratulated the *then* president.

3.5.3 *Where*, *when*, *while* and *why*

The item *where* is an expression for spatial location (example (142) a)). It can also be paraphrased by *in which* in (142) b) and, in this case, would belong to the relative pronoun category. *When* and *while* are used for temporal reference. *While* is used if the reference is to a period of time (example (143)), and can be replaced by *when*, *during which (time)* or *in which (time)*. *When* typically refers to a point in time (example (144)) and also occurs as complement of a temporal preposition, i.e. *since when*, *until when*, *from when* and *by when*. Moreover, it can sometimes be replaced by a preposition plus *which*, e.g. *during which*. Which preposition is used here depends on the context. To give an example, *on which* can replace *when* in (144). Finally, *why* refers to causal expressions (example (145)), where the antecedent mostly involves *reason*. Therefore, *why* is only of limited importance here. It can also be replaced by *for which*, but this use is quite rare (cf. Huddleston, Pullum & Peterson 2010: 1050-1051; Quirk et al. 2012: 1119-1120, 1253-1254).

In addition, the forms listed here have to be differentiated from their non-anaphoric uses. *Where*, *when*, *while* and *why* occur as interrogative pronouns and are then not anaphoric. This is in analogy to the *wh*-forms of relative pronouns, which also have to be distinguished from uses as interrogative pronouns. Moreover, *wh*-elements are non-anaphoric as conjunctions in adverbials and nominal relative clauses[22] (example (146)). Furthermore, *where*, *when*, *while* and *why* occur non-anaphorically as nouns. *While* can furthermore be a non-anaphoric verb, and *why* can occur as a non-anaphoric interjection (example (147)) (cf. “Where” n.d.; Summers 2006: 1569-1570, 1574). Furthermore, the *wh*-items can also be merged with their antecedent, especially in the case of *why* (cf. Quirk et al. 2012: 1053-1059). Example (145) would then read *Why she came was my birthday*. Such instances have to be distinguished from anaphoric uses. Quirk et al. (2012) state:

> Many speakers find their use [i.e. the use of these *wh*-forms] along with the corresponding antecedent somewhat tautologous – especially the type *the reason why* – and prefer the *wh*-clause without antecedent, *ie* a nominal relative clause [...]. (ibid.: 1254)

Additionally, it should be mentioned that the four *wh*-items as adverbs are supplemented by other forms: *whence* and the compounds of *where* plus a preposition, i.e. *whereby*, *wherein* and *whereupon*.[23] The item *whence* (example (148)) is used to denote a relation to a spatial or logical origin and can be replaced by *from which*. *Whence* is archaic, although it can occur in journalism. From the *where*-compounds, only *whereby* (example (149)), and marginally *wherein* and *whereupon*, are still used; other forms such as *wherefrom* are archaic. The three forms *whereby*, *wherein* and *whereupon* are equivalent to *by which*, *in which* and *immediately after which/as a result of which* respectively. All these items can work as anaphors, as the examples below demonstrate. *Whereby*, *whereupon*, and *whence* can also take a clause as antecedent. Due to their restricted use, they are not given much attention here (cf. Summers 2006: 1569; Huddleston, Pullum & Peterson 2010: 1046, 1050-1052).

(142) a) He lives **in the house** where my parents used to live.
b) He lives in the house in which my parents used to live.
(143) **From the beginning of May until the end of October**, while the sun is still shining, the park will be open to visitors.
(144) **It occurred on Friday** when he was at home all by himself.
(145) That is **the reason** why she came to my birthday.

[22] Quirk et al. (2012: 442-444) argue that the *wh*-forms in these uses are not pure conjunctions.

[23] Although *whereupon* is a conjunction, it is listed here because it is a compound involving *where* (cf. Summers 2006: 1569; Quirk et al. 2012: 998).

(146) She does not know *when* the train arrives.
(147) *Why*, he does not know that.
(148) The kitchen is situated next to **the hall**, <u>whence</u> they heard voices.
(149) **The order** <u>whereby</u> we should leave at once reached us yesterday.

3.5.4 Summary

The adverbs *here*, *there* and *then* can all occur as anaphors; they are reductive forms. With regard to the *wh*-forms, *where* and *when* are the two most significant forms, followed by *while* and *why*. *Whence*, *whereby*, *wherein* and *whereupon* are only rudimentarily important. All these items usually show coreference. If the antecedent is a clause or sentence, they belong to the miscellaneous category. Only *here* can take a cataphoric interpretation from these items.

Moreover, the adverbs listed here have to be distinguished from non-anaphoric uses. First, *here*, *there* and *then* can be deictic. Second, *there* is non-anaphoric as existential *there*. Third, *then* is not anaphoric if denoting a temporal sequence or if it can be paraphrased with *also* and similar expressions. Fourth, *here*, *there* and *why* also occur as interjections. Fifth, the *wh*-items have to be differentiated from their interrogative uses, from uses as conjunctions, and, especially in the case of *why*, from forms where the relative adverb is merged with the antecedent. An overview of these features is given in Tables 14 and 15.

Table 14: Anaphoric use of adverbs

	Sense			**Distance**		**Antecedent is clause**	**Cataphor**
	Place	Time	Cause	Proximal	Distal		
here	×	×	-	×	-	×	×
there	×	×	-	-	×	-	-
then	-	×	-	-	×	-	-
where	×	-	-	-	-	-	-
when	-	×	-	-	-	-	-
while	-	×	-	-	-	-	-
why	-	-	×	-	-	-	-
whence	×	-	(×)[1]	-	-	×	-
whereby	-	-	-	-	-	×	-
wherein	-	-	-	-	-	-	-
whereupon	-	×	×	-	-	×	-

[1] i.e. logic

Table 15: Non-anaphoric use of adverbs

	Deixis	**Pointing to tense of preceding clause**	**Sense**		**Form of items**							
			Temporal sequence	Paraphrasing with *also* etc. possible	Inter-jection	Existential *there*	Adjective	Interroga-tive pro-noun	Conjunction		Noun	Verb
									Nominal relative clause	Adverbial clause		
here	×	-	-	-	×	-	-	-	-	-	×	-
there	×	-	-	-	×	×	-	-	-	-	-	-
then	×	×	×	×	-	-	×	-	-	-	-	-
where	-	-	-	-	-	-	-	×	×	×	×	-
when	-	-	-	-	-	-	-	×	×	×	×	-
while	-	-	-	-	-	-	-	×	-	×	×	×
why	-	-	-	-	×	-	-	×	×	×	×	-
whence	-	-	-	-	-	-	-	-	-	-	-	-
whereby	-	-	-	-	-	-	-	-	-	-	-	-
wherein	-	-	-	-	-	-	-	-	-	-	-	-
whereupon	-	-	-	-	-	-	-	-	-	-	-	-

3.6 Noun phrases with a definite article

Noun phrases with a definite article are not such typical anaphors as personal pronouns are, for instance.[24] As linguistic works on anaphors often focus on pro-forms, noun phrases are not considered. As a result, authors such as Stirling & Huddleston (2010) do not discuss anaphoric noun phrases. There are, however, good reasons to consider noun phrases with *the* as one type of anaphor. First, the definite article signals that the entity referred to by this noun phrase was mentioned before or is otherwise contextually clear to the addresser and addressee. Quirk et al. (2012) argue:

> The definite article *the* is used to mark the phrase it introduces as definite, *ie* as 'referring to something which can be identified uniquely in the contextual or general knowledge shared by speaker and hearer'. The 'something' referred to may be any kind of noun phrase referent [...]. (ibid.: 265-266)

Indefinite articles are then used to show that a new idea is brought up. Quirk et al. (2012) explain in more detail:

> *a/an X* will be used where the reference of *X* is not uniquely identifiable in the shared knowledge of speaker and hearer. Hence *a/an* is typically used when the referent has not been mentioned before, and is assumed to be unfamiliar to the speaker or hearer [...]. (ibid.: 272)

Furthermore, Biber et al. (2007) formulate that the definite article

> specifies that the referent of the noun phrase is assumed to be known to the speaker and the addressee. The knowledge could be based on the preceding text, in which case we speak of **anaphoric reference** [...]. In many cases, though, the connection is inferred rather than signalled by repetition, and we speak of **indirect anaphoric reference** [...]. (ibid.: 263)

Thus, noun phrases with a definite article can show anaphoricity and so are a type of anaphor (see also Hoffmann 2012: 78-80; Schubert forthcoming). Moreover, anaphora resolution systems, e.g. Vieira & Poesio (2000) or Markert & Nissim (2005), also deal with noun phrases with *the* as anaphors.

[24] Yet, Lemnitzer & Zinsmeister (2010) count noun phrases with definite articles to prototypical types of anaphors: "Prototypische Beispiele für Anaphern sind Pronomen oder definite Nominalphrasen" (ibid.: 85).

3.6.1 Syntactic characteristics

Noun phrases have a head, which is the central element of noun phrases and which is in most cases preceded by a determinative. For instance, the noun phrase *the report* has the head *report* and the determinative *the*. In addition, noun phrases can be modified, either before the head as premodifier or after the head as postmodifier. The example *the recent report of the project* contains the premodifier *recent* and the postmodifier *of the project* (cf. Aarts & Aarts 1986: 60-63).

What is important for the discussion here is how the determinative and the head are realised. First, the definite article, the indefinite article, possessive pronouns and demonstrative pronouns, for instance, can work as determinative. The result is an indefinite noun phrase in the case of an indefinite article or a definite noun phrase with a definite article, a possessive or a demonstrative determiner. The noun phrases discussed in this category are all definite, and the determinative has to be realised as definite article. Apart from the determinative, the head can also show different realisations. Usually, the head is a noun, as *report* in the above example, or pronoun, as in the noun phrase *he*. In addition, the head can be realised by an adjective, as in *the English*, a participle, as *the affected* and *the defending*, or by a numeral, e.g. in *the two* (cf. Aarts & Aarts 1986: 104-108; Quirk et al. 2012: 863).

Contrary to many papers and books concerning anaphora resolution (e.g. Muñoz, Palomar & Ferrández 2000: 527; Vieira & Poesio 2000: 539; Mitkov 2002: 112-113) and books about natural language processing (e.g. Allen 1995: 359, 440-445), the term “definite descriptions” for anaphoric noun phrases with a definite article is not used here. The term “definite descriptions” also denotes expressions that do not contain a definite article, such as proper nouns (cf. Vater 2005: 108).

3.6.2 Anaphoric/cataphoric and non-anaphoric use

Two subtypes of references can be distinguished if noun phrases with a definite article are used anaphorically. Such noun phrases have either direct or indirect reference to the antecedent. With direct reference, the anaphor and the antecedent have the same head. A good example of direct reference is (150). In addition, this example serves as illustration for an anaphor that takes the definite article, and its antecedent taking the indefinite article because the entity is introduced. Indirect reference means that anaphor and antecedent do not share the same head, as in example (151), in which *the engine* refers to *car*. Quirk et al. (2012) explain:

> INDIRECT ANAPHORA arises when a reference becomes part of the hearer's knowledge indirectly, not by direct mention [...], but by inference from what has already been mentioned [...]. (ibid.: 267)

Furthermore, Halliday & Hasan (2008) argue that the most obvious anaphoric noun phrases with a definite article are "those in which the item is actually repeated" (ibid.: 72), but there are others that avoid repetition and "add lexical variation" (Quirk et al. 2012: 1465), as in example (152).[25] Possible sense relations between anaphor and antecedent in indirect reference are synonymy, hyponymy/hypernymy or meronymy. World knowledge is needed to identify the antecedent in such instances. Example (152) shows hypernymy with *animal* as hypernym of *dog* (cf. Halliday & Hasan 2008: 274-279; Quirk et al. 2012: 860, 1464-1466).

(150) Toby has **a cat** and a dog. The cat is called "Molly".
(151) He went by **car**. After a while, the engine broke down.
(152) Mandy has **a dog**. Her husband does not like the animal.

Quirk et al. (2012: 268-269) argue that certain noun phrases with *the* take a cataphoric interpretation. If a noun phrase is postmodified, this postmodification is the cataphor. As postmodifications belong to noun phrases, these instances are not seen as cataphors here. Even Quirk et al. (2012: 268-269) concede that noun phrases with postmodifications are not truly cataphoric, as an alternative formulation, i.e. paraphrasing, is often possible. Example (153) a) shows cataphoric interpretation in Quirk et al.'s sense because *of Sandy* postmodifies and therefore specifies *the cat*. In (153) b) the paraphrasing is given. Other cataphoric interpretations of noun phrases with a definite article are not possible. At this point Halliday and Hasan (2008) stress:

> Cataphoric or forward reference, with *the*, is limited to the structural type. Unlike the selective demonstratives (*this*, *these* and *here*), *the* can never refer forward cohesively. It can only refer to a modifying element within the same nominal group as itself. (ibid.: 72)

[25] It should be noted that the discourse deictic items *the former* and *the latter* belong to this anaphor type as well (cf. Huddleston 2010c: 1163; Stirling & Huddleston 2010: 1555-1556; Quirk et al. 2012: 1465).

Additionally, no cataphor arises if a noun phrase anaphor and its antecedent are reversed. For example, in (154) an anaphoric proper name is coined by the inversion of anaphor and antecedent (cf. Halliday & Hasan 2008: 17).

(153) a) *The cat* of Sandy ⎫ has fleas.
b) Sandy's cat ⎭

(154) **Betty** repaired the lamp. The girl is only twelve years old.

Noun phrases with a definite article can also be used deictically. For instance, a definite article can occur with noun phrases to refer to the situation (from the surrounding, immediate context to the larger context). Such a use involves general knowledge and does not depend on a specific situation. People then know from their experience to what entity such expressions refer. Pertinent examples are found in (155), which contains two noun phrases with a larger situational reference (cf. Biber et al. 2007: 265-266; Quirk et al. 2012: 266-272).

In addition, we find other non-anaphoric uses. To start with, noun phrases occur with definite article if certain words, e.g. superlative adjectives such as *best* in *the best film*, or ordinals such as *first* are present. Moreover, *the* can show what Quirk et al. (2012: 269) term "sporadic reference". By that they mean noun phrases referring to institutions. They also include related expressions such as those of mass communication and transport, e.g. *the news*. Furthermore, *the* usually accompanies expressions of body parts after prepositions, such as in example (156). Finally, *the* can also occur in generic use. Generic use with singular noun phrases is rather restricted. Examples can be found with musical instruments, such as *play the guitar*. With regard to plural noun phrases, two applications are distinguished: phrases denoting the nationality of people such as *the English*; and other phrases with an adjective as head and which refer to a group as a whole, such as *the rich* (cf. Biber et al. 2007: 265-266; Halliday & Hasan 2008: 70-71; Quirk et al. 2012: 266-272, 282-285).

(155) *The sky* turned red as *the sun* set.

(156) Sue has a tattoo on *the shoulder*.

3.6.3 The relationship between anaphor and antecedent

Noun phrases with *the* are usually coreferential, except for those anaphoric cases that refer to a clause or sentence as antecedent (cf. Halliday & Hasan 2008: 281-282, 304-305; Quirk et al. 2012: 267). Quirk et al. (2012), for example, argue: “Coreference is a general feature of definite noun phrases” (ibid.: 865). Furthermore, noun phrases with *the* are cohesive devices and show lexical cohesion (cf. Halliday & Hasan 2008: 72, 281-282, 304-305; Esser 2009: 14, 42; Quirk et al. 2012: 267). This is different from the items discussed previously, which mostly belong to grammatical cohesion, except for demonstrative pronouns in dependent use (cf. Carter & McCarthy 2006: 245; Halliday & Hasan 2008: 275). As a result, noun phrases with a definite article are different from pronouns in this point because not so much value lies on grammar but on content, for instance. This feature is also important for resolving noun phrase anaphors. Finally, noun phrases with a definite article do not necessarily show reduction. With the exception of repetitions, noun phrases are predominantly used to avoid the same expression and to contribute to variation in utterances (Vater 2005: 47).

3.6.4 Summary

Noun phrases with a definite article can be anaphorically linked to the antecedent directly or indirectly. Noun phrases with *the* show no cataphoric interpretation and are usually coreferential. A number of instances can be listed where noun phrases with a definite article are not anaphoric and so have to be distinguished from anaphoric ones. The anaphoric and non-anaphoric features are illustrated in Tables 16 and 17.

Table 16: Anaphoric/cataphoric use of noun phrases with *the*

	Anaphoric				**Cataphoric**
	Same head	**Different head**			
		Synonymy	Hyponymy/ hypernymy	Meronymy	
the with direct reference	×	-	-	-	-
the with indirect reference	-	×	×	×	-

Table 17: Non-anaphoric use of noun phrases with *the*

	Non-anaphoric				
	Deixis	Superlative adjectives	Sporadic reference	Body parts	Generic
the	×	×	×	×	×

3.7 Proper names

The category of proper names is not a prototypical type of anaphor either. This, for instance, gets obvious when consulting Quirk et al. (2012: 288-297) who do not mention any anaphoric relations with proper names. In the same way, Stirling & Huddleston (2010) do not treat them in their chapter about deixis and anaphora. There are, however, scholars who recognise the anaphoric value of proper names. For example, Halliday & Hasan (2008: 19) speak of an "implicitly anaphoric" relationship if a proper name refers to a preceding proper name in the same form. Example (157) contains the proper name *Betty*, and the second occurrence of it establishes a cohesive link to the first-mentioned item. More importantly, however, are relations where anaphor and antecedent take different forms, but denote the same entity (example (158)). Furthermore, Huang (2000), for instance, also lists proper names, or "names" as he calls them, as one type of anaphor: "Linguistic elements that can be employed as an anaphor include gaps (or empty categories), pronouns, reflexives, names and descriptions" (ibid.: 1).

(157) Betty asked me if you were going swimming today. You could call *Betty* and tell her about your plans.

(158) **Bob Harris** is at a meeting in Berlin today. In urgent cases you can call the secretary there – just ask for <u>Mr Harris</u>.

3.7.1 Proper names and proper nouns

The term "proper name" needs to be distinguished from "proper noun". A proper noun is a single word that usually functions as the head of a proper name, such as in *Great Britain* where *Britain* is the head and so the proper noun. Other examples, such as *University of Passau*, contain *University* as head, but this is a common noun and not a proper noun. This means that proper names can but need not necessarily include proper nouns. A proper noun commonly begins with a capital letter, and the words further describing the proper noun are typically also capitalised. A good example is provided by the proper name *Professor Miller* where the descriptive element *Professor* has an initial capital letter (cf. Halliday & Hasan 2008: 42-43; Quirk et al. 2012: 288, 1637-1638).

Not all words in capital letters are proper names, though. For example, sentences begin with capital letters. Moreover, the item *God* and expressions referring to God are in capital letters. Additionally, capital letters are found in abbreviations, also in those that are not proper nouns such as *PTO* for “Please turn over the page”. These have to be differentiated from capitals marking proper nouns and names (cf. ibid.: 288, 1637-1638).

3.7.2 Syntactic features

Most proper names are noun phrases. Their head has, however, peculiar grammatical features because they frequently show no number contrast: Proper nouns are either singular or plural, which means that singular items do not have a plural form and vice versa. One important exception is if the surname denotes the whole family. For instance, the singular form and surname *Brown* then turns into the plural *the Browns* (cf. Payne & Huddleston 2010: 516, 519-520; Quirk et al. 2012: 288-290).

Different classes can be distinguished with proper names. Quirk et al. (2012: 290-294) list personal, temporal, geographic and other locative names as the main categories. All of these can work anaphorically. Personal names can consist of first name (e.g. *Betty*) or surname (e.g. *Smith*) alone, or of both first name and surname (e.g. *Betty Smith*). First names can also be written as initials (e.g. *B. Smith*). Additionally, first and/or surnames can be accompanied by titles, such as *Ms*, *President*, *Dr* (e.g. *Dr Betty Smith*). Furthermore, temporal names include names of days of the week (e.g. *Sunday*) and of festivals (e.g. *Christmas*), for example. Geographical names cover names of continents (e.g. *Europe*), countries (e.g. *Germany*), towns (e.g. *Passau*), lakes (e.g. *Lake Michigan*), mountains (e.g. *Mount Everest*). Other locative names consist of a proper noun and a descriptive element, such as *river*, *street* and *airport* (e.g. *Hyde Park*). Of course other proper names, such as for newspapers (e.g. *The Times*) occur as well (cf. Glück 2010: 169-170; Payne & Huddleston 2010: 515-518; see also Biber et al. 2007: 245-247).

3.7.3 Anaphoric and non-anaphoric use

Proper nouns are mostly definite, at least as anaphors they have to occur so. Furthermore, if a proper name is anaphoric, it often refers to another proper name, or to a noun phrase with a definite article. If a proper name refers to another proper name, only those instances are regarded as anaphoric relationships, where the two items are not completely identical in form, except for cases where the antecedent is an anaphor itself. The aim to apply anaphora resolution to text retrieval systems seems to justify such a procedure. Establishing links between different expressions is more valuable than examining whether two proper names with the same form are coreferential. Furthermore, the focus is not on coreference resolution but on anaphora resolution. This procedure in fact does not only apply to proper names, but to all other types of anaphors as well. As a result, example (157) does not contain an anaphor, but example (159) does. The second *Ms Smith* [2] in (159) refers anaphorically to the first *Ms Smith* [1], which is itself anaphorically related to *Betty Smith* (cf. Payne & Huddleston 2010: 520).

Furthermore, if a proper name refers anaphorically to an antecedent, this is always a coreferential relationship and constitutes lexical cohesion. To find out about if items are coreferential and therefore anaphors, world knowledge is often needed (example (160)). As proper names can be divided up into different classes, they also show meanings that are specific of the individual class. Such information is helpful in finding the antecedent, which has to share the same or a related meaning. For example, a temporal proper name is likely to have an antecedent that also denotes some type of temporal feature. Finally, a cataphoric reference is not possible with proper names. An inversion of a proper noun and its antecedent, usually a noun phrase, would instead constitute an anaphoric noun phrase with a definite article in most cases (cf. Halliday & Hasan 2008: 42-43; Payne & Huddleston 2010: 520).

Not every occurrence of a proper name is anaphoric, even if it is definite. Contrary to a large number of definite noun phrases that are introduced the first time by a noun phrase with an indefinite article, there is no need for proper names to be introduced. For instance, the first-mentioned proper names at the beginning of examples (157), (158) and (159) are all non-anaphoric. Furthermore, if two proper names have the same form, this does not automatically imply that they denote the same person or entity. Consider example (161) where the conjunction *therefore* signals that the two items are not coreferential (cf. Halliday & Hasan 2008: 281).

(159) **Melanie Smith** has written a book and had it published recently. I am sure that **Ms Smith** [1] will be more successful than Toby Clark. Ms Smith [2] is the better author of the two.

(160) Isabel likes **her new English teacher**. Mr Kennedy is from Reading.

(161) Linda arrived in London last night. Therefore, *Linda* has to prepare the guest bed.

3.7.4 Summary

The distinction between anaphoric and non-anaphoric proper names depends on the surrounding text and is not inherent of an expression. A cataphoric interpretation of proper names is not possible. Anaphoric proper names are coreferential and cohesive (cf. Kübler n.d.: 7). Finally, proper names are not primarily used for reduction but for variation, as is the case with noun phrases with a definite article.

3.8 Indefinite pronouns

The items that belong to indefinite pronouns are *one(s)*, *other(s)*, *another*, *both*, *all*, *each*, *enough*, *several*, *some*, *any*, *either*, *neither*, *none*, *many* & *much/more/most*, *few/fewer/fewest*, *little/less/least* (cf. Stirling & Huddleston 2010: 1512; Quirk et al. 2012: 377, 865, 870-871). *Every* is not included because it is a determiner and cannot work as a pronoun itself (cf. Quirk et al. 2012: 377, 381-383). Additionally, cardinal and ordinal numbers working as indefinite pronouns are not treated here. For instance, number (162) is seen as containing an ellipsis of *guest* after the expression *the second* (cf. Carter & McCarthy 2006: 390).[26]

(162) The first **guest** arrived on time, the second ___ was ten minutes late.

In general, Quirk et al. (2012: 376) divide indefinite pronouns into two subcategories: compound and *of*-pronouns. By compounds they understand indefinite pronouns consisting of two morphemes, the second being *-one*, *-body* or *-thing*. All others belong to the *of*-category because an *of*-phrase can follow, e.g. *few of the examples*. *Of*-pronouns show a substitutional relationship to antecedents, as Quirk et al. (2012) point out: “All the *of*-pronouns can be interpreted as substitutes” (ibid.: 380). They can take noun phrases or nominals as antecedents. *Of*-pronouns are therefore relevant when discussing anaphors; the compound pronouns are not anaphoric.

[26] Quirk et al. (2012: 376-392, 865) also list *half* with indefinite pronouns. However, this item is not included here but rather seen as a case of ellipsis (see chapter 3.11). If it belonged to this category, other forms such as *quarter* would have to be included in this class as well. The category would then expand endlessly. Moreover, Aarts & Aarts (1986: 58) do not add *half* or *quarter* to their category of quantifiers.

There are some more characteristics that unite indefinite *of*-pronouns. For instance, they are per definition indefinite, i.e. "[n]ot referring to, or indicating reference to, an identifiable individual or set of individuals" (Matthews 2007: 188). Their meaning can, however, also be definite if they are used together with definite elements, for example with a definite article, e.g. *the one*. Furthermore, indefinite pronouns have a quantitative meaning.

In addition, the fact that pronouns and determiners show close connections in their morphological form characterises all indefinite pronouns. For instance, the item *few* is a pronoun in the phrase *the last few* and works as a determinative in *a few examples*, but the morphological form is the same in both. Consequently, *of*-pronouns and their corresponding determiners are homomorphs, i.e. they share the same morphological form but take different syntactic functions, either as head or determinative of noun phrases. The only exception is *none*, which is always a pronoun (cf. Halliday & Hasan 2008: 42; Quirk et al. 2012: 70-71, 376, 871). As will become clear in the individual subchapters, most determinative uses of indefinite pronouns are non-anaphoric. All indefinite pronouns that can work anaphorically are now outlined in chapters 3.8.1 to 3.8.7, before some general non-anaphoric uses of these items, their cataphoric use and their status as elliptical forms are discussed.

3.8.1 *One* and *ones*

If *one* works as a substitute, it shows a singular (*one*) and a plural form (*ones*). It substitutes a whole phrase with a count noun as head (example (163)) or a nominal that is a count noun (example (164)).[27] If the antecedent is a whole noun phrase, only the singular form *one* can be used. The anaphor *one* then occurs as head of a noun phrase and does not have any determiners or modifiers. If the antecedent is head of a noun phrase only, both items *one* and *ones* can be used and have to be accompanied by a determiner and/or modifier, of which the latter is usually an adjective. However, *ones* needs not necessarily refer to a plural antecedent and *one* to a singular antecedent. A reference to a singular form with *ones* and plural form with *one* is possible in some cases (example (165)) (cf. Halliday & Hasan 2008: 91). Additionally, the expressions *that of* and *those of* are more usual than *one* and *ones* in formal English. A good illus-

[27] *One* is not used for non-count nouns. Instead, *some* is needed in such contexts, for example, *Would you like some more* ***orange juice*** *or do you already have* <u>*some*</u>? (cf. Carter & McCarthy 2006: 119; Quirk et al. 2012: 870).

tration is given in (166), with *those* and two alternatives with *ones*. Moreover, it is not common to use *one* as a cataphor (cf. Carter & McCarthy 2006: 119-120, 251; Biber et al. 2007: 357; Stirling & Huddleston 2010: 1511-1517; Quirk et al. 2012: 386-388, 869-870).

One working as a substitute has to be distinguished from non-anaphoric uses, such as numerical *one* and generic *one*. Numerical *one* occurs as determinative (e.g. *one book*) and head of a noun phrase (e.g. *one of the books*). Generic *one* is used in formal contexts to denote "people in general" (example (167)). If used specially with a definite article, *one* means "person" (example (168)). The forms of genitive *one's* and reflexive *oneself* can be used generically as well. Generic *one* shows similarity to *you* and *they*, which are also used to stand for "people in general". However, *one* also encompasses the speaker, which *you* and *they* do not do. In addition, *one* occurs in the split construction of reciprocal pronouns. But this was discussed in the chapter about reciprocal pronouns (chapter 3.2), so such cases are excluded in the examination of indefinite pronouns (cf. Stirling & Huddleston 2010: 1513-1516; Quirk et al. 2012: 386-388).

(163) I need **a pen**. Do you have one?

(164) I gave him a yellow **crayon**, but he wanted a green one.

(165) Do you like **apples** because I have a very sweet one for you.

(166) The **theories** of Bühler are as complex as
- those of Chomsky.
- the ones of Chomsky.
- Chomsky's ones.

(167) To be successful *one* has to work at lot.

(168) Paul is the *one* needed here.

3.8.2 *Other*, *others* and *another*

The item *other* is a singular form; the plural form is *others* (example (169)). The item *other* occurs as determiner, e.g. *two other reasons*, and pronoun. It is non-anaphoric as determiner. Furthermore, *others* is used non-anaphorically if it refers to "other people in general" (example (170)). Finally, *other* is found with reciprocal pronouns in the split construction, as illustrated in chapter 3.2. Therefore, the indefinite pronoun *other* needs to be distinguished from such reciprocal pronouns (cf. Stirling & Huddleston 2010: 1517-1518).

Closely connected to *other* is the item *another* because it can be regarded as a combination of *an* and *other*. *Another* can also be seen as an item of *of-*

pronouns (see, for example, Carter & McCarthy 2006: 390).[28] Its anaphoric use (example (171)) has to be distinguished from the reciprocal *one another* in split construction and from determinative use, in which it is non-anaphoric (cf. Payne & Huddleston 2010: 391; Quirk et al. 2012: 388-389).

(169) I have found some of the **documents**, but where are the <u>others</u>?
(170) You do not like cheese, but *others* do.
(171) I lost the **copy** you gave me. Could I have <u>another</u>?

3.8.3 *Both*, *all* and *each*

Both is used with plural count nouns or coordinated nouns if two entities are considered (example (172)). The non-anaphoric uses include *both* as determiner and postposed *both*. Postposed *both* occurs if *both* follows a noun phrase, as in example (173). The expression *the children both* can be reformulated by either *the two children* or *both children*. That is why such instances are not regarded as anaphors. *Both*, as well as *all* and *each*, can have split antecedents (cf. Halliday & Hasan 2008: 156-158).

The item *all* occurs in a similar context. It is used with plural count nouns or coordinated nouns, but now referring to more than two in number, as well as with non-count nouns (example (174)). *All* can also occur as determiner with singular count nouns but then it is non-anaphoric (example (175)). With *of* in example (175), *all* is a pronoun, without *of* it is the determiner of a noun phrase. Such a use, however, is not often encountered: "Before a singular count noun, however, *all* is somewhat formal, and is frequently replaced by a construction with *whole* as an adjective or noun" (Quirk et al. 2012: 381). The expression *all of the village* in example (175) would then read *the whole village* where *whole* is an adjective, or *the whole of the village* in which *whole* is a noun. Moreover, the cases in which *all* is non-anaphoric are the same as with *both*, regarding determinative and postposed uses. A further non-anaphoric use of the pronoun *all* is if it occurs in situations in which it has an equivalent meaning to *everything*. Thus, *everything* in example (176) can replace *all*. Finally, non-anaphoric *all* can

[28] Quirk et al. (2012: 379-380) do not list *other* and *another* with *of*-pronouns as substitutes first. They, however, include them in the detailed description of these pronouns (see ibid.: 388-389, 865). In addition, Carter & McCarthy (2006: 249-250) list them as substitutes with indefinite pronouns.

occur as modifier in the form of an adverb, as in example (177). In such a use it stands for “completely” (cf. Payne & Huddleston 2010: 377; Hornby 2010: 37).

In contrast to *both* and *all*, *each* is only found with singular count nouns, although the pronoun *each* can have a plural count noun or coordinated nouns denoting two people or things as antecedent (example (178)). The antecedent is then interpreted to substitute *each* in the singular. In example (178) the anaphor stands for *each boy*. The item *each* is used non-anaphorically in postposed position and if occurring determinatively or as modifier (cf. Payne & Huddleston 2010: 377; Quirk et al. 2012: 380-383, 870-872). *Both*, *all* and *each* in determinative function, however, can be anaphoric in contexts such as in (179).

(172) The **boys** are already tired. Both got up early.
(173) The children *both* like ice-cream.
(174) We do not have any **butter** at home. Billy used all for the cake.
(175) *All* (of) the village agree with you.
(176) *All* is fine.
(177) His marks are not *all* that bad.
(178) When the two **boys** were asked to show their homework, each looked embarrassed at the teacher.
(179) **Kareem and Nasim** visited us yesterday. Both children are really polite.

3.8.4 *Enough* and *several*

The item *enough* occurs as pronoun in its anaphoric use. It takes plural count nouns and non-count nouns (example (180)). Non-anaphorically, *enough* occurs as determiner and, in the form of an adverb standing for “to the necessary degree”, as modifier (example (181)). The item *several* is used with plural count nouns (example (182)). The item *several* is also non-anaphoric if it functions as determiner or modifier (cf. Hornby 2010: 505; Payne & Huddleston 2010: 391-392, 396-397; Quirk et al. 2012: 388, 870-871, 1140-1142).

(180) We did not buy any **apples**, and still we had enough for the weekend.
(181) You were not friendly *enough* to that customer.
(182) Linda has many **DVDs** but still she is always borrowing several from me.

3.8.5 *Some* and *any*

Some usually occurs with plural nouns (example (183)) and non-count nouns (example (184)). Only determinative *some* is used with singular count nouns. Similarly to *some*, *any* can be used with singular and plural count nouns (example (185)) and non-count nouns. *Some* is an assertive form, i.e. it is linked to positive statements (example (184)), *any* is a non-assertive form, i.e. it is linked to negative statements (example (185)). If *some* and *any* are determiners or modifiers, they are non-anaphoric, as the determiner *some* in example (183). As modifier, *some* stands for "approximately" as in *some twenty people*; *any* is used in place of "at all" (example (186)). Moreover, *some* can occur generically to refer to "people in general", which is a further non-anaphoric use. Example (187) illustrates such a case. *Some* is then interpreted as *some people* (cf. Hornby 2010: 57, 1469; Huddleston 2010c: 1131; Payne & Huddleston 2010: 380-381, 385; Quirk et al. 2012: 83-84, 380, 383-384, 389-392, 870-872).

(183) Timmy found *some* friends last year.
(184) Linda saved the **money I gave her**. She only spent some on new clothes.
(185) This time I took so many **pictures during my trip to China**. Last year I did not take any.
(186) He cannot run *any* faster.
(187) *Some* are really interested in politics, but most are not.

3.8.6 *Either*, *neither* and *none*

It seems that *either* as a pronoun takes antecedents that contain plural count nouns or coordinated nouns (example (188)). Determinative *either* is only used with singular count nouns, in which case it is usually non-anaphoric (cf. Payne & Huddleston 2010: 388-389). *Either* in both uses as pronoun and determiner is restricted to two entities. In this way it contrasts with *any*, which is used if the choice is among three or more entities. Example (188) shows *either* denoting only two people. *Either* could here be paraphrased by *either Jack or Mae*. This means that *either* always substitutes a whole noun phrase, and not a noun as with previous indefinite pronouns.

The item *neither* has similarities to *either* in that it denotes two entities (example (189)). It is used as pronoun in *of*-phrases and takes antecedents that are plural count nouns or coordinated nouns. As with *either*, the item *neither* also works as a substitute for a whole noun phrase. Furthermore, *either* and *neither* can both have split antecedents (cf. Halliday & Hasan 2008: 157-158). *Either* and *neither* can also work as adverbs, such as in (190), in which case they are non-anaphoric (cf. Hornby 2010: 488, 1024; Huddleston, Payne & Peterson 2010: 1308).

None occurs if three or more entities are denoted (example (191)). Anaphoric *none* refers to singular and plural count nouns, and to non-count nouns. *None* differs from indefinite pronouns discussed so far in that it is only used as a pronoun. The corresponding determiner would be *no*, which is not relevant here. *None* is non-anaphoric if it is used as modifier in the form of an adverb (example (192)). In such a use it stands for "not at all" (cf. Hornby 2010: 1036; Quirk et al. 2012: 377, 391-392, 870-872).

(188) If **Jack and Mae** need the car to pick up Cindy, <u>either</u> can use it.
(189) **My two best friends** wanted to visit me at my parents' house, but <u>neither</u> knew that I was in Italy.
(190) He wasn't *either*.
(191) Max has several **cars** but <u>none</u> works.
(192) Although he has read a lot about psychology, he is *none* the wiser.

3.8.7 *Many* and *much/more/most*, *few/fewer/fewest* and *little/less/least*

Quirk et al. (2012) comment on the relationship of these quantifiers to each other as follows: "*Many* ['a large number'] contrasts with *a few* ['a small number'], and *much* ['a large amount'] contrasts with *a little* ['a small amount']" (ibid.: 384).[29] The items *many* and *much* are absolute forms, *more* is the comparative and *most* the superlative form (cf. Herbst, Stoll & Westermayr 1991: 141). *Many* occurs with plural count nouns (example (193)), *much* with non-count nouns (example (194)). The comparative and superlative forms of *many* and *much* are the same. *Much* as well as *more* and *most* occurs non-anaphorically as modifier in the form of an adverb (example (195)). Here *much* stands for "to a great degree".

Few is an absolute form, *fewer* a comparative, *fewest* a superlative. All three items are used with plural count nouns (example (196)). *Few* denotes a bit of a smaller number than *several*. Furthermore, *little* is the absolute, *less* the comparative and *least* the superlative form. All three items occur with non-count

[29] Square brackets cited from the original.

nouns (example (197)). *Less* and *least* are also found with plural count nouns, but only in informal language. Finally, all items discussed here can be used determinatively. *Few* and *little* as well as their comparative and superlative forms are also modifiers, in which cases they are non-anaphoric. An example of a modifier in the form of an adverb is given in (198) (cf. Hornby 2010: 902, 938, 995, 997, 1003; Huddleston 2010c: 1131, 1164-1166; Payne & Huddleston 2010: 394; Quirk et al. 2012: 380, 384-386, 388, 458).

(193) **Students** are usually hard-working, therefore, many achieve high results.
(194) We finished some of **the work** today, but there is still much left to do.
(195) Thank you very *much*.
(196) Last week there were only **fifty people** in the audience. Yesterday there were even fewer.
(197) This year I drank more **alcohol**, while Steven drank less.
(198) These shoes are a *little* bit too small for me.

3.8.8 Further non-anaphoric uses of all *of*-pronouns

All indefinite pronouns discussed so far have some non-anaphoric characteristics in common. For instance, indefinite pronouns can work deictically. Furthermore, if an *of*-phrase follows an indefinite pronoun, the expression is considered to be non-anaphoric. This feature can be exemplified by regarding *all* and *both*. If *all* and *both* precede *the*, there is a choice whether to use *of* or not. As a result, speakers can choose whether *all* and *both* are used as pronouns that are postmodified by an *of*-phrase (example (199) a)) or are used as determiners (example (199) b)) (cf. Quirk et al. 2012: 381). Other items that frequently take an *of*-postmodification are, for instance, *either*, *neither*, *each*, *some*, *any*, *many*, *much*, *few* and *several* (cf. Payne & Huddleston 2010: 413).

In addition, Stirling & Huddleston (2010: 1516) use the terms “implicitly partitive” for expressions without *of*-postmodification, such as example (199) b), and “explicitly partitive” for indefinite pronouns that are postmodified by an *of*-phrase, such as (199) a). This distinction also illustrates that indefinite pronouns with postmodifying *of*-phrases are not anaphoric. The exception is *none*, as this item can only be pronominal. Consequently, indefinite pronouns with *of*-postmodification are not regarded as anaphors.

(199) a) *All* of the restaurants in Marble Street are excellent.
b) *All* restaurants in Marble Street are excellent.

3.8.9 Cataphoric use

A cataphoric interpretation of the items in this category is found much less often than an anaphoric direction. Cataphors can occur across clauses, but not across sentences. Example (200) shows a cataphor (cf. Carter & McCarthy 2006: 248; Halliday & Hasan 2008: 78-79). According to Quirk et al. (2012: 868), the conditions where cataphors occur with forms of substitution are the same as with personal pronouns. However, the anaphor-antecedent relation of substitution is mainly anaphoric and rarely cataphoric (cf. Halliday & Hasan 2008: 145).

(200) If you need <u>one</u>, there is **a ruler** in the drawer.

3.8.10 Indefinite pronouns as elliptical forms

Another point of interest deals with the question of whether the forms except for *one*, *other* and *another* and *none* – as they are only pronouns if used anaphorically – really show substitution or are better seen as cases of ellipsis. For instance, Halliday & Hasan (2008: 155-160, 162) regard indefinite pronouns as instances of ellipsis. Quirk et al. (2012) tend to label them as ellipses as well, even though they concede that this can lead to difficulties and that “many examples will accept a different type of analysis” (ibid.: 871). Examples (201) and (202) illustrate that point. To interpret *each* in example (201) as ellipsis would be acceptable because the full form is *each of Martin and David*. Such an analysis, however, is problematic in example (202). The expression *all of Linda, Bob and Steven* seems questionable (cf. Carter & McCarthy 2006: 250, 252; Quirk et al. 2012: 871-872).

As a result, Payne & Huddleston (2010: 419-421) avoid speaking of substitution and ellipsis with pronouns. Instead, they analyse such instances in terms of a “fused-head construction”. They state:

> Fused-head NPs [i.e. noun phrases] are those where the head is combined with a dependent function that in ordinary NPs is adjacent to the head, usually determiner or internal modifier [...]. (ibid.: 410)

As a result, they regard examples such as (193) as instances in which *many* is the fused form of *many students*. Now, the procedure adapted in this book is different: all items mentioned explicitly in the list above are regarded as substitutions, other instances not included are then treated as cases of ellipses. As indefinite pronouns contribute to a more reduced text than substitutional forms, they are a means of reduction.

(201) **Martin and David** both applied for the job. Each was invited for an interview.

(202) **Linda, Bob and Steven** are watching TV. All like watching cartoons.

3.8.11 Summary

Of indefinite pronouns, only *of*-pronouns can be anaphoric. Of these, pronouns usually function as anaphors; they are reductive and are treated as forms of substitution rather than as elliptical forms. A cataphoric use is very rare. Furthermore, all items here share some non-anaphoric uses, e.g. in determinative function with some exceptions. An overview of the characteristics of indefinite *of*-pronouns is given in Tables 18 and 19.

Table 18: Anaphoric use of indefinite *of*-pronouns (cf. also Quirk et al. 2012: 377)

	Anaphor		**Antecedents**						**Cataphor**
	Determiner	**Pronoun**	**Number**			**Form**			
			Singular count noun	Plural count noun	Non-count noun	Noun phrase	Head of noun phrase	Split antecedents	
one	-	×	×	×	-	×	×	-	(×)
ones	-	×	×	×	-	-	×	-	-
other	-	×	×	-	-	-	×	-	-
others	-	×	-	×	-	-	×	-	-
another	-	×	×	-	-	-	×	-	-
both	(×)	×	(×)[1]	×	-	×	×	×	-
all	(×)	×	(×)[1]	×	×	×	×	×	-
each	(×)	×	(×)[1]	×	-	×	×	×	-
enough	-	×	-	×	×	×	×	-	-
several	-	×	-	×	-	×	×	-	-
some	-	×	-	×	×	×	×	-	-
any	-	×	-	×	×	×	×	-	-
either	-	×	×[2]	×	-	×	-	×	-

neither	-	×	×[2]	×	-	×	-	×	-
none	-	×	×	×	×	×	-	-	-
many	-	×	-	×	-	×	×	-	-
much	-	×	-	-	×	×	×	-	-
more, *most*	-	×	-	×	×	×	×	-	-
few, *fewer*, *fewest*	-	×	-	×	-	×	×	-	-
little, *less*, *least*	-	×	-	(×)	×	×	×	-	-

[1] Only with split antecedents or coordinated nouns
[2] With coordinated nouns

Table 19: Non-anaphoric use of indefinite *of*-pronouns (cf. also Quirk et al. 2012: 377)

	Determiner	**Modifier**	**Non-anaphoric uses of pronouns**			***everything***	**Split construction of reciprocal pronouns**[1]	**Numeral**	**Adverb**
			Deixis	Generic	Post-posed				
one	×	-	×	×	-	-	×	×	-
ones	-	-	×	×	-	-	-	-	-
other	×	-	×	-	-	-	×	-	-
others	-	-	×	×	-	-	-	-	-
another	×	-	×	-	-	-	×	-	-
both	×	-	×	-	×	-	-	-	-
all	×	×	×	×	×	×	-	-	×
each	×	-	×	-	×	-	×	-	-
enough	×	×	×	-	-	-	-	-	×
several	×	×	×	-	-	-	-	-	-
some	×	×	×	×	-	-	-	-	×
any	×	×	×	-	-	-	-	-	×
either	×	×	×	-	-	-	-	-	×
neither	×	×	×	-	-	-	-	-	×
none	-	×	×	-	-	-	-	-	×
many	×	-	×	×	-	-	-	-	-
much	×	×	×	-	-	-	-	-	×
more, *most*	×	-	×	-	-	-	-	-	-
few, *fewer*, *fewest*	×	×	×	×	-	-	-	-	-
little, *less*, *least*	×	×	×	-	-	-	-	-	×

[1] Here seen as "non-anaphoric" because such cases belong to reciprocal pronouns

3.9 Other forms of coreference and substitution: *the same*, *such* and *so*

Some further forms of substitution that are not indefinite pronouns and do not fall into one of the other categories should not be forgotten. These are *the same* and *so*. *Such* is a case of coreference[30] (Quirk et al. 2012: 865).

3.9.1 *The same*

The same is an expression in which *same* can work as the modifier (example (203)) or head of a noun phrase (example (204)). As for a modifier, a noun has to follow *same* and the whole expression refers back to the antecedent. In this case, *same* is an adjective or adverb, not a pronoun (cf. Hornby 2010: 1353; Stirling & Huddleston 2010: 1545).

The item *the same* is a form of substitution, which implies that anaphor and antecedent do not refer to the same entities. For instance, in example (204) Tony and Susan both want a cup of coffee, but will not share one cup, of course. By paraphrasing the second sentence of this example, it is possible to formulate alternatively: *Susan would like one, too*. This "additive element of meaning" (ibid.: 873) is what distinguishes *the same* from other forms of substitution.

The same shows some non-anaphoric uses. First, the expression *all the same* does not work anaphorically. Second, *the same* can be used deictically (cf. Huddleston 2010c: 1138-1140; Stirling & Huddleston 2010: 1545; Quirk et al. 2012: 636). Third, *the same* can be part of the verb phrase *do the same*, which is discussed in a separate category (see chapter 3.10). It is therefore regarded as non-anaphoric with other forms of coreference and substitution.

(203) **Mr. Miller** smashed the window. The same person is responsible for the holes in the road.
(204) Tony would like **a cup of coffee**. Susan would like the same.

[30] An exception is if *such* refers to a whole clause or sentence. In that case, *such* belongs neither to coreference nor substitution.

3.9.2 *Such*

Such is either pronoun (example (205)) or determiner (example (206)). It can work anaphorically in both functions, although the pronoun function is much less frequent. As for a determiner, *such* together with the whole phrase constitutes the anaphor, e.g. *such sacrifices* in example (206). A special case of *such* is if it plays the role of a complement of *as* (example (207)). The antecedent of *such* can be a noun phrase (example (207)) or a whole clause (example (206)). In general, *such* has close similarities to demonstrative pronouns and can often even be replaced by a demonstrative pronoun. For instance, in example (205) *this* or *that* can be used instead of *such*, without any major changes of meaning (cf. Summers 2006: 1390).

Non-anaphoric uses of *such* include a variety of possibilities. First, *such* as degree modifier is not anaphoric, as in (208). Additionally, *such* can occur as part of the subordinators *such ... (that)* and *such ... as*. In example (209), *such* is a degree modifier in the superordinate clause, whereas *that* introduces the subordinate clause. Second, the use as complement of *as* has to be distinguished from the non-anaphoric, metalinguistic function of *as such*, for instance in (210). This example could be paraphrased by *The newspaper is excellent in the strict sense of the term*. The inversion of *such* and *as*, i.e. *such as*, is also non-anaphoric because *such* is here used for exemplification (example (211)). Third, Stirling & Huddleston (2010: 1546) interpret examples such as (212) as anaphors because they regard it as comparative form. The full form of the sentence would therefore have to be: *However, the drills are not such an enjoyment as Joe's games*. Such instances do not show coreference, but rather are cases of ellipses, and therefore are ignored in this category and considered with ellipses (cf. Huddleston 2010c: 1142-1143; Stirling & Huddleston 2010: 1546-1547; Quirk et al. 2012: 76, 257, 376, 999-1000, 1109, 1142-1144, 1315).

(205) I would be happy to **take over the duty of spokesperson** if such is required.

(206) **The room was very small, I did not have many pieces of furniture**, but such sacrifices were necessary in order to go to Harvard.

(207) She only told **true stories**, but they are not seen as such today.

(208) It's *such* a problem to get good waiters.

(209) There was *such* a noise (that) we went out to see what was happening.

(210) The newspaper is excellent as *such* but the layout could be better.

(211) Linguists *such* as Halliday and Hasan contributed much to text linguistics.

(212) **Joe's games** are fun. However, the drills are not *such* an enjoyment ____.

3.9.3 *So*

It depends on how the item *so* is used, whether it is an adverb or a conjunction (cf. Hornby 2010: 1462-1463).[31] There are various anaphoric uses of *so*. First, it can substitute a clause (example (213)). *So* can also be preposed, i.e. occur in the front position of a sentence (example (214)). Here, a whole clause is antecedent again, but the item *so* can also occur at the beginning of clauses and not substitute a whole clause (example (215)). Second, *so* can substitute adjective phrases as in example (216) (cf. Stirling & Huddleston 2010: 1535-1539; Quirk et al. 2012: 879-883, 1323).

Third, *so* can occur as a substitution for the non-finite complement[32] of a lexical verb if it occurs in an *if*-sentence (example (217)). Such a use is only possible with some verbs such as *wish* and *choose*. The non-reduced *if*-clause in example (217) is *if you wish to attend the conference* with *so* substituting the non-finite complement. Similarly, complements of auxiliary verbs can be substituted and again, the antecedent is commonly a non-finite clause. Fourth, *so* occurs in reduced main clauses, frequently in the expression *more so* and *less so* (example (218)). The full sentence of the example reads: *Eleanor needs to practice English, Eleanor needs to practice English even more than Lucy*. Fifth, *so* can substitute expressions where *thus* or *in this way* could be used instead of *so*. Number (219) demonstrates such an instance (cf. Stirling & Huddleston 2010: 1535-1539; Quirk et al. 2012: 879-883).

(213) **Tina will come to the party.** At least I hope so.
(214) **Betty will come over for tea.** So she said, at least.
(215) John was **very happy**. So would Linda be if she had won the race.
(216) They were **very tired**, or at least they seemed so.
(217) Furthermore, you can **attend the conference** if you so wish.
(218) **Eleanor needs to practice English**, even more so than Lucy.
(219) Nowadays, **part-time** employment is increasing steadily. The people so employed earn substantially less than full-time employees.

[31] Summers (2006: 1318-1319) states that *so* can also be an adjective.

[32] The term "complement" here is not used in the sense of clause functions, but in the distinction between "complement" and "adjunct" (cf. Huddleston 2010a: 52-54, 59).

Apart from *so* as anaphor, some non-anaphoric uses need to be mentioned as well. *So* can occur deictically either as degree modifier (example (220)) or as manner adjunct (example (221)). A corresponding indexing act or action has to go with such utterances. As a degree modifier, the demonstrative *this* or *that* could replace *so*; in the case of a manner adjunct, *thus* is possible instead of *so*. Furthermore, if a subordinate clause is introduced by *as*, *so* can occur in the main clause (example (222)). *So* then means *in the same way/time as that*. With regard to their anaphoricity, Stirling & Huddleston (2010) argue: "*So* is here at most only marginally anaphoric, indicating likeness with what has gone before" (ibid.: 1539). Such cases are therefore regarded as being non-anaphoric in this book. Additionally, it is important to mention that *so* can be found in the expression *do so*, which is discussed in chapter 3.10. Furthermore, *so* occurs non-anaphorically as conjunction. Finally, *so* is used idiomatically, e.g. in *or so* and *and so on* (cf. Halliday & Hasan 2008: 139-140; Quirk et al. 2012: 1466).

(220) The crocodile was *so* long.
(221) To make a ship, you have to fold the paper, *so*.
(222) As awful as the story began, *so* it ended.

3.9.4 Cataphoric use

With regard to a cataphoric interpretation, the case is analogous to the substitutional forms of indefinite pronouns with *the same* and *so*. Cataphors are rather infrequent, though possible (example (223)). Another example is (224), where *so* refers to an adjective phrase. *Such* usually takes no cataphoric interpretation (cf. Carter & McCarthy 2006: 248, 254; Halliday & Hasan 2008: 141).

(223) If they said so, **they will take care of your cat**.
(224) Brian was, and still is so, **very aware of his shortcomings**.

3.9.5 Summary

Pointing out the most important aspects of *the same*, *such* and *so*, all three are forms of reduction, either through coreference or through substitution. *Such* and *so* can have cataphoric reference, which, however, is rarely found. Additionally, all items have to be distinguished from various non-anaphoric uses. A summary is given in Tables 20 and 21.

Table 20: Anaphoric use of other forms of coreference and substitution

	Anaphor									**Cata-phor**	**Antecedent**				
	Pronoun	Determiner	Modifier	Head of NP	Complement	Complement of *as*	Preposed	*more so*/ *less so*	Manner adjunct		NP	Clause	Non-finite clause or complement	Adjective phrase	Prepositional phrase
the same	-	-	×	×	-	-	-	-	-	(×)	×	×	-	-	×
such	×	×	-	-	-	×	-	-	-	-	×	×	-	×	-
so	-	-	-	-	×	-	×	×	×	(×)	×	×[1]	×[2]	×	×

[1] If a complement or preposed
[2] After *if*

Table 21: Non-anaphoric use of other forms of coreference and substitution

	Deixis	Conjunction	Degree modifier	Manner adjunct	Comparison	Metalinguistic *as such*	*such as*	*all the same*	*as* in subordinate clause, *so* in main clause	Idiomatic uses	*do so*/ *do the same*
the same	×	-	-	-	-	-	-	×	-	-	×
such	×	-	×	-	×[1]	×	×	-	-	-	-
so	×	×	×	×	-	-	-	-	×	×	×

[1] Constitutes a form of ellipsis

3.10 Verb phrases with *do* and combinations with *so*, *this*, *that*, *it* and *the same (thing)*

The items discussed here are the present tense forms *do*, *does*, the past tense form *did*, the present participle (*-ing*) form *doing* and the past participle (*-ed*) form *done*. With the exception of the non-finite forms *doing* and *done*, the preceding items have corresponding negative forms, either contracted or not, which are *don't/do not*, *doesn't/does not*, *didn't/did not*. Furthermore, the complex forms *do so*, *do this*, *do that*, *do it* and *do the same (thing)* are treated here. With these, *does*, *did*, *doing* and *done* can be used instead of *do* (cf. Huddleston 2010b: 222-223; Quirk et al. 2012: 875).

The verb *do* and the complex forms, in which the main verb *do* is one element, can substitute a verb without auxiliaries, plus any complements and adjuncts (cf. Halliday & Hasan 2008: 113). Additionally, these complex forms have to be distinguished from related categories mentioned above, in which they are not considered if combining with *do*: these are *so* and *the same*, the demonstrative pronouns *this* and *that*, and the personal pronoun *it*. Otherwise, such items would be analysed twice, e.g. as *do so* and also as *so*. Moreover, the complex forms refer to the antecedent and so establish cohesion as a whole, rather than by being split into two parts (cf. ibid.: 126).

3.10.1 Simple forms with *do*

An important point with the simple forms of *do* is the distinction between the main verb *do* (example (225)) and *do* as an auxiliary verb (example (226)). Quirk et al. (2012: 874-875) only treat the main verb *do* as a substitutional form. In case of an auxiliary, they speak of an ellipsis after *do*. They argue:

> A careful distinction must be made, however, between *do* as an auxiliary verb with the status of dummy operator, and *do* as a main verb [...]. It is in the latter function that *do* is a substitute form.
>
> As an auxiliary, on the other hand, *do* is structurally parallel to other operators [such as *can*] [...]. In such cases, we will talk of ellipsis of the predication, rather than of substitution for the predicate. (ibid.: 874)

Yet elsewhere Quirk et al. (2012) concede for examples such as (227):

> This is strictly, however, not ellipsis, but quasi-ellipsis [...], since the insertion of the omitted predication after *did* would result in an unacceptable sentence [...]. [It] is unacceptable

> because DO occurs in a context where the dummy operator cannot occur. But there are other constructions (*eg* clause negation, subject–operator inversion, emphatic operator constructions) in which the operator would occur for independent reasons, and in these cases DO fulfils the conditions of standard ellipsis of the predication [...]. (ibid.: 905-906)

Quirk et al.'s argument is therefore only partly followed in this book. In cases where auxiliary *do* cannot occur together with the unit it refers to, i.e. quasi-ellipsis, these are treated as verb phrase anaphors. Only in contexts in which auxiliary *do* combines with what is left out, ellipsis is assumed. Consequently, example (226) is a case of ellipsis because the full sentence is *I have a dog, but John does not have a dog*. However, the main verb *does* in example (225) is used instead of repeating *speaks English perfectly* and therefore constitutes a case of verb phrase anaphor. The non-reduced sentence so reads: *At least, I think she speaks English perfectly*. *Did* in example (227) is a dummy operator, but constitutes a verb phrase anaphor nevertheless because *Yesterday Tim played volleyball better than Mike did played volleyball* is not possible.

(225) Mary **speaks English perfectly**. At least, I think she does.
(226) I **have a dog**, but John does not ___.
(227) Yesterday Tim **played volleyball** better than Mike did.

Furthermore, some non-anaphoric uses of the simple forms of *do* occur. In general, non-anaphoric uses occur as unreduced main verbs (example (228)). In addition, all uses of auxiliary *do* in combinations with main verbs are non-anaphoric. As auxiliary the present forms *do*, *does* and the past form *did* occur. Consequently, only these items have to be distinguished from uses of substitutional *do*. Moreover, there are certain constructions in which non-anaphoric auxiliary *do* is prone to be confused with substitutional uses of *do*: negation (example (226)), question (example (229)) and emphasis (example (230)). In these constructions, an auxiliary is needed. The auxiliary *do* occurs if no other auxiliary such as *will* or *can* is present. Such uses resemble substitutional *do*, but are forms of ellipsis because the auxiliary combines with the antecedent if inserted, e.g. *Do you play the piano?* in (229) (cf. Stirling & Huddleston 2010: 1523-1524; Quirk et al. 2012: 132-133).

(228) He *did* the washing-up.
(229) I **play the piano**. Do you ___?
(230) Steven insists he did not **watch the film**, but he DID ___.

3.10.2 Complex forms with *do*

Do so, *do this*, *do that* and *do it* are the central complex forms with *do*. Additionally, *do the same (thing)* is included here because the item *the same* has been treated above. To begin with, the items of *do so* are slightly more formal than the simple forms with *do*. The element *so* usually follows *do*, but can precede with the *-ing*-form *doing so* (example (231)). It is not permitted to use *do so* deictically, which means that *do so* is always anaphoric. Whereas *do so* is an idiom, *do this*, *do that*, *do it* and *do the same (thing)* are not. Consequently, the characteristics of these non-idiomatic items are the same as those of their two elements. Stirling & Huddleston (2010) maintain:

> [T]heir meaning and properties can be predicted from those of ***do*** and the NP as used in other combinations. The anaphoric nature of the VPs [i.e. verb phrases] headed by ***do*** [...] is attributable to *it* and the demonstratives, for ***do*** occurs with the same meaning in non-anaphoric VPs [...]. (ibid.: 1532)

These complex forms also show similar non-anaphoric uses as the items on their own. For instance, *do this* and *do that* often occur deictically, in the same way as the demonstrative pronouns do (cf. Carter & McCarthy 2006: 252; Stirling & Huddleston 2010: 1529-1534; Quirk et al. 2012: 876).

Moreover, complex forms differ from each other regarding some additional features. In general, *do so*, *do that* and *do it* can often be used equally, without differences in meaning. However, *do so* is usually the most formal of them. Furthermore, *do that* is used particularly in contexts where contrasts occur and so has an antecedent that fits into this contrast (example (232)). With regard to *do the same (thing)*, it is an alternative to *do that*, especially in comparison. In addition, *do the same (thing)* shows features as *the same*, which means that *do the same (thing)* and its antecedent do not denote identical events, but *do the same (thing)* rather includes the meaning of “too” (cf. Carter & McCarthy 2006: 253-254; Quirk et al. 2012: 878).

(231) He suggested we should **launch an investigation** and we are now in the process of so doing.

(232) The fine weather was perfect for **going shopping**. Elisabeth preferred doing that to studying for the exam.

3.10.3 The form of the antecedent and cataphoric use

For verb phrases with *do* and combinations, the antecedent takes the form of a verb plus any complements and adjuncts (cf. Quirk et al. 2012: 879). The antecedent and the expression that could be inserted instead of the anaphor, however, need not always be morphologically identical. As a result, differences in inflection are generally acceptable. As to complex forms, such differences can even be greater than with simple forms. For instance, the antecedent in (233) is *spoke to my grandmother while she was still alive*, but in place of the anaphor the expression *spoken to my grandmother while she was still alive* has to be inserted. The past form so turns into a past participle form (cf. Stirling & Huddleston 2010: 1525-1526, 1531).

As regards a cataphoric interpretation, it is possible with *do*. Example (234) can serve as illustration. The same goes for complex forms: example (235) shows the item *do so* exemplarily (cf. Halliday & Hasan 2008: 128; Stirling & Huddleston 2010: 1525; Quirk et al. 2012: 875).

(233) I never spoke **to my grandmother while she was still alive**. I now wish I had <u>done so</u>.

(234) If Jenny <u>does</u>, will you also **wear a dress to the party**.

(235) As no scientist has succeeded in <u>doing so</u>, he is striving to **find a solution to the mathematical problem**.

3.10.4 The relationship between anaphor and antecedent

The simple and complex forms of *do* are mostly substitutional forms and show reduction. There is, however, one difference within the complex forms regarding the interpretation of the anaphoric relationship. Stirling & Huddleston (2010) explain this circumstance:

> Anaphoric ***do** it* and ***do** that* characteristically denote specific events, either the same event as that denoted by the antecedent VP or at least the same action involving the same participants as those expressed by the internal complements of the antecedent VP. In contrast, ***do** so* VPs often denote merely the same **kind** of event as the antecedent. (ibid.: 1534)

Consequently, the forms *do it*, *do this* and *do that*, through the influence of their second element, do not show substitution, but usually coreference. They are combinations of the substitute form *do* and the coreferential forms *it*, *this* and *that* (cf. Halliday & Hasan 2008: 125-127; Quirk et al. 2012: 876). If anaphoric items refer to a clause or sentence, they belong to the miscellaneous category.

3.10.5 Summary

The items discussed here are instances of substitution and, with some forms, of coreference. A difference has to be drawn between simple and complex forms, as each of these categories has different characteristics. The most central features are that the simple forms work anaphorically if they occur as main and, in some cases, as auxiliary verbs; the complex forms show close similarities to the features of their second element. The antecedent and the expression that is inserted instead of the anaphor can also differ inflectionally. Finally, a cataphoric interpretation is possible with all items. An overview of the non-anaphoric features is given in Table 22.

Table 22: Non-anaphoric use of verb phrases with *do* and combinations

	As main verb	**As auxiliary verb in:**			**Ellipsis as auxiliary verb**	**See non-anaphoric conditions for *so*, *this*, *that*, *it*, *the same***	**Deixis**
		Negation	Question	Emphasis			
Simple forms							
do	×	×	×	×	×	-	-
does	×	×	×	×	×	-	-
did	×	×	×	×	×	-	-
don't/ do not	-	×	×	×	×	-	-
doesn't/ does not	-	×	×	×	×	-	-
didn't/ did not	-	×	×	×	×	-	-
doing	×	-	-	-	-	-	-
done	×	-	-	-	-	-	-
Complex forms							
do so	-	-	-	-	-	-	-
does so	-	-	-	-	-	-	-
did so	-	-	-	-	-	-	-
doing so	-	-	-	-	-	-	-
done so	-	-	-	-	-	-	-

do this	-	-	-	-	-	×	×
does this	-	-	-	-	-	×	×
did this	-	-	-	-	-	×	×
doing this	-	-	-	-	-	×	×
done this	-	-	-	-	-	×	×
do that	-	-	-	-	-	×	×
does that	-	-	-	-	-	×	×
did that	-	-	-	-	-	×	×
doing that	-	-	-	-	-	×	×
done that	-	-	-	-	-	×	×
do it	-	-	-	-	-	×	×
does it	-	-	-	-	-	×	×
did it	-	-	-	-	-	×	×
doing it	-	-	-	-	-	×	×
done it	-	-	-	-	-	×	×
do the same (thing)	-	-	-	-	-	×	×
does the same (thing)	-	-	-	-	-	×	×
did the same (thing)	-	-	-	-	-	×	×
doing the same (thing)	-	-	-	-	-	×	×
done the same (thing)	-	-	-	-	-	×	×

3.11 Ellipses

The types of anaphors treated so far always had an explicit expression as anaphor. The category of ellipses will now discuss those instances where the anaphor is realised by a gap referring to an antecedent. The items left out, however, have to occur in the text. This excludes ellipsis where the interpretation comes from the situational context. For instance, example (236) can only be clarified if the hearer or reader is familiar to the corresponding situation. Whatever the motivation for uttering the sentence, the missing information cannot be determined from the text (cf. Stirling & Huddleston 2010: 1456; Quirk et al. 2012: 895-896).

The antecedents of ellipses can be inserted in place of the anaphors more or less precisely. Example (237) shows an anaphor whose antecedent is found in the text in exactly the same way as it is required. The expression in example

(238) that has to be filled in instead of the ellipsis, however, does not occur in that form in the text. The alteration that is necessary here involves the grammatical change from a past participle to an infinitive form. Not only antecedents that occur in exactly the required form are regarded in this book, but also where grammatical changes are obligatory. Furthermore, the focus of ellipses lies on those instances that contribute to an informational increase if these are resolved. All in all, ellipses can also be part of anaphoric chains, and they contribute to cohesion (cf. Herbst, Stoll & Westermayr 1991: 183; Stirling & Huddleston 2010: 1457; Quirk et al. 2012: 884-890, 900).

(236) How dare you _ _ _?
(237) He begged me to **go with him**, but I did not want to ___.
(238) He has still not **repaired the machine**, so I will have to ___.

3.11.1 Types of ellipses

The subtypes of ellipses are nominal, verbal and clausal ellipsis.[33] The first category includes instances where a noun is not present in a noun phrase, another item then realises the head (example (239)). The noun can also be accompanied by modifiers, which can be left out together with the noun (example (240)). Such types of ellipses refer anaphorically to parts of a noun phrase. The noun in comparative and superlative noun phrases can also be left out (example (241)) (cf. Halliday & Hasan 2008: 147, 150, 164-166). Additionally, a noun may be elided if a genitive occurs, e.g. *Ms Parkinson's* for *Ms Parkinson's car*. Such genitives are especially used with nouns denoting humans. Coordinate constructions, i.e. clauses connected with *and*, *or* and *but*, for instance, in which a whole noun phrase is missing, are also part of nominal ellipsis (example (242)) (cf. Biber et al. 2007: 296-297, 307).

(239) Normally, twenty **students** are allowed to participate in this course, but for this term we agreed on twenty-five ___.
(240) The last **essay Mary handed in** was better than the first ___.

[33] Quirk et al. (2012: 892-893) distinguish general and special ellipsis. The subtypes nominal, verbal and clausal ellipsis can each contain types that Quirk et al. regard as special ellipsis, e.g. comparative and coordinate constructions. As this book takes the forms and not the constructions in which ellipses occur as basis for the classification, Quirk et al.'s categorisation is not followed.

(241) If you really have to buy a **guitar**, do not get the cheapest ___.
(242) **Cindy** left early in the morning and ___ forgot to wake him up.

Verbal ellipsis covers cases where a verb alone (example (243)), or a verb with object and/or adverbial (example (244) with an object) is missing. Halliday & Hasan (2008) explicitly state that "verbal ellipsis [...] also involve[s] ellipsis that is external to the verb itself, affecting other elements in the structure of the clause" (ibid.: 197). The subject and, in most cases, an auxiliary remain in the sentence in such instances. As mentioned above, *do* as main verb and what Quirk et al. term quasi-ellipsis do not fall into the category of ellipsis, only all other circumstances of auxiliary *do*. Verbal ellipsis can also be found in comparative sentences (example (245)), or in response forms (example (246)). Additionally, verbal ellipsis, together with nominal ellipsis, can occur in coordinations.[34] Example (247) illustrates a case where an auxiliary verb and a noun are left out. Another example that involves nominal ellipsis and a main verb is presented in (248).

Finally, an ellipsis can substitute a whole clause.[35] *To*-infinitive clauses, for instance, can be elided: Take, for example, number (249), where the infinitive clause *(to) play at the concert* has to be inserted. Other types of clausal ellipsis can occur as well: clauses beginning with a *wh*-word, *-ing*-clauses and *that*-clauses. Example (250) shows a *wh*-clause, in which a change in word order takes place: *when Luke will come home* has to be inserted (cf. Halliday & Hasan 2008: 197, 217; Stirling & Huddleston 2010: 1519, 1526, 1542-1543; Quirk et al. 2012: 848-849, 900-911, 1130-1131; see also Swan 2005: chapter 177-182).

(243) You may **drive** if you can ___.
(244) George **likes Mozart**, but Steven does not ___.
(245) She **knows** more than you ___.
(246) Who **called**? – John ___.
(247) **Lucy has** congratulated her brother and ___ given him a present for his birthday.
(248) **Mary invited** her friends but ___ not her neighbours.
(249) You can **play at the concert** if you want to ___.

[34] It is controversial in such cases if the sentence shows ellipsis or a coordination of predications, i.e. of *congratulated her brother* and *given him a present for his birthday* in example (247) (cf. Quirk et al. 2012: 942-945).

[35] Contrary to Halliday & Hasan (2008), clausal ellipsis in this work covers not cases in which "verbal ellipsis and clausal ellipsis go together" (ibid.: 201), but only cases in which a full or nearly full clause is left out.

(250) **When will Luke come home**? Sandy didn’t tell me ___.

An ellipsis can also be interpreted cataphorically. Example (251) demonstrates an antecedent that follows the ellipsis. Cataphoric ellipses underlie the same restrictions as do other types of anaphors. For instance, the elliptical anaphor is in a subordinate clause in (251) (cf. Stirling & Huddleston 2010: 1456, 1523; Quirk et al. 2012: 895).

(251) Do not ask me why ___, but **he has to go see her right now**.

3.11.2 Non-anaphoric ellipsis

Ellipses cannot only be anaphoric but also deictic, or as Carter & McCarthy (2006: 181, 247) term it, not only textual but also situational. Deictic or situational ellipsis is generally found in restricted, conventionalised contexts. For example, non-anaphoric verbal ellipses are possible in instances such as *May I?* if asking for permission to take something. Furthermore, a personal pronoun as subject can be omitted. Such cases are common in informal speech. Usually *I*, prop *it* or existential *there* could be inserted into the non-anaphoric gap (cf. Carter & McCarthy 2006: 181-182, 186; Stirling & Huddleston 2010: 1523, 1540).

3.11.3 Summary

Ellipses contribute to cohesion in that expressions are left out; they are therefore substitutional. Ellipsis is subdivided into nominal, verbal and clausal ellipses in this book. Only anaphoric uses are relevant from these subtypes and these have to be marked off from non-anaphoric uses. Moreover, cataphoric interpretations are also possible. Table 23 shows an overview of ellipses.

Table 23: Characteristics of ellipses

	Anaphoric										**Non-anaphoric: deixis**	**Cataphoric**
	Anaphor – left out:							**Antecedent in:**				
	Noun alone	Noun with modifier	Whole NP	Verb alone	Verb with following elements	Comparative	Superlative	*to-*, *wh-*, *that-*, *-ing*-clause	Coordination	Response forms		
Nominal	×	×	×	-	-	×	×	-	×	-	×	×
Verbal	-	-	-	×	×	×	-	-	×	×	×	×
Clausal	-	-	-	-	-	-	-	×	-	-	×	×

3.12 Non-finite clauses

Non-finite clauses are not prototypical examples of anaphors. In fact, Quirk et al. (2012: 910, 993-995), for instance, regard them as a special type of ellipsis and treat them only rudimentarily in the context of anaphora. As for Stirling & Huddleston (2010), they do not even mention non-finite clauses in their chapter about deixis and anaphora. Nevertheless, non-finite clauses are important for a number of reasons: first, the frequency of non-finite clauses plays a role. Kortmann (2005) states: "English often prefers abbreviated relative clauses (*The man standing at the corner was my uncle*) to finite ones (*The man who stood at the corner was my uncle*)" (ibid.: 180). As will be shown, non-finite clauses also show a high frequency in the corpus. Second, non-finite clauses are reductive devices and so shorten clauses as "[t]hey are more compact" (Biber et al. 2007: 198, 632). As outlined in Kortmann's example, finite relative clauses can usually be used instead of non-finite clauses but these finite clauses are then longer (cf. Quirk et al. 2012: 889). Quirk et al. (2012) explain in more detail:

> Because nonfinite clauses lack tense markers and modal auxiliaries and frequently lack a subject and a subordinating conjunction, they are valuable as a means of syntactic compression. (ibid.: 995)

Third, the consideration of non-finite clauses is the more urgent as they frequently involve subjects as antecedents. Huddleston (2010d) stresses: "The great majority of non-finite clauses have no subject" (ibid.: 1175) and later "but the interpretation requires that we find an 'understood subject'" (ibid.: 1193). Subjects are important because they commonly carry information about the topic of a text, i.e. "there is a significant tendency for the subject to refer to the topic, to what the utterance is primarily about" (Huddleston 2010b: 235). This circumstance can be illustrated by using Kortmann's example. If the two sentences – the first as non-finite clause, the second as its corresponding finite clause – are compared, the relative pronoun – *who* in this case – is only present if the sentence were to be expressed as a finite clause. *Who* is an anaphor and has *the man* – the subject – as antecedent. Additionally, the verb *be* is left out. As a consequence, not considering non-finite clauses would also mean ignoring anaphors such as *who* (cf. Huddleston 2010d: 1211; Quirk et al. 2012: 994-995). Huddleston (2010d) even compares non-finite clauses with personal pronouns. Here he also speaks of anaphora:

> The relation between a missing or covert subject and the controller is a special case of anaphora. It is thus analogous to the relation between a personal pronoun and its antecedent. (ibid.: 1269)

As for the relationship between anaphor and antecedent, *non-finite clause* anaphors show a coreferential aspect as they refer coreferentially to an antecedent. This antecedent would also be the subject of the corresponding finite clause, as seen in the example above. However, the antecedent cannot simply be inserted into the clause but a reformulation is usually necessary. Nevertheless, non-finite clauses are counted as part of the coreferential category.

Non-finite clauses which are characterised by the fact that their verbs are non-finite, i.e. they do not show tense contrast, fall into four different subtypes depending on the form of the verb: *to*-infinitive, bare infinitive, *-ing*-participle and *-ed*-participle. From these, bare infinitive clauses will not be considered here because they are "relatively rare" (Quirk et al. 2012: 993; see also Huddleston 2010d: 1174). A bare infinitive commonly occurs in pseudo-cleft sentences (example (252)) and so is restricted in its use. In such cases, a *to* can be inserted. Additionally, bare infinitive clauses always need their own explicit subject in the function of direct object and then would again be non-anaphoric. Most importantly, clauses containing bare infinitives are hard to detect with an automatic anaphora resolution machine because they are not marked by easily identifiable elements as in comparison to the other three subtypes (*to*, verbs in *-ing*, verbs in *-ed*) (cf. Quirk et al. 2012: 992, 995).

Non-finite clauses in general typically occur as adverbials in complex sentences, or are used in postmodification, usually of noun phrases (cf. Aarts & Aarts 1986: 156; Quirk et al. 2012: 993, 1263-1274). Each of the non-finite clause types will now be discussed in more detail.

(252) What **they** did was (to) repair the pipe.

3.12.1 *To*-infinitive clauses

3.12.1.1 Anaphoric/cataphoric and non-anaphoric use with regard to clause and phrase functions

To-infinitive clauses are clauses beginning with *to*, which are usually followed by a verb. They occur as adverbials in complex sentences without (example (253)) or with a subject introduced by *for* (example (254)). They are relevant as anaphors only if the subject in this non-finite clause is not present (cf. Quirk et al. 2012: 1003-1005). For instance, Biber et al. (2007: 633) maintain that most *to*-infinitive clauses do not include an explicit subject. Non-finite clauses as adverbials can also occur together with subordinators. Quirk et al. (2012) explain:

> Nonfinite […] clauses are subordinate by virtue of the absence of a finite verb as the verb element of the clause. They are, however, sometimes introduced by a subordinator, which generally signals the clause to be adverbial. (ibid.: 1003)

The following subordinators are common for *to*-infinitive clauses: *as if*, *as though*, *in order*, *so as*, *whether … (or)*, *with*, *without*. *With* and *without* have to be followed by a noun phrase, which is not automatically the subject. In the case of *whether … (or)* the subject, if not present in the *to*-infinitive clause, is the same as the subject of the superordinate clause (cf. Quirk et al. 2012: 1003-1005).

If the nominal *to*-infinitive clause itself is the subject or occurs as the extraposed subject, it is frequently non-anaphoric (example (255)). Nominal *to*-infinitive clauses are semantically equivalent to *that*-clauses. They usually express a proposal that would be paraphrased with *should* in *that*-clauses (example (256)), but it is also possible to use them for facts that are considered to be true or for events that have already happened. Furthermore, a *to*-infinitive clause can be found as a postmodification in noun phrases (example (257)) and in adjective phrases, in which they are also nominal clauses (example (258)) (cf. Aarts & Aarts 1986: 86, 117, 121, 160-164; Biber et al. 2007: 198-199; Huddleston 2010d: 1264; Quirk et al. 2012: 150-151, 993, 1061-1063).

(253) To understand the problem, **Toby** needs to gather more information first.
(254) For Cathy *to* be on time, she should be catching the earlier train.
(255) *To* be on time is not necessary.
(256) a) It's natural for mum *to* be with him.
b) It's natural that mum should be with him.
(257) **Peter**'s dream to travel to the moon may soon become reality.
(258) **Dad** is too proud to admit he has made a mistake.

With regard to postmodifications of adjectives, anaphoric as well as non-anaphoric uses are found. Quirk et al. (2012) distinguish between seven types of constructions where a *to*-infinitive clause postmodifies an adjective as subject complement. Five of these constructions are anaphoric, the other three types are usually non-anaphoric. Quirk et al. (2012) state for these constructions:

> In Types (i-iv) the subject of the main clause […] is also the subject of the infinitive clause. […] For Types (v-vii), on the other hand, the subject of the infinitive is unspecified, although the context often makes clear which subject is intended. [However,] [i]n these types it is possible to insert a subject preceded by *for* […]. (ibid.: 1226-1227)

This means that the types v to vii are non-anaphoric, whether the *to*-clause contains a subject (example (259)) or not (example (260) a)). The first type, i.e. type v in Quirk et al., is found in example (260) a). A reformulation with *to handle* at the beginning of the clause ((260) b)) or a paraphrasing resulting in an extraposition ((260) c) is possible with this construction. Adjectives occurring in such constructions "refer to degrees of ease or comfort" (Quirk et al. 2012: 1229). These are the adjectives *awkward*, *convenient*, *difficult*, *easy*, *hard*, *impossible*, *nice*, *pleasant*, *tough*, *tricky* and *unpleasant* (see Table 24).

The next type (type vi) is found in example (261) a). The infinitive clause in this construction can usually be left out ((261) b) or substituted by a passive ((261) c) without any major changes of meaning. Examples of such adjectives are *available*, *fit*, *free*, *frosty*, *ready*, *soft*, *sufficient*. These *to*-infinitive clauses can also end in a preposition (example (262)). Unfortunately, Quirk et al. do not provide a comprehensive list of adjectives concerned.

Finally, the last type (type vii) occurs in extraposition, as in example (263). Here, the anticipatory subject *it* introduces the sentence. Adjectives taking this construction are, for instance, *essential*, *fortunate*, *important*, *lucky*, *possible*, *strange*, *surprising* and *vital*. Additionally, adjectives "chiefly naming evaluative attributes of persons" (Quirk et al. 2012: 1230) are possible with type vii. These are *careful*, *careless*, *crazy*, *foolish*, *greedy*, *mad*, *nice*, *silly*, *unwise*, *wise*, *wonderful* and *wrong*. If an explicit subject is present with these items, the subject is introduced by an *of*-phrase instead of a *for*-phrase (example (264)) (cf. Huddleston 2010d: 1193; Quirk et al. 2012: 1226-1231).

(259) Robin is difficult for my uncle *to* handle.
(260) a) Robin is difficult *to* handle.
b) *To* handle Robin is difficult.
c) It is difficult *to* handle Robin.
(261) a) The towel is soft *to* touch.
b) The towel is soft.
c) The towel is soft *to* be touched.
(262) Susan is easy *to* talk to.
(263) It is important *to* do your homework.
(264) It is wrong of Jane not *to* tell him.

Table 24: *To*-infinitive-clause postmodification of adjectives

Type v: *Robin is difficult to manage.*	*awkward*	*nice*
	convenient	*pleasant*
	difficult	*tough*
	easy	*tricky*
	hard	*unpleasant*
	impossible	
Type vi: *The towel is soft to touch.*	*available*	*ready*
	fit	*soft*
	free	*(in)sufficient*
	frosty	
Type vii: *It is wrong not to tell him.*	*essential*	*careful, careless*
	fortunate	*crazy*
	important	*foolish*
	lucky	*greedy*
	possible	*mad*
	strange	*nice*
	surprising	*silly*
	vital	*unwise, wise*
		wonderful
		wrong

With regard to the antecedent, Huddleston (2010d) argues: “In general, the recovery of the understood subject is determined by semantic principles, not rules of syntax. But in some cases the matter is more grammaticalised” (ibid.: 1260). Later he adds: “[Some non-finite clauses] are syntactically related to the main clause in that the missing subject is controlled by the subject of the main clause” (ibid.: 1266). Because of this, the antecedent can often be derived from the sentence, depending on the function in which the non-finite clause occurs. If the *to*-infinitive clause is direct object without its own subject, the antecedent is always the subject of the superordinate clause (example (265)). If the *to*-infinitive clause follows an indirect object (example (266)), the preceding object is the antecedent of the anaphor. The anaphor *to* in subject complement and appositive use usually refers to the subject of the superordinate clause, often only to the specifying determiner as in examples (267) and (268) (cf. ibid.: 1260).

The antecedent in postmodifications is as follows: as postmodifications of nouns, the anaphors refer to the part of the noun phrase preceding the postmodification. Paraphrasing such *to*-infinitive clauses, a modal interpretation with *should*, for example, might occur. If an adjective is postmodified, however, the antecedent is the subject of the superordinate clause (example (258)), except for cases where this subject is a prop *it*, as in example (269), or existential *there*. *To*-infinitive clauses in the latter cases imply an indefinite subject or an *I* of the speaker, i.e. no referential link can be found in the text (cf. Quirk et al. 2012: 993, 995, 1061-1062, 1185-1187, 1196, 1215-1216, 1226-1229, 1265-1269).

Finally, cataphors frequently occur with these non-finite clauses, especially in adverbial function. Example (253) is an instance with a cataphor (cf. ibid.: 910).

(265) **Anne** likes <u>to</u> read books by Ken Follett.
(266) She told **George** <u>to</u> feed the dog.
(267) **Dave**'s solution would be <u>to</u> go by bus.
(268) **Linda**'s plan, <u>to</u> learn more than others, requires much ambition.
(269) It is not wise *to* leave your luggage unattended.

3.12.1.2 Further non-anaphoric uses

A non-intelligent anaphora resolution system has also to mark off uses of *to* in other contexts than non-finite clauses. First, *to* works as preposition, as the second *to* in example (270). Second, *to* is non-anaphoric as prepositional adverb, e.g. *walk to and fro* (cf. Quirk et al. 2012: 715). Third, *to* is also not anaphoric if it is part of verbs such as *ought to*. This third non-anaphoric use should now be discussed in more detail.

Using the classification of Quirk et al. (2012: 136-148), verbs that include such non-anaphoric *to* are marginal modals (*ought to*, *used to*), modal idioms (*be to*, *have got to*), semi-auxiliaries (e.g. *be able to*, *have to*) and catenatives (e.g. *happen to*). Quirk et al. (2012) do not provide a finite list of verbs concerned with semi-auxiliaries and catenatives but only examples of important verbs (see Table 25). The items given for semi-auxiliaries are: *be able to*, *be about to*, *be apt to*, *be bound to*, *be due to*, *be going to*, *be likely to*, *be meant to*, *be obliged to*, *be supposed to*, *be willing to*, *have to* (see first *to* in example (270)). For practical purposes, as the verbs to be excluded have to be defined explicitly, only the verbs listed above are considered non-anaphoric in the case of semi-auxiliaries.

For catenatives, Quirk et al. (2012: 146-147) list the following items: *appear to*, *come to*, *fail to*, *get to*, *happen to*, *manage to*, *seem to*, *tend to*, *turn out to*. But yet, Aarts & Aarts (1986: 161-164) treat examples listed with catenatives as "predicator complements" where the verb *fail*, for instance, is complemented by a *to*-infinitive clause. Quirk et al. (2012) also acknowledge for catenatives: "Such constructions have meanings related to aspect or modality, but are nearer to main verb constructions than are semi-auxiliaries" (ibid.: 146). Likewise, Huddleston (2010d) explains for catenative verbs: “In many cases the non-finite complement has a finite alternant or near-alternant” (ibid.: 1226), which is a typical feature of anaphoric non-finite clauses. As a result, catenatives are here treated as verbs plus non-finite clauses and therefore as anaphoric.

Table 25: Non-anaphoric *to* with verbs

Marginal modals	*ought to* *used to*		
Modal idioms	*be to* *have got to*		
Semi-auxiliaries	*be able to* *be about to* *be apt to* *be bound to*	*be due to* *be going to* *be likely to* *be meant to*	*be obliged to* *be supposed to* *be willing to* *have to*
Other “fixed expressions”	*to begin (with)* *to cap it (all)* *to conclude*	*to continue* *to recap* *to recapitulate*	*to start (with)* *to summarise* *to sum up*

All items of marginal modals, modal idioms and semi-auxiliaries are, however, only non-anaphoric, if really occurring as such. That means, not every case where, for example, *be* and *to* are used in the combination *be to*, this is automatically a non-anaphoric use. Example (271) shows an instance in which *to*, although preceded by a form of *be*, is anaphoric. Therefore, it is important to mention in which case the following marginal modals, modal idioms and semi-auxiliaries are non-anaphoric. To start with, *used to* “denotes a habit or a state that existed in the past” (Quirk et al. 2012: 140). Furthermore, Quirk et al. (2012) describe the non-anaphoric use of *be to* as “an idiom expressing futurity, with varied connotations of ‘compulsion’, ‘plan’, ‘destiny’, etc, according to context” (ibid.: 143). Finally, *have to* has a similar meaning to *must* (cf. ibid.: 145).

(270) He has *to* walk *to* school every Friday.
(271) **John**’s task is to reorganise the department.

There are also some debatable uses. *To* can be part of multi-word verbs, for example, *listen to*. As such verbs are rarely followed by an infinitive and because no exhaustive list of them exists, they are not considered with non-anaphoricity. Furthermore, some fixed expressions, e.g. *to begin (with)*, *to cap it (all)*, *to conclude*, *to continue*, *to recap*, *to recapitulate*, *to start (with)*, *to summarise*, *to sum up* are non-anaphoric. Finally, instances are excluded if an *-ing-*

participle verb follows *to*, e.g. *prior to receiving*, as potentially anaphoric *to* has to be followed by an infinitive. Such cases could, however, be an example of an *-ing*-participle clause (cf. Quirk et al. 2012: 150, 1069, 1150-1161).

3.12.1.3 Summary

To-infinitive clauses can occur as adverbial, direct object, subject complement, in appositive use and sometimes as subject. *To*-infinitive clauses as postmodifications are possible in noun and adjective phrases. The antecedents are often the nouns preceding, the specifying determiner and the subjects, especially with adjective phrase postmodification. Cataphors are often found if non-finite clauses function as adverbials. Explicit subjects are introduced by *for*, in which case the non-finite clause is non-anaphoric. "Implicit" subjects, i.e. antecedents, are often the subjects of the superordinate clauses. Non-anaphoric uses include instances with understood or unspecified speaker or hearer. Non-anaphoric *to* can occur as preposition, prepositional adverb, and conjunction, or as part of marginal modals, modal idioms, semi-auxiliaries, and of fixed expressions, such as *to sum up*.

3.12.2 *-ing*-participle clauses

3.12.2.1 Anaphoric/cataphoric and non-anaphoric use with regard to clause and phrase functions

Verbs with *-ing*-inflection introduce *-ing*-participle clauses. If a non-finite clause has its own subject, it is non-anaphoric. Subjects are often preceded by a preposition (example (272)). The subject of an *-ing*-participle clause can also be in the genitive case, especially with pronouns denoting a personal reference. This, however, constitutes a formal style (example (273)). An example of an anaphoric *-ing*-participle clause is (274) (cf. Quirk et al. 2012: 150-151, 910, 993).

-ing-participle clauses can function as adverbial in complex sentences (examples (272) and (274)). The conjunctions that are used to introduce such *-ing*-participle clauses are: *although*, *as if*, *as though*, *even if*, *if*, *once*, *though*, *unless*, *until*, *when(ever)*, *whether ... or*, *while*, *whilst*, *with*, *without*. After *with* and *without* a noun phrase has to occur, though, this noun phrase is not automatically the subject of the *-ing*-participle clause. Additionally, *after*, *before* and *since* are used. These items are, according to Quirk et al. (2012: 1005-1006), better regarded as prepositions. As with *to*-infinitive clauses, *-ing*-participle clauses frequently take cataphoric antecedents if they occur in sentence-initial position as

adverbials (example (274)) (cf. Stirling & Huddleston 2010: 1477; Quirk et al. 2012: 1063-1064, 1194).

As direct object, *-ing*-participle clauses contrast with *to*-infinitive clauses. Quirk et al. (2012) explain their difference as follows:

> As a rule, the infinitive gives a sense of mere 'potentiality' for action, as in *She hoped to learn French*, while the [*-ing*-]participle gives a sense of the actual 'performance' of the action itself, as in *She enjoyed learning French.* (ibid.: 1191)

In addition, *-ing*-participle clauses occur as postmodifications in noun phrases (example (275)) (cf. Aarts & Aarts 1986: 117-121, 126, 160-161; Quirk et al. 2012: 1063-1067, 1230-1231). Furthermore, *-ing*-participle clauses only occur in rare use as object complements, in which case they are non-anaphoric (example (276)). Similarly, nominal *-ing*-participle clauses in subject position (example (277)) are also often non-anaphoric. Additionally, *-ing*-participle clauses are not anaphoric if they occur in extraposition (example (278)), although this is mostly restricted to informal speech (cf. Biber et al. 2007: 199-200; Quirk et al. 2012: 1392-1393). They are in this case introduced by *it*.

(272) With the children *watching* TV, we could now talk about this problem you mentioned.
(273) The students were very understanding. I was immensely relieved about their *accepting* my apologies.
(274) <u>Lying</u> in the hammock, **Frank** fantasised about this girl he had met.
(275) **The students** <u>working</u> on their theses attended a course about writing skills.
(276) Susan regarded the birth of her child as *being* the best thing that could have happened to her.
(277) *Walking* alone on the streets at night can be dangerous.
(278) It's no use *arguing* about that.

The antecedent of *-ing*-participle clauses is often the subject of the superordinate clause. *-ing*-participle clauses that occur as direct object and follow an indirect object take the preceding object as antecedent. Noun phrases that are postmodified by *-ing*-participle clauses have the preceding part of the noun phrase as antecedent. Sometimes, the antecedent is found in the determinative of a noun phrase, as in example (279), where the *-ing*-participle clause is the subject.

As with *to*-infinitive clauses, the subject can be left implicit in *-ing*-participle clauses and thus be a case of non-anaphoricity. Similarly, the antecedent with some verbs is not the subject of the superordinate clause, but is indefinite, i.e. non-anaphoric. As a consequence, not Dr Miller eats fruit and vegetables every day in example (280), but he recommends this practice to other people. The verbs belonging to this category are, according to Quirk et al. (2012: 1189-1190), *discourage*, *envisage*, *forget*, *involve*, *justify*, *permit*, *recall*, *recommend*, *remember*, *risk* and *save*. Furthermore, the implied subject can be generic (example (277) and (281)) or arise from the situation. The distinction between anaphoric and non-anaphoric interpretations is not always that clear. In some circumstances, the subject of non-finite clauses can be interpreted in two ways and thus be ambiguous from the sentence structure itself. Example (282) can be interpreted as *Tom detests it when he lies* or alternatively, with an unspecified subject, *Tom detests it when people lie* (cf. ibid.: 1065-1066, 1189, 1194-1195, 1202).

Finally, an *-ing*-participle clause can usually be paraphrased as a finite clause. However, the corresponding finite clause does not necessarily result in progressive aspect. Quirk et al. (2012) state:

> Unlike *-ed* participle [...] clauses, however, these *-ing* participle clauses cannot be regarded as strictly elliptical clauses, since the *-ing* participle does not necessarily represent a progressive form in the equivalent finite clause. The *-ing* participle neutralizes that aspectual distinction [...]. (ibid.: 1005)[36]

Therefore, example (283) paraphrases in a finite clause as *The apparatus which examines / is examining the heartbeat of new-borns attracts the attention of the experts*, which means that both simple as well as progressive aspects are possible without any further context. The tense is that of the clause in which the *-ing*-participle clause is embedded or it has to be inferred from the context. The paraphrase with a finite clause can have a modal or non-modal interpretation. In the previous example it is non-modal, but in example (281) the modal rephrasing *People/We/One could not visit the production hall* is common (cf. ibid.: 1066-1067, 1263-1264).

(279) Insulting the student was not **Sam**'s intention.
(280) Dr Miller recommends *eating* fruit and vegetables every day.
(281) There was no *visiting* the production hall.
(282) **Tom** detests lying.

[36] This argument can also serve as another justification for why non-finite clauses are not seen as ellipses here.

(283) **The apparatus** examining the heartbeat of new-borns attracts the attention of the experts.

3.12.2.2 Further non-anaphoric uses

Non-finite verbs used in anaphoric *-ing*-participle clauses have to be marked off from other verbs ending in *-ing*. These are *-ing*-participles that are part of complex finite verb phrases and follow *be* (example (284)), *have* or modal auxiliary verbs. Furthermore, verb phrases containing *doing* are not considered here as these are items that are part of the category verb phrases with *do* and combinations. Other words ending in *-ing* are especially nouns (e.g. *thing*) or verbs used as gerunds, i.e. "verb forms with a noun-like role" (Matthews 2007: 158) (example (285)); full verbs in base form (e.g. *sing*); and adjectives (e.g. *interesting*). It is sometimes difficult to distinguish between nouns ending in *-ing* and verbs in *-ing*-participle. For instance, *driving* is an anaphoric *-ing*-participle in example (286) a) and a noun that is non-anaphoric in b) (cf. Quirk et al. 2012: 97, 151-152, 1063-1065).

(284) We are *testing* his newest invention.
(285) *Swimming* is one of her favourite pastimes.
(286) a) **Nasim** likes driving fast cars.
b) Sue's *driving* is horrible.

3.12.2.3 Summary

A non-anaphoric interpretation is likely in subject and object complement position. If the *-ing*-participle clause contains an explicit subject, this subject is often introduced by a preposition. Explicit subjects can also occur as genitives in noun phrases. Such clauses are then non-anaphoric. Furthermore, *-ing*-participle clauses in adverbial function take cataphoric antecedents, if they are at the beginning of sentences. In addition, for *-ing*-participle clauses in subordination the antecedent frequently takes the function of the subject in a superordinate clause, and in some cases it is the object. The antecedent is not always a whole phrase but can also be a noun phrase except for its postmodification.

Other non-anaphoric *-ing*-participle clauses are such with unspecified subjects. In addition, the *-ing*-ending of non-finite verbs has to be distinguished from other non-anaphoric uses: *-ing*-participles in complex finite verb phrases; gerunds; and full verbs in base form, nouns and adjectives ending in *-ing*.

3.12.3 *-ed*-participle clauses

-ed-participle clauses are typically characterised by the *-ed*-ending of verbs. Nevertheless, irregular verbs used as past participle do not show this ending. For these irregular forms, Quirk et al. (2012: 114-120) provide a comprehensive list. The items taken from this list are presented in Table 26.[37] It should be pointed out that the forms that are irregular in their past forms but have a regular *-ed*-participle are not included. Furthermore, all regular forms that occur as alternatives to irregular forms are left out because they can be identified with their *-ed*-inflection. Moreover, the following items are excluded from this list although they are irregular: *bled*, *bred*, *fed*, *fled*, *led*, *misled*, *overfed*, *pled*, *shed*, *shred*, *sped*, *wed*. These forms are detected by automatic systems as well, because they end in *-ed*, even if this ending is not an inflection. Finally, *been* and *had* are omitted because it is doubtful if they can work as anaphors in *-ed*-participle clauses (cf. ibid.: 150-151).

Table 26: Irregular *-ed*-participle forms

abode	felt	misdealt	rewritten	stunk
arisen	fought	misgiven	rid	strewn
awoken	found	misheard	ridden	stridden, strid,
borne	fit	mislaid	rung	strode
beaten, beat	flung	misspelt	risen	struck
become	flown	misspent	run	strung
befallen	foreborne[38]	mistaken	sawn	striven
begotten	forbidden, forbid	misunderstood	said	sworn
begun	forecast	mown	seen	sweat
beheld	foreseen	offset	sought	swept
bent	foretold	outbid,	sold	swollen

[37] Huddleston & Pullum (2010) do not include a list of irregular verbs. If consulting dictionaries, some irregular verbs that are not part of Quirk et al. (2012) are found. For instance, the *Oxford Advanced Learner's Dictionary of Current English* (2010) lists the following irregular verbs that do not occur in Quirk et al. (2012): *babysat*, *bespoken*, *breastfed*, *browbeaten*, *dripfed*, *floodlit*, *inlaid*, *input*, *intercut*, *interwoven*, *mishit*, *misread*, *output*, *outsold*, *overdrawn*, *overflown*, *overheard*, *overlaid*, *overlain*, *overpaid*, *oversold*, *overspent*, *overwritten*, *preset*, *proofread*, *redrawn*, *reheard*, *resold*, *resat*, *retaken*, *simulcast*, *spotlit*, *stove*, *sublet*, *typecast*, *typeset*, *undercut*, *underlain*, *underpaid*, *undersold*, *underwritten* (cf. Hornby 2010: reference section 2-4).

[38] The *Oxford Advanced Learner's Dictionary of Current English* (2010) lists *forborne* instead of *foreborne* (cf. Hornby 2010: reference section 2-4).

bereft
besought
beset
bestridden, bestrid, bestrode
bet
betaken
bade, bid, bidden
bound
bitten, bit
blown
broken
brought
broadcast
built
burnt
burst
bust
bought
cast
caught
chidden, chid
chosen
cleft, cloven
clung
come
cost
crept
cut
dealt
deepfrozen
dug
done
drawn
dreamt
drunk
driven
dwelt
eaten
fallen
forgotten, forgot
forgiven
forgone
forsaken
forsworn
frozen
gainsaid
got, gotten
given
gone
ground
grown
hamstrung
hung
heard
hove
hewn
hidden, hid
hit
held
hurt
inset
kept
knelt
knit
known
laid
leant
leapt
learnt
left
lent
let
lain
lit
lost
made
meant
met
miscast
outbidden
outdone
outfought
outgrown
outrun
outshone
overborne
overcast
overcome
overdone
overeaten
overhung
overridden
overrun
overseen
overshot
overslept
overtaken
overthrown
partaken
paid
proven
put
quit
read
rebound
rebuilt
recast
redone
relaid
remade
rent
repaid
reread
rerun
reset
restrung
retold
rethought
rewound
sent
set
sewn
shaken
shaven
shorn
shewn
shone
shod
shot
shown
shrunk
shriven
shut
sung
sunk
sat
slain
slept
slid
slung
slunk
slit
smelt
smitten
sown
spoken
spelt
spent
spilt
spun
spat, spit
split
spoilt
spread
sprung
stood
stolen
stuck
stung
swum
swung
taken
taught
torn
telecast
told
thought
thriven
thrown
thrust
trodden, trod
unbent
unbound
underbid, underbidden
undergone
understood
undertaken
underwritten
undone
unfrozen
unmade
unwound
upheld
upset
woken
waylaid
worn
woven
wept
wet
won
wound
withdrawn
withheld
withstood
wrung
written

3.12.3.1 Anaphoric/cataphoric and non-anaphoric use with regard to clause and phrase functions

-*ed*-participle clauses can include a subject (example (287)) or not (example (288)). As with other non-finite clauses, only non-finite clauses without explicit subjects are anaphoric. With regard to the functions, -*ed*-participle clauses can occur as adverbials, in which case they can be introduced by the following conjunctions: *although*, *as* (for manner), *as if*, *as soon as*, *as though*, *even if*, *if*, *once*, *though*, *unless*, *until*, *when(ever)*, *where(ever)*, *whether … or*, *while*, *whilst*, *with*, *without*. With the last two items a noun phrase has to follow, which is not necessarily the subject. A cataphoric interpretation is also possible (example (288)) (cf. Quirk et al. 2012: 910, 993, 1003-1005).

Furthermore, if -*ed*-participle clauses postmodify noun phrases, they can be paraphrased with finite relative clauses. As a result, example (289) can be reformulated: *The goals that were/have been scored by the team were impressive* (cf. Aarts & Aarts 1986: 117-118; Quirk et al. 2012: 1125, 1264-1265). For the detection of the antecedent, the conditions of *to*- and -*ing*-participle clauses apply. This means that the antecedent of -*ed*-participle clauses is often the subject of the superordinate clause. With regard to noun phrases, the part before the postmodification of the -*ed*-participle clause is the antecedent (cf. Quirk et al. 2012: 1004-1005).

(287) With the members *selected*, the conference could begin.
(288) Prepared for massive criticism, **the chairman** opened the meeting.
(289) **The goals** scored by the team were impressive.

In contrast to -*ing*-participle clauses (example (283)), -*ed*-participle clauses can show progressive aspect. This circumstance is illustrated in example (290). Thus, (290) a) takes simple aspect and is paraphrased as *The bike that was/has been repaired by Frank belongs to me*. Example (290) b) reformulates *The bike that is being repaired belongs to me*. Furthermore, -*ing*- and -*ed*-participle clauses differ from each other in that the -*ing*-participle clause is associated with the active voice, the -*ed*-participle clause with the passive, as is obvious if comparing example (275) with (289) (cf. ibid.: 994-1005, 1264-1265).

(290) a) **The bike** repaired by Frank belongs to me.
b) **The bike** being repaired by Frank belongs to me.

3.12.3.2 Further non-anaphoric uses

A few further non-anaphoric uses need mentioning. Obviously, a simple resolution software would face difficulties differentiating between anaphoric *-ed*-participles and other items taking the ending *-ed* or one of the irregular *-ed*-participle forms. As for regular forms, non-anaphoric verbs with *-ed*-ending are participles in complex finite verb forms. Here, the participle follows modal auxiliary verbs; *have*, which signals perfective aspect (example (291)); or *be*, which signals passive voice. Another non-anaphoric use comes from regular verbs in past form (example (292)) and some regular verbs in base form (e.g. *feed*). Apart from verbs, other words, especially adjectives (e.g. *long-established*, *red*), but also some nouns (e.g. *seed*), can end in *-ed*. With regard to irregular forms, irregular *-ed*-participle forms are also not anaphoric in complex verb forms with modal auxiliaries, *have* and *be* (example (293)), which is closely comparable to regular *-ed*-participle verbs. Finally, if these irregular forms operate as adjectives, they are non-anaphoric as well (cf. Quirk et al. 2012: 96-98, 150-151).

(291) We have *played* tennis before.
(292) The man *carried* the pictures to the new gallery.
(293) The actress is *known* by all fans of the show.

3.12.3.3 Summary

-ed-participle clauses include items ending in *-ed* as well as irregular participle forms. Anaphoric *-ed*-participle clauses as adverbials and direct objects often take the subject of the superordinate clause as antecedent. With regard to noun phrases, the antecedent is the part of the noun phrase preceding the postmodification. Moreover, cataphors are common with sentence-initial *-ed*-participle clauses as adverbials.

Anaphoric participle clauses, whether with regular or irregular *-ed*-form, have to be distinguished from non-anaphoric forms in complex finite verb phrases and as adjectives. Regular *-ed*-ending items with a non-anaphoric use are past forms and, marginally, nouns. In case of an explicit subject within the participle clause or if the clause contains an *-ed*-item as object complement, this clause is also non-anaphoric.

3.12.4 Summary of *to*-infinitive, *-ing*-participle and *-ed*-participle clauses

In general, non-finite clauses are characterised and can thus be identified by verbs that are preceded by *to*, verbs ending in *-ing* or *-ed*. With *-ed*-participles, a number of irregular forms occur. The functions that anaphoric non-finite clauses can take are illustrated in Table 27. The subordinate conjunctions that can introduce non-finite clauses as adverbials are summarised in Table 28.

Table 27: Anaphoric use of non-finite clauses

	Clause function					**Phrase function**		
	Adverbial	Subject complement	Direct object	Object complement	Appositive use	Prepositional complement	Postmodification in noun phrases	Postmodification in adjective phrases
to	×	×	×[1]	-	×	-	×	×
-ing	×	×	×	-	×	×	×	×
-ed	×	-	×	-	×	-	×	-

[1] wh-element can precede *to*

Table 28: Conjunctions introducing non-finite clauses as adverbials

	to	*-ing*	*-ed*
although	-	×	×
as	-	-	×
as if	×	×	×
as soon as	-	-	×
as though	×	-	×
even if	-	×	×
if	-	×	×
in order	×	-	-
once	-	×	×
so as	×	-	-
though	-	×	×

	to	*-ing*	*-ed*
unless	-	×	×
until	-	×	×
when(ever)	-	×	×
where(ever)	-	-	×
whether	×	-	-
whether … or	×	×	×
while	-	×	×
whilst	-	×	×
with	×	×	×
without	×	×	×

As to the antecedent, it is frequently the subject of the superordinate clause. If an indirect object precedes the non-finite clause in object function, this is the antecedent. As for noun phrase postmodification, the part of the noun phrase before the postmodification is usually the antecedent (cf. also Quirk et al. 2012: 994, 1120-1127, 1271-1272). As adverbial at the beginning of a sentence, the antecedent takes a cataphoric interpretation. Finally, *non-finite clause* anaphors show a coreferential relationship with their antecedents and therefore belong to the coreferential category.

With regard to non-anaphoricity, items that look the same as anaphoric verbs of non-finite clauses but are in fact non-anaphoric can be found. Furthermore, a *non-finite clause* item shows a non-anaphoric interpretation if it contains an overt subject, i.e. a subject on its own. Moreover, non-anaphoric items can refer to addresser or addressee, either explicitly to *I*, *you*, *we*, or to an implicit subject, e.g. if items occur in sentences that contain an imperative form in the main clause. These non-anaphoric characteristics are summarised in Table 29.

Table 29: Non-anaphoric use of non-finite verbs and other forms looking like non-finite verbs

	In subject position	In extraposition	Verbs with *to*				Simple finite verb phrase		Complex finite verb phrases		
			Marginal modals	Modal idioms	Semi-auxiliaries	Fixed expressions	Present form	Past form	*be*	*have*	Modal auxiliary verbs
to	×	×	×	×	×	×	-	-	-	-	-
-ing	×[1]	×	-	-	-	-	×	-	×	×	×
-ed	×	-	-	-	-	-	×	×	×	×	×
Irregular *-ed*	×	-	-	-	-	-	×	×	×	×	×

	Overt subjects	*you* etc. as antecedent	Imperative	Postmodification of adjectives			Gerunds	Nouns	Adjectives	Prepositions	Prepositional adverbs	Other fixed expressions
				V	VI	VII						
to	×	×	×	×	×	×	-	-	-	×	×	×
-ing	×	×	×	-	-	-	×	×	×	×	-	-
-ed	×	×	×	-	-	-	-	×	×	-	-	-
Irregular *-ed*	×	×	×	-	-	-	-	×	×	-	-	-

[1] anaphoric use is possible, but rare

3.13 Conclusion

Twelve types of anaphors relevant for computational anaphora resolution have been identified. These types and their items are summarised in Table 30.

Table 30: Anaphor types and their items in English

Anaphor type	Subcategories	Anaphor items
Central pronouns	Personal pronouns	*he*, *she*, *it*, *they*, *him*, *her*, *them*, (*we*), (*us*), *he/she*, *he or she*, *s/he*, *(s)he*, *him/her*, *him or her*
	Possessive pronouns	*his*, *her*, *hers*, *its*, *their*, *theirs*, (*our*), *mine*, *ours*, *yours*, *his/her*, *his/hers*, *his or her*, *his or hers*
	Reflexive pronouns	*himself*, *herself*, *itself*, *themselves*, (*ourselves*), *himself/herself*, *himself or herself*, *themself*
Reciprocal pronouns		*each other*, *one another*
Demonstrative pronouns	Dependent function	*this*, *that*, *these*, *those*
	Independent function	*this*, *that*, *these*, *those*
Relative pronouns		*who*, *whom*, *which*, *whose*, *that*, zero *that*
Adverbs		*here*, *there*, *then*, *where*, *when*, *while*, *why*, *whence*, *whereby*, *wherein*, *whereupon*
Noun phrases with a definite article		*the*
Proper names	Personal proper names Other proper names	
Indefinite pronouns		*one*, *ones*, *other*, *others*, *another*, *both*, *all*, *each*, *enough*, *several*, *some*, *any*, *either*, *neither*, *none*, *many*, *much*, *more*, *most*, *few*, *fewer*, *fewest*, *little*, *less*, *least*
Other forms of coreference and substitution		*the same*, *such*, *so*
Verb phrases with _do_ and combinations with _so_, _this_, _that_, _it_ and _the same (thing)_	Simple forms	*do*, *does*, *did*, *doing*, *done*, *don't*, *do not*, *doesn't*, *does not*, *didn't*, *did not*

	Complex forms	*do so*, *does so*, *did so*, *doing so*, *done so*,
		do this, *does this*, *did this*, *doing this*, *done this*,
		do that, *does that*, *did that*, *doing that*, *done that*,
		do it, *does it*, *did it*, *doing it*, *done it*,
		do the same (thing), *does the same (thing)*, *did the same (thing)*, *doing the same (thing)*, *done the same (thing)*

Ellipses	Nominal ellipsis Verbal ellipsis Clausal ellipsis	
Non-finite clauses		*to*, *-ing*, *-ed*

To round the picture off, the twelve anaphor types are now examined regarding the six features that have been outlined in chapter 2. These six features are: consideration of anaphoric and cataphoric directions; necessity of explicit antecedents; interpretation of anaphors in relation to antecedents; consideration of coreferential, substitutional and other relationships between anaphor and antecedent; anaphors as reductive and/or non-repetitive forms; and anaphors as cohesive devices (see Table 31).

As to the first characteristic, the items classified as anaphors are in a large part anaphoric. A cataphoric interpretation occurs only with some items in specific contexts. The most important cataphoric devices are independent *this* and *here*. Halliday & Hasan (2008) state: "This use of *this* [i.e. in independent function], together with the parallel use of *here* [...], is the only significant instance of cataphoric cohesion in English" (ibid.: 68). Furthermore they stress: "structural cataphora is very common, especially with the definite article [...], but it is simply a realization of a grammatical relationship within the nominal group and has no cohesive, text-forming function" (ibid.: 68). As a result, a cataphoric interpretation is of minor importance here.

Next, all antecedents of the anaphors listed here have to occur in the same text, otherwise they are classified as non-anaphoric. This goes back to the fact that an anaphora resolution system needs to find an antecedent in the text. Furthermore, anaphora resolution systems have to distinguish items working as anaphors from those instances in which these items are not anaphoric. This is not a trivial task. The grammatical features listed with the individual types of anaphors help in categorising items into anaphoric and non-anaphoric. Additionally, these features are of great importance for anaphora resolution systems in finding the correct antecedent of each anaphor.

Furthermore, an anaphor is interpreted in relation to its antecedent. It depends on the type of anaphor as to in how far the antecedent is necessary for determining the referent of the anaphor. Proper names, noun phrases with a definite article and demonstrative pronouns in dependent function are on the one end and carry a lot of information regarding the determination of the referent themselves. On the other end we find ellipses. Between these two poles are central pronouns, reciprocal pronouns, relative pronouns, adverbs, other types of coreference and substitution, demonstrative pronouns in independent function, indefinite pronouns, verb phrases with *do* and combinations with *so*, *this*, *that*, *it*, and *the same (thing)* and non-finite clauses.

A similar classification holds for the fact whether the anaphor is more a case of reduction or occurs in order to avoid repetition. Proper names and noun phrases with definite articles are means to vary and so to avoid repetition. Ellipses are prototypical examples of reduction. The other types fall in between these two poles. Another relevant point here is the distinction between grammatical and lexical cohesion. A large part of anaphor types belong to grammatical cohesion; those that are part of lexical cohesion are proper names and noun phrases with a definite article.

With regard to a coreferential or substitutional relationship between anaphor and antecedent, the following aspects are worth mentioning: reflexive pronouns, reciprocal pronouns and proper names are coreferential. Although non-finite clauses show a special coreferential relationship with their antecedents, they also belong to the category of coreference. Furthermore, dependent demonstrative pronouns, adverbs, noun phrases with a definite article and *such* are coreferential, except for anaphors that refer to clauses. They then belong to the third, miscellaneous category, which includes items that are both or neither coreferential nor substitutional. Similarly, relative pronouns, *such*, the verb phrase anaphors *do this*, *do that* and *do it* are counted as part of the miscellaneous category if referring to a clause or sentence.

Personal pronouns are also coreferential, but fall into the miscellaneous category, if the antecedent includes a quantifier and if *it* refers to a whole clause or one or more sentences. Determinative possessive pronouns are coreferential, but are classified into the micellaneous category if the antecedent includes a quantifier or an interrogative pronoun. Independent possessive pronouns are coreferential and substitutional and so belong to the miscellaneous category, apart from first and second person pronouns, which only show substitution. The independent demonstrative pronouns *that* and *those* can also be substitutional. Additionally, relative pronouns are usually coreferential, except for a reference to antecedents that are direct objects of *have (got)*, in which case they are substitutional. Indefinite pronouns, *the same*, *so* and ellipses show substitution. Finally, verb phrases with *do* and combinations with *so*, *the same (thing)* are also substitutional, but the combinations of *do* with *this*, *that*, *it* are coreferential.

Now that the types of anaphors have been identified and marked off from non-anaphoric instances, the focus can move to analysing the frequency of each of these types in the next chapter.

Table 31: Anaphor types with regard to the six conditions

Condition	**Direction**		**Antecedent**	**Interpretation of anaphor**		**Relationship**			**Anaphor**		**Cohesion**	
Anaphor type	Anaphoric	Cataphoric	**in the same text**	Antecedent absolutely necessary	Antecedent not absolutely necessary	Coreference	Substitution	Miscellaneous category	Reduction	Repetition avoidance	Grammatical	Lexical
Central pronouns	×	×	×	(×)	-	×	×	×	×	(×)	×	-
Reciprocal pronouns	×	(×)	×	(×)	-	×	-	-	×	(×)	×	-
Demonstrative pronouns	×	×	×	(×)	×	×	×	×	(×)	(×)	×	-
Relative pronouns	×	-	×	(×)	-	×	(×)	×	×	(×)	×	-
Adverbs	×	(×)	×	(×)	-	×	-	×	×	(×)	×	-
Noun phrases with a definite article	×	-	×	-	×	×	-	×	-	×	-	×
Proper names	×	-	×	-	×	×	-	-	-	×	-	×
Indefinite pronouns	×	(×)	×	(×)	-	-	×	-	×	(×)	×	-
Other forms of coreference and substitution	×	(×)	×	(×)	-	×	×	×	×	(×)	×	-
Verb phrases with *do* and combinations	×	×	×	(×)	-	×	×	×	×	(×)	×	-
Ellipses	×	×	×	×	-	-	×	-	×	(×)	×	-
Non-finite clauses	×	×	×	(×)	-	×	-	-	×	(×)	×	-

4 Anaphors in hypertexts

This chapter discusses anaphors and their frequency with regard to hypertexts. As anaphors are here investigated from the perspective of text retrieval, it is necessary to examine their use in hypertexts. First, the current state of research towards analysing anaphor frequencies is illuminated. Next, hypertexts are defined and their linguistic characteristics are described, which is followed by a classification of hypertexts. In a further section and with the insights of the previous findings, the corpus established for this book and its design are detailed. Finally, the frequencies of anaphors in the corpus are detailed and discussed.

4.1 Previous research

Only a few studies investigate the frequency of the different types of anaphors. The most comprehensive analysis is part of the Syracuse study (cf. Katzer, Bonzi & Liddy 1986; Bonzi & Liddy 1989; Liddy 1990). In this study, the following types of anaphors are considered: central pronouns; relative pronouns; indefinite pronouns; “definite article” (i.e. noun phrases with a definite article); and “pro-verb”, which seems to correspond to the type “verb phrases with *do*” (cf. Liddy 1990: 44). Furthermore, there are types where it is not clear which anaphors are meant. A first point of criticism so is that some anaphor types are not understandable and that essential explanations of them are missing. These types concern in particular “nominal demonstratives”, which are probably demonstrative pronouns in dependent use, and “subject references”, which are perhaps proper names. Additionally, “nominal substitutes” (e.g. *former*, *one*) and “pro-adjectives” (e.g. *another*, *identical*) are listed. Both are partly included in this book with indefinite pronouns, or seen as cases of ellipses. The type “pro-adverbials”, e.g. *such* and *so*, is treated in a separate category here. Moreover, forms such as *second* are rather classified as nominal ellipses in this book; *where* is not regarded as a relative pronoun but as an adverb.

These anaphor types have been analysed in a corpus of 600 abstracts, which has been established in the Syracuse study. The abstracts were taken half from psychology and half from computer science (cf. Bonzi & Liddy 1989: 431; Liddy 1990: 45). However, there are several drawbacks of the study. One weakness concerns the amount of information given. In order to be able to compare this study to others, details about when these items are regarded as anaphoric or non-anaphoric would be necessary. Furthermore, there is no information

DOI 10.1515/9783110416756

about the number of words the abstracts contained in total. In addition, the Syracuse study treats some items as anaphors where it is questionable in which context these are anaphoric if ever, e.g. *everybody*, *something*. Most significantly, important anaphoric items have not been included, such as: reciprocal pronouns; zero *that*; the adverbs *when*, *while*, *why*, *whence*, *whereby*, *wherein*, *whereupon*; the forms of the verb phrases *don't*, *doesn't*, *didn't* and all forms of the combinations with *do*, i.e. *do so*, *do this*, *do that*, *do it*, *do the same (thing)*; verbal and clausal ellipses; and non-finite clauses. Additionally, cataphoric instances of all anaphor types were not part of the analysis (cf. Katzer, Bonzi & Liddy 1986: 57-66; Liddy 1990: 44, 51). In sum, this study is not adequate to be adopted here for analysing the frequency of the anaphor types identified in chapter 3.

Apart from the Syracuse study, other studies focus mainly on central pronouns. A recent study on central pronouns, for example, was carried out by Mitkov & Hallett (2007). Here, more than one corpus was considered for evaluation. First, Mitkov & Hallett (2007) took technical manuals that were downloaded from the Internet, with a total of 55,444 words. In detail, the corpus encompasses texts from "three on-line Linux technical manuals and in particular Access How-To", "CDRom How-To", "Ethernet How-To", "an extract from an Internet Explorer user manual", "documentation for Panasonic TV" and "Aiwa products" (ibid.: 274). Second, they evaluated newswire texts from a part of the Penn Treebank corpus containing 94,500 words. Third, a corpus consisting of Jules Verne's *From the Earth to the Moon* with 4,965 words was investigated (cf. ibid.: 274-276).

With the two most important studies for this book outlined, it should not be forgotten that there are other, minor studies (cf. also Mitkov 2008: 582-583). A few noteworthy examples are listed here. One example is Hobbs (1986). Back then, he had already carried out a minor examination of the frequency of the personal pronouns *he*, *she*, *it* and *they*. The corpus consists of parts of William Watson's *Early Civilisation in China*, Arthur Haley's *Wheels* and *Newsweek* from July 7, 1975. Hobbs, however, provides no information about the total number of words (cf. ibid.: 344-345).

Another study comes from Vicedo & Ferrández (2000: 346-347), who compared the use of anaphoric central pronouns and *who*, *whose*, *whom* across news from the Time newspaper, medical journals, abstracts and extracts from information science and other computational and technical content. The corpora consist of 57,797 sentences. Furthermore, Barbu (2002) looked at central pronouns in technical manuals. These had been downloaded from the Internet and include "Beowulf HOW TO", "Linux CD-Rom HOW TO", "Linux Access

HOW TO" and "Windows Help file" (ibid.: 276). The corpus contains 28,272 words. A further evaluation was done by Mitkov, Evans & Orasan (2002: 177-178). They selected technical manuals in the field of computer hardware and software with a total number of 247,401 words. Mitkov, Evans & Orasan, however, do not mention what pronouns they considered, i.e. central pronouns or only personal pronouns, or e.g. also relative pronouns.

All these studies mentioned so far as examples of the frequency of anaphor types are mostly valuable per se, but are inadequate to determine the relative frequency of all anaphor types to each other. Most studies about such frequencies focus on central pronouns, other types are rarely examined. Even in the Syracuse study (cf. Liddy 1990), detailed information about anaphoric and non-anaphoric use and the frequency of the items is missing. Consequently, this book introduces its own corpus. A new corpus is also necessary due to the fact that the Syracuse study and others dealing with anaphor types are largely based on technical manuals. These are not suitable here, as they are not hypertexts and/or not prototypical examples of these. Moreover, the goal in this book is to use anaphora resolution for text retrieval on the Internet (cf. chapter 5) and as such there is a need for analysing anaphors in hypertexts. The corpus defined here includes all the items described in chapter 3. The procedure of deciding whether an item belongs to anaphoric or non-anaphoric use also follows the detailed arguments of chapter 3, including cataphoric interpretations.

Before presenting the corpus in more detail, it is discussed what types of hypertexts occur on the Internet and what peculiarities they show in comparison to other written or spoken forms of language use.

4.2 Hypertexts

4.2.1 What are hypertexts?

Texts found on the WWW are typically represented as hypertext documents. Indeed, the largest system of hypertexts is the World Wide Web. The invention and use of the HTML protocol made the hypertext globally accessible (cf. Bolter 2001: 39; Schütte 2004: 71). The word *hypertext* goes back to Ted Nelson, who coined it in 1965 (cf. Nelson 1972: 252-255; Storrer 2000: 222, 225; Endres 2004: 35). Which criteria are obligatory for hypertexts is controversial. To date, no generally accepted definition exists, which might also be a result of the different disciplines that are interested in hypertexts. For example, a hypertext can be viewed from a technological-computational, (text-)linguistic, sociological or psychological perspective (cf. Storrer 2000: 223; Schütte 2004: 27, 68; Żebrowska 2013: 115). Typical characteristics of hypertexts that are usually mentioned are computational realisation, links, non-linearity, interactivity and adaptivity/openness.

A hypertext is originally or "genuinely" of an electronic nature (cf. Schütte 2004: 29). For instance, definitions in dictionaries stress the computational aspect. The *Oxford Advanced Learner's Dictionary* explains "hypertext" as a "text stored in a computer system that contains links that allow the user to move from one piece of text or document to another" (Hornby 2010: 766). Furthermore, Storrer (2000) points out quite comprehensibly:

> Die Charakterisierung von Hypertext als computerverwaltetem Text, als Text, der sich nicht ohne Wertverlust auf Papier ausdrucken lässt, findet sich [...] zurecht in verschiedensten Hypertext-Definitionen wieder. [...] Nur durch diese Eigenschaft lässt sich Hypertext vom gedruckten „Textdesign" einerseits, vom computerverwalteten E-Text andererseits abgrenzen [...]. (ibid.: 229)

The term "Textdesign" is used for printed newspapers and magazines, which also show non-linear principles of information organisation in order to facilitate selective reading (cf. Storrer 2000: 230). Additionally, hypertexts should not be confused with e-texts that are also found on the WWW. E-texts are published electronically, for instance, in PDF-format. Consequently, e-texts are similar to printed, linear texts, but they do not belong to the category of hypertexts (cf. Schubert 2012: 133-134).

Additionally, links are a further characteristic of hypertexts. Dictionaries and encyclopedias (e.g. Agnes 2007: 702; Hornby 2010: 766; Glück 2010: 275) mention that the links of hypertexts are one of the most important features. These links furthermore influence the way texts are written and read. They also influence the cohesion and coherence of texts (cf. Storrer 2004: 14), as will be discussed in chapter 4.2.2. Definitions from computer scientists, such as Henrich (2007a: 346-347), for instance, also stress the importance of links. However, the linking of elements is also important from a textlinguistic point of view (cf. Żebrowska 2013: 116; Fix 2014: 22).

Another feature of hypertexts that is typically mentioned in definitions is their non-linearity. It means that there is no predefined sequence among different hypertexts. As a result, the user can select which hypertext he or she reads first; there is no clear beginning or end. A hypertext is usually also not read completely but only in parts. This stems from the nature of hypertexts because it is not possible to prescribe a defined order of reading within the World Wide Web as a huge network of hypertexts. The principle of non-linearity in hypertexts is expressed prototypically by links. Hypertext documents are linked to each other, leading to a network of documents (cf. Schmitz 2004: 43; Schütte 2004: 29; Storrer 2007: 211-212; Storrer 2008: 325-326). Levene (2010) states in this context:

> What differentiates the Web from a mere collection of documents is its hypertextual nature. The hyperlinks that are embedded in web pages allow us to surf the web by following the links that transport us from one page to another. (ibid.: 108)

Apart from that, definitions of "hypertext" list features such as interactivity. This generally means that by clicking on links a system reacts in a specific way, which allows users to select the path of reading themselves, for instance. As a result, hypertext designers can create a compressed text with only the most important bits of information first. In such cases, by clicking on links the users can then decide themselves where they would like to have further details, according to the principle "detail on demand" (cf. Schütte 2004: 68; Storrer 2007: 217). Furthermore, interactivity implies that users can enter search terms, select a particular item from a list or that hypertexts encourage the reader to make a contribution such as to give his or her opinion (cf. Storrer 2012: 286).

A further feature of hypertexts is their adaptivity or openness. For instance, the storing of information and the presentation of this information on a particular screen is not fixed but can rather be adapted to the needs of the users and to the devices used, e.g. a computer screen, smart phone or tablet computer (cf. Storrer 2008: 319; Storrer 2012: 287). Hypertexts are also characterised by the fact that they are open, i.e. changeable or dynamic. Therefore, hypertexts can be extended or updated quickly and easily (cf. Schütte 2004: 68; Storrer 2008: 321; Storrer 2012: 287; Fix 2014: 23).

A hypertext has traditionally been distinguished from hypermedia. Per definition, hypermedia involves not only text, but also images, sounds and videos, i.e. visual and audio information. As hypertexts nowadays hardly consist of only textual information, this difference has become neutralised. Therefore, the term "hypertext" is also used to denote multimedia content, i.e. pages with visual and audio media (cf. Claus & Schwill 2006: 301; Storrer 2007: 212; Storrer 2008: 320-321; Ince 2012: hypertext; Storrer 2012: 286).

Although the computational nature is usually taken as one important characteristic of a hypertext, some current researchers also apply "hypertext" to printed texts (e.g. Jucker 2005: 286; Schubert 2012: 131; Hoffmann 2012: 46, 52-55). Indeed, there are some similarities between hypertexts and printed texts. For example, dictionaries and encyclopedias also show non-linear principles of organisation as they are not designed to be read from beginning to end. Furthermore, all cross-references and footnotes are close to the structures found in hypertexts. Nevertheless, there is a qualitative difference between hypertexts and printed texts: Links in hypertexts can be accessed much more quickly, more precisely and more immediately than links in printed texts, due to their computational nature (cf. Schütte 2004: 69). Bolter (2001) stresses:

> Although in a printed book it would be intolerably pedantic to write footnotes to footnotes, in the computer we have already come to regard this layered writing and reading as natural. Furthermore, the second page is not necessarily subordinate to the first. [...] All the individual pages may be of equal importance in the whole text, which becomes a network of interconnected writings. (ibid.: 27)

It should also be mentioned here that a printed text cannot be copied one to one in order to form a hypertext. Jakob Nielsen (cited in Storrer 2007: 216) points out quite explicitly: "Anything that is a great print design is likely to be a lousy web design."

In sum, the features of computational realisation and non-linearity with the necessity of links seem to be the most important features in many definitions of hypertexts (cf. Storrer 2008: 318). The principle of adaptivity or openness is rather an optional criterion. Huber (2002) offers a textlinguistic definition of "hypertext": "Hypertexte sind im elektronischen Medium realisierte, tendenziell nicht-lineare und potentiell multimedial ausgerichtete Texte" (ibd.: 45).[1] For this book it is important that a hypertext is realised in a computational environment and that it is not confused with e-texts or "Textdesign". Attempts to classify hypertexts are presented in chapter 4.2.4, but before that the question of how cohesion and coherence are represented in hypertexts will be discussed.

4.2.2 Cohesion and coherence in hypertexts

As hypertexts differ from printed texts in a number of ways, consequences for cohesion and coherence in hypertexts arise. Three important consequences are now outlined.

First, as hypertexts are non-linear, writers cannot anticipate the sequence in which recipients read hypertexts (cf. Storrer 2000: 228; Jakobs & Lehnen 2005: 160). This also affects the use of cohesive means. As each hypertext should be coherent, the antecedents of pronouns, for example, have to occur in the same hypertext as the anaphoric pronouns (cf. Giltrow & Stein 2009: 12). Nevertheless, the number of cohesive elements in hypertexts is not generally

[1] Underlining of "Texte" removed.

lower than in printed texts, but it is different (cf. also Schütte 2004: 102). Schubert (2012) argues that lexical cohesion occurs frequently: “Andererseits ist die *lexikalische* Kohäsion in Links stark vertreten” (ibid.: 134). And a bit later: “Dadurch, dass die lexikalische Kohäsion stark vertreten ist, können also grammatische Mittel wie Pronomina [...] in Hypertexten reduziert werden, ohne dass die Kohärenzherstellung ernsthaft gefährdet ist” (ibid.: 135).

Second, text boundaries are not easily perceptible as there is no clear beginning and end. Only parts are perceived at once, rather than the full contents of related hypertexts. Books, on the contrary, are clearly marked off from other text documents in their form. A printed page, for instance, can easily be related to a whole book, which is not the case with a website (cf. Storrer 2004: 35-38; Storrer 2007: 220; Schubert 2012: 130-131).

Third, as readers choose the sequence of which hypertexts they read after each other individually, topics may change from one hypertext to the next. Consequently, the degree of coherence is lower than in printed texts. In some cases the links give clues about what is described in the document and to what it is linked, such as *about the company*. These are “semantically filled links” (Schubert 2012: 134). In other cases the links themselves do not make explicit what linked document should be expected, only the context then gives hints. These are called “semantically empty links” (ibid.: 135), such as *click here*. To identify such expressions in semantically empty links can be of importance for anaphora resolution in order to distinguish non-anaphoric from anaphoric items because items in such links are usually non-anaphoric (cf. Storrer 2004: 35-38; Schubert 2012: 130-131).

Storrer (2004) finally concludes for coherence: “[D]ie hypertextuelle Organisationsform [erleichtert] die Prozesse der Kohärenzbildung im Vergleich zum linear und thematisch kontinuierlich aufgebauten Text nicht unbedingt” (ibid.: 38). Nevertheless, there are also devices that help the user to establish coherence. To give an example, contextualisation devices of navigation show the user the path of the homepage (cf. Schubert 2012: 136). For instance, the website text WS41 (cf. chapter 4.3.4) shows the path *Life & Style > Women > Families*.

4.2.3 Linguistic characteristics of hypertexts

Hypertexts show a use of language that partly differs from traditional written or spoken forms. These features are now described in more detail. Crystal (2008: 31) maintains that the language on the Internet has features of both written and spoken forms (see also Bittner 2003: 53).[2] For Crystal, typical aspects of writing are that it is space-bound, that a time-lag between producing and receiving the message occurs and that a visual contact is lacking between the participants. The last aspect leads to a reduction of deictics such as *here* in writing, as these could be ambiguous. Speech is time-bound, spontaneous and face-to-face. Moreover, the sentences and structures are more complex in writing. They are organised in lines and paragraphs and use other graphic means, such as punctuation. In speech, sentences are shorter and have a looser structure and forms are contracted, such as *they're*. Errors and interruptions can be corrected in writing and are then not visible, but they are noticed in speech. Finally, writing is more suitable to communicate facts, speech more able to establish social functions, such as to build and strengthen relationships (cf. Crystal 2008: 27-31).

Applying these features of writing and speaking to hypertexts, the language on Web pages – although it varies depending on what Web page is analysed – is close to features of writing. Crystal (2008) maintains that "most varieties of written language can now be found on the Web with little stylistic change other than an adaption to the electronic medium" (ibid.: 31). Blogging hypertexts, however, show more features of speech and so deviate from traditional writing conventions. Nevertheless, different blogs vary in their style, as is the case with different Web pages. In Crystal's own words: "Some blogs are highly crafted; others are wildly erratic, when compared with the norms of the standard written language" (Crystal 2011: 21). Blogs from companies and newspapers are usually edited and display more formal language; comments and personal blogs tend to use more informal language (cf. Crystal 2008: 246).

Another feature of hypertexts is that they frequently show multi-authorships, especially in Wikipedia, but also in blogs with comments from different contributors. The linguistic consequences of this are, for example, that the intentions of the contributions or formatting conventions of the writers may vary. Further consequences if various writers are involved are that language changes between a formal and informal style, or that different varieties are used within one hypertext (cf. Jakobs & Lehnen 2005: 165; Crystal 2011: 30-32; Fix 2014: 24).

In addition, although certain Web pages are checked for errors, such as in journalistic writing and with homepages of companies and institutions, both Web and blog hypertexts often do not undergo any editing. As a result, a high quality of these hypertexts cannot be guaranteed. Errors such as *recieve* instead of *receive* occur frequently, specifically in blogs (cf. Crystal 2008: 215-216). In general, two types of errors can be distinguished: typographical and cognitive

[2] For approaches to speech and writing differences see Durant & Lambrou (2009: 11-12).

errors. With the latter, the writer does not know the correct spelling of the word. The majority of errors are typographical errors, however. Among typographical errors, the following categories commonly occur: insertion, deletion, substitution and transposition. Insertion means that an additional character is added, e.g. *arre* instead of *are*. Deletion occurs if one character is left out, e.g. *ar* instead of *are*. Substitution means that one character is replaced by another, e.g. *arw* instead of *are*. Finally, transposition might occur, i.e. if two characters are changed in their position, e.g. *aer* instead of *are* (cf. Jurafsky & Martin 2009: 106-107; Croft, Metzler & Strohman 2010: 198-199).

To correct such errors automatically before processing such texts (e.g. for indexing, see chapter 5.5.2) is not an easy task and requires a lot of resources. For instance, that *form* instead of *from* is an error can only be clarified by looking at the surrounding words (cf. Stock 2007: 306-307). Other errors that do not lead to a different word are more easily detected as these are not included in a dictionary. This strategy has its limitations as well because not all words are found in a dictionary, due to language change and the finite capacity of a dictionary. To give an example, *Merkollande*, a blend of *Merkel* and *Hollande*, is only of recent origin and is not present in a printed dictionary shortly after its first use (cf. Jurafsky & Martin 2009: 113).[3]

4.2.4 Classifications of hypertexts

With regard to hypertexts, different types can be distinguished. A “text type” is defined by Brinker (2010) as follows:

> Ganz allgemein gesprochen können Textsorten als komplexe Muster sprachlicher Kommunikation verstanden werden, die innerhalb einer Sprachgemeinschaft im Laufe der historisch-gesellschaftlichen Entwicklung aufgrund kommunikativer Bedürfnisse entstanden sind. (ibid.: 135)

Different text types can be distinguished in traditional offline as well as in online media. New media, however, do not take over all text types found in traditional media, but rather form new text types by adding unique features as mentioned in chapter 4.2.1. One website can even compass texts that belong to more than just one hypertext type (cf. Bittner 2003: 269; Żebrowska 2013: 102-103, 138; Fix 2014: 28-29).

[3] For more information on error correction see Siddiqui & Tiwary (2008: 71-76) and Fliedner (2010).

Rehm (2007) defines hypertext types generally as “Teilmengen von *Hyper*texten, die sich durch bestimmte relevante gemeinsame Merkmale beschreiben und von anderen Teilmengen abgrenzen lassen“ (ibid.: 7). He argues that it is arbitrary which features are selected for classifying text types. It is indeed controversially discussed which features are necessary for the classification of text types (cf. Jakobs 2003: 234; Rehm 2007: 47). Rehm (2007: 64) takes the operationalisability through computerlinguistic processes as the main criterion of his classification of hypertext types. Consequently, he distinguishes among the following hypertext types: institutional homepages, personal homepages, online newspapers, online encyclopedias, hotlists, weblogs or blogs, guest books and other, miscellaneous types of hypertexts. Hotlists are lists of Web links and were common until 1998. Today they are found only occasionally and therefore they are not relevant here (cf. Rehm 2007: 195-196). Furthermore, Rehm mentions guest books as one hypertext type. Diekmannshenke (2000) characterises guest books as having “Formular- und Listencharakter” (ibid.: 142). Rehm (2007) describes the function of guest books as follows:

> Wesentliche Intentionen der Einträger sind Selbstdarstellung, Kontaktpflege sowie Klatschen und Tratschen, die gruppen- und beziehungskonstituierend wirken, d.h. Gästebücher werden zur Verfolgung der eigenen sozialen und kommunikativen Interessen eingesetzt. [...] Die Nachrichten sind häufig intendiert als unverbindliches Angebot zur Kommunikation und Kontaktaufnahme [...]. (ibid.: 199)

Additionally, Rehm’s other, miscellaneous hypertext types fall into two groups: further interactive types of hypertexts (“Weitere interaktive Hypertextsorten”, ibid.: 200) and hypertext- and Web server-related hypertext types (“Hypertext- und Webserver-bezogene Hypertextsorten”, ibid.: 201). The former include, for example, input masks of search engines, discussion forums and lonely hearts ads. The latter have as a determining feature that they concern the meta-level of a website. They include, for instance, error messages, sitemaps, notes that a website has changed its address (cf. Rehm 2007: 200-204).

Storrer (1999: 6-8) provides another classification. She distinguishes institutional homepages (universities, authorities etc.) from commercial homepages (e.g. companies). This difference is also adopted here. Apart from institutional and commercial homepages, she identifies theme-related homepages (e.g. from online newspapers), private homepages and personal homepages. Rehm subsumes the last two types in his category of personal homepages.

It should be mentioned that there is also a classification by Crystal (2008), who considers the whole Internet and not only hypertexts. He differentiates between seven Internet situations: e-mail, synchronous chatgroups, asynchronous chatgroups, virtual worlds, World Wide Web, instant messaging, blogging (cf. ibid.: 10-18). E-mail is used to exchange messages between users, whereby the messages are sent to private mailboxes. Chatgroups fall into the synchronous type, i.e. if the interaction happens in real time, or asynchronous type, i.e. if the interaction is postponed in time. As Crystal (2008) explains for both types: "Chatgroups are continuous discussions on a particular topic, [...] in which computer users interested in the topic can participate" (ibid.: 11). In synchronous chatgroups, users enter a chat room to communicate, in asynchronous chatgroups, such as message boards, the communication is stored (cf. Claridge 2007: 87). Virtual worlds are "imaginary environments which people can enter to engage in text-based fantasy social interaction" (Crystal 2008: 12) and where they can "remake themselves" (Bell 2009: 33). Current virtual world games are not just text-based, such as the Internet-based game *Second Life* (cf. Greiffenstern 2010: 39). The World Wide Web is then for Crystal (2008) "the full collection of all the computers linked to the Internet which hold documents that are mutually accessible through the use of a standard protocol (the HyperText Transfer Protocol, or HTTP)" (ibid.: 13).[4] Wikipedia or online encyclopedias are not counted as a single Internet situation in Crystal (2008: 14), but as part of the variety World Wide Web. Furthermore, instant messaging means that people who know each other communicate in real time, for instance, with the system ICQ. Here, people send messages that appear on the computer screen (cf. Greiffenstern 2010: 39). Finally, Crystal (2008) describes blogging as follows:

> It takes the form of a personalized web page where the owner can post messages at intervals. Many blogs are personal diaries, ranging in length from brief notes to extended essays; many are on topics of general interest or concern, such as a hobby or political issue. Some blogs are monologues; some have shared authorship; some are interactive. (ibid.: 15)

It is important to remember that these seven situations are increasingly blended and some are hardly found in a "pure" version today. Additionally, Internet access is not limited to computers any more, but can be established by mobile phones as well (cf. Crystal 2011: 2). What Crystal (2008) does not mention – and this might be due to the publication date – are social networks, e.g. Facebook, that are presently taking a vital role in Internet communication (cf. Hoffmann 2012: 1). Social networks, however, cannot be seen as hypertexts and so cannot be treated here.

[4] For the distinction between "World Wide Web" and "Internet" see also Ince (2012: Internet) and Waltinger & Breuing (2012: 534).

Comparing Rehm's (2007) and Crystal's (2008) classifications, Rehm only considers the situations World Wide Web and blogging. This means that Rehm only discusses the World Wide Web and here regards blogging as one hypertext type among others. The analysis of anaphors will be restricted to most of Rehm's hypertext types. Rehm's classification fits well because texts and documents from the World Wide Web (including blogging) are used for text retrieval, rather than e-mails that are included in Crystal's classification, for instance.

A further classification comes from Farkas & Farkas (2002), cited in Jakobs (2003). It is a functional-pragmatic approach and classifies hypertexts according to their intention. Eight categories are distinguished: education; entertainment; providing news, public information, and specialised information; e-commerce: promotion, selling, support; web portals; persuasion; building and sustaining community; and personal and artistic expression. However, some hypertexts fall into more than one category as the intentions may overlap. As is obvious, hypertext type classifications such as from Farkas & Farkas are very much determined by historical features, i.e. the chronological development of hypertexts (cf. Jakobs 2003: 238).

In sum, there are not many classifications of hypertexts. Some also consider only a particular part of the World Wide Web or see hypertexts as one text type of the Internet among others, e.g. Crystal (2008). From the approaches mentioned above, Rehm's (2007) classification principally fits best for the aim of investigating anaphors in hypertexts and relating anaphora resolution to text retrieval. The concrete classification is presented now.

4.3 Corpus of hypertexts

4.3.1 Corpus design

From the classification of Rehm (2007) and the discussion above, the corpus is constructed. Only hypertexts are represented in the corpus, all other forms of the Internet will not be examined and are therefore not part of the corpus. A corpus can be defined as "a [systematic] collection of texts that has been compiled for a particular reason" (Cheng 2012: 3; cf. also Matthews 2007: 83). The intention for the corpus here is to serve as the basis for analysing the frequency of anaphors in hypertexts. As it will be used for computational goals, it is important that the classification is based on operationalisability through computer-linguistic processes. This is why the classification of Rehm (2007) and not other classifications will be taken as basis for the corpus here.

As for the corpus design, guest books as well as the two categories of miscellaneous hypertext types are not part of the corpus here because they usually contain little information that would be relevant for text retrieval (cf. chapter 5). This leaves the hypertext types institutional homepages, personal homepages, online newspapers, online encyclopedias and blogs. In Rehm's (2007) classification, the category institutional homepages includes pages from companies and all types of organisations. This book, however, makes a distinction between "real" institutional pages and pages from companies (cf. chapter 4.3.4). The historically most recent and socially increasingly important of Rehm's hypertexts types are online encyclopedias with Wikipedia as the best-known example (cf. Levene 2010: 403) and blogs (cf. chapter 4.3.2 and 4.3.3). Therefore, it is important that these two hypertext types are included in the corpus. Additionally, institutional and personal homepages and online newspapers are also regarded, although each of them has a minor importance in the corpus.

The corpus so consists of three hypertext types or categories: an example of an online encyclopedia ("Wikipedia"), texts of different blogs ("blogs") and a category including the other hypertext types, which are called "traditional websites": institutional websites, commercial websites i.e. websites of companies, personal websites and online newspapers. Table 1 shows an overview of these hypertext types and the number of words involved. Each of these three types, their detailed features and the texts chosen as representations of these hypertext types are discussed in the chapters 4.3.2 to 4.3.4.

Table 1: Corpus

Hypertext type	Number of texts	Number of words
Wikipedia	19 pages	25,533 words
Blogs	24 pages	25,312 words
Traditional websites	60 pages	25,129 words
Hypertexts in total	103 pages	75,974 words

4.3.2 "Wikipedia"

As for the representation of online encyclopedias, Wikipedia texts have been chosen. Wikipedia was founded in 2001 and currently comprises more than 4,660,000 English articles (July 2014) (cf. "Wikipedia: About" 2014). The characteristic feature of Wikipedia is that every reader can change or add articles. Bell (2009: 35) speaks of "user-generated content", which also applies to blogs, by the way. The prerequisite for people who would like to contribute to Wikipedia articles is only to have basic skills in word processing programmes; there is no need to know HTML or other programming languages (cf. Gauntlett 2009: 40; Pscheida 2010: 351).

Schuler (2007: 77), for example, states that the quality of Wikipedia articles varies: one problem of some articles is that they do not achieve objectivity, although desired, because individuals seek to promote their interests (cf. ibid.: 94-96). But yet Wikipedia articles undergo review processes by Wikipedia's editors. Anyone who has Internet access, however, can become an editor, irrespective of their qualification (cf. Waltinger & Breuing 2012: 552; "Wikipedia: About" 2013). Those articles that these editors consider best are called "featured articles". Such articles are marked by an icon in the form of a bronze star. They amount to 4,400 articles and currently only occupy a small proportion of all Wikipedia articles (cf. "Wikipedia: Featured articles" 2014). Articles that do not achieve such high quality but are better than others are termed "good articles". Currently, about 21,300 articles enjoy this status. They are marked by a green plus in a circle (cf. "Wikipedia: Good articles" 2014). Rehm (2007), for instance, also appreciates the quality of the articles with regard to another aspect:

> Sie [d.h. die Artikeln] besitzen einen sehr großen Abdeckungsgrad und eine inhaltliche Qualität, die über vergleichbare kommerzielle Produkte hinausgeht [...] Ein Vergleichstest hat gezeigt, dass die Wikipedia bezüglich ihrer Inhalte und insbesondere hinsichtlich der Aktualität den meisten kommerziellen Produkten überlegen ist [...]. (ibid.: 241-242)

In a study of the magazine *Nature* in 2005, 50 articles of Wikipedia and the Encyclopædia Britannica were compared. The results showed that Wikipedia is as reliable as printed encyclopedias and that Wikipedia does not contain more errors. Furthermore, a big advantage is that Wikipedia is accessible more quickly and more easily than printed books (cf. Wikimedia Deutschland 2011: 234). Another clear difference between articles from printed encyclopedias and articles from Wikipedia is that Wikipedia is designed to be changed constantly, for instance, in order to include more up-to-date information, whereas a printed encyclopedia is designed to be complete (cf. Storrer 2012: 293). The popularity of Wikipedia even forced publishers to cease the printing of their traditional encyclopedias, as was the case with the German Brockhaus Enzyklopädie or with Microsoft Encarta, for instance (cf. Wikimedia Deutschland 2011: 276-277).

Levene (2010) summarises the benefit of Wikipedia as follows: "In many cases, it [i.e. Wikipedia] gives a quick and mostly accurate description [...] that can be verified with other sources if necessary" (ibid.: 403). The coverage and actuality of information Rehm and Levene address is surely one reason why

Wikipedia is so popular. Another reason is that search engines such as Google display articles from Wikipedia among the first few hits in their result pages (cf. Levene 2010: 403; Pscheida 2010: 332). Wikipedia counted more than 9.6 billion page views per month in July 2014 (cf. "English Wikipedia" 2014). Consequently, Wikipedia articles are prototype examples if Internet users search for information.

Furthermore, various studies, for example, those Pscheida (2010: 337-338) or Rehm (2007: 194-195) mention, have shown that the writing style in Wikipedia generally does not differ from printed encyclopedias and so is formal. Rehm summarises the results: "Eine qualitative Analyse der Beiträge belegt, dass die Wikipedia keine umgangssprachlichen oder informellen Ausdrücke enthält [...] [und] von stilistischer Homogenität geprägt ist" (ibid.: 195).

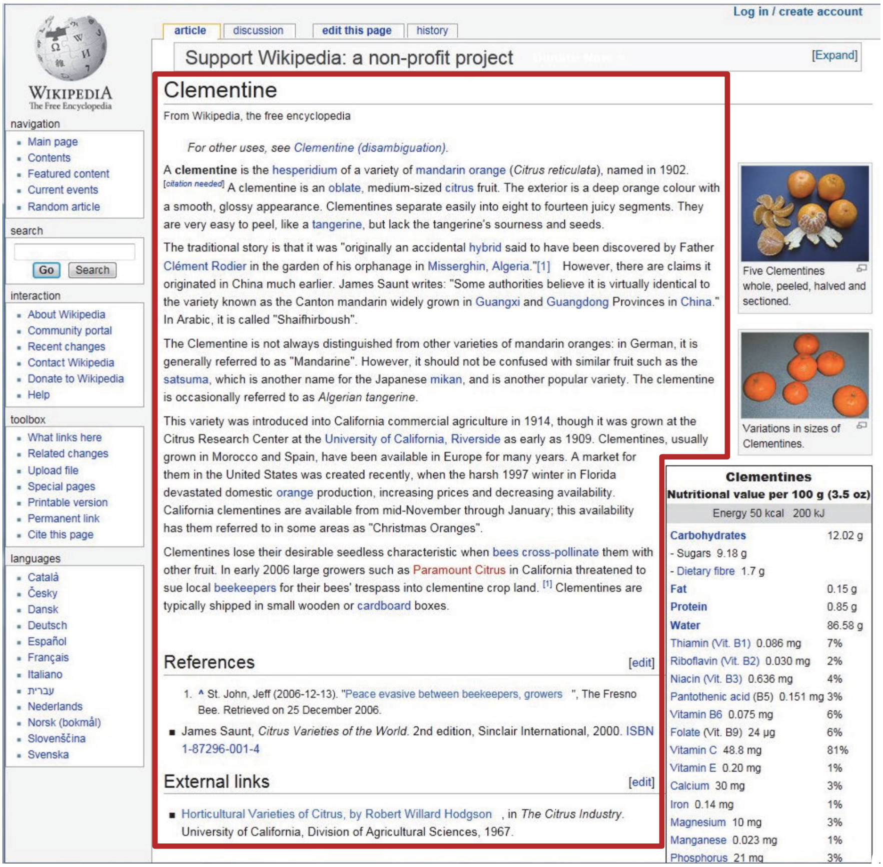
The Wikipedia articles chosen for the corpus have been selected from Wikipedia's main topic classifications. This page of Wikipedia lists twenty-six major topics such as *Education*, *Health* or *Nature* and by clicking on one category, subcategories are loaded.[5] This process is repeated until articles of a specific subcategory within one main topic classification are listed. Nineteen articles with 25,533 words in total were selected from the twenty-four main topic classifications. In many cases an article belongs to more than one of these topic classifications, e.g. the article *Clementine* is part of the classifications *agriculture*, *life* and *nature*. The articles chosen are shown in Table 2. Each article has an abbreviation for easier reference. The number of words each article contains is listed as well. Some articles have been shortened at boundaries of one (sub-)chapter to another, so that not one article is too prominent in the corpus in terms of the number of words.[6] This was necessary with W1, W7, W10, W12, W13, W14, W15, W17 and W19. All shortened articles start at the beginning because otherwise anaphors could not be resolved correctly, for instance, an antecedent might be missing. Furthermore, pictures and their captures in invisible tables as well as other tables on the right side of Wikipedia's pages have been deleted if they were not mentioned in the text itself, as was the usual case (see Figure 1 for an example).

[5] See http://en.wikipedia.org/wiki/Category:Main_topic_classifications (date of last access: 17/11/2013).

[6] Other important principles of corpus construction that have been considered are found, for instance, in Nelson (2010: 53-65), Reppen (2010: 31-32), Cheng (2012: 3-4, 30-31), Hoffmann (2012: 214) and McEnery & Hardy (2012: 59).

Table 2: Wikipedia texts

Abbre- viation	Words	Article	Abbre- viation	Words	Article
W1	1,749	Acupuncture	W11	616	Michael Halliday
W2	1,637	Australia (continent)	W12	1,527	Money
W3	1,235	Cha-cha-cha (dance)	W13	1,704	Mormon
W4	1,184	Civil and political rights	W14	1,037	Movie studio
W5	401	Clementine	W15	1,656	Occam's razor
W6	973	Earl Grey tea	W16	1,318	Pragmatics
W7	1,664	Family	W17	1,411	Rococo
W8	1,352	Information Age	W18	1,707	Soprano
W9	1,731	James Shirley	W19	1,160	Through the Looking-Glass
W10	1,471	Lobster			



Clementine

From Wikipedia, the free encyclopedia

For other uses, see Clementine (disambiguation).

A **clementine** is the hesperidium of a variety of mandarin orange (*Citrus reticulata*), named in 1902. [*citation needed*] A clementine is an oblate, medium-sized citrus fruit. The exterior is a deep orange colour with a smooth, glossy appearance. Clementines separate easily into eight to fourteen juicy segments. They are very easy to peel, like a tangerine, but lack the tangerine's sourness and seeds.

The traditional story is that it was "originally an accidental hybrid said to have been discovered by Father Clément Rodier in the garden of his orphanage in Misserghin, Algeria."[1] However, there are claims it originated in China much earlier. James Saunt writes: "Some authorities believe it is virtually identical to the variety known as the Canton mandarin widely grown in Guangxi and Guangdong Provinces in China." In Arabic, it is called "Shaifhirboush".

The Clementine is not always distinguished from other varieties of mandarin oranges: in German, it is generally referred to as "Mandarine". However, it should not be confused with similar fruit such as the satsuma, which is another name for the Japanese mikan, and is another popular variety. The clementine is occasionally referred to as *Algerian tangerine*.

This variety was introduced into California commercial agriculture in 1914, though it was grown at the Citrus Research Center at the University of California, Riverside as early as 1909. Clementines, usually grown in Morocco and Spain, have been available in Europe for many years. A market for them in the United States was created recently, when the harsh 1997 winter in Florida devastated domestic orange production, increasing prices and decreasing availability. California clementines are available from mid-November through January; this availability has them referred to in some areas as "Christmas Oranges".

Clementines lose their desirable seedless characteristic when bees cross-pollinate them with other fruit. In early 2006 large growers such as Paramount Citrus in California threatened to sue local beekeepers for their bees' trespass into clementine crop land. [1] Clementines are typically shipped in small wooden or cardboard boxes.

References [edit]

1. ^ St. John, Jeff (2006-12-13). "Peace evasive between beekeepers, growers ", The Fresno Bee. Retrieved on 25 December 2006.

- James Saunt, *Citrus Varieties of the World*. 2nd edition, Sinclair International, 2000. ISBN 1-87296-001-4

External links [edit]

- Horticultural Varieties of Citrus, by Robert Willard Hodgson , in *The Citrus Industry*. University of California, Division of Agricultural Sciences, 1967.

Five Clementines whole, peeled, halved and sectioned.

Variations in sizes of Clementines.

Clementines		
Nutritional value per 100 g (3.5 oz)		
Energy 50 kcal 200 kJ		
Carbohydrates		12.02 g
- Sugars 9.18 g		
- Dietary fibre 1.7 g		
Fat		0.15 g
Protein		0.85 g
Water		86.58 g
Thiamin (Vit. B1)	0.086 mg	7%
Riboflavin (Vit. B2)	0.030 mg	2%
Niacin (Vit. B3)	0.636 mg	4%
Pantothenic acid (B5)	0.151 mg	3%
Vitamin B6	0.075 mg	6%
Folate (Vit. B9)	24 μg	6%
Vitamin C	48.8 mg	81%
Vitamin E	0.20 mg	1%
Calcium	30 mg	3%
Iron	0.14 mg	1%
Magnesium	10 mg	3%
Manganese	0.023 mg	1%
Phosphorus	21 mg	3%

Fig. 1: Example of a Wikipedia text, W5 (only the part in the box has been taken for the analysis)

4.3.3 “Blogs”

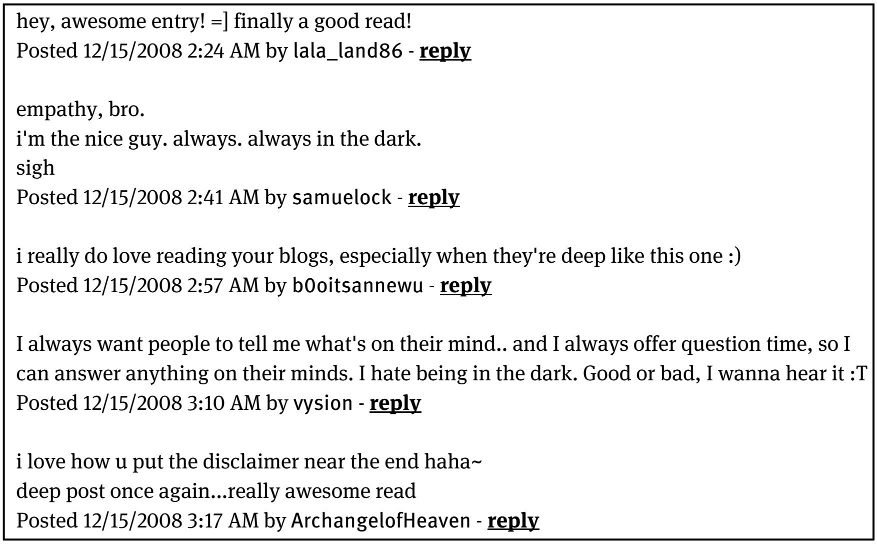
The word *blog* is an abbreviation of “weblog” or “web log”. *Blog* first appeared in 1999, *weblog* or *web log* was supposedly coined by John Barger in 1997. At that time, the term was used for pages that contained links (cf. Crystal 2008: 239; Hoffmann 2012: 14). For more recent blogs, Crystal (2008) explains the character as follows:

> [A] blog is a web application which allows the user to enter, display, and edit posts at any time. It is essentially a content-management system – a way of getting content onto a web page. Most users think of blogging as a genre akin to diary-writing or bulletin-posting, and add posts with some regularity, usually daily, often several times a day. The posts then appear on the site in chronological order, identified by date and time, typically with the most recent at the top. (ibid.: 240)

However, blogs today are not diaries in digital form that are operated by one person any more. Current blogs are usually interactive as they welcome comments from readers. Hoffmann (2012: 3) here speaks of a dialogicity of blogs. Blogs differ regarding the people or group that operate the blog. There are personal blogs, in which one person gives information about himself or herself; corporate blogs, which are in the responsibility of institutions; and other blogs, for instance, from journalists, politicians or interest groups (cf. Crystal 2008: 240-242; Pedersen 2010: 3; Ince 2012: blog). Journalistic blogs promote participation from the audience (cf. Pedersen 2010: 35) and they “tend to be associated more with the opinion side of newspapers than with the provision of breaking news” (ibid.: 27). In 2011, blogging was a hobby for 60 % of all bloggers. The other 40 % of the bloggers write for companies, organisations, or to make money. Additionally, Technorati’s analysis of the state of the blogosphere shows that about 70 % of the bloggers in 2011 are college graduates or have a graduate degree (cf. Technorati 2011a).

Crystal (2008: 244-246) states that different styles are found in blogs, from formal to informal (cf. also Hoffmann 2012: 2). This depends on the type of blog, whether it is a corporate or personal blog, for example. Blogs that undergo editorial processes, as corporate blogs, usually show a more formal style. Personal blogs, however, vary more in their style and are frequently informal. Blogs with informal language often display unconventional punctuation, spelling and grammar (cf. Crystal 2008: 31, 213-214; Crystal 2011: 19-20). Crystal (2008) states in this context: “There are several features of informal written English which would be eliminated in a copy-edited version of such texts for publication” (ibid.: 244). However, it should also be mentioned that our language has generally become more informal in the last decades (cf. Wawra 2011: 103).

To give examples, comments from the blog text B16 as illustrated in Figure 2 show informal features: no capital letters; contracted forms (e.g. *I'm*); the use of emoticons (e.g. *:)* in the third comment); and the use of two or three dots as in the fourth and fifth comment, which symbolises a pause. Furthermore, there are abbreviations such as *US* for *the United States* and clippings such as *phone* as the short form of *telephone*. Pronouns as subjects and function words are also omitted (e.g. *always in the dark* instead of *I'm always in the dark*) (cf. Crystal 2008: 244-245; Greiffenstern 2010: 27-28, 45-47; Schubert 2012: 139-140).



hey, awesome entry! =] finally a good read!
Posted 12/15/2008 2:24 AM by lala_land86 - **reply**

empathy, bro.
i'm the nice guy. always. always in the dark.
sigh
Posted 12/15/2008 2:41 AM by samuelock - **reply**

i really do love reading your blogs, especially when they're deep like this one :)
Posted 12/15/2008 2:57 AM by b0oitsannewu - **reply**

I always want people to tell me what's on their mind.. and I always offer question time, so I can answer anything on their minds. I hate being in the dark. Good or bad, I wanna hear it :T
Posted 12/15/2008 3:10 AM by vysion - **reply**

i love how u put the disclaimer near the end haha~
deep post once again...really awesome read
Posted 12/15/2008 3:17 AM by ArchangelofHeaven - **reply**

Fig. 2: Example of comments on a blog (B16)

The comments on blogs can also differ considerably from the style of the blog itself, i.e. the blog entry might be formal, the comments on the blog quite informal (cf. Yus 2007: 132; Hoffmann 2012: 3). The style in blogs is nevertheless in total more informal than that in Wikipedia texts. In sum, blogs are not a homogeneous hypertext type. They deal with different topics and vary in their target audience. Therefore, blogs are better viewed as a hybrid with characteristics of e.g. diaries, editorials, letters to the editor and travel reports (cf. Rehm 2007: 196-198).

Turning to the comparison of blogs and websites, blogs are similar to personal websites, but also show differences. For instance, not so much computing skills are necessary for establishing blogs as for designing personal websites.

Moreover, blogs are updated more often, are fairly text-based and contain fewer images than websites (cf. Yus 2007: 121, 133). The character of blogging is also different: "personal home pages present a medium for self presentation, whereas blogs normally present a medium for self disclosure" (ibid.: 121).

With regard to the amount of blogs in English, the number can only be estimated. In 1997, only 23 websites were considered blogs (cf. Pedersen 2010: 4). At the beginning of the year 2010, the estimates for blogs in the English language were around 450 million. Technorati's directory counts about 1,290,000 blogs at the end of the year 2011 (cf. Technorati 2011c). However, the number also includes "dead blogs", i.e. blogs that are not used and updated anymore. As many as 45 % of all blogs could probably be abandoned (cf. ibid.: 4). In a similar way, the total number of blog readers is uncertain. Estimates are in the hundreds of millions of people, perhaps 500 to 600 million (cf. "So How Many Blogs are There, Anyway?" n.d.; Crystal 2008: 246).

Turning to the blog texts of the corpus, 24 blogs with 25,312 words were selected randomly for the corpus by using Technorati's blog directory (http://technorati.com/blogs/directory, date of last access: 08/02/2013) and Google's blogsearch (http://www.google.com/blogsearch, date of last access: 08/02/2013). Half of the blogs are blogs from the online newspapers *CNN* (http://edition.cnn.com, date of last access: 08/02/2013), *The Guardian* (http://www.guardian.co.uk, date of last access: 08/02/2013), *The New York Times* (http://www.nytimes.com, date of last access: 08/02/2013) and *The Telegraph* (http://www.telegraph.co.uk, date of last access: 08/02/2013).[7] The other half consists of blogs predominantly from companies, but there are also some personal blogs and blogs from organisations. Some blogs have been shortened in order not to give one example text too much weight. As the full blog entries frequently consist of a blog text to which readers can add commentaries, the shortening so concerns only the number of comments analysed for each blog text. Shortening was carried out for the blogs B4, B5, B6, B8, B9, B14, B15, B16, B21, B22, B23 and B24. The source of the blogs as well as the type of each blog text are listed in Table 3.

[7] The reason for the predominance of blogs from online newspapers is that these can be easier found and selected on the Internet. An alternative could have been to select blogs from the current top 100 or so blogs, for example, from Technorati's list. The popularity there may, however, change quickly and a blog can only be popular for a short time. Technorati's list is, for example, updated daily (cf. Technorati 2011b).

Table 3: Blog texts

Abbreviation	Words	Text	Source
B1	446	Why YouTube Makes Sense for Corporate Blogs	Direct2Dell (company)
B2	369	Dell's Green Efforts Highlighted in Ceres Report	Direct2Dell (company)
B3	441	Wrapping Up the Year in Storage and Looking Ahead With EMC	Direct2Dell (company)
B4	1,038	Merry Christmas from the Blairs	Guardian: Books Blog (online newspaper)
B5	1,041	Protection is the name of the game	Guardian: Word of Mouth Blog (online newspaper)
B6	1,865	Suddenly the Big Four look fallible in these frugal times	Guardian: Sport Blog (online newspaper)
B7	1,415	George Bush shoe attack an acute symbol of disrespect	Guardian: News Blog (online newspaper)
B8	1,757	Google's 'Treat All Rich Companies the Same' Vision of Net Neutrality	The New York Times: Bits Blog (online newspaper)
B9	1,193	Lied About Any Good Books Lately?	The New York Times: Paper Cuts Blog (online newspaper)
B10	700	Three Men in a Tub	The New York Times: Wheels Blog (online newspaper)
B11	426	Politico Ad Network Gets A Boost From Reuters	paidContent.org (company)
B12	255	BitTorrent Renegotiates Third Round; Takes $10 Million Less Than Before	paidContent.org (company)
B13	1,402	Pirates vs. Ninjas: Who would win?	Technorati (company)
B14	1,296	Obama: My Administration Will Value "Science" and "Facts"	Talking Points Memo (company)
B15	1,013	How Much Muscle Is Too Much?	healthkicker (personal)
B16	1,479	You really wanna know the truth?	wongfu (company)
B17	1,448	Grammar Attacks	The Blog Herald (company)
B18	559	Israel Launches Its Own Arabic-Language Channel	CNN: Middle East Blog (online newspaper)
B19	600	Political Hot Topics	CNN: politicalticker Blog (online newspaper)
B20	1,826	Held over a barrel when it comes to home heating	CNN: business360 Blog (online newspaper)
B21	1,339	Newsweek's ghoulish cover of Diana and Kate Middleton is a disgrace	The Telegraph: Royal family Blog (online newspaper)
B22	963	Social media: where do these billion dollar valuations come from?	Telegraph: Internet Blog (online newspaper)
B23	1,335	Mother-in-Law Won't Listen	Dog Blog (club)
B24	1,106	Building self esteem one step at a time	the Self Improvement Blog (personal)

4.3.4 "(Traditional) Websites"

"Traditional websites" comprise four subtypes in this book: websites from companies, personal websites, websites from online newspapers and institutional websites.[8] Personal homepages split up into two types: private pages and professional pages. Private pages are under the control of the individual and usually give information about hobbies and interests. Professional pages, for instance, detail the professional setting and tasks of a person (cf. Rehm 2007: 172-174). Institutional pages are homepages e.g. from governments, authorities and further organisations, whereas pages from companies are commercial pages that usually offer information about the company itself and their products (cf. Rehm 2007: 162-166). Finally, the online newspapers that have been considered and from which texts have been chosen randomly are: *The Guardian*, *The New York Times*, *The Times* (http://www.thetimes.co.uk, date of last access: 08/02/2013).

In establishing the corpus, the search engine Google (http://www.google.com, date of last access: 08/02/2013), the open directory http://www.dmoz.org (date of last access: 08/02/2013) and the website http://www.gksoft.com/govt/en/gb.html (date of last access: 08/02/2013) for institutional homepages were referred. The open directory says of itself that it is "the largest, most comprehensive human-edited directory of the Web" ("About the Open Directory Project" 2011) and the website for selecting institutional homepages describes itself as "[c]omprehensive database of governmental institutions on the World Wide Web" (Anzinger n.d.). To select texts keywords such as "company", "personal homepage" have been searched for with Google and the open directory. Afterwards, the examples have been chosen randomly. As representations of traditional websites, 60 texts with 25,129 words have been analysed. Table 4 shows the distribution of these texts across the subtypes.

Table 4: Traditional website subtypes

Subtype	Number of pages	Number of words
Companies	22 pages	5,993 words
Personal websites	15 pages	6,228 words
Institutional websites	15 pages	6,330 words
Online newspapers	8 pages	6,578 words

[8] Numbers for the frequency of these traditional websites on the Internet were not found to be available.

The texts of traditional websites are shorter, resulting in a larger number of texts for this category. As a consequence, a shortening of these texts was not necessary. The details of the texts chosen are given in Table 5. Those abbreviations of texts that share the same number (e.g. *1* in WS1a and WS1b) originate from one website, but the texts themselves are found on different pages.[9] The numbers 1-4 stand for websites of companies, 21-35 for personal websites, all of which are professional pages. The websites with the numbers 31-34 are themselves dedicated to a specific person. The numbers 41-47 are websites of online newspapers and numbers 61-74 constitute institutional websites. The third column in the table shows the text's name. In the case of the numbers 2-4 and 61-74, there are arrows to represent the path, i.e. where the individual text can be found on the website. Furthermore, italic words in brackets are comments and serve as further information about the text. The rightmost column displays the source or context to which the article belongs.

Table 5: Traditional website texts

Abbreviation	Words	Text	Source
WS1a	361	Company	Infineon
WS1b	457	Automotive, Industrial & Multimarket (AIM) *(link from 1a)*	Infineon
WS1c	606	Communication Solutions *(link from 1a)*	Infineon
WS2a	277	About us → Culture & diversity	Clarks
WS2b	286	About us → Our recruitment process	Clarks
WS2cI	212	About us → History & heritage → In the beginning	Clarks
WS2cII	456	About us → History & heritage → 1825-1900	Clarks
WS2cIII	310	About us → History & heritage → 1900-1946	Clarks
WS2cIV	201	About us → History & heritage → 1946-1990	Clarks
WS2cV	332	About us → History & heritage → 1990-present	Clarks
WS2d	114	About us → Recruitment agencies	Clarks
WS2e	506	About us → Social responsibility	Clarks
WS3a	59	*(welcome page)*	Smart
WS3b	90	UK → Information & service → After Sales	Smart
WS3c	62	UK → Information & service → More about smart → Business → The brand	Smart
WS3d	67	UK → Information & service → More about smart → Business → Safety	Smart

[9] A "website" consists of one or more "Web pages" (cf. "Website" n.d.; Agnes et al. 2007: 1622; Ince 2012: Web site). A "Web page" is defined technically as "a single window of scrollable material" (Ince 2012: Web page).

WS3e	227	UK → Information & service → More about smart → Business → Practicality	Smart
WS3f	191	UK → Information & service → More about smart → Business → Environment	Smart
WS3g	267	UK → Information & service → More about smart → Business → Economics	Smart
WS4a	412	About Abu Dhabi → Abu Dhabi	ADAC
WS4b	171	About us → ADAC Mandate	ADAC
WS4c	329	Airports → Abu Dhabi Airport → Abu Dhabi International Airport	ADAC
WS21	358	George Brown, Professor Emeritus	Stanford University
WS22	116	Dr Bert Vaux *(university lecturer)*	University of Cambridge
WS23	270	Erin L. O'Bryan, Ph.D., CF-SLP *(research specialist)*	University of Arizona
WS24	61	Nadja Stern, Chief Executive	Rambert Dance Company
WS25	62	Sarah Cooper, Staff Nurse	Institute for Innovation and Improvement
WS26	173	Dr John Mitchell *(senior lecturer)*	University College London
WS27	433	Dr Eddy Donnelly *(Seear Fellow)*	London School of Economics and Political Science
WS28	125	Samuel Moon *(research officer)*	Overseas Development Institute
WS29	254	Professor Paul Ward	University of Huddersfield
WS30	508	Dr Richard Kirkham *(lecturer)*	The University of Manchester
WS31	683	Katherine Wood-Jacobs *(for Lancaster Prothonotary)*	-
WS32	344	Dr. Susan Blackmore *(writer, lecturer, broadcaster)*	-
WS33	469	Ann Parker *(writer)*	-
WS34	452	Linda Eder *(singer)*	-
WS35	1,920	Biography for Margot Kidder *(actress)*	The Internet Movie Database
WS41	1,706	Are you afraid of teenagers?	The Times
WS42	163	Inflation tumbles to 4.1 % on fuel price fall	The Times
WS43	490	Energy groups ordered to speed up price cuts	The Times
WS44a	990	Caroline Kennedy Is Seeking Seat Held by Clinton *(page 1)*	The New York Times
WS44b	712	Caroline Kennedy Is Seeking Seat Held by Clinton *(page 2)*	The New York Times
WS45	965	For Runners, Soft Ground Can Be Hard on the Body	The New York Times
WS46	1,139	Make or break week	The Guardian
WS47	413	European car sales slump adds to pressure for rescue	The Guardian
WS61	666	Role → Role of Air Power	Royal Air Force

WS62	355	Home → About Ofcom → What is Ofcom?	Ofcom
WS63	713	About → About the NHS	NHS
WS64	864	Travel and Transport → Cycling → Cycling safely	Directgov
WS65	334	Budget → Guide to the Budget	HM Treasury
WS66a	250	About the Bank → Relationship with Parliament	Bank of England
WS66b	246	Monetary policy	Bank of England
WS67	331	Collection & Exhibitions → Exhibitions → Power of Making	Crafts Council
WS68	317	About us	Council for Science and Technology
WS69	246	Opportunities & Advice → How we market Britain	VisitBritain
WS70	316	Government Efficiency - overview	Cabinet Office
WS71	662	Widening participation → Working with institutions to embed WP → Student retention and success	Higher Education Funding Council for England
WS72	394	What we do → Key issues → Governance and conflict → Democratic governance	Department for International Development
WS73	290	IP Enforcement → What is IP crime? → Our role in IP crime	Intellectual Property Office
WS74	346	About the CAA → Diversity	Civil Aviation Authority

4.4 Results of the corpus analysis

4.4.1 Frequency of anaphors in the corpus

4.4.1.1 Distribution of the twelve anaphor types

The analysis of the hypertext corpus in terms of the twelve categories of anaphors as defined in chapter 3 yields surprising results. The most frequent type of anaphor is non-finite clauses (29.2 %), which occur more often than central pronouns (27.5 %). Together they comprise more than half of all anaphors. Of further importance are also proper names, relative pronouns, noun phrases with a definite article, demonstrative pronouns and ellipses. Only marginally important are adverbs, indefinite pronouns, verb phrases with *do*, other forms of coreference and substitution, and reciprocal pronouns. The distribution of the twelve anaphor types in the corpus is illustrated graphically in Figure 3. Furthermore, Table 6 gives detailed results in terms of absolute numbers, i.e. how many items can be found in the hypertext types and the whole corpus of a certain anaphor type. The relative frequency in per mille (‰), i.e. how many items in 1,000 words are anaphors, is also given in brackets. This relative (or normalised) frequency helps in comparing results because the hypertext types very slightly in terms of the number of words (cf. McEnery & Hardy 2012: 48-51).

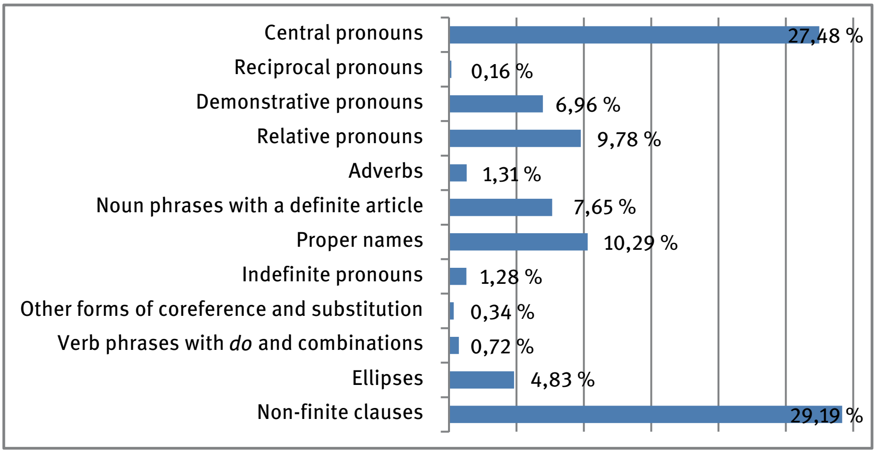


Fig. 3: Distribution of anaphor types across the whole corpus of hypertexts

Table 6: Absolute numbers of anaphors and, in brackets, the relative frequency of anaphors in per mille

	Central pronouns	**Reciprocal pronouns**	**Demonstrative pronouns**	**Relative pronouns**	**Adverbs**	**Noun phrases with a definite article**	**Proper names**
Wikipedia	415 (16.25)	4 (0.16)	139 (5.44)	182 (7.13)	22 (0.86)	228 (8.93)	233 (9.13)
Blogs	640 (25.28)	2 (0.08)	174 (6.87)	216 (8.53)	25 (0.99)	80 (3.16)	165 (6.52)
Traditional websites	619 (24.63)	4 (0.16)	111 (4.42)	198 (7.88)	33 (1.31)	158 (6.29)	229 (9.11)
Hypertexts in total	**1,674** **(22.03)**	**10** **(0.13)**	**424** **(5.58)**	**596** **(7.84)**	**80** **(1.05)**	**466** **(6.13)**	**627** **(8.25)**

	Indefinite pronouns	**Other forms of coreference and substitution**	**Verb phrases with *do* and combinations**	**Ellipses**	**Non-finite clauses**	**Anaphors in total**
Wikipedia	29 (1.14)	13 (0.51)	8 (0.31)	109 (4.27)	561 (21.97)	**1,943** **(76.10)**
Blogs	37 (1.46)	7 (0.28)	28 (1.11)	88 (3.48)	574 (22.68)	**2,036** **(80.44)**
Traditional websites	12 (0.48)	1 (0.04)	8 (0.32)	97 (3.86)	643 (25.59)	**2,113** **(84.09)**
Hypertexts in total	**78** **(1.03)**	**21** **(0.28)**	**44** **(0.58)**	**294** **(3.87)**	**1,778** **(23.40)**	**6,092** **(80.19)**

Turning to the distribution of anaphor types across the hypertexts of Wikipedia, blogs and traditional websites, the analysis has revealed that anaphors are similarly frequent in each of these three hypertext types. The mean value of anaphors lies at 80.2 ‰ with a standard deviation of 4.0. Although anaphors are found in a similar frequency whatever hypertext type is chosen, the anaphor types are distributed unevenly across these three types. Table 7 shows in which hypertext type each anaphor type is the most and least frequent. These figures are now discussed in more detail.

To start with, central pronouns are used in about the same frequency in blogs and traditional websites, but they are distinctly less frequent in Wikipedia texts and here only occupy 24.8 %. Reciprocal pronouns occur only ten times in the corpus so that there is no general tendency for the different hypertext types. Demonstrative pronouns are by far the most numerous in blogs and here occupy 41.0 %. Noun phrases with a definite article and also proper names are distributed unequally across the hypertext types, with blogs containing considerably the fewest in both cases (17.2 % and 26.3 %). Additionally, noun phrases with a definite article are frequent in Wikipedia texts and here nearly comprise half of all anaphors with 48.9 %.

As adverbs, indefinite pronouns, other forms of coreference and substitution and verb phrases with *do* and their combinations are relatively rare in the corpus, an interpretation of the results for the different hypertexts sorts has to be treated with caution. In the corpus, adverbs are the most frequent in traditional websites. Moreover, indefinite pronouns and verb phrases with *do* and their combinations are the most frequent in blogs, whereas 61.9 % of all instances of other forms of coreference and substitution are found in Wikipedia. Non-finite clauses, relative pronouns and ellipses show approximately the same frequency across the three text types.

Table 7: Frequency of each anaphor type in the three hypertext types

	Central pronouns	**Reciprocal pronouns**	**Demonstrative pronouns**	**Relative pronouns**	**Adverbs**	**Noun phrases with a definite article**	**Proper names**
Wikipedia	24.8 %	40.0 %	32.8 %	30.5 %	27.5 %	48.9 %	37.2 %
Blogs	38.2 %	20.0 %	41.0 %	36.2 %	31.3 %	17.2 %	26.3 %
Traditional websites	37.0 %	40.0 %	26.2 %	33.2 %	41.3 %	33.9 %	36.5 %
Hypertexts in total	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**

	Indefinite pronouns	**Other forms of coreference and substitution**	**Verb phrases with *do* and combinations**	**Ellipses**	**Non-finite clauses**	**Anaphors in total**
Wikipedia	37.2 %	61.9 %	18.2 %	37.1 %	31.6 %	**31.9 %**
Blogs	47.4 %	33.3 %	63.6 %	29.9 %	32.3 %	**33.4 %**
Traditional websites	15.4 %	4.8 %	18.2 %	33.0 %	36.2 %	**34.7 %**
Hypertexts in total	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**	**100 %**

Based on Figure 3, Table 6 and Table 7, the frequency and importance of each anaphor type in the whole corpus and in each hypertext type can be derived. The ranking of the twelve anaphor types with the most frequent on the top is presented in Table 8. Generally, the ranking does not differ much. Most variations are found in blog texts. Finally, non-finite clauses are the most frequent in all hypertext types, except for blog texts. However, it has to be kept in mind that the aim of this book is not primarily to investigate the use of anaphors in different hypertext types, but rather to examine the relative frequency of different anaphor types in hypertexts in general.

Table 8: Ranking of the anaphor types according to the frequency

Hypertexts in total	**Wikipedia texts**	**Blog texts**	**Traditional website texts**
1. Non-finite clauses	1. Non-finite clauses	1. Central pronouns	1. Non-finite clauses
2. Central pronouns	2. Central pronouns	2. Non-finite clauses	2. Central pronouns
3. Proper names	3. Proper names	3. Relative pronouns	3. Proper names
4. Relative pronouns	4. Noun phrases with a definite article	4. Demonstrative pronouns	4. Relative pronouns
5. Noun phrases with a definite article	5. Relative pronouns	5. Proper names	5. Noun phrases with a definite article
6. Demonstrative pronouns	6. Demonstrative pronouns	6. Ellipses	6. Demonstrative pronouns
7. Ellipses	7. Ellipses	7. Noun phrases with a definite article	7. Ellipses
8. Adverbs	8. Indefinite pronouns	8. Indefinite pronouns	8. Adverbs
9. Indefinite pronouns	9. Adverbs	9. Verb phrases with *do* and combinations	9. Indefinite pronouns
10. Verb phrases with *do* and combinations	10. Other forms of coreference and substitution	10. Adverbs	10. Verb phrases with *do* and combinations

11. Other forms of coreference and substitution	11. Verb phrases with *do* and combinations	11. Other forms of coreference and substitution	11. Reciprocal pronouns
12. Reciprocal pronouns	12. Reciprocal pronouns	12. Reciprocal pronouns	12. Other forms of coreference and substitution

As to cataphoric interpretations, these are included in the corresponding anaphor types. If now considering the distribution of anaphoric versus cataphoric interpretations, the following observations are worth mentioning: in sum, anaphoric interpretations occupy 98.8 % of all anaphors and cataphoric interpretations only 1.2 %. The cataphoric direction is, with 64.5 % of all cataphors, about three to four times more common in traditional website texts than in blog and Wikipedia texts (see Table 9). Furthermore, it is interesting to see in which anaphor types these cataphoric interpretations occur. As Figure 4 shows, a cataphoric interpretation is found the most often with *non-finite clause* anaphors: about three quarters of all cataphoric interpretations are *non-finite clause* items. 10 cataphoric items belong to central pronouns, the remainder of the cataphoric interpretations occur four times or less in other anaphor types (see Table 10). When looking at the distribution across the three hypertext types, it turns out that one reason for the high frequency of cataphoric interpretations in traditional websites is the high frequency of non-finite clauses there. Especially -*ing*-participle clause anaphors are cataphoric (28 items). -*ed*-participle clause anaphors count 7 cataphoric items, *to*-infinitive clause anaphors 6 items.

Table 9: Distribution of anaphoric and cataphoric interpretations (absolute numbers and, in brackets, relative per mille numbers)

	Items in total	Anaphoric	Cataphoric
Wikipedia	**1,943**	1,931	12
	(76.10)	(75.63)	(0.47)
Blogs	**2,036**	2,021	15
	(80.44)	(79.84)	(0.59)
Traditional websites	**2,113**	2,064	49
	(84.09)	(82.14)	(1.95)
Hypertexts in total	**6,092**	**6,016**	**76**
	(80.19)	**(79.18)**	**(1.00)**

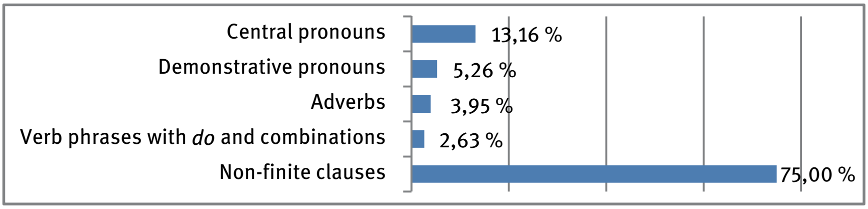


Fig. 4: Distribution of cataphoric interpretations across anaphor types

Table 10: Cataphoric interpretations across anaphor types (absolute numbers and, in brackets, relative numbers in per mille)

	Cataphoric	Central pronouns	Demonstrative pronouns	Adverbs	Verb phrases with *do* and combinations	Non-finite clauses
Wikipedia	**12** **(0.47)**	2 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	10 (0.39)
Blogs	**15** **(0.59)**	1 (0.04)	4 (0.16)	2 (0.08)	2 (0.08)	6 (0.24)
Traditional websites	**49** **(1.95)**	7 (0.28)	0 (0.00)	1 (0.04)	0 (0.00)	41 (1.63)
Hypertexts in total	**76** **(1.00)**	**10** **(0.13)**	**4** **(0.05)**	**3** **(0.04)**	**2** **(0.03)**	**57** **(0.75)**

With regard to the most frequent items across all anaphor types, a visualisation is given in Figure 5. Generally, the most frequent item across all types is *to*, closely followed by *-ing* in the corpus. They represent 11.7 % and 11.4 % respectively in relation to all individual anaphor items. Noun phrases with a definite article hold the third rank with 7.6 %, personal proper names are in fourth position with 6.4 % and *it* with 5.6 % is on the fifth rank. As for the distribution of these high-scoring anaphors across the three hypertext types, not all of them are similarly frequent in each type. Noun phrases with a definite article are the most frequent items in Wikipedia texts, *-ing* is the most commonly found in traditional website texts and *to* is the most frequent item in blogs (see Figure 6).

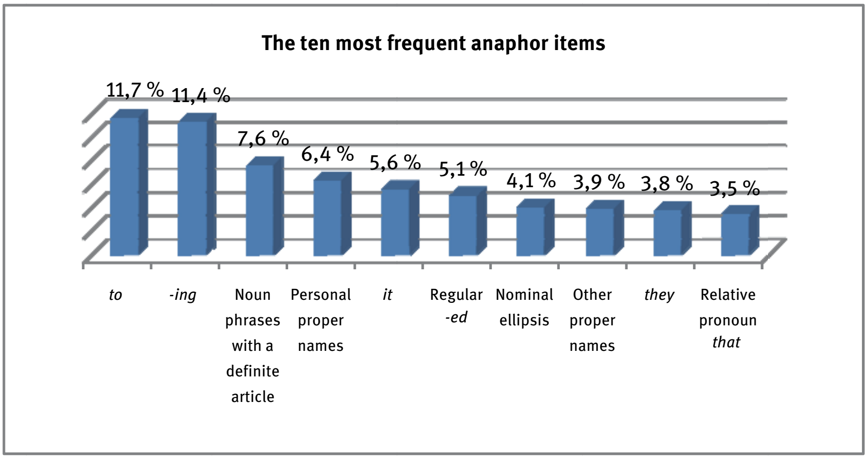


Fig. 5: Anaphor items relative to all anaphors in the corpus

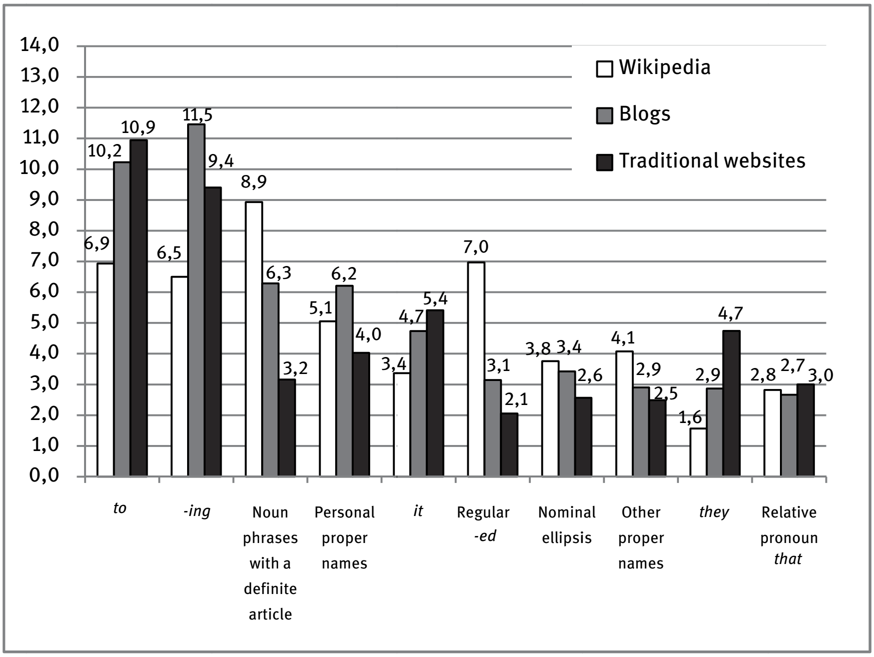


Fig. 6: Distribution of frequent anaphor items in each hypertext type (numbers in per cent and relative to all anaphors of one hypertext type)

Chapters 4.4.1.2 to 4.4.1.12 will now give more details about the distribution of items within each of the anaphor types. Noun phrases with a definite article anaphors do not need further analysis, as their items all consist of a head and a definite article as determinative, plus optional pre- and postmodifications.

4.4.1.2 Central pronouns

Central pronouns fall into the subtypes personal, possessive and reflexive pronouns. In terms of their distribution within central pronouns, personal pronouns lead with 64.1 %, followed by possessive pronouns with 33.3 % and reflexive pronouns occupy a mere 2.6 %. Previous studies show similar results. For instance, the distribution of central pronouns in Barbu (2002: 276) is as follows: personal pronouns amount to 82.2 %, possessive pronouns 16.4 % and reflexive pronouns 1.4 %. Furthermore, Mitkov & Hallett (2007: 275-276) report the following results: 85.5 % for personal pronouns, 12.5 % for possessive pronouns and 2.0 % of reflexive pronouns occur in the corpus of technical manuals; 64.3 % for personal pronouns, 34.3 % for possessive pronouns and 1.4 % for reflexive pronouns in the Penn Treebank corpus; and 50.7 % for personal pronouns, 43.4 % for possessive pronouns and 5.9 % for reflexive pronouns in Jules Verne's text (see also Biber et al. 2007: 344).

Within each of the subtypes, in the corpus *it* leads with 31.9 % in personal pronouns. The item *their* with 33.0 % is the most frequent item of the possessive pronouns and *themselves* with 34.9 % leads within reflexive pronouns. The distribution of all occurring items within central pronouns, irrespective of their membership to the subtype, is visualised in Figure 7. It shows that *it* is the most frequent item of all central pronouns. The absolute and the per mille numbers are then given in Table 11.

Other studies show the following relative per mille numbers, i.e. how many anaphoric central pronouns occur relative to the number of words: Barbu (2002: 276) arrives at 12.9 ‰ (i.e. 366 anaphoric central pronouns in 28,272 words); Mitkov, Evans & Orasan (2002: 178) at 9.1 ‰ (i.e. 2,263 anaphors in 247,401 words); and Mitkov & Hallett (2007: 275-276) at 9.8 ‰ for technical manuals (i.e. 545 anaphoric central pronouns in 55,444 words), 21.8 ‰ for newswire texts (i.e. 2,063 anaphors in 94,500 words) and 41.3 ‰ for Jules Verne's text (i.e. 205 anaphors in 4,965 words). It can readily be seen that the numbers vary according to the type of texts analysed.

As for the three hypertext types, personal pronouns are the most frequent in blog texts (41.1 %). What is striking from the numbers is also that *he* occurs more often in Wikipedia texts than *she*. This can be explained by the selection of the texts in Wikipedia. For example, all texts about people are about male people, namely the texts W9 *James Shirley* and W11 *Michael Halliday*. Another explanation comes from Biber et al. (2007: 333-334) who found that pronouns denoting men are generally more frequent than those denoting women. Furthermore, Biber et al. state that the nominative forms of pronouns are more frequent than their accusative forms, which is consistent with the findings of the corpus analysis. Another outcome of the analysis is that *they*, *them*, *their* and probably *themselves* occur proportionally more often in blogs than in the other two text types. Relative to the other two hypertext types, 51.7 % of all items of *they* occur in blogs. Furthermore, 49.0 % of *them*, 43.5 % of *their* and 66.7 % of *themselves* are found in blogs.

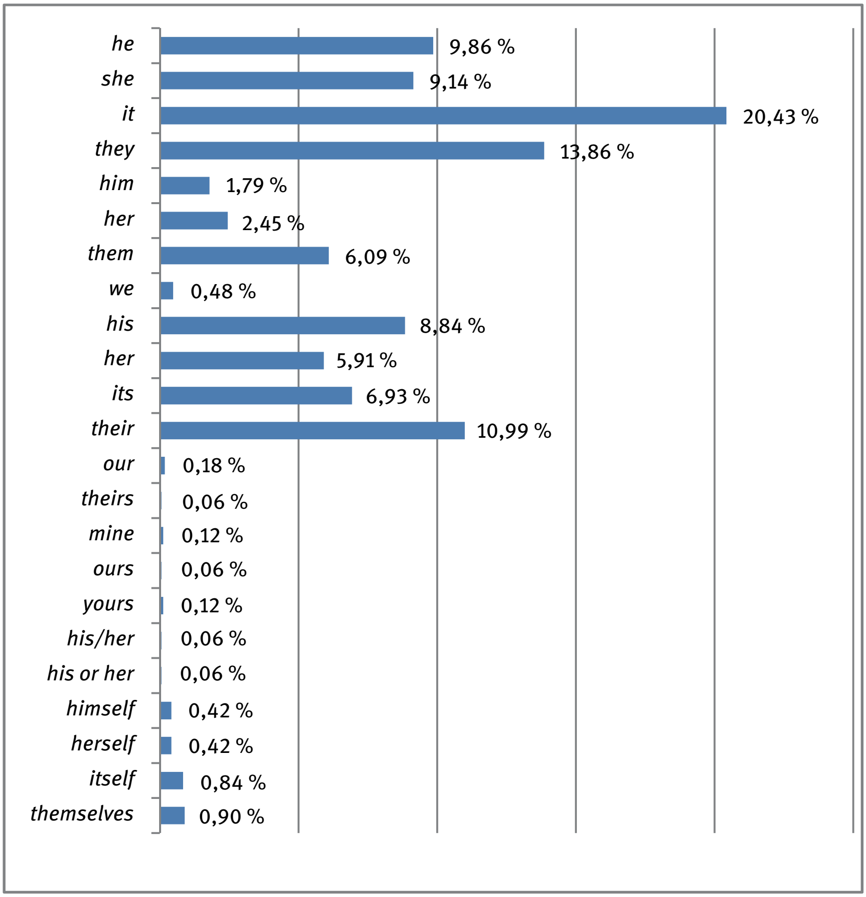


Fig. 7: Distribution of all occurring items within central pronouns

Table 11: Absolute distribution and – in brackets – relative distribution of central pronouns (in per mille)

	Central pronouns	**Personal pronouns**	***he***	***she***	***it***	***they***	***him***	***her***	***them***
Wikipedia	**415**	230	59	14	86	40	5	5	21
	(16.25)	(9.01)	(2.31)	(0.55)	(3.37)	(1.57)	(0.20)	(0.20)	(0.82)
Blogs	**640**	441	44	42	137	120	22	18	50
	(25.28)	(17.42)	(1.74)	(1.66)	(5.41)	(4.74)	(0.87)	(0.71)	(1.98)
Traditional websites	**619**	402	62	97	119	72	3	18	31
	(24.63)	(16.00)	(2.47)	(3.86)	(4.74)	(2.87)	(0.12)	(0.72)	(1.23)
Hypertexts in total	**1,674**	**1,073**	**165**	**153**	**342**	**232**	**30**	**41**	**102**
	(22.03)	**(14.12)**	**(2.17)**	**(2.01)**	**(4.50)**	**(3.05)**	**(0.39)**	**(0.54)**	**(1.34)**

	we	***us***	***he/she***	***he or she***	***s/he***	***s(he)***	***him/her***	***him or her***
Wikipedia	0	0	0	0	0	0	0	0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Blogs	8	0	0	0	0	0	0	0
	(0.32)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Traditional websites	0	0	0	0	0	0	0	0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Hypertexts in total	**8**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
	(0.11)	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**

	Possessive pronouns	***his***	***her***	***hers***	***its***	***our***	***their***	***theirs***
Wikipedia	171	66	13	0	38	0	52	1
	(6.70)	(2.58)	(0.51)	(0.00)	(1.49)	(0.00)	(2.04)	(0.04)
Blogs	186	54	15	0	31	3	80	0
	(7.35)	(2.13)	(0.59)	(0.00)	(1.22)	(0.12)	(3.16)	(0.00)
Traditional websites	201	28	71	0	47	0	52	0
	(8.00)	(1.11)	(2.83)	(0.00)	(1.87)	(0.00)	(2.07)	(0.00)
Hypertexts in total	**558**	**148**	**99**	**0**	**116**	**3**	**184**	**1**
	(7.34)	**(1.95)**	**(1.30)**	**(0.00)**	**(1.53)**	**(0.04)**	**(2.42)**	**(0.01)**

	mine	***ours***	***yours***	***his/her***	***his or her***	***his/hers***	***his or hers***
Wikipedia	0	0	0	0	1	0	0
	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)
Blogs	0	0	2	1	0	0	0
	(0.00)	(0.00)	(0.08)	(0.04)	(0.00)	(0.00)	(0.00)
Traditional websites	2	1	0	0	0	0	0
	(0.08)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Hypertexts in total	**2**	**1**	**2**	**1**	**1**	**0**	**0**
	(0.03)	**(0.01)**	**(0.03)**	**(0.01)**	**(0.01)**	**(0.00)**	**(0.00)**

	Reflexive	***him-***	***her-***	***itself***	***them-***	***our-***	***himself/***	***himself***	***themself***
	pronouns	***self***	***self***		***selves***	***selves***	***herself***	***or herself***	
Wikipedia	14	5	1	6	2	0	0	0	0
	(0.55)	(0.20)	(0.04)	(0.23)	(0.08)	(0.00)	(0.00)	(0.00)	(0.00)
Blogs	13	0	0	3	10	0	0	0	0
	(0.51)	(0.00)	(0.00)	(0.12)	(0.40)	(0.00)	(0.00)	(0.00)	(0.00)
Traditional	16	2	6	5	3	0	0	0	0
websites	(0.64)	(0.08)	(0.24)	(0.20)	(0.12)	(0.00)	(0.00)	(0.00)	(0.00)
Hypertexts	**43**	**7**	**7**	**14**	**15**	**0**	**0**	**0**	**0**
in total	**(0.57)**	**(0.09)**	**(0.09)**	**(0.18)**	**(0.20)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**

Central pronouns have some cataphoric items, which are listed in Table 12. Most of these are personal pronouns (6 items), *its* is a possessive and *themselves* a reflexive pronoun.

Table 12: Cataphoric interpretations with central pronoun items (absolute and relative per mille numbers)

	Cataphoric	***she***	***it***	***they***	***them***	***its***	***themselves***
	central pronouns						
Wikipedia	**2**	1	0	1	0	0	0
	(0.08)	(0.04)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)
Blogs	**1**	0	1	0	0	0	0
	(0.04)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
Traditional	**7**	0	2	0	1	3	1
websites	**(0.28)**	(0.00)	(0.08)	(0.00)	(0.04)	(0.12)	(0.04)
Hypertexts	**10**	**1**	**3**	**1**	**1**	**3**	**1**
in total	**(0.13)**	**(0.01)**	**(0.04)**	**(0.01)**	**(0.01)**	**(0.04)**	**(0.01)**

4.4.1.3 Reciprocal pronouns

Reciprocal pronouns are the rarest anaphor type in the corpus. Within this type, *each other* is found more often than *one another*. This is consistent with the findings of Biber et al. (2007: 346) who also found that *each other* is much more frequent and that both are not common compared to personal pronouns. The details of reciprocal pronouns are given in Table 13.

Table 13: Reciprocal pronouns in absolute and relative numbers (per mille numbers in brackets)

	Reciprocal pronouns	***each other***	***one another***
Wikipedia	**4** **(0.16)**	3 (0.12)	1 (0.04)
Blogs	**2** **(0.08)**	1 (0.04)	1 (0.04)
Traditional websites	**4** **(0.16)**	4 (0.16)	0 (0.00)
Hypertexts in total	**10** **(0.13)**	**8** **(0.11)**	**2** **(0.03)**

4.4.1.4 Demonstrative pronouns

Demonstrative pronouns fall into dependent and independent function. Demonstrative pronouns in dependent function account for 55.0 % of all demonstrative pronouns, those in independent function for 45.0 %. A detailed distribution of the demonstrative pronoun items is presented in Figure 8. The abbreviation *d.* stands for pronouns in “dependent function”, *ind.* for “independent function”. Additionally, Table 14 shows the absolute and relative numbers. If dependent and independent function items are taken together, *this* with 47.2 % is the most frequent, followed by *that*, *these* and finally *those*. Split into dependent and independent function, dependent *this* with 48.5 % of all dependent demonstrative pronouns and independent *this* with 45.5 % of all independent demonstrative pronouns are the most frequent items.

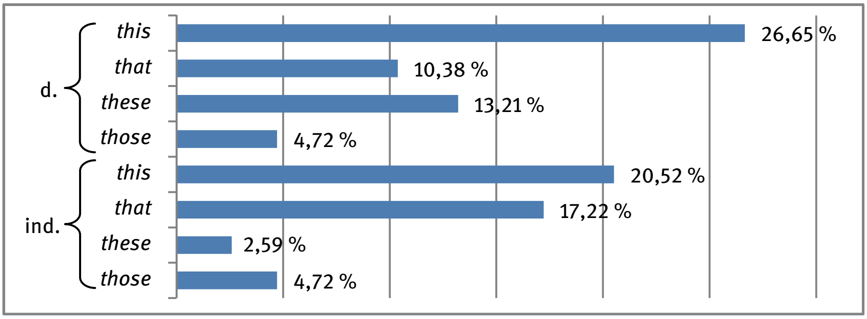


Fig. 8: Distribution of items within demonstrative pronouns

In addition, there are a few noteworthy results across the hypertext types. The distribution of items in dependent and independent function is more or less equal in both blog and traditional website texts. Yet, demonstrative pronouns in dependent function occur nearly twice as often in Wikipedia texts as those in independent function. Finally, Table 15 shows the few cataphoric demonstrative pronoun items. These items are only from demonstrative pronouns in independent function.

Table 14: Demonstrative pronouns in absolute numbers (in brackets the relative per mille numbers)

	Demonstrative pronouns	**Dependent function**	***this***	***that***	***these***	***those***
Wikipedia	**139** **(5.44)**	91 (3.56)	52 (2.04)	9 (0.35)	26 (1.02)	4 (0.16)
Blogs	**174** **(6.87)**	83 (3.28)	33 (1.30)	27 (1.07)	13 (0.51)	10 (0.40)
Traditional websites	**111** **(4.42)**	59 (2.35)	28 (1.11)	8 (0.32)	17 (0.68)	6 (0.24)
Hypertexts in total	**424** **(5.58)**	**233** **(3.07)**	**113** **(1.49)**	**44** **(0.58)**	**56** **(0.74)**	**20** **(0.26)**

	Independent function	***this***	***that***	***these***	***those***
Wikipedia	48 (1.88)	29 (1.14)	7 (0.27)	5 (0.20)	7 (0.27)
Blogs	91 (3.60)	32 (1.26)	49 (1.94)	2 (0.08)	8 (0.32)
Traditional websites	52 (2.07)	26 (1.03)	17 (0.68)	4 (0.16)	5 (0.20)
Hypertexts in total	**191** **(2.51)**	**87** **(1.15)**	**73** **(0.96)**	**11** **(0.14)**	**20** **(0.26)**

Table 15: Cataphoric interpretation with demonstrative pronoun items (absolute and relative per mille numbers)

	Cataphoric demonstrative pronouns	**Independent *this***	**Independent *that***
Wikipedia	**0** **(0.00)**	0 (0.00)	0 (0.00)
Blogs	**4** **(0.16)**	3 (0.12)	1 (0.04)
Traditional websites	**0** **(0.00)**	0 (0.00)	0 (0.00)
Hypertexts in total	**4** **(0.05)**	**3** **(0.04)**	**1** **(0.01)**

4.4.1.5 Relative pronouns

One high-scoring item of relative pronouns is *that*, which represents 36.1 % of all relative pronouns, other items are zero *that* with 9.6 % and the *wh*-items with 54.4 % in sum. The *wh*-items considered individually, *which* with 29.7 % relative to all relative pronoun items is the most frequent. More details can be seen in Figure 9. Table 16 shows the absolute and relative numbers for relative pronouns. It demonstrates that zero *that* is the most frequent in blogs with 61.4 %. As zero *that* is usual in informal contexts, the low occurrence in Wikipedia accounts for its more formal style and the high numbers in blogs underline their informal style. With regard to the *wh*-items, *who* is only half as frequent in blogs as in Wikipedia texts, for example. Additionally, *which* is used less often in blogs than in Wikipedia.

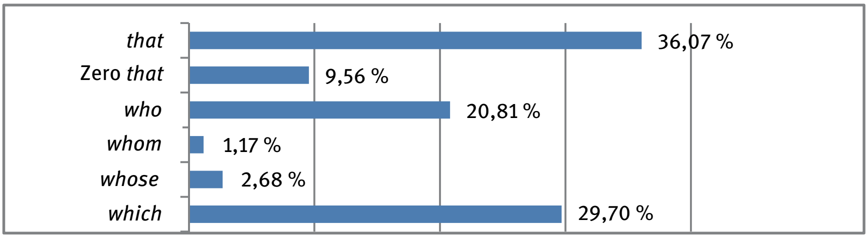


Fig. 9: Distribution of relative pronoun items

Table 16: Absolute numbers and, in brackets, relative numbers in per mille

	Relative pronouns	***that***	**Zero *that***	***who***	***whom***	***whose***	***which***
Wikipedia	**182** **(7.13)**	72 (2.82)	7 (0.27)	26 (1.02)	2 (0.08)	5 (0.20)	70 (2.74)
Blogs	**216** **(8.53)**	76 (3.00)	35 (1.38)	52 (2.05)	3 (0.12)	5 (0.20)	45 (1.78)
Traditional websites	**198** **(7.88)**	67 (2.67)	15 (0.60)	46 (1.83)	2 (0.08)	6 (0.24)	62 (2.47)
Hypertexts in total	**596** **(7.84)**	**215** **(2.83)**	**57** **(0.75)**	**124** **(1.63)**	**7** **(0.09)**	**16** **(0.21)**	**177** **(2.33)**

4.4.1.6 Adverbs

The corpus analysis reveals that *where* is by far the most frequent item within adverbs (56.3 % of all adverbs). *When* and *there* take the second and third position in terms of frequency. All other items occur rarely in the corpus, i.e. five times or less. In sum, the *wh*-items account for 80.0 % of all adverbs. Figure 10 and Table 17 show the details. The distribution is relatively even across the hypertext types. Moreover, Table 18 lists the cataphoric items. Only *here* is used cataphorically.

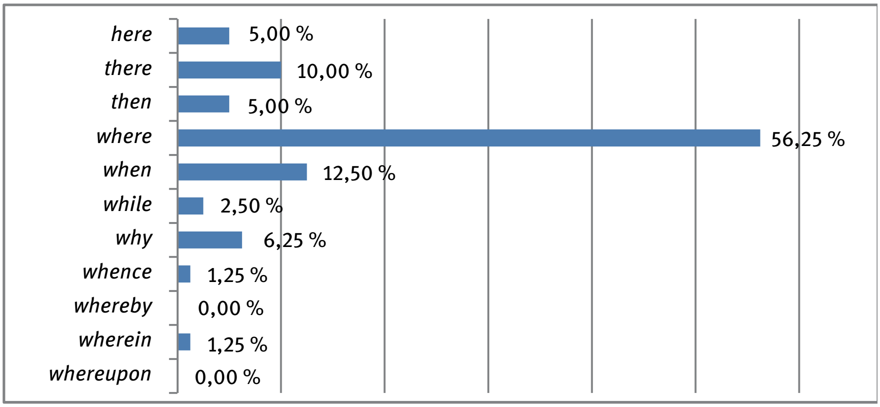


Fig. 10: Distribution of items within adverbs

Table 17: Adverbs in absolute and relative numbers (per mille)

	Adverbs	***here***	***there***	***then***
Wikipedia	**22** **(0.86)**	0 (0.00)	2 (0.08)	2 (0.08)
Blogs	**25** **(0.99)**	2 (0.08)	5 (0.20)	2 (0.08)
Traditional websites	**33** **(1.31)**	2 (0.08)	1 (0.04)	0 (0.00)
Hypertexts in total	**80** **(1.05)**	**4** **(0.05)**	**8** **(0.11)**	**4** **(0.05)**

	where	***when***	***while***	***why***	***whence***	***whereby***	***wherein***	***whereupon***
Wikipedia	15 (0.59)	2 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.04)	0 (0.00)
Blogs	11 (0.43)	2 (0.08)	1 (0.04)	2 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Traditional websites	19 (0.76)	6 (0.24)	1 (0.04)	3 (0.12)	1 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)
Hypertexts in total	**45** **(0.59)**	**10** **(0.13)**	**2** **(0.03)**	**5** **(0.07)**	**1** **(0.01)**	**0** **(0.00)**	**1** **(0.01)**	**0** **(0.00)**

Table 18: Cataphoric items with adverbs (absolute and relative per mille numbers)

	Cataphoric adverbs	***here***
Wikipedia	**0** **(0.00)**	0 (0.00)
Blogs	**2** **(0.08)**	2 (0.08)
Traditional websites	**1** **(0.04)**	1 (0.04)
Hypertexts in total	**3** **(0.04)**	**3** **(0.04)**

4.4.1.7 Proper names

With proper names, the corpus analysis distinguishes between personal with 61.7 % and all other proper names with 38.3 %. Additionally, 40.3 % of all personal proper names are found in traditional website texts, although proper names are slightly more often used in Wikipedia texts (see Table 19 for more details).

Table 19: Proper names in absolute and relative per mille numbers

	Proper names	**Personal**	**Other**
Wikipedia	**233** **(9.13)**	129 (5.05)	104 (4.07)
Blogs	**165** **(6.52)**	102 (4.03)	63 (2.49)
Traditional websites	**229** **(9.11)**	156 (6.21)	73 (2.91)
Hypertexts in total	**627** **(8.25)**	**387** **(5.09)**	**240** **(3.16)**

4.4.1.8 Indefinite pronouns

The most frequent items within indefinite pronouns are *one* (25.6 %) and *both* (12.8 %). All other indefinite pronoun items occur seven times or less in the corpus. The singular and plural forms taken together, *one* and *ones* account for 32.1 % and *other* and *others* for 16.7 %. The items *enough*, *several*, *either*, *most*, *fewest*, *little*, *less* and *least* are not found in the corpus. All in all, the items are distributed relatively evenly across the hypertext types (see also Figure 11 and Table 20).

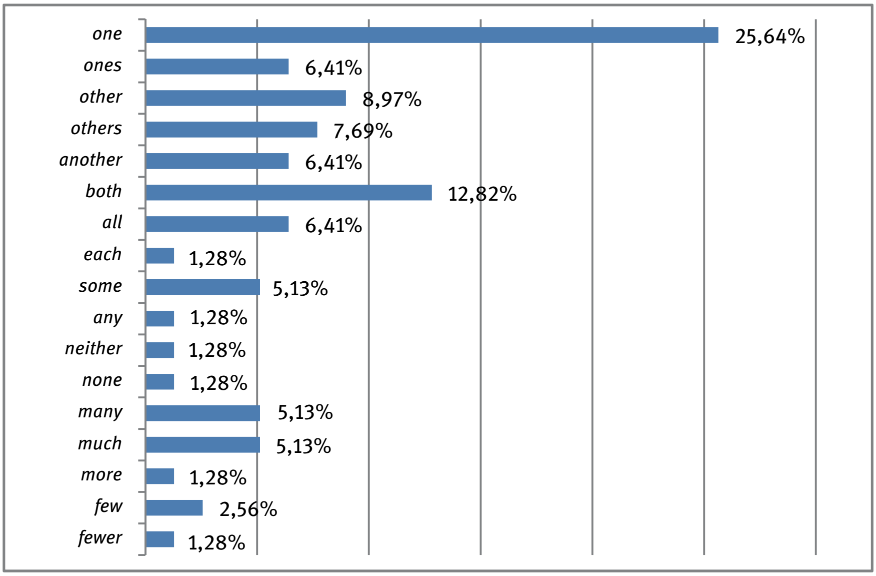


Fig. 11: Distribution of occurring items within indefinite pronouns

Table 20: Indefinite pronouns in absolute and relative per mille numbers

	Indefinite pronouns	***one***	***ones***	***other***	***others***	***another***	***both***	***all***	***each***
Wikipedia	**29** **(1.14)**	8 (0.31)	3 (0.12)	3 (0.12)	1 (0.04)	1 (0.04)	4 (0.16)	2 (0.08)	0 (0.00)
Blogs	**37** **(1.46)**	9 (0.36)	1 (0.04)	2 (0.08)	3 (0.12)	3 (0.12)	6 (0.24)	3 (0.12)	0 (0.00)
Traditional websites	**12** **(0.48)**	3 (0.12)	1 (0.04)	2 (0.08)	2 (0.08)	1 (0.04)	0 (0.00)	0 (0.00)	1 (0.04)
Hypertexts in total	**78** **(1.03)**	**20** **(0.26)**	**5** **(0.07)**	**7** **(0.09)**	**6** **(0.08)**	**5** **(0.07)**	**10** **(0.13)**	**5** **(0.07)**	**1** **(0.01)**

	enough	***several***	***some***	***any***	***either***	***neither***	***none***
Wikipedia	0 (0.00)	0 (0.00)	2 (0.08)	0 (0.00)	0 (0.00)	1 (0.04)	1 (0.04)
Blogs	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)
Traditional websites	0 (0.00)	0 (0.00)	2 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Hypertexts in total	**0** **(0.00)**	**0** **(0.00)**	**4** **(0.05)**	**1** **(0.01)**	**0** **(0.00)**	**1** **(0.01)**	**1** **(0.01)**

	many	*much*	*more*	*most*	*few*	*fewer*	*fewest*	*little*	*less*	*least*
Wikipedia	1	0	0	0	1	1	0	0	0	0
	(0.04)	(0.00)	(0.00)	(0.00)	(0.04)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)
Blogs	3	4	1	0	1	0	0	0	0	0
	(0.12)	(0.16)	(0.04)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Traditional	0	0	0	0	0	0	0	0	0	0
websites	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Hypertexts	**4**	**4**	**1**	**0**	**2**	**1**	**0**	**0**	**0**	**0**
in total	**(0.05)**	**(0.05)**	**(0.01)**	**(0.00)**	**(0.03)**	**(0.01)**	**(0.00)**	**(0.00)**	**(0.00)**	**(0.00)**

4.4.1.9 Other forms of coreference and substitution

Other forms of coreference and substitution are realised by the items *the same*, *such* and *so*. *The same* occurs only three times. *Such* is the most frequent in Wikipedia texts with eleven cases. Table 21 shows the detailed results.

Table 21: Other forms of coreference in absolute and relative per mille numbers

	Other types of coreference and substitution	***the same***	***such***	***so***
Wikipedia	**13**	2	11	0
	(0.51)	(0.08)	(0.43)	(0.00)
Blogs	**7**	1	2	4
	(0.28)	(0.04)	(0.08)	(0.16)
Traditional	**1**	0	1	0
websites	**(0.04)**	(0.00)	(0.04)	(0.00)
Hypertexts	**21**	**3**	**14**	**4**
in total	**(0.28)**	**(0.04)**	**(0.18)**	**(0.05)**

4.4.1.10 Verb phrases with *do* and combinations

The ratio between verb phrases with *do* and their combinations is 47.7 % to 52.3 %. Considering all verb phrase forms, the most common item is *do* with 29.5 %. Additionally, verb phrases with *do* and their combinations are used considerably more often in blogs than in the other hypertext types, taking 63.6 % of all items. The detailed numbers are given in Figure 12 and Table 22. Finally, the two cataphoric items are listed in Table 23.

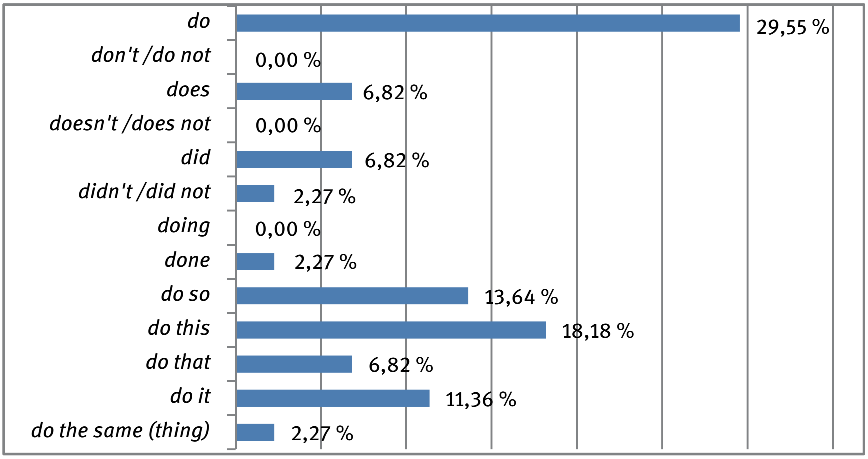


Fig. 12: Distribution of the items within the anaphor type verb phrases with *do* and combinations

Table 22: Verb phrases with *do* and combinations in absolute and relative numbers (numbers in brackets in per mille)

	Verb phrases with *do* and combinations	***do* forms**	***do***	***don't/ do not***	***does***	***doesn't/ does not***	***did***	***didn't/ did not***	***doing***	***done***
Wikipedia	**8** **(0.31)**	7 (0.27)	5 (0.20)	0 (0.00)	1 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	1 (0.04)
Blogs	**28** **(1.11)**	11 (0.43)	7 (0.28)	0 (0.00)	0 (0.00)	0 (0.00)	3 (0.12)	1 (0.04)	0 (0.00)	0 (0.00)
Traditional websites	**8** **(0.32)**	3 (0.12)	1 (0.04)	0 (0.00)	2 (0.08)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Hypertexts in total	**44** **(0.58)**	**21** **(0.28)**	**13** **(0.17)**	**0** **(0.00)**	**3** **(0.04)**	**0** **(0.00)**	**3** **(0.04)**	**1** **(0.01)**	**0** **(0.00)**	**1** **(0.01)**

	Combinations with *do*	***do so***	***do this***	***do that***	***do it***	***do the same (thing)***
Wikipedia	1 (0.04)	0 (0.00)	1 (0.04)	0 (0.00)	0 (0.00)	0 (0.00)
Blogs	17 (0.67)	5 (0.20)	6 (0.24)	3 (0.12)	2 (0.08)	1 (0.04)
Traditional websites	5 (0.20)	1 (0.04)	1 (0.04)	0 (0.00)	3 (0.12)	0 (0.00)
Hypertexts in total	**23** **(0.30)**	**6** **(0.08)**	**8** **(0.11)**	**3** **(0.04)**	**5** **(0.07)**	**1** **(0.01)**

Table 23: Cataphoric verb phrases with *do* and combinations (absolute numbers and relative numbers in per mille)

	Cataphoric verb phrases with *do* and combinations	***do this***	***do the same thing***
Wikipedia	**0** **(0.00)**	0 (0.00)	0 (0.00)
Blogs	**2** **(0.08)**	1 (0.04)	1 (0.04)
Traditional websites	**0** **(0.00)**	0 (0.00)	0 (0.00)
Hypertexts in total	**2** **(0.03)**	**1** **(0.01)**	**1** **(0.01)**

4.4.1.11 Ellipses

With ellipsis, nominal, verbal and clausal ellipsis are distinguished. As is apparent from Figure 13, it is nominal ellipsis that occupies the most of all elliptical cases with 90.5 %. The distribution according to the hypertext types, as shown in Table 24, illustrates that ellipses are slightly the most frequent in Wikipedia texts with 37.1 % of all ellipses. Furthermore, verbal ellipsis is the most common in blogs, although ellipses generally are the least frequent there.

A combination of nominal and verbal ellipsis can occur as well. Such cases were counted as part of both nominal and verbal ellipsis in Table 24. The detailed numbers of how often the combination of nominal and verbal ellipsis is found in the corpus are given in Table 25. On closer examination, verbal ellipsis predominantly occurs together with nominal ellipsis. In sum, 19 out of 26 verbal ellipses are cases where verbal ellipsis is combined with nominal ellipsis. Verbal ellipsis on its own is only more frequent in blogs: 5 of 14 instances here are uses without nominal ellipsis.

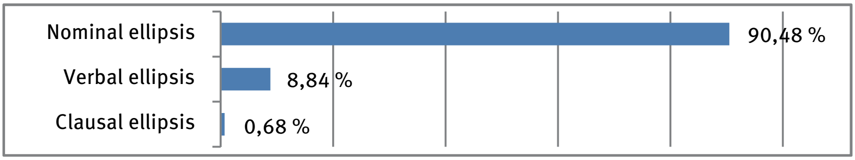


Fig. 13: Distribution of nominal, verbal and clausal ellipsis

Table 24: Absolute and relative (per mille) numbers of nominal, verbal and clausal ellipsis

	Ellipses	Nominal ellipsis	Verbal ellipsis	Clausal ellipsis
Wikipedia	**109**	102	7	0
	(4.27)	(3.99)	(0.27)	(0.00)
Blogs	**88**	74	14	0
	(3.48)	(2.92)	(0.55)	(0.00)
Traditional	**97**	90	5	2
websites	**(3.86)**	(3.58)	(0.20)	(0.08)
Hypertexts	**294**	**266**	**26**	**2**
in total	**(3.87)**	**(3.50)**	**(0.34)**	**(0.03)**

Table 25: Frequency of nominal and verbal ellipsis combined and the use of nominal and verbal ellipsis in general (absolute and relative per mille numbers)

	Nominal and verbal ellipsis combined	Nominal ellipsis without combined forms	Verbal ellipsis without combined forms
Wikipedia	6	96	1
	(0.23)	(3.76)	(0.04)
Blogs	9	65	5
	(0.36)	(2.57)	(0.20)
Traditional	4	86	1
websites	(0.16)	(3.42)	(0.04)
Hypertexts	**19**	**247**	**7**
in total	**(0.25)**	**(3.25)**	**(0.09)**

4.4.1.12 Non-finite clauses

The subtypes *to*, *-ing* and *-ed* are distributed as follows: *to* is the most frequent, followed closely by *-ing*. *-ed* is only about half as common as the other two items (see Figure 14). This finding is in accordance with Quirk et al. (2012: 993), who also states that *to*-infinitive and *-ing*-participle clauses are the most frequent of the three non-finite clauses. Most *non-finite clause* items are found in traditional website texts (36.2 % of all items), followed by blog and Wikipedia texts, which both contain a similar amount of *non-finite clause* items. However, *non-finite clause* items are distributed unevenly across the hypertext types. The items *to* and *-ing* are rarer in Wikipedia than in the other two hypertext types, but *-ed*-items are by far the most frequent in Wikipedia texts. 58.1 % of all *-ed*-items occur in Wikipedia (see Table 26).

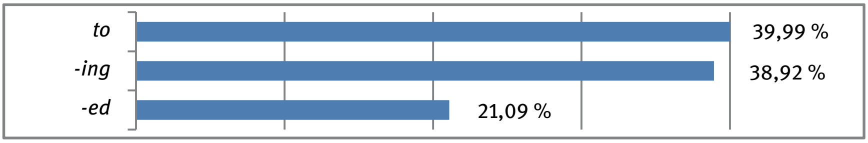


Fig. 14: Distribution of *non-finite clause* anaphors

Table 26: Absolute and relative (per mille) numbers of non-finite clauses

	Non-finite clauses	***to***	***-ing***	***-ed***
Wikipedia	**561** **(21.97)**	177 (6.93)	166 (6.50)	218 (8.54)
Blogs	**574** **(22.68)**	277 (10.94)	238 (9.40)	59 (2.33)
Traditional websites	**643** **(25.59)**	257 (10.23)	288 (11.46)	98 (3.90)
Hypertexts in total	**1,778** **(23.40)**	**711** **(9.36)**	**692** **(9.11)**	**375** **(4.94)**

-ed-non-finite clause anaphors distinguish further between regular forms that have *-ed*-inflection and irregular forms as listed in chapter 3.12.3. The relation between regular and irregular forms is 82.4 % versus 17.6 % in the corpus (Figure 15). The distribution is slightly different in blogs, where irregular forms take up only 11.9 % of all *-ed*-forms. Furthermore, only 23 of the 268 irregular forms listed above occur in the corpus. These items are given in Table 27. The most frequent of the irregulars, with 15 times, is *known*, making up 22.7 % of all irregular forms, followed by *made* (12 items), *written* (7) and *found* (6). The other items occur three times or less.

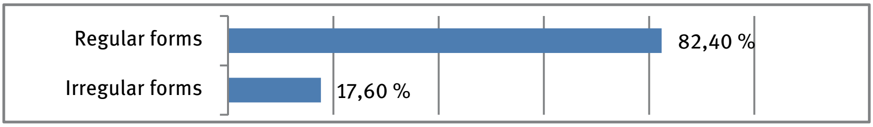


Fig. 15: Distribution of regular and irregular *-ed* forms

Table 27: Absolute and relative (per mille) numbers of regular and irregular *-ed* forms

	-ed	**Regular**	**Irregular**	***brought***	***built***	***drawn***	***eaten***	***found***
Wikipedia	**218**	178	40	1	1	0	0	5
	(8.54)	(6.97)	(1.57)	(0.04)	(0.04)	(0.00)	(0.00)	(0.20)
Blogs	**59**	52	7	0	0	0	1	0
	(2.33)	(2.05)	(0.28)	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)
Traditional	**98**	79	19	0	0	1	0	1
websites	**(3.90)**	(3.14)	(0.76)	(0.00)	(0.00)	(0.04)	(0.00)	(0.04)
Hypertexts	**375**	**309**	**66**	**1**	**1**	**1**	**1**	**6**
in total	**(4.94)**	**(4.07)**	**(0.87)**	**(0.01)**	**(0.01)**	**(0.01)**	**(0.01)**	**(0.08)**

	grown	***hung***	***held***	***known***	***laid***	***led***	***left***	***made***	***met***
Wikipedia	2	1	1	12	1	1	0	5	1
	(0.08)	(0.04)	(0.04)	(0.47)	(0.04)	(0.04)	(0.00)	(0.20)	(0.04)
Blogs	0	0	0	1	0	0	0	3	0
	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)	(0.12)	(0.00)
Traditional	0	0	1	2	0	2	1	4	0
websites	(0.00)	(0.00)	(0.04)	(0.08)	(0.00)	(0.08)	(0.04)	(0.16)	(0.00)
Hypertexts	**2**	**1**	**2**	**15**	**1**	**3**	**1**	**12**	**1**
in total	**(0.03)**	**(0.01)**	**(0.03)**	**(0.20)**	**(0.01)**	**(0.04)**	**(0.01)**	**(0.16)**	**(0.01)**

	overseen	***paid***	***said***	***sold***	***spent***	***split***	***taken***	***taught***	***written***
Wikipedia	0	0	1	0	0	0	1	2	5
	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)	(0.04)	(0.08)	(0.20)
Blogs	0	0	0	1	0	0	0	0	1
	(0.00)	(0.00)	(0.00)	(0.04)	(0.00)	(0.00)	(0.00)	(0.00)	(0.04)
Traditional	1	1	0	1	1	1	1	0	1
websites	(0.04)	(0.04)	(0.00)	(0.04)	(0.04)	(0.04)	(0.04)	(0.00)	(0.04)
Hypertexts	**1**	**1**	**1**	**2**	**1**	**1**	**2**	**2**	**7**
in total	**(0.01)**	**(0.01)**	**(0.01)**	**(0.03)**	**(0.01)**	**(0.01)**	**(0.03)**	**(0.03)**	**(0.09)**

As for cataphoric interpretations of *non-finite clause* items, the most frequent item here is *-ing* with 73.2 % of all instances. It is therefore more than five times as frequent as *-ed* or *to*. Traditional websites contain the most *-ing*-items, with 68.3 % relative to the two other hypertext types. However, *-ing*-items are generally found the most often in traditional websites. The *-ed*-items fall into regular (6 items) and irregular (2 items) forms, the latter are the forms *known* and *made* (see Table 28).

Table 28: Cataphoric *non-finite clause* items (absolute and relative per mille numbers)

	Cataphoric non-finite clauses	***-ing***	***-ed***				***to***
			in total	regular	*known*	*made*	
Wikipedia	**10** **(0.39)**	8 (0.31)	1 (0.04)	0 (0.00)	1 (0.04)	0 (0.00)	1 (0.04)
Blogs	**5** **(0.20)**	5 (0.20)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)	0 (0.00)
Traditional websites	**41** **(1.63)**	28 (1.11)	7 (0.28)	6 (0.24)	0 (0.00)	1 (0.04)	6 (0.24)
Hypertexts in total	**56** **(0.74)**	**41** **(0.54)**	**8** **(0.11)**	**6** **(0.08)**	**1** **(0.01)**	**1** **(0.01)**	**7** **(0.09)**

4.4.2 Ratio of items occurring anaphorically and non-anaphorically

Apart from anaphors, the corpus has also been examined with regard to how often the items listed as anaphors occurred non-anaphorically. This is important because anaphora resolution systems have to distinguish anaphoric from non-anaphoric uses. It is essential to keep in mind that the non-anaphoric counting has been based on how primitive systems distinguish between the anaphoric and non-anaphoric use of items. For instance, all items ending in *-ing* are potential anaphors. Consequently, all forms taking the *-ing*-ending and not functioning as anaphors are counted as non-anaphoric, even though a large part of them are not *-ing*-participles. Similarly, the items of relative pronouns, such as *who*, may function as interrogative pronouns, or *do* in the type verb phrases with *do* and combinations as auxiliary for forming questions. This means that the non-anaphoric items are listed in the corresponding anaphor types not because of their word class membership or function, but due to their identical form with the items of the anaphor type.

The general frequency of the non-anaphoric uses of each anaphor type is illustrated in Figure 16. Most non-anaphoric items appear as non-finite clauses with 32.3 %, followed by proper names (25.6 %) and noun phrases with a definite article (19.7 %). Indefinite pronoun items make up 8.3 %. More rare, with less than 5 % are non-anaphoric items in the anaphor types of central pronouns, adverbs, relative pronouns, verb phrases with *do* and combinations, demonstrative pronouns, other forms of coreference and substitution, and reciprocal pronouns. Ellipses are not included, as this type per definition does not contain specific items but spaces.

The distribution of the non-anaphoric items across the hypertext types is almost equal (see Table 29). In sum, non-anaphoric items are the most frequent in blog and Wikipedia texts, here constituting 34.5 % and 34.0 % of all items respectively and the rarest in traditional website texts. Most anaphor types are the most frequent in blogs, except for noun phrases with a definite article and proper names. The distribution across the three hypertext types so reveals that blogs contain about twice as many non-anaphoric demonstrative pronouns (58.2 %) and adverbs (49.8 %) as the other two hypertext types. Additionally, non-anaphoric verb phrases with *do* and combinations occur about four and a half times more often in blogs (65.3 % of all these verb phrases) than in Wikipedia and about three times more often than in traditional websites. Finally, it is interesting that non-anaphoric *non-finite clause* items show about the same frequency in blogs and traditional websites, but are rarer in Wikipedia texts. This might go back to the fact that anaphoric non-finite clauses are also rarer in Wikipedia.

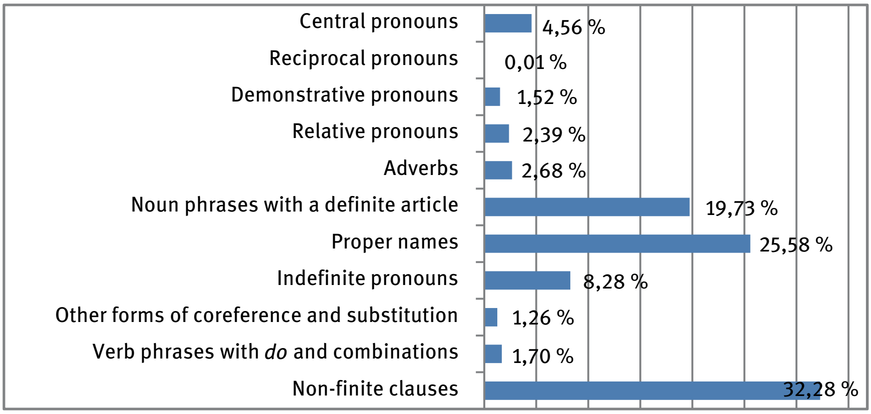


Fig. 16: Non-anaphoric uses of all anaphor type items

Table 29: Absolute and relative per mille numbers of all non-anaphoric items

	Non-anaphoric items in total	Central pronouns	Reciprocal pronouns	Demonstrative pronouns	Relative pronouns	Adverbs
Wikipedia	**5,717** **(223.91)**	54 (2.11)	0 (0.00)	52 (2.04)	126 (4.93)	104 (4.07)
Blogs	**5,796** **(228.98)**	377 (14.89)	0 (0.00)	149 (5.89)	156 (6.16)	224 (8.85)
Traditional websites	**5,280** **(210.12)**	335 (13.33)	2 (0.08)	55 (2.19)	119 (4.74)	122 (4.85)
Hypertexts in total	**16,793** **(221.04)**	**766** **(10.08)**	**2** **(0.03)**	**256** **(3.37)**	**401** **(5.28)**	**450** **(5.92)**

	Noun phrases with a definite article	Proper names	Indefinite pronouns	Other forms of coreference and substitution	Verb phrases with *do* and combinations	Non-finite clauses
Wikipedia	1,280 (50.13)	1,848 (72.38)	452 (17.70)	60 (2.35)	41 (1.61)	1,700 (66.58)
Blogs	915 (36.15)	1,283 (50.69)	539 (21.29)	97 (3.83)	186 (7.35)	1,870 (73.88)
Traditional websites	1,119 (44.53)	1,164 (46.32)	400 (15.92)	55 (2.19)	58 (2.31)	1,851 (73.66)
Hypertexts in total	**3,314 (43.62)**	**4,295 (56.53)**	**1,391 (18.31)**	**212 (2.79)**	**285 (3.75)**	**5,421 (71.35)**

Turning now to the ratio between anaphoric and non-anaphoric items across anaphor types, Table 30 shows the details and Figure 17 provides a visual illustration. In these, the frequency of anaphoric items is compared with that of the non-anaphoric items. In sum, about 27 % of all relevant items are anaphoric and about 73 % are non-anaphoric. Furthermore, the corpus analysis shows that the likelihood that items are anaphoric is highest for reciprocal pronouns, followed by central pronouns. However, for reciprocal pronouns it can only be a tendency due to the low frequency. The highest percentage of non-anaphoric items relative to their anaphoric items is found with indefinite pronouns and other forms of coreference and substitution.

Table 30: Relation of anaphoric and non-anaphoric items (absolute numbers and, in brackets, relative numbers in per cent)

	Items in total	Central pronouns	Reciprocal pronouns	Demonstrative pronouns	Relative pronouns	Adverbs
Wikipedia (anaphoric)	**1,943 (25.37)**	415 (88.49)	4 (100.00)	139 (72.77)	182 (59.09)	22 (17.46)
Wikipedia (non-anaphoric)	**5,717 (74.63)**	54 (11.51)	0 (0.00)	52 (27.23)	126 (40.91)	104 (82.54)
Blogs (anaphoric)	**2,036 (26.00)**	640 (62.93)	2 (100.00)	174 (53.87)	216 (58.06)	25 (10.04)
Blogs (non-anaphoric)	**5,796 (74.00)**	377 (37.07)	0 (0.00)	149 (46.13)	156 (41.94)	224 (89.96)
Traditional websites (anaphoric)	**2,113 (28.58)**	619 (64.88)	4 (66.67)	111 (66.87)	198 (62.46)	33 (21.29)
Traditional websites (non-anaphoric)	**5,280 (71.42)**	335 (35.12)	2 (33.33)	55 (33.13)	119 (37.54)	122 (78.71)
Hypertexts in total (anaphoric)	**6,092 (26.62)**	**1,674 (68.61)**	**10 (83.33)**	**424 (62.35)**	**596 (59.78)**	**80 (15.09)**
Hypertexts in total (non-anaphoric)	**16,793 (73.38)**	**766 (31.39)**	**2 (16.67)**	**256 (37.65)**	**401 (40.22)**	**450 (84.91)**

	Noun phrases with a definite article	**Proper names**	**Indefinite pronouns**	**Other forms of coreference and substitution**	**Verb phrases with *do* and combinations**	**Non-finite clauses**
Wikipedia (anaphoric)	228 (15.12)	233 (11.20)	29 (6.03)	13 (17.81)	8 (16.33)	561 (24.81)
Wikipedia (non-anaphoric)	1,280 (84.88)	1,848 (88.80)	452 (93.97)	60 (82.19)	41 (83.67)	1,700 (75.19)
Blogs (anaphoric)	80 (8.04)	165 (11.40)	37 (6.42)	7 (6.73)	28 (13.08)	574 (23.49)
Blogs (non-anaphoric)	915 (91.96)	1,283 (88.60)	539 (93.58)	97 (93.27)	186 (86.92)	1,870 (76.51)
Traditional websites (anaphoric)	158 (12.37)	229 (16.44)	12 (2.91)	1 (1.79)	8 (12.12)	643 (25.78)
Traditional websites (non-anaphoric)	1,119 (87.63)	1,164 (83.56)	400 (97.09)	55 (98.21)	58 (87.88)	1,851 (74.22)
Hypertexts in total (anaphoric)	**466 (12.33)**	**627 (12.74)**	**78 (5.31)**	**21 (9.01)**	**44 (13.37)**	**1,778 (24.70)**
Hypertexts in total (non-anaphoric)	**3,314 (87.67)**	**4,295 (87.26)**	**1,391 (94.69)**	**212 (90.99)**	**285 (86.63)**	**5,421 (75.30)**

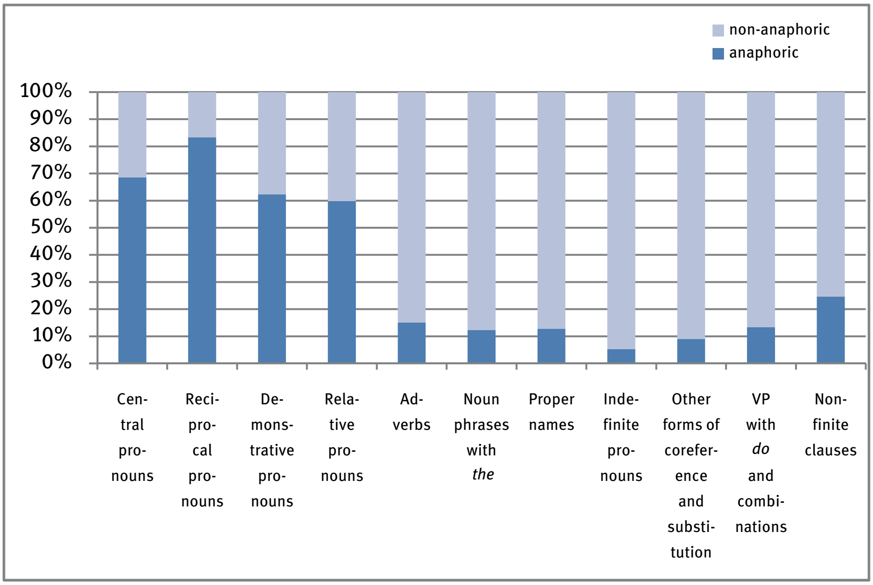


Fig. 17: Percentage distribution of anaphoric and non-anaphoric items

As to previous studies, the Syracuse study (e.g. Liddy 1990: 45) comes close to the findings of the hypertext corpus. It reports 2,204 anaphoric instances and 4,780 non-anaphoric cases of all items, i.e. 31.6 % anaphoric items and 68.4 % non-anaphoric items. Other studies such as Mitkov & Hallett (2007: 275), for instance, found that 545 central pronouns are anaphoric and 108 non-anaphoric in the corpus of technical manuals, i.e. 83.5 % are anaphoric. Unfortunately, no numbers for non-anaphoric pronouns are given for Mitkov & Hallett's other two corpora. Similarly, Barbu (2002: 276) reports 86.7 %, i.e. 366 anaphoric items out of 422 central pronouns. Moreover, Mitkov, Evans & Orasan (2002: 178) found 84.7 % anaphoric central pronouns, i.e. 2,263 anaphoric pronouns were identified and 408 non-anaphoric items. These higher numbers in those corpora might result from not including the items *we*, *us*, *our*, *ourselves*. If they were left out in the hypertext corpus, 87.5 % of all central pronouns would be anaphoric. This number would then be in accordance with previous findings. Incidentally, Hobbs (1986) and Vicedo & Ferrández (2000) do not give any details about the relation of anaphoric and non-anaphoric items. In Liddy (1990: 45), noun phrases with a definite article were found to have the most non-anaphoric items; only 13.5 % are anaphoric there. Furthermore, 78.5 % of central pronouns, 54.9 % of relative pronouns and 24.6 % of indefinite pronouns are anaphoric. Except for indefinite pronouns, the numbers are similar to the analysis of the hypertext corpus. This deviation in indefinite pronouns might result from exactly what items are included with indefinite pronouns.

As for the hypertext types in the corpus, most non-anaphoric items relative to the anaphoric items with central pronouns, demonstrative pronouns, relative pronouns, adverbs, noun phrases with a definite article and non-finite clauses are found in blogs. As obvious from the findings of the corpus analysis, noun phrases with a definite article are more non-anaphoric than anaphoric. This is consistent with the statement of Halliday & Hasan (2008: 73-74). Reciprocal pronouns, indefinite pronouns, other forms of coreference and substitution and verb phrases with *do* and combinations have the most non-anaphoric items in traditional websites. Additionally, it turned out that proper names in the corpus are used more often anaphorically in traditional websites than in the other two hypertext types. Finally, other forms of coreference and substitution and verb phrases with *do* and combinations are used less frequently non-anaphorically in Wikipedia than in blog and traditional website texts.

4.5 Conclusion

Previous studies as outlined in chapter 4.1 do not reveal the relative frequency of all the anaphors described in chapter 3. Studies up to now focus primarily on central pronouns and generally are restricted in the diversity of hypertexts analysed. This means that the studies lack important anaphor types and are not representative concerning the hypertexts considered. Therefore, a new study was carried out. The corpus established for this analysis encompasses three hypertext types: Wikipedia texts, blog texts and traditional website texts containing websites from companies, personal websites, institutional websites and online newspapers. These types represent a variety of texts found on the Internet that are relevant for text retrieval systems (cf. chapter 5).

The corpus analysis revealed that the most frequent anaphor types are non-finite clauses, closely followed by central pronouns. This has important consequences for linguistic research in anaphora and computational anaphora resolution: so far *non-finite clause* anaphors have never been considered as one anaphor type in classifications of anaphors, let alone in text retrieval systems. Further relevant anaphor types are proper names, relative pronouns, noun phrases with a definite article, demonstrative pronouns and ellipses. The other anaphor types are of minor importance because they only occupy about 1 ‰ or less.

Comparing the results of the corpus analysis to previous studies outlined in chapter 4.1, the Syracuse study, for example, found that noun phrases with a definite article are the most frequent, with 21.2 % of all anaphors in the corpus containing abstracts. In second position are relative pronouns, which are represented as 19.0 %, central pronouns take the third position with 17.6 %. The significance of this study and its comparability to the corpus analysis of this book, however, is limited because the Syracuse study focuses on abstracts and it does not detail which items in which contexts are considered anaphoric. Furthermore, the items that are meant by anaphor types such as “subject references” is not clear. Most importantly, *non-finite clause* anaphors are not part of the Syracuse study (cf. Liddy 1990: 45). In the *Longman Grammar of Spoken and Written English*, Biber et al. (2007: 237) maintain that the main anaphoric items are personal pronouns and definite noun phrases. With definite noun phrases, not only noun phrases with a definite article, but also demonstrative pronouns in dependent function and determinative possessive pronouns, for instance, are meant. Not only is the category of definite noun phrases too broad but Biber et al. also do not make any mention of *non-finite clause* anaphors. These comparisons show the necessity of a comprehensive corpus analysis as carried out in this book.

As for the individual items, *to* is the most common of all anaphor items in the corpus and is closely followed by *-ing*. With regard to cataphoric items, text retrieval systems should consider such interpretations with non-finite clauses and probably for personal pronouns. Cataphors rarely occur with other anaphor types. The knowledge of how anaphor types are distributed in terms of their frequency brings substantial benefits to anaphora resolution systems because text retrieval systems then can focus on those anaphor types that are more frequent and therefore more important.

Another discovery of the study is that anaphors are distributed in about the same frequency across the three hypertext types, but differ in the distribution of anaphor types. These results can be helpful in deciding which anaphor type is more important in which hypertext type (cf. Table 8). Moreover, some differences across the hypertext types can be explained straightforwardly. For instance, central pronouns are less common in Wikipedia texts. Blogs and traditional website texts contain websites from online newspapers, which often discuss news involving people. As a consequence, the need for referring to people is higher and this is done by using central pronouns (cf. Biber et al. 2007: 333). Other findings, e.g. the distribution of reciprocal pronouns across the hypertext types, would need a larger amount of texts for overall conclusions in order to verify whether current deviations in the hypertext types are caused by the texts chosen or rather represent a general trend.

Finally, text retrieval systems also have to distinguish between items that are anaphoric and items that are non-anaphoric or have the same form as the corresponding anaphoric item. Generally, most non-anaphoric items in the corpus fall on *non-finite clause* anaphors. They occupy 32.3 % of all non-anaphoric items. The ratio between anaphoric and non-anaphoric items is, however, worst for indefinite pronouns and other forms of coreference and substitution. Proper names and noun phrases with a definite article also show a high number of non-anaphoric items. Across the hypertext types, most non-anaphoric items are found in Wikipedia texts. These results are useful for anaphora resolution systems in order to decide how likely an item is an anaphor. It should be kept in mind that the comparison of anaphoric versus non-anaphoric items does not suggest that an item with a high number of non-anaphoric instances is difficult to resolve for anaphora resolution systems. It rather suggests the need for more pre-processing stages in which non-anaphoric items are excluded. However, the amount of computational effort does not only depend on the ratio of anaphoric versus non-anaphoric items. Detecting the non-anaphoricity might be easy for some items, while it might be more demanding for others (cf. chapter 7.1).

Apart from that, the characteristics of language on the Internet have to be taken into account when analysing anaphors. Language can show aspects not found in traditional texts, e.g. no capital letters, limited or unusual punctuation and errors. Furthermore, analysing such features can help to identify non-anaphoric items, for example, in the case of semantically empty links.

In sum, the results of the corpus analysis do not only reveal the relative frequency and in consequence the importance of the twelve anaphor types. The findings are also valuable for text retrieval and anaphora resolution systems, for instance, in deciding which anaphor types are the most important to be resolved and which anaphor types are only of minor importance due to their low frequency.

From the insights obtained so far, the following two research questions for computational anaphora resolution arise: first, how far anaphora resolution can improve the effectiveness of text retrieval systems needs to be explored. As this book analyses hypertexts and as these are mostly accessed with the help of text retrieval systems, it is essential to consider anaphora resolution in the light of text retrieval. For this purpose, it is first necessary to outline how text retrieval systems work and to what extent they draw on anaphora resolution systems (see chapter 5). Afterwards, the benefit of anaphora resolution in text retrieval can be demonstrated (see chapter 6.1.5). As will be discussed, research in the application of anaphora resolution in text retrieval systems is limited, even if it would be essential to find out about the content of a text (see chapter 6.4).

Second, as *non-finite clause* anaphors have proven to be the most frequent anaphor type, they should not be ignored in anaphora resolution systems. The question is, however, whether rules that also show a satisfactory degree of effectiveness can be devised for their resolution. Consequently, we first need to know about how anaphora resolution systems work (see chapter 6.2 and 6.3). In a further step, the linguistic foundations of *non-finite clause* anaphors from chapter 3 are taken as a basis for developing the rules. To show their effectiveness, these rules are then evaluated on the corpus (see chapter 6.5 and 7). Only if the rules are accurate enough, are they deemed useful to be implemented in anaphora resolution systems.

5 Text retrieval and its handling of anaphors

The previous chapter detailed anaphors and their frequency in hypertexts. As hypertexts are part of the Internet and typically accessed by using text retrieval systems, the functionality of text retrieval will now be described.

5.1 What are text retrieval systems?

Text retrieval systems are used to search for and to find written information. The information that is needed is already stored somewhere. The task of text retrieval systems is then to make sure that this information is accessible in a suitable form when needed. The information cannot only be of a textual, but also of a visual or an oral nature, e.g. images, videos. In this context, the terms “information retrieval” and “text retrieval” occur. By definition, “information retrieval” concerns all types of media, e.g. written, audio and visual forms; “text retrieval” is restricted to the written forms of language (cf. Holzinger 2002: 16; Henrich 2007a: 16-20; Meadow et al. 2007: 2-5; Stock 2007: 9-10, 95). Consequently, Büttcher, Clarke & Cormack (2010) define “information retrieval” as follows: “Information retrieval (IR) is concerned with representing, searching and manipulating large collections of electronic text and other human-language data” (ibid.: 2). This book focuses on text retrieval, due to the relevance for anaphora resolution systems. However, various facts presented here are not specific of text retrieval systems but are also valid for information retrieval systems. Consequently, the term “information retrieval” is often used in discussions that only focus on text retrieval (e.g. Stock 2007: 295; Siddiqui & Tiwary 2008: 301-303; Baeza-Yates & Ribeiro-Neto 2011: 159).

Text retrieval systems have already been used before the advent of computers, although the term “retrieval” is a more recent innovation: it was first recorded in 1958 (cf. Simpson & Weiner 1989: 794). Since humankind has established libraries, text retrieval systems, i.e. systems for the detection of desired information in such libraries, have been in use. However, text retrieval systems have experienced unprecedented demand and boom since people use the World Wide Web in everyday life. Search engines such as Google are prototype examples of text retrieval systems (cf. Stock 2007: 38, 47; Büttcher, Clarke & Cormack 2010: 2).

Looking more closely at how text retrieval systems work, a generalised procedure of such systems is explained as a starting point. An illustration of a text retrieval procedure is shown in Figure 1. On the one side there is the information

DOI 10.1515/9783110416756

need of the user (left side in the Figure). In order to satisfy this need he or she formulates a query in natural language. On the other side there are documents (right side in the Figure), which are available to satisfy this information need of the user.[1] Both the query and the documents are represented in a particular form, so that they can be processed more easily. The query and the documents in these representations are then compared ("matching"). Finally, those documents matching the query are returned as results (cf. Henrich 2007a: 38-39).

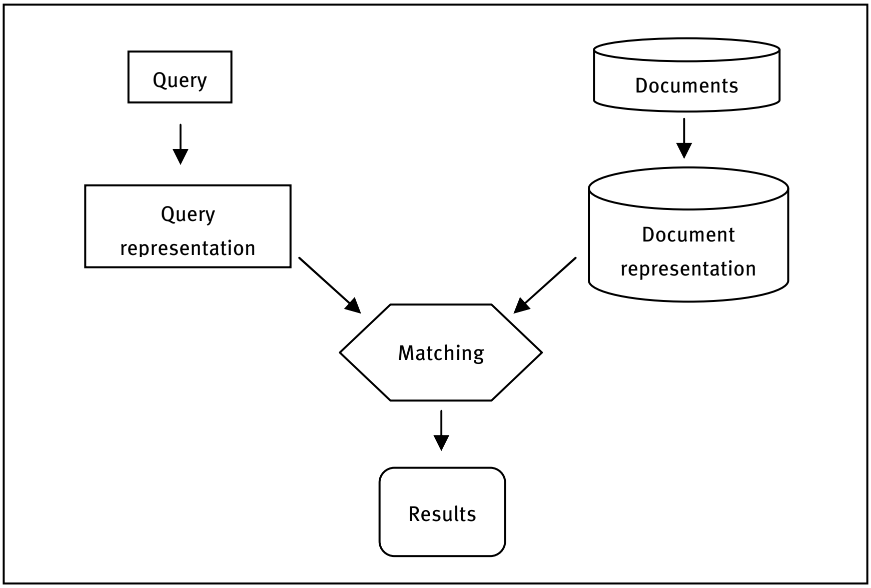


Fig. 1: Generalised model of text retrieval (adapted from Henrich 2007a: 38)

5.2 Text retrieval models

The basic processes of text retrieval systems outlined above can now be realised in different ways leading to various types or models of text retrieval systems. The most common and classic ones are the Boolean, the vector space and the probabilistic models (cf. Henrich 2007a: 39; Stock 2007: 102-108). Weber (2006:

[1] The term "document", according to Jurafsky & Martin (2009), "refers generically to the unit of text indexed in the system and available for retrieval. [...] In Web-based applications, document can refer to a Web page, a part of a page, or an entire website" (ibid.: 801-802).

1-45), for instance, also mentions the Latent Semantic Indexing Model for text retrieval. These models are now described in more detail.

The traditional Boolean model is based on Boolean logic and basically uses three operators: “AND”, “OR”, “AND NOT”. If the user, for example, needs documents in which the two expressions *university* and *Passau* should appear, “AND” is used, i.e. he or she enters *university AND Passau*. If only one of these two terms[2] needs to occur in the document, “OR” is used, i.e. *university OR Passau*. In this case, documents that either contain the term *university* or the term *Passau* are returned. The operator “AND NOT”, for example, in *Passau AND NOT university* delivers documents including the term *Passau*, but not the term *university*.

A major disadvantage of the traditional Boolean model is that documents are not sorted by relevance, either a document is returned or not, but no further steps in sorting the matching documents are taken. Furthermore, morphological variants of words with one and the same root are not considered, i.e. if searching for *universities*, the term *university* is ignored. Additionally, this model does not take into account where the word appears in a document, for instance, whether in headlines or the body of a text. Terms in headlines can be more revealing about the content of a document and therefore could be regarded as more important. The retrieval quality is worse as with, for instance, the vector space and probabilistic models. Despite these disadvantages, the Boolean models are – with various adaptions – the most commonly used models in text retrieval (cf. Weber 2006: 1-28, 1-29; Henrich 2007a: 39-43; Stock 2007: 104; Jurafsky & Martin 2009: 799).

In vector space models, the documents and the query are represented as vectors. The smaller the angle between a document and the query, the more relevant the document is to the query. The vector is made up by the terms occurring in the query or document: for each term the vector takes one dimension. Figure 2 illustrates a simplified case where only two dimensions (x- and y-axis), i.e. two terms, are involved. Here, Document 2 (D2) with a smaller angle β comes closer to the query and is more relevant for the query than Document 1 (D1) with a larger angle α. Most weighting schemes that are used in the vector space models include two factors for representing documents: the within document frequency weight (WDF) and the inverse document frequency weight (IDF). The

[2] *Term* is used instead of *word*, as the item entered in a query does not necessarily need to be a word. A term can also be a phrase or an item such as *inform** that represents the words *inform* and *information*, for example (cf. Büttcher, Clarke & Cormack 2010: 6; see also Jacquemin & Bourigault 2004: 600-607).

first factor gives the frequency of a term in a document relative to the whole number of terms or relative to the most frequent term in a document. The more frequent a term is in a document, the higher is the WDF of this term. The IDF considers the ratio between the number of the documents in a collection that contain a specific term and the number of all documents. The more documents contain this term in a collection, the lower is the IDF of this term, which is why it is called "inverse" (cf. Manning, Raghavan & Schütze 2008: 108-109; Croft, Metzler & Strohman 2010: 22, 245-246). Tzoukerman, Klavans & Strzalkowski (2004) sum up the advantages of this model: "The vector space model is simple, fast, and popular" (ibid.: 534). Nevertheless, there are disadvantages as, for instance, the model assumes that index terms[3] are independent of each other (cf. Baeza-Yates & Ribeiro-Neto 2011: 79).

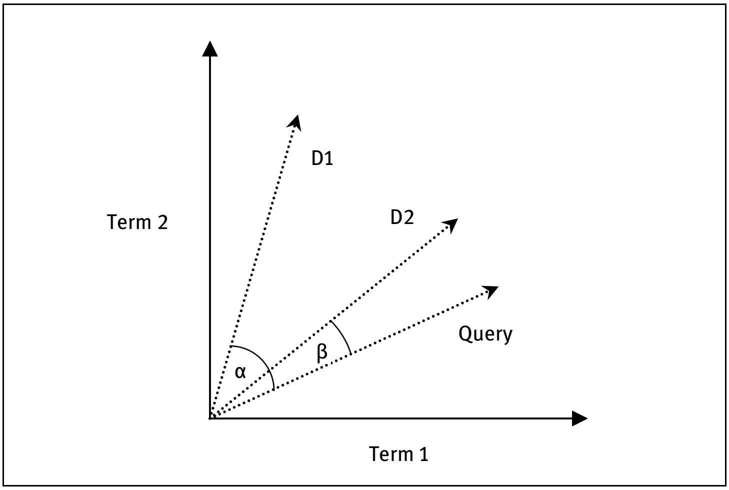


Fig. 2: Vector representation in the vector space model (cf. Croft, Metzler & Strohman 2010: 242-244)

Third, probabilistic models are based on probability theory, i.e. how likely a document matches the query. The main advantage of these models is that "documents are ranked in order of their probability to be relevant" (Tzoukerman, Klavans & Strzalkowski 2004: 534). As one disadvantage, this method

[3] Croft, Metzler & Strohman (2010) define an index term as being "the representation of the content of a document" (ibid.: 75). This means that these terms are listed in the index of a search engine (ibid.: 132).

does not consider how frequent index terms are in each document (cf. Tzoukerman, Klavans & Strzalkowski 2004: 534; Weber 2006: 1-39; Henrich 2007a: 43-44; Meadow et al. 2007: 63-64; Stock 2007: 104-105; Baeza-Yates & Ribeiro-Neto 2011: 86). At present, probabilistic models are more important than vector space models (cf. Büttcher, Clarke & Cormack 2010: 55). Croft, Metzler & Strohman (2010) argue:

> [P]robabilistic retrieval models [...] are the dominant paradigm today. These models have achieved this status because *probability theory* is a strong foundation for representing and manipulating the uncertainty that is an inherent part of the information retrieval process. (ibid.: 247-248)

Finally, the Latent Semantic Indexing model is similar to the vector space models. As the vector space models do not find documents that contain synonyms of query terms, alternatives have been searched for. For example, if the query contains the term *big*, the documents including the term *large* are not found. The Latent Semantic Indexing model can help here because its aim is not to represent individual terms but a number of terms, i.e. a concept. It assumes that each document has an inherent semantic structure. In order to represent that structure, the vectors of documents are transformed and clustered to new vector dimensions by using statistical methods. Although the model returns good retrieval quality, it is in sum not better than the quality the other models show. One major disadvantage is its "[e]xtremely expensive computation" (Weber 2006: 1-63), which means that it takes much longer in this model to calculate the values in order to compare a document and a query. Fast algorithms are not available for this model (cf. Krüger-Thielmann & Paijmans 2004: 365-366; Tzoukerman, Klavans & Strzalkowski 2004: 534-535; Weber 2006: 1-45, 1-52; Henrich 2007a: 277).

There are further models predominantly used on the World Wide Web. For instance, the link-based model ("linktopologische Modell", Stock 2007: 105) counts how many links are contained on a specific website and how many links on the Internet refer to this website. Network models are based on the principle of clusters. This means that certain documents or names, for instance, are more central and therefore important. The ranking of the documents can be inferred from this information. Other models focus on the user: if websites document user behaviour, this can be used for ranking the documents. Additionally, information about the user can be incorporated for the text retrieval process. For instance, if the user searches for the nearest restaurant, geographical information about the position of the user is helpful (cf. Kleinberg 1999: 604; Stock 2007: 105-106).

In sum, Zhai (2009) points out for text retrieval models: “We do not yet have a clear single winner among all the models that can consistently outperform all other models” (ibid.: 25). Consequently, an advantage of the models described above is that they are not mutually exclusive. Stock (2007) states: “Es sei betont, dass sich die genannten Retrievalmodelle keineswegs gegenseitig ausschließen. Insofern ist es in praktischen Anwendungen sinnvoll, mehrere Modelle zusammen zu implementieren” (ibid.: 106).

5.3 Evaluation of text retrieval results

In order to evaluate the results of text retrieval systems, various measures can be used. There are measures for effectiveness, i.e. about the quality of results and measures for efficiency, i.e. about accomplishing the text retrieval processes with as few resources as possible, such as time, space and cost. To start with efficiency, the time required from entering a query in a search engine to displaying the results can be measured, for instance. Text retrieval systems should not take too long to respond. Besides, effectiveness is by far the most important aspect when dealing with text retrieval systems because what predominantly counts is the result a text retrieval system returns. Among effectiveness measures, precision and recall are traditionally used (cf. Rijsbergen 1979: 10-11; Henrich 2007a: 59; Meadow 2007: 335; Büttcher, Clarke & Cormack 2010: 406-407, 468).

In more detail, precision is used to measure the percentage of relevant documents in the results returned. This means how many relevant documents are included in the results versus how many documents of the result are not relevant. The goal for text retrieval systems is to return, at best, only relevant documents. Recall gives the per cent number of completeness, i.e. how many relevant documents were found in the whole document collection (see Figure 3). The aim of text retrieval systems is of course to find as many, at best all, relevant documents in a collection (cf. Henrich 2007a: 53-54; Stock 2007: 63-64). Here precision and recall is illustrated visually:

$$\text{Precision} = \frac{\text{Relevant documents found}}{\text{Documents found in total}}$$

$$\text{Recall} = \frac{\text{Relevant documents found}}{\text{Relevant documents in total}}$$

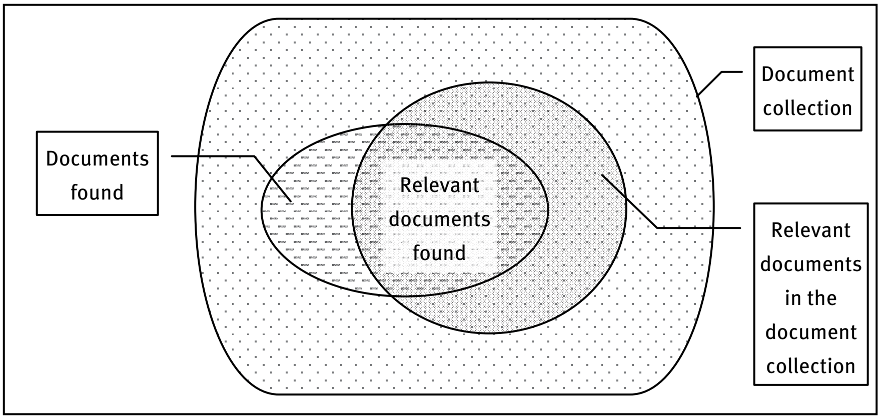


Fig. 3: Visualisation of the components needed for precision and recall (adapted from Baeza-Yates & Ribeiro-Neto 2011: 135)

A value of 80 % in recall then means that the text retrieval system has found 80 % of all relevant documents in the collection; 60 % in precision indicates that 60 % of the results returned are documents that are relevant. Optimal text retrieval systems would achieve 100 % for recall and for precision. Current systems on the Web, however, achieve only a number around 40 %, both for recall and for precision. As users look at the first few pages of the results, it is the precision within these pages that is more important in practice. Recall and precision cannot be separated in reality, as practice has shown that higher recall leads to worse precision and vice versa. Additionally, it should be kept in mind that recall and precision simplify the classification process of documents. These values presuppose that documents can always be classified binary, i.e. whether they are relevant or not, which is not always the case (cf. Henrich 2007a: 53-54, 60-61; Stock 2007: 63-64; Kowalski 2011: 8). Moreover, what is relevant and not relevant in the list of results is also subjective to a more or less extent. As a result, Stock (2007: 56) differentiates objective information adequacy, i.e. relevance, from subjective information adequacy, i.e. pertinence.

Apart from that, the calculation of recall needs further explanation. To calculate the value for recall, the number of all relevant documents in a collection is needed. Yet, it is difficult to define which and how many documents that a text retrieval system has not found are relevant, especially if document collections are as large as on the World Wide Web. In such cases it is not possible to count all relevant documents. As a result, different approaches have been developed to estimate the number of relevant documents not found.

The most common approach is query expansion. With that method a query is formulated and the text retrieval system returns its results first. After that, the query is expanded by including synonyms in the query or substituting connected terms with “AND” for “OR”. It is assumed that the results returned with such an extended query contain all results that have been included with the first result list and also further queries that were relevant in the first query but that have not been found. From that, recall can be calculated. A specific form of query extension is the so-called “pooling”. It is used if, for example, different text retrieval systems are compared. It is assumed that at least one system finds a relevant document, so the combined result lists of all systems would include all relevant documents. From that, recall can be calculated (cf. Voorhees 1998: 295; Henrich 2007a: 66-69; Stock 2007: 63; Büttcher, Clarke & Cormack 2010: 73-74; Croft, Metzler & Strohman 2010: 307; Gödert, Lepsky & Nagelschmidt 2012: 332).

The values of recall and precision can also be combined and represented in one figure, the so-called “F-measure”. It is possible to give recall or precision more weight in this measure. This is represented by α. If both take the same weight, α is 0.5 (cf. Henrich 2007a: 65). The equation for calculating F-measure is as follows[4]:

$$\text{F-measure} = \frac{\text{Precision} \times \text{Recall}}{(1 - \alpha) \times \text{Recall} + \alpha \times \text{Precision}}$$

In order to compare and assess different algorithms and methods and their success in retrieving as many relevant documents as possible and few non-relevant documents, Text REtrieval Conferences (TREC)[5] can be referred to. TREC is one of the best-known test collections where text retrieval approaches are tested on one and the same document collection (cf. Krüger-Thielmann & Paijmans 2004: 369-370; Büttcher, Clarke & Cormack 2010: 23-26; Croft, Metzler & Strohman 2010: 5-6, 307). There are different tracks, each focusing on a different type of text retrieval task. Commercial Web search engines, however, have so far not taken part in any of these tests (cf. Levene 2010: 31-32).

[4] In this chapter, the symbol “×” stands for multiplication.

[5] For more information see http://trec.nist.gov (date of last access: 08/02/2013). See also Henrich (2007a: 77-87) and Baeza-Yates & Ribeiro-Neto (2011: 158-167) for TREC and other, however smaller, test collections.

5.4 Natural language processing methods in text retrieval

In order to represent the content of a query and of documents, natural language processing (NLP) methods can be used. Jackson & Moulinier (2002) define NLP as follows:

> The term 'Natural language processing' (NLP) is normally used to describe the function of software or hardware components in a computer system which analyze or synthesize spoken or written language. The 'natural' epithet is meant to distinguish human speech and writing from more formal languages, such as mathematical or logical notations, or computer languages, such as Java, LISP, and C++. (ibid.: 2-3)

There are two approaches to NLP, one is "symbolic", the other "empirical". The first, i.e. symbolic approach, "consists largely of rules for the manipulation of symbols, e.g., grammar rules [...] Symbolic NLP tends to work top-down by imposing known grammatical patterns and meaning associations upon texts" (Jackson & Moulinier 2002: 7). Empirical NLP looks at statistical distributions, i.e. the quantitative analysis of a language. It "tends to work bottom-up from the texts themselves, looking for patterns and associations to model, some of which may not correspond to purely syntactic or semantic relationships" (ibid.: 7). In order to achieve an efficient analysis of language with natural language processing methods, it is advisable to combine these two approaches (cf. ibid.: 8). As will be shown in chapter 7, this book draws on symbolic, i.e. syntactic features, as well as empirical approaches, i.e. insights from the corpus analysis, to define rules for anaphora resolution.

Yet, natural language processing methods are used to a different extent: some text retrieval systems use more, some fewer of these methods. Generally, an analysis starts at the largest units, i.e. the documents, and goes on step by step to the smallest units, i.e. individual words. This means that documents are divided up into paragraphs and these into sentences and sentences into words. Words can then be analysed in terms of their word class membership (cf. Jackson & Moulinier 2002: 9).

From that, different levels can be outlined for processing texts with NLP techniques. Siddiqui & Tiwary (2008: 4-6, 302-304) speak of the phonological, morphological, lexical, syntactic, semantic, discourse and pragmatic level. To start with, the phonological level analyses phonemes of a language but is hardly relevant for text retrieval. The morphological level considers morphemes. One of its uses in text retrieval is stemming (cf. chapter 5.4.3). Furthermore, the lexical level deals with words as units. Frequent uses involving this level are tokenisation, POS-tagging and stop word detection (cf. chapters 5.4.1, 5.4.2 and 5.4.4). The syntactic level takes into consideration the structure of sentences, that is, the constituents and how these are related grammatically (cf. chapter 5.4.5). The semantic level is about the meaning of words and sentences and is used in word sense disambiguation, for instance (cf. ibid.: 156). The discourse level usually deals with larger units than sentences, such as paragraphs and whole documents. One use of the discourse level is anaphora resolution, which is, however, hardly considered in text retrieval. Finally, the pragmatic level is about how the context influences the meaning of units and requires world knowledge. This level is not used in text retrieval so far. From all these levels, only the morphological level is frequently used in text retrieval, all others are hardly ever considered. Henrich (2007a), however, points out the importance of such techniques:

> Ziel des Information Retrieval ist es nun, die Semantik – also den Inhalt – der Dokumente zu adressieren. Dabei sollte von der konkreten Begriffswahl und Formulierung abstrahiert werden. Hierzu sind im Laufe der Jahre zahlreiche Techniken entwickelt worden. (ibid.: 91)

It is therefore important to use language processing methods in order to obtain better text retrieval results. In the following chapters, these approaches are described. Central methods are stop word detection and a morphological analysis of the terms identified, for instance, by using stemmers. Finally, natural language processing methods are discussed in relation to the Web.

5.4.1 Sentence delimitation and tokenisation

In order to segment entities into sentences, sentence delimiters are in use. Sentence boundaries can be detected by paying attention to punctuation marks. However, such marks are often ambiguous because a full stop, for example, is not always an indication of a sentence boundary. It can occur together with an abbreviation, as a decimal point in English or in enumerations. Another approach is to analyse words starting with a capital letter as this is a feature of words at the beginning of sentences in English. Again, not all capitalised words mark the beginning of a sentence. This is the case with the word *English*, which is always capitalised. Consequently, different rules have to be established or empirical information has to be gathered to account for such exceptions (cf. Jackson & Moulinier 2002: 9-10).

Tokenisers, also called “word segmenters” or “lexical analysers”, divide a string of characters into tokens[6], in this case words. Generally, blank spaces indicate word boundaries. There are again several exceptions, as with hyphens, e.g. *blue-green* and compounds with and without a blank space, e.g. *tongue twister*, *downhill* (cf. Jackson & Moulinier 2002: 10; Krüger-Thielmann & Paijmans 2004: 357). Approaches to tokenisation are either symbolic or statistic. The first method relies on heuristics[7] that are established on information about what tokens look like. This leads to rules that are applied for identifying the tokens. Systems with a statistic method “learn” such rules and either use a corpus for training or unsupervised data (see chapter 6.3.2) (cf. Hering, Gutekunst & Dyllong 2000: 275-276; Hagenbruch 2010: 267-271).

5.4.2 Stop word detection

Stop words are items that carry little or hardly any semantic content and/or are unimportant for the semantics of a document. Items that belong to stop words are, for example, prepositions such as *in*, *of* and articles such as *the*. There is no definite list of the types of items that are considered stop words. Nor can stop words be defined as belonging to certain word classes. For example, not only function words are stop words; it rather depends on the documents that are considered. To give an example, if each document in a collection contains the term *computer* several times because these documents all discuss topics in information technology, it makes sense to regard this term as a stop word. The user then would not search for *computer* because probably all documents in the collection contain this item. The case is different if a collection deals with different topics and if the term is then not included in all or most documents (cf. Henrich 2007a: 93-96; Meadow et al. 2007: 141-142; Stock 2007: 224-225).

Nevertheless, a system using stop word detection needs pre-defined stop word lists. These can either be established manually or automatically. If devised

[6] A “token” is “[a]n instance of a unit” (Matthews 2007: 409) and is distinguished from “type”. To give an example, the sentence *Your happiness is my happiness* contains one type of *happiness*, but two word tokens of *happiness*. A *token* therefore counts each single item (cf. McEnery & Wilson 2001: 82; Jurafsky & Martin 2009: 120; Gibbon 2010: 520; Herbst 2010: 96).

[7] Rothlauf (2011) defines *heuristics* as follows: “Heuristics are problem-specific and exploit known rules of thumb, tricks or simplifications to obtain a high-quality solution for the problem. Heuristics do not guarantee finding an optimal solution but are usually faster than approaches returning an optimum.” (ibid.: 83). And later: “Heuristics are usually designed for a particular problem and try to exploit problem-specific knowledge” (ibid.: 85).

manually, a list has to define all items that are treated as stop words. There are already several stop word lists that can be edited and adapted for specific uses. Examples are found at http://members.unine.ch/jacques.savoy/clef/index.html (date of last access: 04/01/2013) and http://www.textfixer.com/resources (date of last access: 04/01/2013). Automatic stop word lists consider the specific document collection and here extract the most frequent terms. For instance, terms that appear in more than 20 % of all documents can be categorised as stop words. Where the boundary of stop word versus non-stop word is drawn is treated differently. Henrich (2007a) maintains that terms occurring in more than 10 % of all documents in a collection as well as terms that are found in only one document can usefully be treated as stop words. Such a procedure in defining a lower boundary in addition to an upper boundary also excludes items with spelling mistakes. Moreover, terms with a low frequency are rarely used in queries and are therefore not that important (cf. Weber 2006: 1-13; Henrich 2007a: 94-96).

Stop word detection can now be used in such a way that the terms identified as stop words in the document collection are "deleted", i.e. not considered any further in the representation of the documents. By applying stop word detection, certain positive effects are achieved. Memory capacity is reduced in the area of 30 to 50 % and the matching process of query and document collection is faster and better, as a smaller amount of data is processed. Finally, stop word detection can also improve recall and precision. This can be explained by the fact that highly frequent terms, e.g. *the*, which are unimportant, are left out. As a consequence, stop words do not mask semantically much more expressive and less frequent terms anymore. Such less frequent terms are then rated higher, usually with the overall effect of being more representative of a document (cf. Henrich 2007a: 94; Stock 2007: 100, 222; Croft, Metzler & Strohman 2010: 20; Baeza-Yates & Ribeiro-Neto 2011: 226).

On the other hand, it is dangerous to eliminate and exclude stop words from the entire text retrieval representation and process. The quotation by Hamlet *To be, or not to be*, for example, would then not be found. And as *the* is most likely treated as a stop word, the abbreviation *THE* for "Times Higher Education", for example, is not found if the system does not additionally consider capitalisation. Finally, stop words are language dependent. If the language of a document is not taken into account, stop words that are actually not stop words in the respective language might be excluded, e.g. the German article *die* takes the form of the English verb *die*. Stop word elimination is not always unproblematic in a further case, namely with anaphora resolution. Some anaphoric items may be deleted as stop words and so would not be considered any further. Alternatives to stop word deletion are to mark stop words and consider them only if searched for explicitly, or to keep stop word lists small (cf. Henrich 2007a: 94; Stock 2007: 100, 222, 298-299; Croft, Metzler & Strohman 2010: 20; Baeza-Yates & Ribeiro-Neto 2011: 226; Gödert, Lepsky & Nagelschmidt 2012: 260).

To sum up, it should be pointed out once more that stop word lists are rather based on statistics about word frequencies than on linguistic foundations. The idea behind stop word detection is that not all words in a text are equally important for representing the content of a document (cf. Gödert, Lepsky & Nagelschmidt 2012: 257-258). Gödert, Lepsky & Nagelschmidt (2012), for instance, even argue that stop word lists are not necessary:

> Die Konsequenz für die Gestaltung von Retrievalsystemen ist eindeutig: es gibt keinen Grund für das Anlegen von Stoppwortlisten. Das ist interessanterweise den Suchmaschinenanbietern klar – *Google*, *Ask*, *Yahoo* kennen keine unerwünschten Indexterme [...]. (ibid.: 260)

However, the usefulness or hindrance of stop word detection also depends on the text retrieval system in action and in how far it considers linguistic aspects in the representation of documents.

5.4.3 Stemming and lemmatisation

Words with the same root but in different morphological forms are a further issue in natural language processing. Words occur with different inflectional and derivational forms. Herbst (2007) explains: "Derivation is one way of forming new words, whereas inflection distinguishes different grammatical forms of the same word" (ibid.: 86). Words taking a derivational affix that is a prefix, i.e. at the beginning of words, or suffix, i.e. at the end of words, usually change the word class: e.g. adding *-dom* to *free* (adjective) results in *freedom* (noun). Inflections in English are always at the end of words, i.e. suffixes, for example, the plural *-s*. Inflectional suffixes do not change the word class as they encode grammatical features, e.g. *player* (noun) and *players* (noun) (cf. Jackson & Moulinier 2002: 11).

The term used for the form that is left when removing inflectional and derivational suffixes is defined differently in the literature. It encompasses the terms "root", "base" and "stem". Yule (2010: 68), for instance, only uses the term "stem", which he defines as "the base form to which **affixes** are attached in the formation of words" (ibid.: 295). Furthermore, the process of removing any suffixes automatically for computational applications is termed "stemming" (cf. Jackson & Moulinier 2002: 11; Fromkin, Rodman & Hyams 2011: 397). Bauer (1983) points out: "'Root', 'stem' and 'base' are all terms used in the literature to designate that part of a word that remains when all affixes have been removed. Of more recent years, however, there has been some attempt to distinguish consistently between these three terms" (ibid.: 20).

The problem with words taking different derivational or inflectional affixes is as follows: if entering a term, such as *linguistics*, in a query, only documents with exactly that form are returned. Documents with terms that take other affixes, such as *linguist*, *linguists*, *linguist's*, *linguistic*, *linguistically*, are not returned, although they are most likely also relevant. In such situations stemming or lemmatisation can help. The difference between stemming and lemmatisation is as follows: stemming uses rule-based (heuristic) approaches, lemmatisation uses approaches based on lexicons (cf. Krüger-Thielmann & Paijmans 2004: 357; Henrich 2007a: 96-98). Sometimes, both approaches are also subsumed under the term "stemming" (see e.g. Jackson & Moulinier 2002: 11-12; Tzoukerman, Klavans & Strzalkowski 2004: 531).

As lexicon-based approaches are time-consuming and expensive in their building and maintenance and draw heavily on linguistic knowledge, stemmers are predominantly used. Stemmers operate without a lexicon; they only define rules that determine which affixes are removed from terms. Moreover, stemmers also have rules, but not necessarily linguistically correct ones. As a result, when using such stemmers it is accepted that these return incorrect roots, i.e. forms that are not words anymore. To give an example, if *writing* undergoes a stemming, the *-ing*-suffix is identified as an affix and the false root *writ* is returned instead of *write*. Lemmatisation always returns correct roots because word forms are reduced to a lemma (cf. Jackson & Moulinier 2002: 11-12; Stock 2007: 227, 228-232). Sinclair (1991) defines a "lemma"[8] as follows:

> A lemma is what we normally mean by a 'word'. Many words in English have several actual word-forms [...], for example, the verb *to give* has the forms *give*, *gives*, *given*, *gave*, *giving*, and *to give*. [...] So 'the word *give*' can mean either (i) the four letters **g, i, v, e,** or (ii) the six forms listed above. [...] [T]he composite set of word-forms is called the lemma. (ibid.: 173)

[8] A "lemma", as defined with Sinclair (1991), typically comprises word forms differing in inflections. Similarly, Bauer (2004) states that a lemma "is a term used particularly by lexicographers and corpus linguists to refer to a word in all its inflectional and spelling forms" (ibid.: 61). It is controversially discussed, however, if the term can also be applied to word forms taking derivational affixes (e.g. Knowles & Mohd Don 2004: 71-72; Fitschen & Gupta 2008: 552-554).

According to the distinction between derivation and inflection, there are derivational and inflectional morphological stemmers/lemmatisers. The most common are inflectional stemmers/lemmatisers. The cutting off of prefixes generally does not lead to a much improved text retrieval because the meaning of the form resulting from this process changes. Consequently, stemming or lemmatisation is usually restricted to suffixes. But even here improvements in quality might be small (cf. Stock 2007: 232-233; Croft, Metzler & Strohman 2010: 93-95). Although different studies in Tzoukerman, Klavans & Strzalkowski (2004: 532) have shown that stemming or lemmatisation improves the results in computational applications, the increase was only slight and lies between 1 and 3 % (see also Baeza-Yates & Ribeiro-Neto 2011: 226-227).

One well-known stemmer is the Porter stemmer. It is popular "because of its simplicity and elegance" and nevertheless "yields results comparable to those of the more sophisticated algorithms" (Baeza-Yates & Ribeiro-Neto 2011: 227). As a result, it is frequently cited and used. The Porter stemmer applies heuristics and considers inflectional and derivational suffixes (cf. Porter 1980: 132-137; Levene 2010: 95). Generally, approaches based on rules that are not too time-consuming and complex work well for languages that are less inflected, such as English. With languages such as German, which are strongly inflected, approaches that are based more on lexicons are used (cf. Jackson & Moulinier 2002: 12; Henrich 2007a: 107-108; Stock 2007: 234-235).

For employment in text retrieval, two approaches are distinguished. A morphological analysis can first be used to extend the query in text retrieval systems. This is an approach that adds all possible word forms to the form found in the query so that not only the form entered in the query is looked for but also relevant other forms. The second approach reduces both words of query and documents to their roots. In comparison, if taking into account advantages and disadvantages, the second approach outweighs the first approach as it is more easily accomplished and computationally less time-consuming (cf. Henrich 2007a: 98-99).

Finally, it is worth mentioning that stemming and lemmatisation does not always work without problems. Generally, two types of errors might occur with stemming: overstemming and understemming. Overstemming means that two words that are morphologically unrelated are reduced to one form e.g. *magnesia* and *magnetic* to *magnes*. Understemming is if two words should be reduced to one form but the stemmer retains two forms: e.g. *acquire* and *acquisition* are reduced to *acquir* and *acquis* (cf. Tzoukerman, Klavans & Strzalkowski 2004: 532; Gödert, Lepsky & Nagelschmidt 2012: 312-313). Such errors also occur with the Porter stemmer. Similarly, lemmatisation produces errors. Some word forms are not reduced to one lemma, especially adjectives, e.g. *young*/*younger*/*youngest*; other words are falsely reduced to one lemma, e.g. the verb *sit* and the noun *site* (cf. O'Keeffe, McCarthy & Carter 2007: 32-33; O'Keeffe & McCarthy 2010: 125). Jurafsky & Martin (2009) conclude: "Most modern Web search engines therefore need to use more sophisticated methods for stemming" (ibid.: 806).

5.4.4 Part-of-speech tagging

Part-of-speech (POS) taggers assign each token its tag – i.e. word class – for example, *computer* is a "noun". Aside from parts of speech, POS taggers often also return other types of information, e.g. inflectional information or whether nouns are common or proper nouns. POS taggers need tokenisers and sentence delimiters before they can be used. Most taggers include their own tokenisers and stemmers (cf. Mitkov 2002: 40; McEnery & Hardie 2012: 31).

POS tagging is again a task that cannot always be carried out unambiguously. There are tokens that are part of more than one word class, e.g. *play* belongs to the word class "verb" as well as "noun". Due to the nature of language, i.e. as new words enter a language regularly, a POS tagger has to deal with unknown words as well. Algorithms that consider unknown words can, for instance, pay attention to how a word is spelt and thereby use morphological information. To give an example, if a word ends in *-ed*, it is likely to be a verb (cf. Mikheev 2004: 212, 215; Ule & Hinrichs 2004: 219; Jurafsky & Martin 2009: 167, 169, 192).

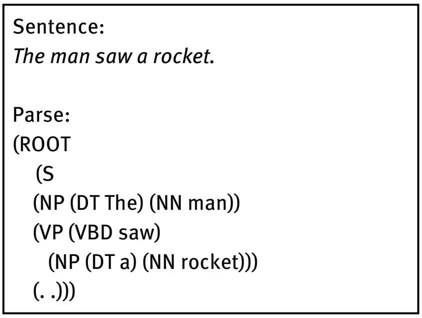
There are two approaches for POS taggers: rule-based and stochastic taggers. The first approach uses linguistic knowledge for analysing tokens into parts of speech. For instance, if *play* is preceded by an article, it is a noun. Stochastic taggers rely on information about probability to clarify ambiguous tokens.[9] These can be based on the frequency of certain word classes in a corpus or in a context. For example, *play* is more frequently used as a verb and therefore it is more likely to be a verb than a noun (cf. Jackson & Moulinier 2002: 12-13; Voutilainen 2004: 220; Siddiqui & Tiwary 2008: 77-87; Lemnitzer & Zinsmeis-

[9] Sachs (2006) explains the term "stochastics" as follows: "Die **Stochastik** beschreibt zufällige Vorgänge mit Hilfe mathematischer Modelle und entwickelt Verfahren, um daraus für die Praxis verwertbare Folgerungen zu ziehen. […] „stochastisch" bedeutet „dem Zufall ausgesetzt". Die Stochastik befasst sich mit den **mathematischen Gesetzmäßigkeiten des Zufalls**, mit zufälligen Ereignissen" (ibid.: 20).

ter 2010: 73). Various tagger algorithms can also be combined. Current POS taggers achieve an accuracy rate of about 97 %. If, however, the sentence accuracy is taken into consideration, i.e. how many complete sentences a tagger can tag correctly and not how many tokens are assigned with the right tag, this reduces accuracy to 55 to 57 % (cf. Damascelli & Martelli 2003: 21; Voutilainen 2004: 223; Jurafsky & Martin 2009: 157-158, 189, 197; Manning 2011: 171).

5.4.5 Parsing

Parsers analyse the parts of speech according to their syntactic relations. They determine the form and function of an item in a clause and also its function within a phrase. Fromkin, Rodman & Hyams (2011) explain: “a parser in a computer uses a grammar to assign a phrase structure to a string of words” (ibid.: 399). The syntactic relations in a sentence are then represented by a tree diagram or, more commonly, by using parentheses (cf. Jackson & Moulinier 2002: 15-17; Damascelli & Martelli 2003: 21-22). Figure 4 shows a representation that the Stanford parser returns:



Sentence:
The man saw a rocket.

```
Parse:
(ROOT
  (S
  (NP (DT The) (NN man))
  (VP (VBD saw)
    (NP (DT a) (NN rocket)))
  (. .)))
```

Fig. 4: Representation of syntactic relations

The sentence was entered in the online interface http://nlp.stanford.edu:8080/parser (date of last access: 05/02/2013). The sentence (S) is divided up into the noun phrase (NP) *the man* with a determiner (DT) and a noun as head (NN = singular or mass noun), the verb phrase (VP) *saw*, which is a verb in past tense (VBD) and another noun phrase. The end is marked by a full stop (.).

Parsing is carried out by using grammatical rules. However, all current grammars for parsers cannot account for all constructions of sentences. A parser also has to be able to manage unknown phrase structures. Besides, automatic parsing is more difficult than part-of-speech tagging. As a consequence, the results returned are less correct than with POS tagging, ranging around 85 to 87 % for both precision and recall. An exact number for average techniques is not available due to great differences among parser methods, among the evaluation methods and data used. Parsers often need information from sentence delimiters, tokenisers, stemmers and POS taggers. Many parsers already have their own tools for POS tagging, such as the Stanford parser (cf. Jackson & Moulinier 2002: 15-17; Damascelli & Martelli 2003: 21-22; Paroubek 2007: 108, 116-117; Tsujii 2011: 60).

Parsers vary in what depth they analyse texts. There are so-called "shallow parsers" or "partial parsers", which are sufficient for some applications. One type of partial parser divides a text into chunks, i.e. phrases. Each chunk is assigned a syntactic label, e.g. noun phrase, and the element within this chunk that is the head of the phrase is marked. For example, *the man* in the sentence above is identified as one chunk with *man* as head. Other shallow parsers only recognise noun phrases and proper names. Name recognisers – also called "named entity recognisers" – identify names of people or companies, for example. A POS tagging is often not sufficient because it would split up names such as *Barack Obama* into two entities and not identify them as first and last name. Some name recognisers also recognise categories of names, i.e. names of people, place, companies, for instance. Noun phrase parsers only carry out a partial syntactic analysis as they only extract noun phrases. Noun phrase and name recognisers can again work symbolically, i.e. rule-based, or statistically, i.e. considering the frequency of items (cf. Jackson & Moulinier 2002: 13-14; Carroll 2004: 233-234; Jurafsky & Martin 2009: 484-485, 761-768; Neumann 2010: 596-599).

In addition, there are other, so-called "deep parsers", also termed "complete parsers" or "full parsers", which are also able to categorise the function of constituents, e.g. subject, object. Some parsers can even identify semantic aspects in the constituents identified, for instance, semantic roles such as agent, patient or whether a noun is animate or inanimate (cf. Jackson & Moulinier 2002: 15-16).

5.4.6 Natural language processing on the Web

As has been shown, sentence delimitation and tokenisation, stop word elimination, stemming and lemmatisation, POS tagging and parsing are common approaches that are used when analysing texts computationally. Which approaches are used and to what extent depends, however, on the application and the accuracy needed. The tendency is to use only those approaches that are absolutely necessary, because all analysing steps require processing time.

As to natural language processing methods on the Web, Jackson & Moulinier (2002) state:

> The primary application of language processing on the Web is still *document retrieval*: [...] the finding of documents that are deemed to be relevant to a user's query. [...] One can perform document retrieval without doing significant NLP, and many search engines do, but the trend in the 1990s has been towards increasing sophistication in the indexing, identification and presentation of relevant texts [...]. (ibid.: 8)

Siddiqui & Tiwary (2008) maintain that "nowadays, information retrieval (IR) has emerged as one of the most important applications of NLP" (ibid.: iii). Looking closer at text retrieval on the World Wide Web, current search engines typically use tokenisers, but not all include POS taggers (cf. Jackson & Moulinier 2002: 9; Baeza-Yates 2004: 446-447). According to Baeza-Yates & Ribeiro-Neto (2011: 226), a lot of Web search engines do not use stemming algorithms. In the following chapter, the focus of text retrieval turns more to its use in the World Wide Web.

5.5 Text retrieval and the Internet

If searching for information on the Web, search engines are used for retrieving relevant documents. Search engines have to represent as many homepages as possible and store these in a suitable way for text retrieval in order to return good results. The various processes that lie behind such search engines are discussed in the following chapters.

5.5.1 Crawling

First of all, a search engine begins by storing as many documents from the Internet as possible, ideally all. This is done by crawling the Web. A crawler, also called "robot" or "spider", starts with a pool of websites and then continues with the links contained on those websites. The websites visited are then downloaded for further analysis. Crawling the Web is not an easy task due to the characteristics of the Web. The Web is a network of numerous servers that store a vast amount of homepages. As a result, the number of homepages that are stored on the servers is not known. So, the number of websites can only be estimated. Levene (2010: 10-13) assumes that about 600 billion pages existed in 2010.

In this context, the so-called “shallow Web”, which is accessible through links, has to be distinguished from the “deep Web”.[10] Examples of the deep Web are databases that can only be accessed through an interface and are usually invisible to crawlers. Furthermore, pages that are secured by passwords cannot be crawled, for instance. Additionally, administrators can block websites from crawlers (cf. Jackson & Moulinier 2002: 57; Henrich 2007a: 341-344, 350-355; Büttcher, Clarke & Cormack 2010: 511; Baeza-Yates & Ribeiro-Neto 2011: 520-521). Jackson & Moulinier (2002: 57-58) and Henrich (2007a: 391) believe that crawlers can access only one third of the whole Web.

If a website has been crawled, this does not mean that the process of crawling is finished with this website. As the content of homepages changes all the time, crawlers have to keep the collection of websites up to date and so crawl the Web constantly. Crawlers usually learn to know which homepages change frequently and which homepages hardly ever change. As a result, crawlers visit homepages whose content changes regularly such as online newspapers more often than homepages that hardly ever change. However, no search engine so far can return the utmost current state of the Web. Due to the vast amount of websites, crawlers need weeks to crawl the Web and to detect changes on websites. Thus, the index of search engines is always one of the past (cf. Henrich 2007a: 362; Levene 2010: 82-84).

5.5.2 Indexing

Crawlers have as output a large collection of websites that need to be processed further in order to be searchable. From these documents, the key terms are identified, i.e. feature extraction is carried out and then an index is built as a representation of the documents. Indexing can be carried out manually by human indexers, or automatically, as it is done with large databases such as on the Web and will be explained in more detail here (cf. Stock 2007: 542; Croft, Metzler & Strohman 2010: 404). Before such an index can be established, several processes have to take place.

[10] The estimate of the number of websites given above does not include the deep Web.

5.5.2.1 Processes of feature extraction

First, the structure is eliminated from each document. This means that text on the one hand and layout and navigation on the other hand have to be distinguished. Texts on homepages are usually encoded in HTML, which Rehm (2010: 160) calls the lingua franca of the World Wide Web (see also Ince 2012: World Wide Web). HTML-tags serve to lay down the type of fonts, type size, colour, title and other formatting options, i.e. how the document should be shown to the user. Metadata and comments can also be included. An index, however, should only include elements that users can see. An example of HTML-tags of a website is given in Figure 5. Other types of codes, for instance, XML websites, PDF documents and documents in Microsoft Office-format have also to be cleared from such structures and/or converted to text (cf. Damascelli & Martelli 2003: 51-52; Weber 2006: 1-6, 1-7; Henrich 2007a: 357-358, 361-362; Stock 2007: 100-101; Büttcher, Clarke & Cormack 2010: 9-13; Croft, Metzler & Strohman 2010: 18). Additionally, advertisements, banners and other elements on the page have to be removed from the main content. As Figure 1 in chapter 4 has shown, only the squared part is taken for analysis. The remainder, apart from this “content block”, is called “noise”, which is removed before proceeding (cf. Croft, Metzler & Strohman 2010: 63-68).

```
<html>
<head>
<title>Definitions</title>
<meta name="description"
content="This page is about some important definitions">
</head>
<body>

</body>
</html>
```

Fig. 5: Example of an HTML code

Second, tokens are identified by tokenisation. After that, stop word detection can be carried out. The details of this process have already been described (cf. chapter 5.4.2). Stop word detection saves space in indices (cf. Stock 2007: 222; Croft, Metzler & Strohman 2010: 19-20). The remaining text is then, in a third step, mapped to terms. The majority of search engines use words or phrases as terms. Another possibility is to use so-called “n-grams”. Here the text is split up into word fragments. For instance, if the sequence *computer* is divided into 3-grams, these are *com*, *omp*, *mpu*, *put*, *ute*, *ter*. When using n-grams, stemming and other language specific tools are not necessary anymore. As the splitting of a text into words is much more frequent, the following explanations focus on how the index is established here. The terms remaining from stop word detection enter an index. This is done as follows: text retrieval systems often represent terms together with their position in the document. Furthermore, the term frequency, i.e. how often one term occurs in a document, is used for ranking the documents. The more common a term is in a document, the more relevant it might be to the user's query containing that term (cf. Weber 2006: 1-2, 1-4, 1-15, 1-16; Meadow et al. 2007: 142-143; Jurafsky & Martin 2009: 117; Büttcher, Clarke & Cormack 2010: 92-93).

In the fourth step, stemming is carried out, as has been described in more detail above. Which terms then undergo stemming varies from search engine to search engine (cf. chapter 5.4.3). Google uses partial stemming, in which only the most common suffixes, such as plural forms, are removed (cf. Clark 2009: 168; Levene 2010: 96). Finally in step five, two terms apart from their differences in inflections and derivations can show semantic relationships, for example, if these are synonyms or hypernyms. Such relations could also be considered in the index. To give an example, if a document contains the term *cat*, the hypernym *animal* can also be assumed to occur in the document. One resource that is suitable for semantic relationships in the English language is WordNet. WordNet consists of databases for nouns, verbs, adjectives and adverbs. Other word classes, i.e. function words, are not included. WordNet represents the sense relations synonymy, hypernymy and hyponymy, meronymy and antonymy. As a result, items showing lexical cohesion can profit from WordNet (cf. Tengi 1998: 105-106; Endres-Niggemeyer 2004: 418; Stock 2007: 276-281; Siddiqui & Tiwary 2008: 372; Jurafsky & Martin 2009: 651-652).

However, not all search engines, e.g. Google, implement[11] step five. Step four is also not always found in search engine algorithms. Furthermore, the order of the five steps can vary from search engine to search engine and some steps can be combined or further divided up (cf. Weber 2006: 1-4, 1-17, 1-24, 1-25; Stock 2007: 271, 276-281; Levene 2010: 95).

[11] The verb *implement* means that a "theory or algorithm [...] has been formulated as a program that runs on a computer" (Mitkov 2004b: 743).

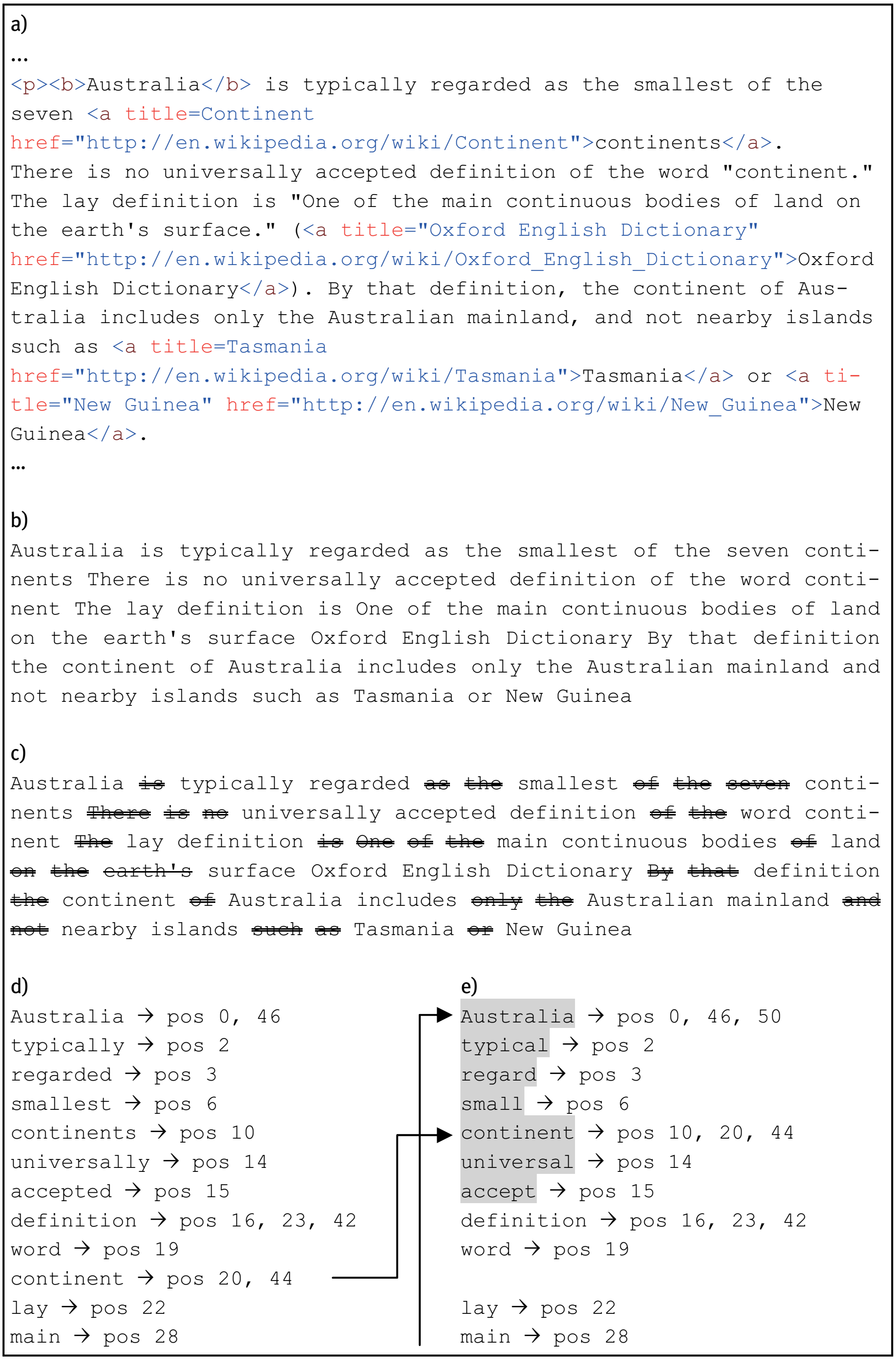
5.5.2.2 Example of a feature extraction

The process of how a document is analysed in terms of feature extraction is demonstrated below. The sequence has been taken from the text W2. Figure 6 part a) shows the sequence as it is embedded in the body of the HTML code, together with HTML-tags. It shows only a few sentences and their HTML encoding and so is only a part of the full HTML code of the document. For instance, the meta-tags as well as the tag for the end of the paragraph i.e. "</p>" are not included. As to the HTML code in part a), only the black printed parts constitute the text that a user sees on the website. This text is formatted by using tags, for example by using tags such as *<b>* and *</b>* for bold printed parts (cf. Born 2011: 40-41, 59-61, 139-145). The end for the formatting is usually marked, namely by the same tag including a slash. To give an example, the item *Australia* appears in bold print because the corresponding tag encloses this sequence. Additionally, a few links are given, for example, for the item *continents*.

Part b) then shows what is left if the structure has been deleted. Afterwards, stop word detection is carried out. Which items are left away as stop words is shown in part c). The decision about what to delete is based on the full corpus of hypertexts and the frequency of the items in this corpus. As a consequence, the items doubly crossed out are found 70 times or much more in the hypertext corpus. The stop words identified tend to be part of stop word lists and, in fact, are mentioned, for instance, in at least one of the lists provided by http://www.ranks.nl/resources/stopwords.html (date of last access: 13/12/2012). The only item singly crossed out occurs only once in the corpus. It is deleted due to the infrequency and because misspellings should so be filtered out, at least to some extent.

In the next step, as illustrated in part d), the terms that are left are listed in the index together with the positions where they occur in the document.[12] Words are represented as terms in the index below, rather than phrases. Consequently, the item *New Guinea* is split up into two terms if no named entity recognition is carried out (cf. Stock 2007: 248-250). As shown in part e), the terms undergo stemming afterwards. This has a consequence for the terms shaded in grey. Some terms can even be reduced to one form, as is the case for *continent* and *continents* and *Australia* and *Australian*. Finally, part f) illustrates the outcome if semantic relationships are considered. The synonyms in part f) are taken from "The Free Dictionary" (2012).

[12] The first position in a document is not 1, but 0. Moreover, terms are counted here. An alternative is to count characters, e.g. *Australia* is in position 1, *typically* in position 14, because the term begins with its initial character in that position (cf. Baeza-Yates & Ribeiro-Neto 2011: 341).



a)

```
…
<p><b>Australia</b> is typically regarded as the smallest of the
seven <a title=Continent
href="http://en.wikipedia.org/wiki/Continent">continents</a>.
There is no universally accepted definition of the word "continent."
The lay definition is "One of the main continuous bodies of land on
the earth's surface." (<a title="Oxford English Dictionary"
href="http://en.wikipedia.org/wiki/Oxford_English_Dictionary">Oxford
English Dictionary</a>). By that definition, the continent of Aus-
tralia includes only the Australian mainland, and not nearby islands
such as <a title=Tasmania
href="http://en.wikipedia.org/wiki/Tasmania">Tasmania</a> or <a ti-
tle="New Guinea" href="http://en.wikipedia.org/wiki/New_Guinea">New
Guinea</a>.
…
```

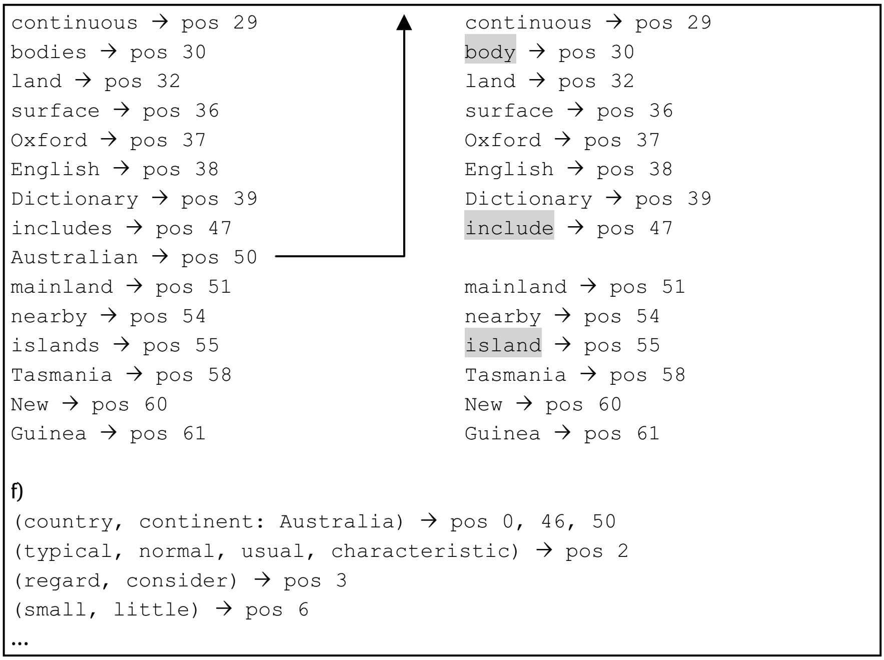
b)

Australia is typically regarded as the smallest of the seven continents There is no universally accepted definition of the word continent The lay definition is One of the main continuous bodies of land on the earth's surface Oxford English Dictionary By that definition the continent of Australia includes only the Australian mainland and not nearby islands such as Tasmania or New Guinea

c)

Australia ~~is~~ typically regarded ~~as~~ ~~the~~ smallest ~~of~~ ~~the~~ ~~seven~~ continents ~~There~~ ~~is~~ ~~no~~ universally accepted definition ~~of~~ ~~the~~ word continent ~~The~~ lay definition ~~is~~ ~~One~~ ~~of~~ ~~the~~ main continuous bodies ~~of~~ land ~~on~~ ~~the~~ ~~earth's~~ surface Oxford English Dictionary ~~By~~ ~~that~~ definition ~~the~~ continent ~~of~~ Australia includes ~~only~~ ~~the~~ Australian mainland ~~and~~ ~~not~~ nearby islands ~~such~~ ~~as~~ Tasmania ~~or~~ New Guinea

d)	e)
Australia → pos 0, 46	Australia → pos 0, 46, 50
typically → pos 2	typical → pos 2
regarded → pos 3	regard → pos 3
smallest → pos 6	small → pos 6
continents → pos 10	continent → pos 10, 20, 44
universally → pos 14	universal → pos 14
accepted → pos 15	accept → pos 15
definition → pos 16, 23, 42	definition → pos 16, 23, 42
word → pos 19	word → pos 19
continent → pos 20, 44	
lay → pos 22	lay → pos 22
main → pos 28	main → pos 28



```
continuous → pos 29              continuous → pos 29
bodies → pos 30                  body → pos 30
land → pos 32                    land → pos 32
surface → pos 36                 surface → pos 36
Oxford → pos 37                  Oxford → pos 37
English → pos 38                 English → pos 38
Dictionary → pos 39              Dictionary → pos 39
includes → pos 47                include → pos 47
Australian → pos 50
mainland → pos 51                mainland → pos 51
nearby → pos 54                  nearby → pos 54
islands → pos 55                 island → pos 55
Tasmania → pos 58                Tasmania → pos 58
New → pos 60                     New → pos 60
Guinea → pos 61                  Guinea → pos 61

f)
(country, continent: Australia) → pos 0, 46, 50
(typical, normal, usual, characteristic) → pos 2
(regard, consider) → pos 3
(small, little) → pos 6
...
```

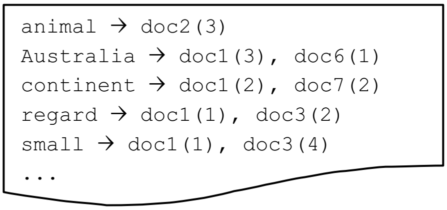
Fig. 6: Example of indexing (adapted from Weber 2006: 1-5)

5.5.2.3 Index

As the relevant features are now identified for each document, a file for the whole document collection can be established so that the documents that are relevant for a query are found quickly. One common method for text retrieval is to store either inverted indexes, also called "inverted files", or full inverted indexes.[13] To start with inverted indexes, each term that is present in one of the documents is stored in a file. In addition to that, information in which documents this very term occurs is stored. That way, access is more efficient. The lists are termed "inverted lists" because the terms are the starting point and not the documents. This is because a search does not begin by looking at each document, but by looking at one or more terms (cf. Krüger-Thielmann & Paijmans 2004: 358; Henrich 2007a: 138-139, 146-147; Henrich 2007b: 11). In addition, the term frequency is stored together with the document number in in-

[13] A different approach is to store so-called "signatures" instead of terms. *Signatures* are sequences of numbers that can represent one word or even a whole paragraph (cf. Henrich 2007a: 147-148; Baeza-Yates & Ribeiro-Neto 2011: 357-359).

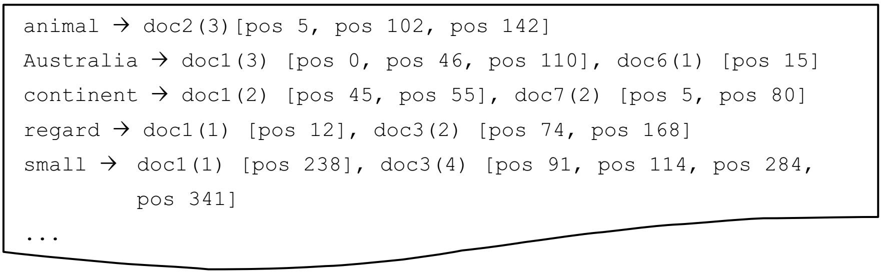
verted indexes. Alternatively, the weights for the term in the specific document can be stored instead of the term frequency. Terms in an inverted list are then represented as Figure 7 shows. These terms are here ordered alphabetically, although other criteria can be applied as well (cf. Stock 2007: 132-134; Baeza-Yates & Ribeiro-Neto 2011: 341).



```
animal → doc2(3)
Australia → doc1(3), doc6(1)
continent → doc1(2), doc7(2)
regard → doc1(1), doc3(2)
small → doc1(1), doc3(4)
...
```

Fig. 7: Example of an inverted index

Full inverted indexes, however, additionally contain further information about the position in the relevant documents. This is necessary if the distance or proximity between two terms needs to be measured. Figure 8 shows an example of a full inverted index. Here, not only the terms, together with the documents and term frequency, are stored, but also the word positions within the documents (cf. Weber 2006: 0-28, 1-2, 1-66; Baeza-Yates & Ribeiro-Neto 2011: 340-342).



```
animal → doc2(3)[pos 5, pos 102, pos 142]
Australia → doc1(3) [pos 0, pos 46, pos 110], doc6(1) [pos 15]
continent → doc1(2) [pos 45, pos 55], doc7(2) [pos 5, pos 80]
regard → doc1(1) [pos 12], doc3(2) [pos 74, pos 168]
small →  doc1(1) [pos 238], doc3(4) [pos 91, pos 114, pos 284,
         pos 341]
...
```

Fig. 8: Example of a full inverted index

Current numbers of how many documents search engines include in their index are not available. Google, for example, published the last number of the documents they had indexed in 2005. At that time, Google said that it had 8 billion documents in its index (cf. Weber 2006: 0-9). Estimates, however, are gained by searching for items that are likely to occur in nearly all documents. If searching for *a*, *1* or *www* with Google (http://www.google.com, date of research: 20/01/2012), 25.27 billion results were returned in each case.

With regard to the use of NLP tools for establishing an automatic index, some tools can be used to improve the index. One tool is part-of-speech tagging, which can be applied to identify items as function words, i.e. belonging to word classes such as pronouns, conjunctions or auxiliary verbs and in consequence leave them out of the index. Such a method differs from stop word detection, as not the most frequent terms irrespective of their word class are ignored, but items that are part of specific word classes (cf. Voutilainen 2004: 221).

The model of text retrieval systems as outlined in Figure 1 now has to be expanded by tools characteristic of the Web: crawlers are necessary to find new websites and updated websites found previously. A further tool – the so-called “indexer” – is required, which extracts the terms of these websites. These terms are then stored in the index. The index corresponds to the document representation; the matching is done by the so-called “query engine” and is explained in more detail in the next chapter. The query entering and the displaying of the results is carried out by the search interface (cf. Henrich 2007a: 342-344; Levene 2010: 78-81).

5.5.3 Queries and search results

Having an index allows search engines to “answer” questions posed by the user through terms he or she enters as a query in the search interface. To do that, the query is processed similarly as the documents, as far as this is necessary. This means that terms need to be identified if the query consists of more than one item. Stop word detection and stemming with queries is also necessary if the terms in the index have undergone such techniques. If stemming has not occurred in the index, the query has to be expanded by variants in inflection, e.g. if *cat* is entered in the query, *cats*, *cat’s* are relevant as well (cf. Croft, Metzler & Strohman 2010: 194-195). Afterwards, a matching of the query on the one hand and the document collection represented in an index on the other hand can be carried out. Levene (2010) maintains in this context: “The *query engine* is the algorithmic heart of the search engine. The inner working of a commercial query engine is a well-guarded secret” (ibid.: 80). The results of this matching process are then displayed to the user via the search interface (cf. Henrich 2007a: 342-343; Levene 2010: 80-81). The search for Web documents and the processes lying behind this, as described here, are illustrated in Figure 9.

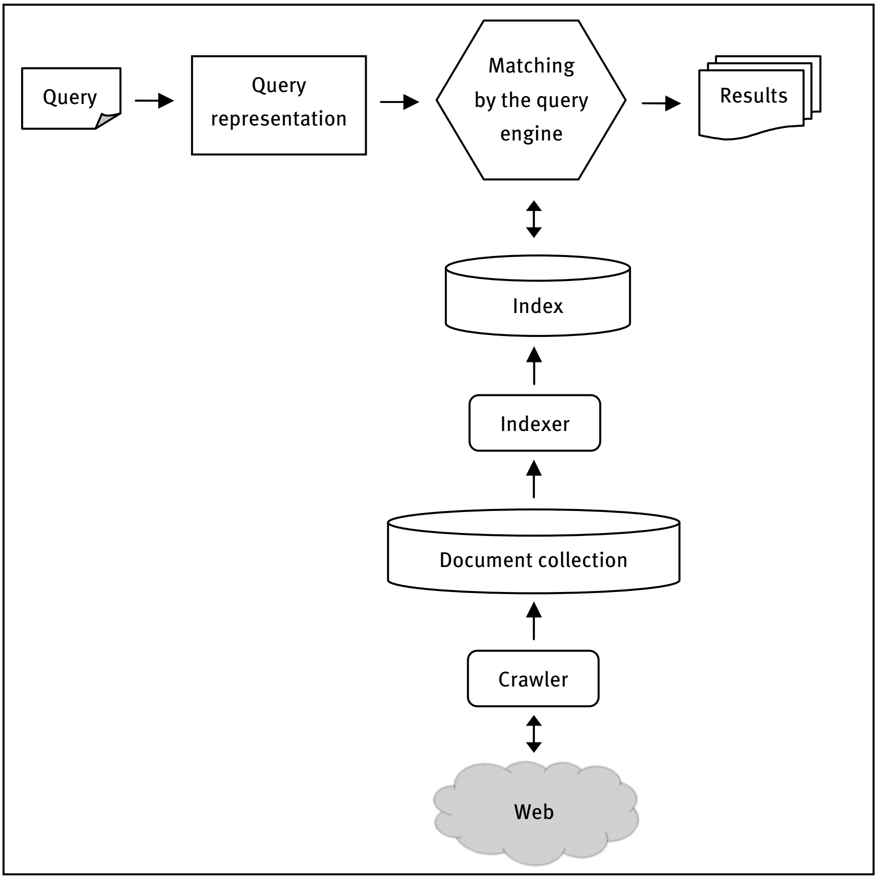


Fig. 9: Procedure of Web retrieval systems (adapted from Jurafsky & Martin 2009: 802)

What now needs further discussion is how the query engine is realised. To begin with, the text retrieval models used with search engines for retrieving relevant documents vary. Google as well as Yahoo use Boolean text retrieval; AltaVista the vector space model. Boolean text retrieval and vector space models are most common with search engines (cf. Weber 2006: 1-91; Henrich 2007a: 367).

Additionally, it is important to rank the documents on the Web because a vast amount of documents is often returned and the user then can only examine and read a small fraction. Fuhr (2011: 4) mentions that 90 % of all users only consider the first ten documents returned. In this context it should also be mentioned that precision is usually more important in text retrieval than recall (cf. Henrich 2007a: 380).

One basic approach to rank the documents according to their relevance is to use methods that are based on the individual query. To do that, a score, also called “retrieval status value” (RSV) is computed when a query is posed. This means that the document collection is examined and the degree of closeness between query and each document is calculated. Documents can then be ranked on this measure by starting at the document with the highest RSV and going down to that document with the lowest RSV (cf. Weber 2006: 0-33; Meadow et al. 2007: 241-242; Zhai 2009: 11; Büttcher, Clarke & Cormack 2010: 7).

Standard means to rank relevant documents are the term frequency (TF) and the inverse document frequency (IDF). The TF measures how frequent one term is in one document; it should also consider the length of the documents so that the TFs of different documents can be compared reliably. Here, the measures TF and WDF are distinguished. TF is the absolute frequency of a term in one document. As a result, the TF is higher in longer documents than in shorter ones, leading to distortions because the length of documents is not taken into account. The WDF, in contrast, considers the length of documents by dividing the TF by the number of terms occurring in one document. The IDF counts how many documents contain one term and then the number of documents in the collection is divided by this number. The IDF is therefore higher for less frequent terms. For ranking documents, the TF is multiplied with the IDF and documents are then ranked according to this value (cf. Manning, Raghavan & Schütze 2008: 108-109; Levene 2010: 96-102; Gödert, Lepsky & Nagelschmidt 2012: 291-292).

A further technique for ranking the documents returned is the so-called “PageRank” algorithm, which Google uses, for example, together with the TF×IDF measure (cf. Weber 2006: 1-91; Henrich 2007a: 367; Levene 2010: 110). It is based on the relationship of websites to each other. Websites usually contain hyperlinks. Hyperlinks occur in two types: forward links and back links. Forward links are links that are given on the website itself and link to different further websites. Back links are links that link to the very website considered and are included in different other websites. The idea is that the number of back links gives evidence of the quality of a website. The more back links a website has the more likely it is to be of a high quality. PageRank then ranks those relevant documents highest that have the most back links (cf. Henrich 2007a: 381-382; Meadow et al. 2007: 160-161; Stock 2007: 382-385). This is, however, only one method to rank documents. Stock (2007) states:

> Es wäre ein großer Irrtum anzunehmen, dass die Gewichtungswerte nach PageRank etwas mit der Qualität einer Webseite zu tun haben [...]. PageRank gibt der Stellung einer Seite im Web einen quantitativen Ausdruck – und nichts mehr. (ibid.: 385)

Other query-insensitive methods for ranking documents are, for instance, to measure the popularity of a website. To give an example, the more often users visit a website, the more important it might be. Furthermore, the state of the information regarding its recentness can be considered, i.e. the more up-to-date a website is, the more relevant it can be (cf. Stock 2007: 371; Levene 2010: 104-105, 130-132).

Yet other approaches are to use text positions and structural characteristics of terms within a document, which can be exploited for deciding the status in the ranked list. To give an example, if a term occurs in the title or is marked in a specific way, e.g. in bold print, it is likely to be more important in the whole document. Moreover, documents that are in the same language as the query can be preferred and also documents that are geographically closer to the query poser, e.g. if the user comes from Great Britain, English documents from Australia might not be so relevant. Additionally, if more than one term is used for the query, the order of these terms might reveal the relevance, i.e. the first term could be more important. If queries contain more than one term, the documents can also be ranked on how close these terms are to each other in a document, i.e. according to the distance of the terms. As a result, the closer the terms appear to each other, the more relevant the document is (cf. Stock 2007: 323, 370-371).

A method that is common with Web search engines is to analyse the query terms for misspellings. Levene (2010: 105-106) mentions that about 10 to 15 % of all queries contain misspellings. As these are not part of the inverted file, the system can either correct the misspelling or give suggestions back to the user. The type of corrections and suggestions that are selected can depend on how close a misspelling is to a term in the inverted file; additionally, correct terms tend to be more frequent in the inverted file than a misspelling (cf. Baeza-Yates & Ribeiro-Neto 2011: 32).

5.6 Conclusion

In this chapter, text retrieval has been examined. The focus was on text retrieval on the Web, due to the relevance for the discussion of anaphora resolution. As text retrieval systems return documents, which are written in natural language, analysing the document collection with natural language processing tools can be useful. NLP tools encompass sentence delimitation, tokenisation, stop word detection, stemming, part-of-speech tagging and partial or full parsing. Web retrieval systems, however, vary to what degree they include such tools. Usually, tokenisation is carried out, but hardly ever POS-tagging or full parsing.

In order to return those documents that are relevant to the query of a user, a number of processes take place. To begin with, the entered query is transformed into a representative form. This query representation is then compared to the index of the document collection. Matching documents are finally returned to the user. Building such an index from the document collection needs further steps. First a crawler combs the Web. The indexer then transforms the websites found and any new websites found afterwards into an index, usually a (full) inverted index, so that the matching can be done quickly (cf. Croft, Metzler & Strohman 2010: 23). Establishing an index requires the elimination of the structure of documents and term mapping. Often, stop word detection is carried out before term mapping. Afterwards, stemming and the identification of semantic relationships might be performed. The resulting index can then be used for the matching process. How the matching process works is determined by the models used. Common text retrieval models on the Web are the vector space and the Boolean text retrieval models. On the other hand, matching by the query engine does not only compare the query with an index, but also ranks the relevant documents. Well known ranking methods are the TF×IDF weights and PageRank. Finally, in order to evaluate text retrieval systems, tests with test collections such as TREC, or specific measures can be used. Standard measures for evaluating the effectiveness of text retrieval systems are precision and recall.

After this outline of basic processes in text retrieval systems on the Internet, we can now focus on anaphora resolution, which is the topic of the next chapter.

6 Approaches to and uses of anaphora resolution

This chapter will first examine those fields of natural language processing in which anaphora resolution is predominantly used. Additionally, the potential of anaphora resolution in text retrieval will be discussed. It will be seen that anaphora resolution plays a central role in systems that process natural language. Afterwards, the general structure and functionality of anaphora resolution systems will be outlined. In a further step, current anaphora resolution systems in general and their use in text retrieval will be detailed. Finally, approaches to evaluate anaphora resolution systems as well as the performance of current anaphora resolution systems are presented.

6.1 Uses of anaphora resolution systems

Anaphora resolution is important whenever text understanding is required or desired. Consequently, anaphora resolution is mostly used in the following fields of natural language processing: machine translation, information extraction, question answering and text summarisation (cf. Mitkov 2002: 123-125; Baeza-Yates 2004: 447-448; Harabagiu & Moldovan 2004: 562; Mitkov 2004a: 275-277). These fields are interrelated: Neumann (2010) subsumes information extraction, question answering, text summarisation and text retrieval under the term "text-based information management". Furthermore, he maintains:

> [F]ür die zukünftige Forschung gilt, dass IR [i.e. information retrieval], IE [i.e. information extraction], QA [i.e. question answering] und TZ [i.e. text summarisation] sich immer stärker aufeinander zu bewegen und verschmelzen werden." (ibid.: 614)

All of these fields also consider coreference resolution, in which coreferential relationships of pronouns are the most important. Apart from that, anaphora resolution is of interest in dialogue systems that process natural language. Here, anaphors have to be resolved when a machine interacts with the user, e.g. in speech dialogue systems[1] (cf. Becker 2010: 629-630; Neumann 2010: 576-577; Strube 2010: 399). In the following subchapters, the fields of machine translation, information extraction, question answering, text summarisation and text

[1] As the book focuses on written forms of language, anaphora resolution in dialogue systems is not further examined.

DOI 10.1515/9783110416756

retrieval as well as their applications of anaphora resolution are outlined in more detail.

6.1.1 Machine translation

Machine translation means “the use of computers to automate translation from one language to another” (Jurafsky & Martin 2009: 895). Automatic machine translation is difficult because languages differ from each other on various levels. For instance, some languages such as German are synthetic, i.e. they mainly use inflections to express the grammatical relation of words; other languages such as English are analytic, i.e. they basically rely on word order to express the relation of words in sentences (cf. Baugh & Cable 2002: 56). Another example concerns elements that can or are typically omitted in a language. In English and German, pronouns in subject position cannot usually be left out in simple sentences, whereas this is common in Spanish (see example (294)).

(294) *Spanish:* ___ Tiene un coche.
English: She has a car.

Furthermore, the lexicon differs across languages. One word does not always have a counterpart in another language. This becomes especially apparent where cultural differences are concerned, such as cooking traditions, e.g. German *Schnitzel*. Even if words have counterparts, this does not mean that the translated word has the same range of meaning. Words might be more specific or general in one language, e.g. for English *wall* the German language offers *Mauer* and *Wand* among others, depending on the context (cf. Hutchins 2004: 505). Furthermore, if a word in one language is polysemous or homonymous (cf. chapter 2.2.4), this does not have to apply for the word in the target language. In such situations, a word sense disambiguation has to be carried out in order to find the exact translation for the specific context (cf. Siddiqui & Tiwary 2008: 229; Jurafsky & Martin 2009: 895-902).

In some cases, a very rough translation is already sufficient, e.g. if people want to acquire information on the Web and only need to understand the gist of a text. So here automatic machine translation is already helpful. One example of an online translation service is Google Translate (http://translate.google.com, date of last access: 12/01/2013). Furthermore, a text that has undergone machine translation can be corrected by humans in order to arrive at a better translation. Here, machine translation helps translators because the text then only has to be post-edited. This is also referred to as “computer-aided human translation” (cf. Jurafsky & Martin 2009: 897). Finally, it should also be mentioned that machine learning already achieves a high quality in specific domains, e.g. in weather forecasts, software manuals, recipes. These domains use limited vocabulary and structures, which is why automatic machine translation is feasible (cf. ibid.: 897-898; Way 2010: 558).

Moreover, three basic or classical translation architectures are distinguished: direct translation, transfer approaches and interlingua approaches. In direct translation, a text in one language is translated word-by-word to another language by using bilingual dictionaries. Transfer approaches analyse the text by parsing it (cf. chapter 5.4.5), then transfer this structure to another language by using rules and finally generate sentences in the target language. Interlingua approaches also analyse a text, but create an abstract representation of its meaning that is language-independent, e.g. by using semantic roles (cf. chapter 5.4.5). From that representation, the target language text is generated. Transfer and interlingua approaches are rule-based, i.e. rules and dictionaries are exploited for translation. Other approaches are data-based and, for example, use statistical machine translation. These approaches “[build] probabilistic models of faithfulness and fluency and then combin[e] these models to choose the most probable translation” (Jurafsky & Martin 2009: 911). For these approaches to work, they first need to be trained on parallel corpora, which consist of texts in two languages (cf. ibid.: 897, 903-906, 921; Jekat & Volk 2010: 644-646).

Comparing rule-based and data-based approaches, the rules are deduced from linguistic knowledge with rule-based approaches; data-based approaches are inductive because knowledge is inferred. In terms of performance, data-based approaches cannot reach rule-based ones up to now (cf. Jekat & Volk 2010: 651; Way 2010: 554-555). That is probably why most commercial systems use rule-based machine translation. Research from the last few years focused on data-based approaches, but a recent trend for machine translation is to consider rule-based approaches again. As Way (2010) remarks: “it is widely agreed that more linguistic knowledge can indeed play a role in improving today’s statistical systems, in all phases of the process” (ibid.: 568).

As to anaphora resolution in machine translation, it is necessary because when a text is translated from one language to another, a different type of anaphor item might be needed. Especially pronouns are prone to be used differently in other languages. *It*, for example, translates into German as *es*, *er* or *sie*, depending on the context. If *it* is used as an anaphor referring to “the moon” in English, the corresponding German anaphor is *er*, but in the case of “the sun” it is *sie*. Similarly, if pronouns are omitted in one language, this might not be possible in another language (see example (294)). In this case the correct pronoun has to be inferred (cf. Somers 2004: 518-520; see also Eberle 2003: 216-217). Whereas machine translation is concerned with two or more languages, the rest of the four fields only consider one language at the same time.

6.1.2 Information extraction

Information extraction is "the automatic identification of selected types of entities, relations, or events in free text" (Grishman 2004: 545). This means that not a whole text is analysed, but only selected types of information. Information-extraction systems are designed for specific domains, which is why domain features, e.g. how information is structured and presented in a text, can be used for the development of information-extraction systems. Information extraction is especially used in the news domain. First systems were rule-based, current approaches are predominantly corpus-based (cf. Jurafsky & Martin 2009: 759; Grishman 2010: 517-518). Information extraction, for instance, is used by specialised search engines, e.g. ZoomInfo (http://www.zoominfo.com, date of last access: 12/01/2013), which only searches for people and companies (cf. Neumann 2010: 598).

Information-extraction systems cover different tasks. One of the first uses was name extraction, which means that names are identified in a text and then classified according to the type of name, e.g. name of a person, name of an organisation (cf. Jurafsky & Martin 2009: 759; Grishman 2010: 517). Identifying proper names is achieved by named entity recognition (cf. chapter 5.4.5). Another task in information extraction is entity extraction. Here, all expressions that refer to one and the same entity are identified. These expressions can be proper names, noun phrases or pronouns. Anaphora resolution can help in entity extraction as it, for instance, can clarify which pronouns refer to one and the same entity (cf. Siddiqui & Tiwary 2008: 337-338, 342; Grishman 2010: 522-523). A further information extraction task is relation extraction. The goal here is to identify two entities that are related. For instance, in *The American president, Barack Obama, announced…* the system has to find out that the expressions *the American president* and *Barack Obama* are related. With relation extraction, information about coreferential entities and therefore, for instance, the resolution of anaphoric pronouns is vital (cf. Siddiqui & Tiwary 2008: 342; Grishman 2010: 517, 523-524).

Finally, information extraction can be used for event extraction. This means that events, their type and their details, e.g. date, time, location, are identified.

For that purpose, a template can be specified, which has then to be filled with arguments. One example is to define a template that extracts all company names of a text (see Figure 1). The template then does not only define that company names are extracted, but for each company, the website of the company and its telephone number have to be filled in (cf. Jurafsky & Martin 2009: 786; Grishman 2010: 517, 526-527).

Company name: Clarks Limited

Website: http://www.clarks.com
Phone number: 0049-0851/420-01

Fig. 1: Simple example of a template

In contrast to other natural language processing methods, it is worth mentioning that with information extraction, no terms are entered in a query as is the case with text retrieval, but a template is designed with event extraction. This template is then filled by searching the text. Only those parts of a text that are needed in the template are returned, all other parts of the text are discarded. Information extraction has to be contrasted to question answering (see chapter 6.1.3) because no "answers" are returned in information extraction, but only the pre-defined slots of information are filled. These facts are delivered in a fixed format and the task is therefore more simple than, for instance, question answering. Neither do information extraction systems return a summary of a text, as is the case with text summarisation (see chapter 6.1.4) (cf. Siddiqui & Tiwary 2008: 337).

6.1.3 Question answering

A question-answering system "attempts to find the precise answer or at least the precise portion of text in which the answer appears" (Siddiqui & Tiwary 2008: 358). This means that a user asks a question through a query and the system returns a short answer in the form of a passage rather than numerous documents. Users might often prefer a short answer to whole documents. The system so has to "understand" the text in order to extract the answer from the documents. Therefore, question-answering systems need more natural language processing methods and semantic knowledge than, for instance, text retrieval systems. One example of a question-answering system on the Web is Ask (http://www.ask.com, date of last access: 13/01/2013) (cf. Siddiqui & Tiwary 2008: 358; Jurafsky & Martin 2009: 799).

The simplest questions for systems are so-called "factoid questions". The answers for such questions are words or short phrases containing simple facts that a system can identify rather easily in texts. Factoid questions are introduced by *who*, *what*, *where* and probably *when*. For instance, the question *What is the capital of Austria?* returns the answer *Vienna*. The answer so delivers the proper name of a capital. Other answers might return proper names of countries, sights or people. Other, more complex questions are introduced by *why* and *how*, which involve longer answers. Such complex questions have not been researched much (cf. Neumann 2010: 800-801; Webber & Webb 2010: 634, 644).

Factoid question answering can be split up into three stages: question processing, passage retrieval or document processing and answer processing. In the first stage, the question is analysed and the terms that are entered in a query are identified. In the easiest cases, such as in the question above, the question already contains relevant terms (in our case: *capital*, *Austria*), in more difficult cases, a reformulation has to take place. In this stage, the type of answer that is expected is already determined, e.g. a name of a place. In the passage retrieval stage, the documents returned by a text retrieval system are analysed and relevant passages identified. Finally, in the answer processing stage, an answer is generated, which is then returned to the user. This answer has been extracted from the document passages (cf. Siddiqui & Tiwary 2008: 359-364; Jurafsky & Martin 2009: 813-820).

A problem for question answering is that there might be more than one correct answer. For instance, the answer to the question *Where is the University of Passau?* might be more specific or general, depending on the user's home town or country. A user from Passau might want to know the exact street, a user from Northern Germany might want to know that it is situated in Bavaria, a user from North America might probably only want to know that Passau is in Germany. In order to account for that, the location of the user has to be considered. Some questions might also require further information about the user, e.g. the user's age. Other strategies for delivering better answers use interaction, which is called "interactive question answering". For instance, interaction can help if questions are ambiguous because the system can then present the user a list from which he or she can select the intended question (cf. Webber & Webb 2010: 647).

It should be clear from this short outline that question answering also involves processes of text retrieval. The difference between question answering and text retrieval, however, is that with question answering links to entire documents are not returned, but instead a word, a short phrase or passage. In the case of cross-lingual question answering, machine translation is also required. Question answering differs from information extraction in that there is no template, i.e. the precise type of the fact that is being returned is not known beforehand (cf. Harabagiu & Moldovan 2004: 562; Siddiqui & Tiwary 2008: 337, 358; Neumann 2010: 606; Webber & Webb 2010: 630).

Anaphora resolution, or rather coreference resolution, can help in question-answering systems because it can "[establish] coreference links between entities or events in the query and those in the documents" (Mitkov 2004a: 276). This means that coreference resolution is used to clarify to what extent items in the query and in a sentence of a document denote the same entity. The different sentences of the documents are finally ranked and those ranked highest are returned to the user (cf. ibid.: 276-277).

6.1.4 Text summarisation

Text summarisation is "the process of distilling the most important information from a text to produce an abridged version for a particular task and user"[2] (Jurafsky & Martin 2009: 821). This reduced version should then reflect the content of a text. The user should thus be able to assess if a document is relevant to him or her.

With summaries, different types can be distinguished. First, summaries vary as to whether they are indicative or informative. Indicative summaries detail what topics a text addresses, whereas informative summaries give an overview of the content of a text. Second, generic and user-oriented summaries are distinguished. User-oriented summaries provide summaries for the information need of a particular user, whereas the user in generic summaries is not considered. User-oriented summaries are used with search engines such as Google, which return a so-called "snippet" for each link. A snippet is a user-oriented summary of a retrieved text, where the terms of the query are highlighted (cf. Jurafsky & Martin 2009: 822, 836; Neumann 2010: 608-612).[3] Third, a

[2] Original in italics.

[3] Text summarisation can also be used for question answering, especially for complex tasks. In such cases, the summary has to be user-oriented. Furthermore, answers then are not so short as with traditional question answering, but longer units are returned (cf. Jurafsky & Martin 2009: 836).

summary can be created from a single document or from multiple documents. On the Web, multi-document summarisation is especially important due to the vast amount of texts. Multi-document summarisation might also need to use information extraction methods in order to be able to extract specific types of information from different texts (cf. Jurafsky & Martin 2009: 837-838). Finally, summaries can be extracts or abstracts. Generating extracts is the simpler task because here a summary is produced by using phrases and sentences from a text. In abstracts, the content of a text is analysed and the summary is then generated by using other words than found in the text. Current text summarisation systems mostly produce extracts and here mainly extract sentences (cf. Siddiqui & Tiwary 2008: 347-349; Jurafsky & Martin 2009: 822, 830, 842).

Approaches to text summarisation are knowledge-poor or knowledge-rich. Knowledge-poor approaches involve only the syntactic level, whereas knowledge-rich approaches additionally use the semantic and discourse level for a text analysis. The first type usually produces extracts; the latter approach commonly generates abstracts (cf. Siddiqui & Tiwary 2008: 349).

Anaphora resolution can be useful in text summarisation because extracted sentences might contain anaphors. If the antecedent of such sentences is not part of the summary, this is a case of a “dangling anaphor” (cf. Siddiqui & Tiwary 2008: 351). Dangling anaphors can lead to misunderstandings or at least to incorrect language use. Consequently, anaphors need to be taken into account in text summarisation (cf. Neumann 2010: 611-613).

6.1.5 Text retrieval

Apart from the fields mentioned, anaphora resolution is also vital in a further field of application: text retrieval (cf. chapter 5). The use of anaphora resolution and text retrieval is, however, only rarely discussed, let alone investigated in more depth. Meadow et al. (2007) state that “IR [i.e. information retrieval] systems permitting natural-language queries tend to ignore anaphora” (ibid.: 96). Stock (2007: 295-299) argues that anaphora resolution in text retrieval is important due to two reasons. First, not resolving anaphors makes proximity searching more difficult and second, term frequency is distorted without anaphora resolution. These two arguments are now discussed in more detail.

With proximity searching, two or more terms have to occur within a specific distance. Two realisations are common: one solution is that the system counts how many words or characters occur between the two terms and defines, for instance, that only up to ten words are allowed in between. If searching for

Passau and *university* and also counting stop words, the text containing *The university that is located in Passau is beautiful* is appropriate for the search, the passage *The university has about 10,000 students. It is a rather small institution, embedded in the beautiful city Passau*, however, is not considered, although it is relevant. Another approach exploits the structure of a text. Here, the terms have to appear, for example, within one sentence or one paragraph. If setting the distance to one sentence, the first example would be retrieved, the second not because *university* and *Passau* occur in two sentences. It is also possible to define the order of the terms in proximity searching, e.g. *university* has to precede *Passau* (cf. Meadow et al. 2007: 215-217; Stock 2007: 147-150; Kowalski 2011: 16). For proximity searching to work, the position of all the terms (cf. chapter 5.5.2) has also to be stored in the index (cf. Berry & Browne 2005: 68).

Anaphora resolution can now help greatly in proximity searching: "Suchen mit Abstandsoperatoren sind fehleranfällig, da in den Texten nicht stets die selben Begriffe vom Autor benutzt werden, sondern Umschreibungen, elliptische Weglassungen und Anaphora" (Stock 2007: 149). Consequently, if the anaphors *it* and, most importantly, the non-finite *-ed*-item *embedded* were resolved in the second example above, both approaches – word counting and structure considerations – would retrieve this example.

Proximity searching can be carried out by using extended Boolean operators. Some search engines, such as Exalead on the website http://www.exalead.com/search (date of last access: 19/01/2013), for example, offer the NEAR-operator (cf. "Exalead: Web Search Syntax" 2012). This usually means that one to ten words, depending on the search engine, are allowed to occur between two terms (cf. "Beyond Boolean Search: Proximity and Weighting" 27/06/2011). Google seems to offer the operator AROUND(n) where users can define themselves how many words might occur between two terms: *university AROUND(1) Passau* would return results with one word between the two terms (cf. Agarwal 06/02/2012).

In addition, anaphora resolution influences term frequency. Terms and their frequency are stored in an index of a text retrieval system. If anaphors are not resolved, expressions that are related anaphorically cannot be reduced to one term.[4] Consequently, term frequency is not represented optimally in the index. Here, anaphora resolution can help in better representing the content of a document (see chapter 6.4) (cf. Stock 2007: 225, 298-299). Moreover, some anaphoric items might be deleted in the process of stop word detection (cf.

[4] Reducing items to one term is also the aim of other approaches, such as stemming (cf. chapter 5.4.3).

chapter 5.4.2) and these items are then not considered any further when establishing the index. Yet, if anaphora resolution is carried out, stop words must not be deleted, otherwise anaphors cannot be resolved anymore (cf. ibid.: 225, 298-299).

6.2 Structure of anaphora resolution systems

Anaphora resolution systems carry out three steps in order to find the correct antecedent for each anaphor. First, the anaphora resolution system has to identify potential anaphors. This means that the system searches for items that can be used anaphorically. From these, non-anaphoric uses and forms looking like an anaphor have to be detected and eliminated. A fine example is the personal pronoun *it*, which shows anaphoric as well as non-anaphoric use. All fully automatic systems include a detection of non-anaphoric items. Some research is even solely concerned with this task, e.g. Boyd, Gegg-Harrison & Byron (2005).

Second, possible candidates for antecedents are identified. A simple example is again the case of *it*, which can refer back to a noun phrase or a sentence. As a result, noun phrases and the preceding sentence are possible candidates. Consider, for instance, example (295). Here, candidates for the antecedent of *it* are the noun phrases *Caroline*, *me*, *a postcard from her holidays*, *her holidays* and the sentence *Caroline sent me a postcard from her holidays* (see also Table 1). How many units are considered, i.e. the search scope, also depends on the type of anaphor. It is usually small in the case of central pronouns, where two or three sentences before the anaphor are examined. The scope is larger with other anaphor types such as noun phrases with a definite article because they can refer further back. Here, up to ten sentences back could be analysed (cf. Mitkov 2002: 18-19). Biber et al. (2007: 239-240) found that the anaphoric distance can be highest with noun phrases with a definite article that repeat the antecedent's noun, followed by those noun phrases with a definite article where the noun is a synonym to the antecedent's noun. A slightly lower distance is allowed for dependent demonstrative pronouns that repeat the antecedent's noun, followed by dependent demonstrative pronouns where the noun is a synonym to the antecedent's noun. Still lower is the distance with personal pronouns and the lowest distance is allowed for independent demonstrative pronouns.

(295) Caroline sent me **a postcard from her holidays**. <u>It</u> shows London's Tower Bridge.

The most likely antecedent from this list is then chosen in the third step. More or less all anaphora resolution systems use constraints and preferences, whether explicitly or implicitly. Constraints exclude anaphor candidates, whereas preferences help to choose the correct antecedent from the remaining candidates. Yet, preferences are tendencies and so not appropriate in all situations. With preferences, the candidates still left are ordered according to their salience. Strube (2010) explains: "Mit Salienz bezeichnet man, wie prominent oder aktiv eine Diskursentität im Diskursmodell ist" (ibid.: 400).

For establishing constraints and preferences, various types of information can be considered. One example of a constraint is if anaphor and antecedent have to agree in number and gender (cf. chapter 3.1.1), e.g. in the case of central pronouns. This discards *Caroline*, *me* and *her holidays* in (295) (see also Table 1). Another instance is if the binding theory is applied for central pronoun anaphors within a sentence. Due to its rules, the item *Sue* is excluded as antecedent candidate of the anaphor *her* in *Sue admires her*. Other constraints take into account syntactic or semantic restrictions. A good example of a preference is if candidates in subject positions are ranked higher than those candidates that are not in such positions. Additionally, if anaphors occur in main clauses, these are also preferred. Another preference is to consider the distance between anaphor and antecedent and to prefer the unit nearest to the anaphor. This preference is usually used if all other preferences have been considered in order to arrive at a solution if more than one candidate is still left (cf. Mitkov 2002: 33-47, 57-62; Jurafsky & Martin 2009: 735-738; Strube 2010: 399-400).

Table 1: Steps in anaphora resolution

Steps	**Anaphor resolution of example sentence (295)**
1st step: detection of anaphors	*it* (anaphoric)[5]
2nd step: detection of antecedent candidates	noun phrases: *Caroline*, *me*, *a postcard from her holidays*, *her holidays* sentence: *Caroline sent me a postcard from her holidays.*

[5] A further anaphor in the example sentence is *her*. As illustration, only the procedure for the anaphor *it* is detailed in the table, however.

3rd step: selection of the most likely antecedent	constraint (number): noun phrases: *Caroline*, *me*, *a postcard from her holidays*, ~~*her holidays*~~ sentence: *Caroline sent me a postcard from her holidays.* constraint (gender): noun phrases: ~~*Caroline*~~, ~~*me*~~, *a postcard from her holidays* sentence: *Caroline sent me a postcard from her holidays.* preference (the nearest noun phrase is preferred to an entire sentence): antecedent: *a postcard from her holidays*

For an anaphora resolution system to work efficiently, it needs manifold types of information: a system requires morphological and lexical knowledge, e.g. which word class an item takes, or the number and gender of items. Such information is, for example, provided by POS-taggers and dictionaries. Furthermore, syntactic knowledge such as word, clause and phrase boundaries is necessary. Here, tokenisers and parsers can help, for instance. Semantic knowledge such as sense relations and animacy helps to find the correct antecedent, e.g. by incorporating WordNet or dictionaries. Discourse knowledge, e.g. what is the central topic of a paragraph or text, can also be included. This, for instance, is used by the centering theory, which is based on cognitive principles and on methods of artificial intelligence (cf. chapter 6.3.1). Finally, real-world knowledge such as logic and information about facts, e.g. that David Cameron is the present Prime Minister of the United Kingdom, is necessary with some anaphors to determine the correct antecedent. World knowledge can also overrule preferences if it is incorporated in a system. For instance, in the sentence *The cat caught a mouse and it died* the anaphor *it* refers logically more likely to *a mouse* than to *the cat*, although *the cat* would be chosen as antecedent if the subject of the sentence was preferred (cf. Mitkov 2002: 28-34, 46-49, 53-57, 62-66).

Which types of resources and information are used differs from one system to the other and depends on the strategy for anaphora resolution that is adopted. Generally however, all systems need certain basic pre-processing steps such as POS-tagging or NP extraction (cf. ibid.: 28-34, 48-49).

6.3 Anaphora resolution approaches

Up to now, a large number of anaphora resolution methods can be found. The first anaphora resolution systems date back to the 1960s; since the 1990s, research has intensified (cf. Mitkov 2002: 68-69; Mitkov 2004a: 277). It is the task of this chapter to outline the most important anaphora resolution approaches, starting with anaphora resolution systems in general and then focusing on anaphora resolution methods for text retrieval in the next chapter.

Each anaphora resolution system uses different approaches and some systems consider more, others fewer anaphor types. In the same way as natural language processing methods (cf. Siddiqui & Tiwary 2008: 3) generally fall into two categories (see chapter 5.4), anaphora resolution systems can be divided up into two basic approaches: rule-based approaches that apply rules and (linguistic) knowledge versus data-based approaches that learn such knowledge and normally use some machine learning. Further terms for similar distinctions are found in other natural language processing methods and encompass, for instance, symbolic vs. empirical, symbolic vs. statistical, deep vs. shallow, rule-based vs. corpus-based, deductive vs. inductive, knowledge-rich vs. knowledge-poor (cf. chapters 5.4.1, 5.4.2, 5.4.5, 6.1.1, 6.1.2, 6.1.4). At the beginning of anaphora resolution research, mainly rule-based methods were developed. Since the mid-1990s, the tendency shifted to devising data-based methods (cf. Mitkov 2002: 95). The two approaches, including important and frequently cited algorithms and methods, will be discussed in detail in the following subchapters.

6.3.1 Rule-based approaches

Rule-based methods are more labour-intensive than data-based ones. Rule-based methods can use linguistic knowledge, such as syntactic rules, extensively, or heuristics, i.e. rules that are based on heuristics (see footnote 7 in chapter 5). Heuristics are then usually expressed by weights that are manually fixed for each factor. One important factor for assigning a central pronoun anaphor the right antecedent, for example, is the distance between anaphor and antecedent. Consequently, antecedent candidates that are nearer to the anaphor are more likely the correct antecedent. Another factor is that nouns are preferred to whole sentences for determining what is likely to be an antecedent. Here, the first factor, i.e. nouns, might be given more weight than, for instance, the second factor, i.e. sentences, because the first one might be more important (cf. Strube 2010: 400-407).

One of the earliest algorithms that is still referred to in the literature is Hobbs's naive algorithm from 1976 (cf. Hobbs 1976). It resolves personal and possessive pronouns with noun phrases as antecedents. Additionally, Hobbs's algorithm deals with split antecedents, coordinated noun phrases and it also considers noun phrases as antecedents that have first to be reconstructed if they are left away through verb phrases with *do*. It does not treat *it* in “extended reference”, i.e. if *it* relates anaphorically to a clause or sentence. Hobbs’s algorithm is one method that takes a linguistic approach. Anaphora resolution is carried out by using the syntactic representation of sentences. Consequently, fully parsed sentences are necessary (cf. Hobbs 1986: 340-344; Mitkov 2002: 72-77; Jurafsky & Martin 2009: 738-740).

Another linguistic approach, less popular than Hobbs’s algorithm, is the centering model (Grosz, Joshi & Weinstein 1983). Mitkov (2002) explains the concept of centering theory as follows: “certain entities mentioned in an utterance are more central than others” and “each utterance features a topically most prominent entity called the center”[6] (ibid.: 53). The theory distinguishes between forward-looking and backward-looking centres. It states that each utterance contains one backward-looking centre. Furthermore, an utterance can have more forward-looking centres. But the forward-looking centre that is preferred most, because it takes the highest rank in the utterance, is called “the preferred centre” and is likely to become the backward-looking centre in the following utterance. Preferred centres are prone to be pronominalised in the following utterances, or, to put it differently, pronouns are usually backward-looking centres.

In (296), the first sentence contains two forward-looking centres, *Susan* and *the piano*. The preferred centre is *Susan* because expressions in subject position are preferred to those in object position. In the second sentence, the preferred centre and, at the same time, the backward-looking centre is *Susan*, which is realised by a pronoun. The same is also valid for the third sentence. The centering theory can now account for the coherence of utterances. Example (296) is coherent because the centre does not change. The theory, however, does not say, how to resolve anaphors (cf. Mitkov 2002: 53-57; Strube 2010: 400-402). One computational realisation of the centering theory in English has been carried out by Brennan, Friedman & Pollard (1987). This realisation uses centring to resolve anaphoric personal pronouns (ibid.: 155). A more recent use of the centering theory is found with Tetreault (2001), for instance.

(296) Susan plays the piano. She likes music. She started taking piano lessons when she was six years old.

[6] Bold printing of *utterance* and *center* removed.

Another algorithm that is frequently cited is Lappin and Leass's Resolution of Anaphora Procedure (RAP) (1994), which is a heuristic approach. Mitkov (2002) acknowledges:

> Lappin and Leass's work is one of the most influential contributions to anaphora resolution in the 1990s: it has served as a basis for the development of other approaches [...] and has been extensively cited in the literature. (ibid.: 105)

In comparison to Hobbs's algorithm, RAP considers all subtypes of central pronouns, i.e. it also includes reflexive pronouns. In addition, reciprocal pronouns are regarded. In order to distinguish anaphoric from non-anaphoric *it*, it contains a procedure for identifying pleonastic pronouns (cf. chapter 3.1.1.3). RAP uses full parsing and for identifying antecedents, morphological and syntactic filters are used. Afterwards, salience measures are applied to select the most likely antecedent. Comparing Hobbs's and Lappin and Leass's algorithms, Lappin and Leass's RAP performs slightly better in sum (cf. Lappin & Leass 1994: 535-536, 544; Mitkov 2002: 99-105; Mitkov & Hallett 2007: 265, 272-273, 279-281).

Apart from these two algorithms, there are other approaches that are not based on full parsing but on partial parsing. Among those are Kennedy & Boguraev's algorithm (1996), Baldwin's CogNIAC (1997), Mitkov's approach (1998) and his newer version MARS[7] (cf. Mitkov, Evans & Orasan 2002). All these are heuristic approaches. To start with, Kennedy & Boguraev's algorithm (1996) is an adaption and extension of Lappin & Leass's (1994), but carries out partial parsing instead of full parsing. Baldwin's approach only deals with third-person personal, possessive and reflexive pronouns. Its peculiarity is claimed to be as follows: "What distinguishes CogNIAC from algorithms that use similar sorts of information is that it will not resolve a pronoun in circumstances of ambiguity" (Baldwin 1997: 38). This leads to high precision, at the cost of low recall, however (cf. Mitkov 2002: 110).[8]

Mitkov's algorithm (1998) resolves pronouns whose antecedents are noun phrases. Unfortunately, no details are given concerning the types of pronouns that are treated (cf. also Mitkov 2002: 145-176). Mitkov's new implementation MARS in 2002 identifies non-anaphoric pronouns such as pleonastic *it* automatically. Mitkov's algorithm in 1998 did not handle that, it had to be done manually instead. In order to find the correct antecedent both approaches as-

[7] For the classification of *it*, MARS uses an instance-based machine learning approach (see chapter 6.3.2) (cf. Mitkov, Evans & Orasan 2002: 173-174).

[8] For evaluation measures in anaphora resolution see chapter 6.5.1.

sign positive and/or negative scores for antecedent candidates, ranging from minus one to plus two. These scores are preferences that help to determine which candidate is the more likely antecedent. Which of these scores is used depends on the specific indicators such as "lexical reiteration", i.e. noun phrases that are repeatedly mentioned, which are more likely to be correct antecedents (cf. Mitkov 1998: 870; Mitkov, Evans & Orasan 2002: 169-171).

Further rule-based algorithms also deal with other anaphor types, apart from or together with central pronouns. However, such algorithms have only been established years after dealing mainly with central pronouns. Especially in the last two decades, research on various types of anaphors has increased considerably. The most investigated of these other types are noun phrases with a definite article. This anaphor type is often referred to as "definite descriptions", a term that is problematic (see chapter 3.6.1). One approach to noun phrases with a definite article is Vieira & Poesio's (2000). They use partial parsing, which extracts noun phrases, and WordNet. Apart from rules relying on heuristics, a version that uses an automatic decision tree has also been implemented (cf. chapter 6.3.2). The results were similar for both. Furthermore, they subclassify noun phrases with a definite article into "direct anaphora", where the anaphor has the same head as the antecedent (e.g. *a dog – the dog*), "bridging descriptions", i.e. the antecedent takes a different head (e.g. *a dog – the animal*) and "discourse-new" where anaphor and antecedent are "not related by shared associative knowledge" (Vieira & Poesio 2000: 542) (e.g. *David Cameron – the Prime Minister of the United Kingdom*) (cf. Poesio & Vieira 1998: 185-191; Vieira & Poesio 2000: 539-546, 556, 575-576, 581-584; Mitkov 2002: 112-113). Another approach towards "associative anaphora" is found with Meyer & Dale (2002).

Treatises about verb phrases and verbal ellipses are found in Asher, Hardt & Busquets (2001). They apply a rule-based approach with a discourse representation theory, using semantic and real world knowledge. Hobbs & Kehler (1997) discuss verb phrases with regard to parallelism to find antecedents; Kehler (2002: 35-79) brings in some ideas about verb phrases and coherence (cf. Lappin 2005: 7-8). Hardt (1997) uses a parsed corpus and applies heuristics to resolve antecedents of verb phrases and verbal ellipses.

Byron (2002) is among the few who treat demonstrative pronouns and personal pronouns. Furthermore, there are systems that include several types of anaphors. Stuckhardt's ROSANA (2001: 491-492) considers central pronouns, reciprocal pronouns, relative pronouns, noun phrases with a definite article and proper names. It uses a parser for pre-processing. A further investigation has been carried out by Markert & Nissim (2005), who examine noun phrases with a definite article as well as *other*, *another* and *such* used for comparison. They examine what resources of knowledge could be used to enhance anaphora resolution with the anaphor types mentioned previously.

A more recent approach comes from Haghighi & Klein (2009), who, however, focus more on coreference than anaphora resolution. They consider central pronouns, noun phrases with a definite article and proper names. The tools used for pre-processing are a tagger and parsers. They use constraints and filters to find whether items are coreferential or not.

6.3.2 Data-based approaches

Data-based systems usually need a training corpus that is annotated with anaphoric relations. Other systems do not rely on a corpus but on unsupervised data. From these data, machine learning systems then infer, i.e. "learn", rules that are consequently used to resolve anaphors in unannotated texts. Mitkov (2002) explains:

> [M]achine learning methods offer the promise of automating the acquisition of this knowledge [i.e. "knowledge about morphology, syntax, semantics, discourse and pragmatics and general knowledge about the real world"] [...] by learning from a set of examples (patterns). (ibid.: 113)

There are different types of machine learning approaches, e.g. decision tree or instance-based methods. Decision trees are "techniques for solving classification problems" (Schmid 2010: 180) and are established from training data. Instance-based methods "simply remember past training instances and make a decision about a new case based on its similarity to specific past examples" (Mooney 2004: 377). Mooney (2004: 391) states that both decision tree and instance-based methods have been used for anaphora resolution.

Advantages of data-based approaches are that they are often domain- and language-independent. Furthermore, machine learning methods are more robust than rule-based approaches and may detect connections of factors that human beings might not notice (cf. Mitkov 2002: 113; Mooney 2004: 376-377; Schmid 2010:180-181; Strube 2010: 400-407). However, Mitkov & Hallett (2007: 271) state that machine learning algorithms cannot reach rule-based ones up to now.

One of the most important machine learning algorithms is Soon, Ng & Lim's (2001). It has been frequently cited with machine learning algorithms. However, it deals with coreference resolution and not anaphora resolution in the first place. Soon, Ng & Lim's algorithm considers the resolution of central and demonstrative pronouns, noun phrases with a definite article and proper names. The algorithm mainly uses tokenisation, sentence segmentation, a POS tagger and partial parsing including noun phrase identification and named entity recognition. Furthermore, they apply a decision tree. A disadvantage of this algorithm is that it does not consider the context because not much linguistic, semantic and world knowledge is included in the algorithm (cf. Soon, Ng & Lim 2001: 521-526, 542; Mitkov 2002: 116-117; Strube 2010: 406).

Soon, Ng & Lim's algorithm has often been extended by other researchers. Among these extensions are, for instance, Ng & Cardie (2002b), who consider noun phrases with *the* as well as proper names and pronouns and they use a decision tree. A more recent algorithm is Versley et al.'s BART (2008: 10), which also builds on and extends Soon, Ng & Lim's approach. BART uses machine learning and focuses on coreference resolution rather than anaphora resolution proper. Additionally, BART uses a tagger and chunker, a named entity recogniser and a parser for pre-processing. BART also draws on information from Poesio & Kabadjov's (2004: 663-664) GuiTAR, another algorithm. GuiTAR treats personal pronouns and noun phrases with a definite article and uses heuristics in the pre-processing stage. In its essence, GuiTAR incorporates information of a number of other systems in turn: Mitkov's MARS for pronoun resolution and Vieira & Poesio's approach to noun phrases with a definite article. Due to the modular structure of GuiTAR, anaphora resolution algorithms with additional anaphor types can be tested in the frame of the system.

Other recent algorithms, also based on coreference resolution and built according to the machine learning algorithm from Soon, Ng & Lim (2001) are Stoyanov et al.'s Reconcile (2010) and Uryupina's Corry (2010). They resolve coreferential noun phrases, however, they do not say which types. Reconcile uses a classifier as machine learning and it also uses clustering. In addition, a tokeniser, POS tagger, parser and named entity recogniser are included during the pre-processing stage. Corry uses a tagger, parser and WordNet.

Among the less important algorithms using machine learning approaches are algorithms focusing on the detection of pleonastic *it*, e.g. Evans (2001) and Boyd, Gegg-Harrison & Byron (2005), both of which are instance-based. Such methods are also needed for an automatic resolution system of anaphoric personal pronouns.[9] Detecting non-anaphoric *it* is also part of Lappin & Leass's algorithm, but there the approach is rule-based. As to the comparison of rule-

[9] A larger empirical analysis on the pronoun *it*, including cases where it is anaphoric and takes antecedents other than noun phrases and non-anaphoric uses, has been carried out by Gundel, Hedberg & Zacharski (2005), however, on spoken language.

based systems and machine learning systems for the detection of non-anaphoric *it*, Evans (2000: 239) comes to the conclusion that both approaches achieve about the same results (see also Boyd, Gegg-Harrison & Byron 2005: 44). Moreover, the distinction between anaphoric and non-anaphoric uses of central pronouns, demonstrative pronouns, noun phrases with a definite article and proper names has been an issue of Ng & Cardie (2002a). Additionally, Bean & Riloff (1999) examine anaphoric and non-anaphoric noun phrases with a definite article. They, however, use heuristics.

Furthermore, there are some approaches for less investigated types of anaphors: relative pronouns are the focus of Cardie (1992), using an instance-based method. The algorithm employs partial parsing; extracting noun, verb and prepositional phrases. Kolhatkar & Hirst (2012) concentrate on the resolution of demonstrative pronouns and base their algorithm on machine learning. Ng et al. (2005) present a decision tree learning algorithm that focuses on the resolution of the indefinite pronoun *one*. A data-based approach on verb phrases and verbal ellipses comes from Nielsen (2004), using parsers and different machine learning algorithms. More details about algorithms on verb phrases and verbal ellipses are given in Lappin (2005: 4-9). In addition, Kehler (2002: 67-71) presents an overview of methods towards resolving verb phrases and verbal ellipses.

6.3.3 Comparison of anaphora resolution approaches

The previous outline shows that anaphora resolution systems differ greatly with regard to various parameters. Each approach, whether rule-based or data-based, uses different resources and (pre-)processing stages. Consequently, comparing such approaches is difficult, if not impossible.

Most importantly, from a linguistic point of view, all these approaches so far are restricted to specific types of anaphors. Some algorithms consider more, some fewer types of anaphors. No system, however, includes all types of anaphors. A further difficulty is that some approaches do not provide all the necessary information about the anaphor types that they consider and most lack information about which items are being considered as anaphors and in which contexts such items are categorised as anaphoric or non-anaphoric. Finally, important recent approaches to anaphora resolution are restricted to determining coreferential relations between anaphor and antecedent.

Table 2 gives an overview of all approaches mentioned, according to which anaphor types each approach considers. It is not always possible, however, to find out exactly which anaphoric items and anaphor types a system treats because with some methods the anaphor types that are regarded are not stated.

Table 2: Overview of important anaphora resolution methods of rule- and data-based approaches (* not mentioned which anaphor types are treated in detail)

Anaphor type	Rule-based approaches	Data-based approaches
Central pronouns	Hobbs (1976) Brennan, Friedman & Pollard (1987)* Lappin & Leass (1994) Kennedy & Boguraev (1996) Baldwin (1997) Hobbs & Kehler (1997) Mitkov (1998)* and (2002)* Stuckhardt (2001) Tetreault (2001)* Byron (2002) Haghighi & Klein (2009)	Evans (2001) Soon, Ng & Lim (2001) Ng & Cardie (2002a) Ng & Cardie (2002b)* Poesio & Kabadjov (2004) Boyd, Gegg-Harrison & Byron (2005) Versley et al. (2008)* Stoyanov et al. (2010)* Uryupina (2010)*
Reciprocal pronouns	Lappin & Leass (1994) Kennedy & Boguraev (1996) Stuckhardt (2001)	-
Demonstrative pronouns	Byron (2002)	Soon, Ng & Lim (2001) Ng & Cardie (2002a) Ng & Cardie (2002b)* Versley et al. (2008)* Stoyanov et al. (2010)* Uryupina (2010)* Kolhatkar & Hirst (2012)
Relative pronouns	Stuckhardt (2001)	Cardie (1992)
Adverbs	-	-
Noun phrases with a definite article	Bean & Riloff (1999) Vieira & Poesio (2000) Stuckhardt (2001) Meyer & Dale (2002) Markert & Nissim (2005) Haghighi & Klein (2009)	Vieira & Poesio (2000) Soon, Ng & Lim (2001) Ng & Cardie (2002a) Ng & Cardie (2002b)* Poesio & Kabadjov (2004) Versley et al. (2008)* Stoyanov et al. (2010)* Uryupina (2010)*
Proper names	Stuckhardt (2001) Haghighi & Klein (2009)	Soon, Ng & Lim (2001) Ng & Cardie (2002a) Ng & Cardie (2002b)* Versley et al. (2008)* Stoyanov et al. (2010)* Uryupina (2010)*

Indefinite pronouns	Markert & Nissim (2005)	Ng et al. (2005)
Other forms of coreference and substitution	Markert & Nissim (2005)	-
Verb phrases with *do* and combinations	Hardt (1997) Hobbs & Kehler (1997) Asher, Hardt & Busquets (2001) Kehler (2002)	Nielsen (2004)
Ellipses	Hardt (1997) Hobbs & Kehler (1997) Asher, Hardt & Busquets (2001) Kehler (2002)	Nielsen (2004)
Non-finite clauses	-	-

6.4 Anaphora resolution in text retrieval

There are not many papers that address anaphora resolution in the context of text retrieval. One important contribution is the Syracuse study, e.g. Katzer, Bonzi & Liddy (1986), Bonzi & Liddy (1989) and Liddy (1990). The anaphor types and their resolution are only formulated as rules but not implemented in the Syracuse study. The study investigates in what way anaphora resolution – based on their classification – has an influence on term frequency of text retrieval systems. The increase of term weights ranges from 138 % to 154 %, or 54 % to 82 % if document length is taken into account (cf. Liddy 1990: 49).

Another study of Pirkola & Järvelin (1996a) and Pirkola (1999) examines anaphora resolution in text retrieval concerning proximity searching. Proximity searching in text retrieval makes use of the proximity operator, which is “a specific case of the and-operator” (Pirkola 1999: 19) and considers the distance between two terms (cf. chapter 6.1.5). They found out that anaphora resolution leads to improvements in precision and recall. For instance, recall increases at around 18 % if anaphors are resolved and users search for proper names within a sentence and a 29 % increase of recall is shown for searching within a paragraph. However, Finnish newspapers are analysed in the study and the results may not be the same for the English language (cf. Pirkola & Järvelin 1996b: 459; Stock 2007: 149-150).

Ferrández, Palomar & Moreno’s SUPAR (1999), which is also used in Ferrández, Rojas & Peral (2007), also deals with anaphora resolution in text retrieval. They claim to resolve pronouns, noun phrases with *one* as head and noun phrases such as *the former*, *the latter*, *the first/second*, which they term "surface-count anaphora". Here, resolution is based on having information about the sequence of coordinated expressions. They do not give further information on what items their algorithm exactly treats. SUPAR works on a POS-tagger and a partial parser and stores coordinated noun and prepositional phrases as well as verb phrases and conjunctions (cf. Ferrández, Palomar & Moreno 1999: 7; Ferrández, Rojas & Peral 2007: 79-80).

A more recent contribution to anaphora resolution and text retrieval is found in Do Carmo Pereira, Seibel Júnior & de Freitas (2009). They use a new model for text retrieval, the so-called "Discourse Nominal Structure model", which identifies the most important entities in a text. By resolving anaphors, relations in texts can be better seen and then these items are ranked as more important. However, this approach is applied to the Spanish language.

In sum, research so far has shown improvements if anaphora resolution is considered in text retrieval. This is not totally surprising because anaphors are cohesive devices (cf. chapter 2.4 and 2.5) that refer, for example, to items in subject position. Therefore, anaphors are usually used to denote the topic and consequently represent key terms of a text.

6.5 Evaluation of anaphora resolution systems

6.5.1 Measures for evaluation

If evaluating anaphora resolution systems, the measures precision and recall can be adopted from text retrieval systems. In order to calculate precision and recall, the measures true positives, true negatives, false positives and false negatives are needed. These four quantities measure how many anaphors are actually identified by a program and whether they are classified correctly or not. In the context of anaphora resolution, a true positive means that an item is correctly identified as anaphor; a true negative is an item that is rightly labelled as non-anaphoric. A false positive occurs if a system classifies an item that is not anaphoric as an anaphor; a false negative is an item that a system classifies as non-anaphoric, but which is in fact anaphoric. Ideally, a system returns few items that are false positives or false negatives (cf. Büttcher, Clarke & Cormack 2010: 332; Levene 2010: 402). For measuring how well a system has classified items as anaphoric or non-anaphoric, precision and recall can be calculated as shown here (cf. Olson & Delen 2008: 137-138):

$$\text{Precision} = \frac{\text{True positives}}{\text{True positives + False positives}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{True positives + False negatives}}$$

Consequently, if a system can exclude all non-anaphoric items, there are no false positives. If it does not exclude any correct anaphors, there are not any false negatives. To say it differently, if all anaphors in the corpus are found, recall is 100 %. If all items that are found are anaphors, precision is 100 %. So precision and recall can be defined specifically for anaphor detection as shown here:

$$\text{Precision} = \frac{\text{Anaphors found}}{\text{Items found in total}}$$

$$\text{Recall} = \frac{\text{Anaphors found}}{\text{Anaphors in total}}$$

For evaluating a system concerning its rate with resolving anaphors, i.e. how well the correct antecedent is assigned to an anaphor, precision and recall can be calculated as follows. The value for precision is returned if the number of anaphors that were resolved correctly is divided by the number of anaphors the algorithm tried to resolve. Recall is defined by the number of anaphors correctly resolved divided by the number of anaphors that the algorithm found. An illustration of precision and recall is shown here:

$$\text{Precision} = \frac{\text{Number of anaphors resolved correctly}}{\text{Number of anaphors attempted to be resolved}}$$

$$\text{Recall} = \frac{\text{Number of anaphors resolved correctly}}{\text{Number of all anaphors}}$$

These definitions follow Baldwin (1997: 41-42). However, a different calculation of recall has been proposed by Aone & Bennett (1995: 126). They define recall as the number of anaphors resolved correctly, divided by the number of all anaphors identified by the program. The problem with such a definition of recall is that if algorithms always returned an antecedent, the value for precision and recall would be the same.

In addition, a number of other measures have been suggested for evaluation. For instance, Mitkov proposed a success rate, which he defined as the number of anaphors resolved correctly divided by the number of anaphors in the texts (cf. Mitkov & Hallett 2007). Mitkov's success rate measure corresponds to Baldwin's measure of recall (cf. Mitkov 2002: 178, 190). Furthermore, Mitkov (2002: 179-185, 189-190) lists a few more measures that can be used to evaluate anaphora resolution systems.

6.5.2 Comparing different algorithms

Apart from evaluations of individual systems, comparing different algorithms and their successfulness in resolving anaphors is desirable. Generally, comparing algorithms from single implementations is difficult if not impossible, because each implementation uses different sources of knowledge, such as POS-tagger, parser, both POS-tagger and parser, or none (cf. also chapter 6.3). Furthermore, there are different POS-taggers and parsers available. Apart from that, each algorithm is tested on its own corpus, which makes a comparison with other algorithms difficult. Not all proposed algorithms work in a fully automatic mode, i.e. "that there is no human intervention at any stage" (Mitkov & Hallett 2007: 262). Some algorithms edit the pre-processing results and correct errors that these pre-processing tools make. Additionally, pre-processing steps can be simulated. Consequently, such a procedure leads to better results with anaphora resolution. Some systems that have been mentioned above carry out a manual error correction after the pre-processing stage. Pleonastic pronouns, for example, that had not been identified by the algorithm have been manually deleted in Kennedy & Boguraev (1996). Lappin & Leass (1994) and Mitkov (1998) corrected errors of their parser manually and Ferrández, Palomar & Moreno (1998) postedited the results from their POS tagger and partial parser. Yet, it is frequently not mentioned in the papers to what extent the systems are subjected to manual editing (cf. Mitkov & Hallett 2007: 262-263).

As a result, benchmark systems can be established for evaluation purposes (cf. Mitkov 2002: 181-182). A benchmark is "something that is used as a standard by which other things can be judged or measured" (Mayor 2009: 141) and is "an established way for evaluating automatic systems which tackle the same task" (Chu Min Xian, Zahari & Lukose 2011). One benchmark has been established by Mitkov & Hallett (2007). They included some well-known algorithms: Hobbs's (1976), Lappin & Leass's (1994), Kennedy & Boguraev's (1996), Baldwin's (1997) and Mitkov's (1998) algorithm. Mitkov & Hallett (2007: 262-263) compare and evaluate the chosen algorithms on three corpora, one consisting of technical manuals, the second of newswire texts and the third of literary texts. Furthermore, the algorithms were re-implemented in a fully automatic way. This means that there is not any human intervention, although some algorithms corrected the output of pre-processing tools in their original implementation.

As to the performances of these anaphora resolution systems, Mitkov & Hallett found out that the resolution rate was much lower than in the original results. Mitkov & Hallett discovered that algorithms with full parsing generally performed better than those with partial parsing. Within full-parser algorithms, Lappin & Leass's algorithm (1994) outperformed Hobbs's algorithm (1976) with 50.4 % versus 48.4 % for both precision and recall. Among the three partial parser algorithms (not considering MARS from 2002), Kennedy & Boguraev's algorithm (1996) scored best with 44.7 % precision and recall. Although Baldwin's algorithm in sum did worse than algorithms using full parsing, Mitkov & Hallett (2007: 280) showed that it did better on anaphors where the antecedent was not found in the same, but another sentence (intersentential anaphors) than Lappin and Leass's RAP (56.8 % versus 52.7 % precision/recall). In this intersentential evaluation, however, non-anaphoric pronouns and pronouns referring to other antecedents than noun phrases were excluded. All these measures are given for the corpus of technical manuals. However, the results go in the same direction with the other two corpora (cf. Mitkov & Hallett 2007: 277-280, 290-291). As the output of pre-processing tools, especially parsers, contains errors, this also influences the performance of anaphora resolution systems. Mitkov & Hallett (2007: 283-284) have shown that an elimination of errors leads to a performance increase of anaphora resolution systems between 4 and 10 %.

The analysis carried out by Mitkov & Hallett, however, is restricted to central pronouns. Another system with benchmarking efforts is, for instance, Reconcile (Stoyanov et al. 2010), which provides different benchmark data sets and scoring metrics for coreference resolution but not for anaphora resolution proper. In conclusion, benchmarking in anaphora resolution concentrates on central pronouns and on coreference resolution for noun phrase entities.

As there are no benchmarks for other anaphor types, new approaches frequently re-implement existing algorithms in order to compare their approach to a state-of-the-art system such as Soon, Ng & Lim's (2001) algorithm, or to similar approaches (cf. Mitkov 2002: 181; Strube 2010: 406). A different approach is to use baselines. For instance, Hobbs's (1978) algorithm is often used as a baseline (cf. Jurafsky & Martin 2009: 738). A baseline is defined by Hagenbruch (2010) as follows: "Damit [d.h. mit einer Baseline] bezeichnet man denjenigen Grad an Genauigkeit, der erreicht wird, wenn man den einfachsten möglichen Algorithmus auf das Problem anwendet" (ibid.: 267). Mitkov (2002) further remarks on the importance of baselines:

> The **evaluation against baseline models** is important to provide information as to how effective an approach is, by comparing it with unsophisticated, basic models. This type of evaluation justifies the usefulness of the approach developed: however high the success rate may be, it may not be worth while developing a specific approach unless it demonstrates clear superiority over simple baseline models. (ibid.: 181)

At the opposite end, there is the “gold standard”. It is the best solution to a problem, which can usually be achieved by manual annotation. For instance, anaphora resolution systems can be compared to a gold standard, i.e. texts annotated manually with anaphoric relations (cf. Menke 2012: 305).

6.5.3 Annotated corpora for evaluation

6.5.3.1 Current annotated corpora

Whether evaluating systems individually or in comparison, annotated corpora are needed, where anaphors and their antecedents are marked in a machine-readable way. Up to now, many researchers have established their own corpus for anaphora resolution and its evaluation, or at least adapted an existing one (cf. Mitkov 2002: 195-196). However, these corpora are usually not made available. More importantly, most corpora so far are restricted to specific anaphor types. For instance, corpora annotated with central pronouns can be found. One of the few corpora publicly available is Mitkov et al.’s corpus on newswire texts, which marks noun phrase coreference within and across documents. This corpus contains about 55,000 words in total and deals with the topic security/terrorism (cf. Mitkov et al. n.d.; Hasler, Orasan & Naumann 2006: 1168; Ng 2010: 1397). It can be downloaded from the website http://clg.wlv.ac.uk/projects/NP4E/#corpus (date of last access: 16/01/2013).

In addition, an increase in the number of corpora annotated for coreference resolution can be observed in the last few years. Coreferential noun phrases are central and demonstrative pronouns, noun phrases with a definite article and proper names, all of which have to show a coreferential relationship to their antecedents. One of the first corpora annotated for coreferential relations has been the MUC-6 and MUC-7 corpora (cf. Grishman & Sundheim 1996: 468; Hirschman & Chinchor 1997). These corpora are only downloadable for a fee from the Linguistic Data Consortium (http://www.ldc.upenn.edu, date of last access: 16/01/2013). Another corpus annotated with coreferential noun phrases is OntoNotes (cf. “OntoNotes: Coreference” 2012). A part of OntoNotes was used in SemEval-2010, a workshop on semantic evaluation. In 2010, one task of SemEval was coreference resolution and a corpus annotated with coreferential pronouns, noun phrases with a definite article and proper names was released. The corpus amounts to 120,000 words, deals with newswire and broadcast news and can be downloaded from the Linguistic Data Consortium or from the website http://semeval2.fbk.eu/semeval2.php?location=data (date of last access: 16/01/2013). A larger corpus amounting to one million words is available from OntoNotes (Release 4.0).[10] It also contains texts from broadcast conversation, newsgroups and blogs. The corpus is available for free from the Linguistic Data Consortium, only registration is required.

Another task in SemEval-2010 focused on the resolution of verb phrases and verbal ellipses. A part of the OntoNotes corpus has been annotated, consisting of texts from the Wall Street Journal. In total, 500 items in about 53,600 sentences have been annotated (cf. Bos & Spenader 2011: 481; "SemEval 2010: VP Ellipsis Processing" 2011). This corpus can be freely downloaded from http://semeval2.fbk.eu/semeval2.php?location=data (date of last access: 16/01/2013).

Other corpora are, for example, the Lancaster Anaphoric Treebank, the GNOME corpus and the ARRAU corpus, none of which is, however, publicly available (cf. Botley & McEnery 2000: 26-27; Mitkov 2001: 115; Mitkov 2008: 582; "Anaphoric Bank Data" 2009). Moreover, an important contribution of corpora to text retrieval is found with Text REtrieval Conference (TREC), as outlined on http://trec.nist.gov (date of last access: 17/11/2012). The corpora found here can be accessed for a fee and/or with registration.

All of these existing corpora are insufficient to evaluate the anaphor type classification proposed in chapter 3. Some are not available or only for a fee and – most importantly – all focus on some text types, but not specifically on different hypertext types (cf. Rehm 2007, see also chapter 4.1). To overcome this deficiency, the hypertext corpus described in chapter four has been annotated with anaphoric relations. In that way, it can be of use for computational processing, e.g. evaluation purposes of anaphora resolution systems or as training data for machine learning approaches, as Mitkov (2001: 114-116 and 2002: 192-196) demanded. The annotation procedure in the hypertext corpus will now be described.

[10] This number only applies to texts from English. OntoNotes also contains texts from Arabic and Chinese. In addition, the SemEval-2010 corpus includes Catalan, Dutch, German, Italian and Spanish texts.

6.5.3.2 An annotated corpus for all anaphor types

In general, the annotations in the hypertexts are encoded as XML-tags (cf. McEnery & Hardie 2012: 29-30). Each anaphor has been marked with two types of information: type of anaphor ("anaType") and type of relationship between anaphor and antecedent ("anaSubType"). With the former, not only the 12 anaphor types, but also their subtypes have been differentiated and given a separate label, if considered reasonable. The tags used for annotation are shown in Table 3.

Table 3: Tags of anaphor types

Tag	Anaphor (sub)type
pers	Personal pronoun
poss	Possessive pronoun
refl	Reflexive pronoun
recip	Reciprocal pronoun
demZ	Demonstrative pronoun in dependent function
demE	Demonstrative pronoun in independent function
rel	Relative pronoun
adv	Adverb
defNom	Noun phrase with a definite article
eigP	Personal proper name
eigG	Other proper name
ers	Indefinite pronoun or other form of coreference and substitution
vp	Verb phrase with *do* and combinations
elln	Nominal ellipsis
ellv	Verbal ellipsis or clausal ellipsis
ellnv	Nominal and verbal ellipses combined
to	*Non-finite clause* anaphor in the form of *to*
ing	*Non-finite clause* anaphor in the form of *-ing*
ed	*Non-finite clause* anaphor in the form of *-ed*

Furthermore, three types are distinguished and marked for the anaphor-antecedent relationship, as introduced in chapter 2: coreference (with the tag "coreference"), substitution (tag "substitute") and the miscellaneous category comprising items that are neither coreferential nor substitutional (tag "coref/subst"). The antecedent is marked by "source"-tags. The connection between anaphor and its antecedent(s) is defined by ID-numbers. Those anaphors and one or more antecedents match that have the same numbers. Finally, the text has been broken up into sentences by using the Stanford parser. Here an example of the corpus annotations is shown, originating from W5:

```
[...]
  <sentence senID="6">
    <source srcID="4">Clementines</source> separate easily into eight to fourteen juicy segments.
  </sentence>
  <sentence senID="7">
    <anaphora refID="4" anaSubType="coreference" anaType="pers">They</anaphora> are very easy to peel
[...]
```

In some circumstances there is not only one correct antecedent but the same word or phrase that can also function as antecedent linguistically is found more than once in a text. Here, anaphora resolution systems have different options and can choose among more correct antecedents. In order to account for such cases, all possible antecedents have been annotated if that seemed justified. An example is shown below. Here, *it* in line 4 can have *the Internet* in line 1 or 2 as antecedent, which is why both have been annotated:

The Internet was originally conceived as a distributed, fail-proof network that could connect computers together and be resistant to any one point of failure; the Internet can't be totally destroyed in one event, and if large areas are disabled, the information is easily re-routed. **It** was created mainly by DARPA; its initial software applications were email and computer file transfer.

6.6 Conclusion

Anaphora resolution can be of help in fields of natural language processing because here text understanding is required. Up to now, anaphora resolution approaches have been extensively studied in machine translation, information extraction, question answering and text summarisation. Only a few studies so far focus on the benefit of anaphora resolution in text retrieval, although anaphora resolution would better represent term frequency in an index and improve proximity searching.

With that in mind, the structure of anaphora resolution systems was detailed. Generally, systems accomplish three steps: detection of anaphors, identification of anaphor candidates, and selection of the most likely antecedent based on constraints and preferences. Anaphora resolution systems here can use various types of information, where some are easier applied and others are more complex and time-consuming if they are integrated. On the usefulness of different kinds of resources for anaphora resolution, Garnham (2001) states:

> Psycholinguists have studied many factors that influence the interpretation of anaphoric expressions. [...] Strictly linguistic factors should win out over general knowledge [...]. General knowledge in turn should win out over heuristic strategies, such as parallel function [...]. (ibid.: 93-94)

This could then mean that anaphora resolution systems would work more correctly if linguistic factors rather than heuristics were considered. However, as was shown in subchapter three, recent anaphora resolution systems rather use approaches that rely on minimal linguistic input than on more expensive and complex linguistic information. This is particularly true for data-based approaches that involve even less linguistic knowledge than rule-based approaches.

Apart from that, important methods and algorithms of rule-based and data-based approaches were outlined. It has been shown that no system so far considers all anaphor types. Moreover, research in the application of anaphora resolution in text retrieval was described. Previous studies demonstrated that anaphora resolution can indeed lead to improvements in text retrieval, which is why anaphora resolution should here be examined in more depth in the future.

A further part illustrated how anaphora resolution systems can be evaluated. Here, important measures were explained, the problem of comparing algorithms discussed and annotated corpora for evaluation described. So far, some corpora are freely available; however, most of them focus on coreference resolution rather than on anaphora resolution. As a result, only certain anaphoric items and relations are annotated. Instances, for example, where an anaphor does not show a coreferential relationship, are not considered, e.g. all anaphoric items that show only a substitutional relationship between anaphor and antecedent, such as indefinite pronouns, are usually ignored. Additionally, these corpora do not contain a systematic representation of hypertext types. To overcome these weaknesses, a hypertext corpus has been compiled for this book. Furthermore, it has been annotated with all anaphoric relations that have been proposed in the categorisation of chapter 3.

From that discussion, the first research question posed in chapter 4 can now be answered: with anaphora resolution, proximity searching and, most importantly, the effectiveness of text retrieval systems is improved because terms are better represented in the index.

7 Development of extensive linguistic rules for anaphora resolution: the example of *non-finite clause* anaphors

Current systems frequently lack a sound linguistic analysis of rules for anaphora resolution. For this purpose, the following section will discuss rules that should be implemented in computational anaphora resolution systems. To discuss linguistic rules for all anaphor types would go beyond the scope of this book, so only one anaphor type has been chosen as an example. It would, however, be possible to define rules for all other anaphor types with the help of the detailed classification and description of anaphors in chapter 3. As no system up to now considers *non-finite clause* anaphors and because this anaphor type is also the most frequent in the hypertext corpus, it is the resolution of non-finite clauses that will be examined here in detail from a computational point of view.

In chapter 7.1, linguistic rules for identifying *non-finite clause* anaphors will be examined before rules for detecting the correct antecedent of each *non-finite clause* anaphor will be discussed in 7.2.[1] These rules are derived from information in standard grammar books (e.g. Huddleston & Pullum 2010, Quirk et al. 2012; see also chapter 3) and, especially in the second subchapter, from a thorough analysis of the hypertext corpus.

7.1 Identifying anaphoric items

A system first has to distinguish between items that are anaphoric and those that are non-anaphoric. For *non-finite clause* items, identifying anaphors can be achieved as follows: a system searches for *to* and all items ending in *-ing* and *-ed*. As this search will return anaphoric as well as non-anaphoric items, all non-anaphoric instances then have to be excluded by rules. Such rules can be inferred from the characteristics of non-anaphoric items that have been thoroughly discussed in chapter 3.12. The characteristics have been summed up in

[1] In terms of the threefold structure of anaphora resolution systems (cf. chapter 6.2), chapter 7.1 discusses the first step, which is the detection of anaphors. Chapter 7.2 comprises steps two and three of anaphora resolution systems: it first outlines the detection of antecedent candidates (step two), which is quite simple for *non-finite clause* anaphors, before the selection of the correct antecedent (step three) is examined in depth.

DOI 10.1515/9783110416756

Table 29 of chapter 3, which is reprinted here in Table 1, together with the detailed frequency of non-anaphoric *-ing-*, *-ed-* and *to*-items.

The analysis of non-anaphoric *-ing*-items in the corpus reveals 1,375 items. They are distributed across the categories as shown in Table 1. The most frequent non-anaphoric *-ing*-items are nouns ending in *-ing*, followed by adjectives and participles that are part of verb phrases with *be*. Other non-anaphoric *-ing*-items of the corpus that do not belong to one of the mentioned classes are summed up in the “context” category. These items make up 6.3 % of all non-anaphoric *-ing*-items. This number is important as anaphora resolution systems will probably have difficulties detecting the items of the “context” category as non-anaphoric because a deeper analysis, i.e. the inclusion of context or world knowledge, is required in the majority of cases. But yet it is also possible – through the analysis of the hypertext corpus – to formulate some rules for items of the context category.

Table 1: Non-anaphoric *-ing*-items in the corpus

	In subject position	In extraposition	Simple finite verb phrase	Complex finite verb phrases			Overt subject	*you* etc. as antecedent	Imperative
			Present form	*be*	*have*	Modal auxiliary verbs			
-ing	40	5	18	206	17	9	27	42	25

	Gerunds	Nouns	Adjectives	Prepositions	Context	**Non-anaphoric items in total**
-ing	53	595	224	28	86	**1,375**

With regard to *-ed*-items, regular and irregular forms are distinguished. Table 2 gives the details. Most non-anaphoric items are past forms, followed by *be* in complex finite verb phrases. Eight items, i.e. 0.3 %, fall into the “context” category.

Table 2: Non-anaphoric *-ed*-items in the corpus

	In subject position	Simple finite verb phrase		Complex finite verb phrases			Overt subject	*you* etc. as antecedent	Imperative
		Present form	Past form	*be*	*have*	Modal auxiliary verbs			
Regular *-ed*	1	34	665	483	248	103	2	2	0
Irregular *-ed*	0	106	247	132	138	54	0	2	10
***-ed* in total**	**1**	**140**	**912**	**615**	**386**	**157**	**2**	**4**	**10**

	Nouns	Adjectives	Context	**Non-anaphoric items in total**
Regular *-ed*	41	497	3	**2,079**
Irregular *-ed*	56	107	5	**857**
***-ed* in total**	**97**	**604**	**8**	**2,936**

Non-anaphoric *to*-items are distributed as shown in Table 3. The most frequent *to*-items are prepositions. Furthermore, the number of items whose non-anaphoricity is determined by the context lies at 7.2 % of all non-anaphoric *to*-items.

Table 3: Non-anaphoric *to*-items in the corpus

	In subject position	Verbs with *to*				Overt subject	*you* etc. as antecedent	Imperative
		Marginal modals	Modal idioms	Semi-auxiliaries	Fixed expressions			
to	10	7	9	81	0	11	116	27

	Postmodification of adjectives			Prepositions	Prepositional adverbs	Other fixed expressions	Context	**Non-anaphoric items in total**
	V	VI	VII					
to	17	1	0	743	0	8	80	**1,110**

Turning to how these non-anaphoric categories can be used for selecting anaphors, the following aspects have to be considered. To start with, how many non-finite item categories such as nouns or prepositions can be excluded automatically depends on what natural language processing tools are used in the system. The approach that was used as a starting point is to search for *to*, items ending in *-ing* and *-ed* and items that are irregular past participle forms such as *bought*, which do not end in *-ed* (cf. Table 26 in chapter 3.12.3). This is a procedure any simple system should be capable of. For that, only a tokeniser is needed. Furthermore, how many categories can be implemented also depends on what types of information an anaphor detection system incorporates. Consequently, if a system does not contain information about subjects of sentences, for instance, some items that need this type of information cannot be detected as non-anaphoric. It should, however, also be kept in mind that the more information that is used for detection, the more processing time the system needs. Tools that return satisfactory types of information for the non-anaphoric categories above are parsers. The rules for identifying anaphoric/non-anaphoric items will here be evaluated on the Stanford parser.[2] Nevertheless, the rules outlined are not specifically adapted to the Stanford parser but can be used with other parsers as well.

7.1.1 Rules for *-ing*-items and their evaluation

Theoretically, parsers such as the Stanford parser are able to identify nouns, adjectives and prepositions ending in *-ing*. If a parser worked without errors, this would mean that from the 1,375 non-anaphoric items, 847 items (595 nouns, 224 adjectives and 28 prepositions) would be excluded and no anaphoric items would be categorised as false negatives. This number is important, because to achieve the exclusion of these items, no additional rules – and with that no additional processing time – are required theoretically. The precision rate of identifying anaphors with these approaches is 56.7 % and the recall rate 100 %.[3]

Parsers here do not always work perfectly. For instance, "some scholars believe that the bloodletting for which these stones were likely used presages certain acupuncture techniques." (W1) contains the noun *bloodletting*, which the Stanford parser categorises as verb ending in *-ing*. To overcome this problem, parsers could be improved by incorporating additional rules, such as: if a word ending in *-ing* is preceded by *the*, *a*, *an*, or a determinative possessive pronoun, this item is non-anaphoric. Similarly, cases where *the*, *a*, *an*, or a determinative possessive pronoun is followed by an adjective and a word ending in *-ing* represent non-anaphoric items.

Apart from nouns, adjectives and prepositions, the Stanford parser automatically detects verbs in simple verb phrases that have *-ing* as final letters. These are not non-finite *-ing*-forms because the *-ing* is not a suffix here. Some examples are *sing* and *bring*. As a result, the precision rate rises slightly to 57.6 % because 18 items are identified.

After that, further rules that can be implemented relatively easily are as follows. If the verb ending in *-ing* is part of a verb phrase and not a verb phrase itself, i.e. if it is the main verb and accompanied by one or more auxiliary verbs, it can be excluded as non-anaphoric. Either the primary verbs *be* or *have*, or

[2] This standard parser can be tested online on the website http://nlp.stanford.edu:8080/parser (date of last access: 20/01/2013). For more information on this parser see, e.g. "The Stanford Parser: A Statistical Parser" (n.d.); Klein & Manning (2003); Marneffe, MacCartney & Manning (2006); Jurafsky & Martin (2009: 506-507, 521-522).

[3] For the calculation of precision and recall see chapter 6.5.1.

some modal auxiliary verbs can be used here. Modal auxiliaries are the forms *can*, *could*, *may*, *might*, *shall*, *should*, *will*, *would*, *must* and the contractions *'ll*, *'d* (cf. Quirk et al. 2012: 135, 151). If these rules are implemented, 232 items (206 for *be*, 17 for *have* and 9 for modal auxiliary verbs) are excluded. With 278 non-anaphoric items left, the precision rate rises to 71.3 %.

In the case of gerunds, the Stanford parser does not mark them specifically. Actually, not only gerunds but all verbs ending in *-ing* are labelled "VBG", i.e. a verb used either as gerund or as present participle (cf. Marcus, Marcinkiewicz & Santorini 1993: 317; Marneffe & Manning 2011: 2). Consequently, these items cannot be excluded automatically. As will be seen, some of the rules below can detect some gerunds.

Apart from that, it is also possible to consider clause functions in order to rule out non-anaphoric items. If an item in *-ing* appears in subject position and a finite verb follows in the same clause, it is non-anaphoric, as, for instance, in *Freezing the lobster may toughen the meat* (W10). As is apparent, the finite verb does not necessarily have to follow immediately after the *-ing*-item. With that approach, a 74.4 % precision rate is achieved because 40 items occur in the corpus. Consequently, 238 non-anaphoric items are left.

Furthermore, the syntactic structure of sentences where the non-anaphoric items occur can be exploited for detection. If a non-finite clause contains an overt, i.e. explicit, subject, the *-ing*-item is non-anaphoric. Here, 11 out of 27 items are identified if the following rule is considered: before the *-ing*-item, a noun phrase in a prepositional complement of *with* is preceded by a comma. However, three anaphors are excluded as well. This leads to a precision value of 75.2 % and a recall of 99.6 %. In addition, *-ing*-items can refer to addresser or addressee, either explicitly to "antecedents" such as *you* or implicitly if an imperative is used in the main clause. Items that have *you* or related expressions as "antecedent" cannot be excluded so far, but only when an antecedent is searched for with each anaphor. Other non-anaphoric items – apart from non-anaphoric pronouns such as *you* – are *it* and *there*. Extrapositions such as "but it's worth bearing in mind that we were in North London" (WS41) can again only be detected if antecedent searching information is used because the subject here is pleonastic *it*.

Imperative constructions in the main clause should be detectable because there is no subject (25 items in the corpus). Here, the verb in *-ing* is typically preceded by a verb in imperative form. An example is *Please consider using...* (W16). The imperative form and the verb in *-ing* can, however, also be further apart, although belonging to the same main clause (applying to 7 out of 25 items). Through that, a precision rate of 77.3 % is achieved.

What is still left are the non-anaphoric items that were categorised as requiring context information. With these, further aspects can be taken into consideration for an exclusion. Some non-anaphoric items are detected if they appear in constructions that are "chunks" or "fragments". These consist of a verb only or are followed by a noun phrase. Other phrases might follow as well, except verb phrases. This is relevant for headlines, titles of books and so on. Such a rule also helps in excluding gerunds. For instance, *Analysing the books behind the rictus smiles* (B4) is a case where this rule applies. However, if such chunks occur within brackets or are part of a list, they could also be anaphoric. In the corpus, 4 anaphoric and 2 non-anaphoric items with brackets and 4 anaphoric and 12 non-anaphoric items within a list occur. Therefore, it is only useful – at least for this corpus – to not exclude chunks that are part of brackets and where the expression within brackets consists of more than only the *-ing*-item. It is helpful here if brackets and inverted commas are regarded and expressions within them treated as sentences. However, items are not non-anaphoric if they occur after colons. With these types of information, 60 non-anaphoric items are detected in total. Precision rises to 82.8 %, at the expense of lower recall of 99.0 % due to four more false negatives. Still left are 142 non-anaphoric items, i.e. false positives.

In addition, if *-ing* occurs at the beginning of a sentence and is optionally preceded by a negative particle, conjunction and/or an adverb, it can be non-anaphoric. That is the case if no matrix clause, which would contain the antecedent, follows (see also chapter 7.2.1). This either means that no clause follows or that a coordinated clause is used. In both cases, the *-ing*-item is non-anaphoric. One exception is if the *-ing*-item is followed by a noun phrase that contains a determinative in form of a genitive because then this genitive could be the antecedent (cf. chapter 3.12.2.1). However, if the sentence is within brackets, this must not be subject to this rule. With this information, 6 items (including four gerunds) where no clause follows and 4 items in a coordinated clause but without a matrix clause are excluded. Furthermore, if the *-ing*-item is preceded by a noun in genitive, it is also non-anaphoric. There is, however, only 1 item that occurs in that way in the corpus. In sum, 83.9 % precision and 99.0 % recall is achieved with 131 non-anaphoric items left.

Additionally, other structures with non-anaphoric *it* and *there* could be excluded. Typical cases are *it is* and *there is/are* if shortly afterwards a verb in *-ing* occurs in the same clause, for example *It's about making a difference by the practices they adopt…* (WS2e). In the corpus, 9 non-anaphoric items with *it* and three with *there* occur. However, with that strategy 6 false negatives with *it* and 5 false negatives with *there* have to be accepted. Consequently, the precision rate with 85.0 % is slightly better, but recall is 97.4 %. This rule is therefore not that useful, at least for this corpus. Another rule could be to eliminate structures where a noun phrase is followed by the preposition *of* and a verb in *-ing*. 10 items account for this structure. However, this would also lead to 18 false negatives. Consequently, precision lies at 84.6 %, but recall with 96.4 % is 2.6 % worse than without the rule. Again, this rule is better not implemented.

In sum, precision lies at 83.9 % and recall amounts to 99.0 %. Remaining are 131 non-anaphoric items and of these 47 are items where the consideration of information about antecedents could lead to a further reduction of non-anaphoric items.

7.1.2 Rules for *-ed*-items and their evaluation

Anaphors of *-ed*-items fall into regular and irregular forms. First, regular forms are examined. As outlined above with *-ing*-items, the Stanford parser usually detects nouns and adjectives ending in *-ed*. Furthermore, abbreviations for nouns ending in *-ed*, followed by a full stop, which have been counted as part of the noun category, can be identified, such as *ed.* for *edition*. Thus, figures amount to 41 nouns, including 12 abbreviations and 497 adjectives. If all these word classes are identified, the precision rate lies at 16.7 % with a recall of 100 %. Apart from that, the Stanford parser is able to exclude certain verbs that are always non-anaphoric. These are verbs in present and past tense simple. With 34 and 665 items occurring in the corpus respectively, their exclusion results in a precision rate of 26.8 %.

From that on, a few rules can be formulated in order to increase precision. Past participles are part of complex verb forms, namely they either serve to produce the passive together with *be* or they form the perfective together with *have*. Furthermore, past participles can occur in combination with modal auxiliaries such as *can* or *may*. Such constructions can also enter a combination with *be*, such as *could be guessed* for the passive and *have*, such as *could have played* for the perfective, or *could have been guessed* for the perfective and passive simultaneously (cf. Quirk et al. 2012: 135, 151). As with *-ing*-items, it is relatively easy to define rules when *-ed*-items are non-anaphoric if they occur in such complex verb phrases. That means, if *-ed*-items occur together with *be*, *have* and/or a modal auxiliary, they are non-anaphoric. This in consequence leads to an exclusion of 483 *be*, 248 *have* and 103 modal auxiliary items. Precision in turn rises to 97.5 % and 8 non-anaphoric items are left.

As discussed with *-ing*-items, *-ed*-items are non-anaphoric if they appear in subject position followed by a finite verb. In the corpus, only 1 item like that occurs. Moreover, the two overt subjects are excluded if they precede the *-ing*-item, if they are realised as a noun phrase in a prepositional complement of *with* and if this phrase then is preceded by a comma. In addition, 2 items could be ruled out when the anaphors are resolved because they refer to *you* and other non-anaphoric items. The context category with 3 items cannot be excluded, as they depend on information that varies from one use to the other. As a result, 5 items are left, of which 2 items need information about antecedents. Precision therefore lies at 98.4 %, recall at 100 %.

The situation for irregular verbs is as follows. After the 56 nouns and 107 adjectives are excluded, the precision rate for irregular forms lies at 8.7 % and recall at 100 %. Detecting verbs in present and past tense, which are 106 and 247 items respectively, leads to precision of 16.2 %. If rules for items in complex verb phrases with *be* (132 items), *have* (138) and modal auxiliaries (54) are formulated, a precision rate of 79.5 % can be achieved. As with regular forms, irregular forms occurring in subject position and followed by a finite verb can be excluded. However, no examples can be found in the corpus. Apart from that, imperative forms can be excluded amounting to 10 in the corpus, as defined with *-ing*-items. Antecedent information is needed with items that have "antecedents" such as *you*. Yet, only 2 items occur as such in the corpus. In sum, these 2 items plus 5 further items are left. This then leads to a precision rate of 90.4 % and a recall of 100 %.

Combining regular and irregular forms, a precision value of 96.9 % and a recall of 100 % is achieved if all the above rules are implemented, leaving 12 non-anaphoric items, of which 4 could be resolved later with information about antecedents.

7.1.3 Rules for *to*-items and their evaluation

To-items behave differently from *-ing*- and *-ed*-items. This also becomes evident with the Stanford parser, which always returns the same tag if it encounters *to*, irrespective of its word class membership or function (cf. Marcus, Marcinkiewicz & Santorini 1993: 317). The only way to rule out non-anaphoric instances of *to* is therefore to pay attention to the surrounding text.

To start with, a *to*-item is non-anaphoric if it works as preposition. Anaphoric *to*-items are always followed by a verb, which can also be "interrupted" by adverbs or certain punctuation marks such as commas or colons. In the corpus, adverbs in between *to* and the verb were counted four times, punctuation marks twice. If a noun phrase follows *to*, it is non-anaphoric. This rules out 710 of 743 prepositional items. Furthermore, prepositional *to* occurs if it is followed by a verb in *-ing* (16 items), a preposition such as *in* or *about* (13), or an adjective (1). In between such structures, punctuation marks can occur, which applies to 3 items. Such a procedure leads to precision of 66.0 % and recall of 100 %.

Apart from that, non-anaphoric *to* is preceded by certain adjectives and verbs. In more detail, type v adjective constructions amount to 17 items in the corpus, type vi occurs with 1 item, type vii does not occur. In addition, 7 marginal modals, 9 modal idioms and 81 semi-auxiliaries appear in the corpus. Besides, there are items that belong to the category "fixed expressions". These are *have much/nothing to do* (4 items), *in order to* (2 items), *come to be/terms with* (2 items). There are not any prepositional adverbs in the corpus. If all of these items are excluded, a precision rate of 74.5 % is achieved, leaving 244 non-anaphoric items.

Furthermore, there are cases where *to* appears in subject position (10 items) and where *to* occurs in sentences in which an imperative is found in the main clause (27 items). This is analogous to what was discussed with *-ing*- and *-ed*-items. Moreover, if an overt subject occurs in the non-finite clause (11 items), *to* can be defined as being non-anaphoric. This is the case if a noun phrase is a prepositional complement of *for* and this phrase is preceded by a comma. However, this works at the cost of one anaphor that is also being excluded. Identifying such elements leads to a precision value of 78.4 % and a slightly lower recall of 99.9 %.

Additionally, other structural information can be used for detecting items of the context category. For instance, chunks that only consist of *to* and are followed by another element and/or where no matrix clause comes afterwards can be excluded. In these cases, the *to*-item is at the beginning of a sentence, optionally preceded by a negative particle, conjunction and/or adverb. This holds for 8 items in the corpus, leading to precision of 79.1 %. Additionally, if *how to* or *what to* introduces a clause (cf. Quirk et al. 2012: 840), *to* is non-anaphoric in cases where no noun phrase, or noun phrase plus verb phrase precede *how/what to*. However, if the preceding noun is in the genitive, *how/what to* is also non-anaphoric. With this rule, 1 item with *what to* and 12 items with *how to* are excluded. Furthermore, if a verb is preceded by *to* and these two elements occur on their own, i.e. within brackets or inverted commas, they are used to denote an infinitive of a verb and so are non-anaphoric. One item occurs like that.

Further, 116 non-anaphoric items are detected if resolving the antecedent. Here, antecedents are items that were categorised as non-anaphoric such as *you*. The precision rate without considering information about antecedents lies at 80.2 % with a recall of 99.9 %. Apart from 116 items needing information about antecedents, 59 non-anaphoric items are left undetected.

The results of applying the linguistic rules for anaphor detection on *-ing*-, *-ed*- and *to*-items are presented as an overview in Table 4. These numbers are not yet the final results because they do not consider antecedent information from the second step, i.e. some non-anaphoric items currently categorised as false positives are detected as non-anaphoric in the process of assigning each anaphor its antecedent.

Table 4: Results of the anaphor detection

	-ing	***-ed***	***to***	**Non-finite items in total**
True positives	685	375	710	**1,770**
False negatives	7	0	1	**8**
True negatives	1,244	2,924	935	**5,103**
False positives	131	12	175	**318**
Precision	83.9 %	96.9 %	80.2 %	**99.6 %**
Recall	99.0 %	100 %	99.9 %	**84.8 %**

In sum, precision for identifying anaphoric *non-finite clause* items lies at 84.8 % with a recall of 99.6 %. This means that there are 8 false negatives and 318 false positives, of which 167 could be detected with information about antecedents. The attention will now turn to resolving the antecedent for the anaphors identified.

7.2 Identifying antecedents

Generally, antecedents of *non-finite clause* anaphors are nouns and noun phrases. Therefore, antecedent candidates are all nouns and noun phrases that occur in the same sentence as the *non-finite clause* anaphor. Units within brackets or after colons that are sentences on their own can have antecedents in the previous sentence. As will be discussed, it is also possible for an anaphor to take an entire clause as antecedent, at least in some cases.

In the following parts, the selection of the antecedent is discussed. In all rules, non-finite-clause items in coordinated phrases are treated the same, i.e. the second item is resolved analogically to the first one.

7.2.1 Rules for *-ing* and their evaluation

If assigning antecedents, different cases have to be distinguished. The discussion here breaks down the rules for anaphoric *-ing*-items into six cases that are outlined below.

1st case: if an anaphor occurs at the beginning of a sentence, the antecedent is the subject of the following matrix clause.

If a sentence begins with a subordinate clause in the form of a non-finite *-ing*-clause, which is usually an adverbial (cf. chapter 3.12.2), the antecedent is the subject of the matrix clause. An example where this rule applies is (297). Such anaphors occur at the beginning of a sentence, either without any other element preceding (example (297)) or with conjunctions (example (298)) and/or adverbs/prepositions and/or occasionally with negative particles (cf. Quirk et al. 2012: 67-68). Consequently, the antecedent is always a cataphoric element, occurring right of the anaphor. For the frequency in the corpus, 45 anaphors are found. Of these, 19 appeared sentence-initial. However, there are 3 false positives that are resolved. Additionally, 3 false positives that need information about antecedents can be excluded.

(297) <u>Having</u> taught Chinese for a number of years, **he** changed his field of specialisation to linguistics [...]. (W11)

(298) Upon <u>leaving</u> the house, **she** enters a garden, where the flowers speak to her and mistake her for a flower. (W19)

2nd case: if a noun phrase precedes the anaphor, the antecedent is this noun phrase.

Generally, if a noun phrase immediately precedes the anaphor, it is usually the antecedent.[4] The noun phrase can also consist of a pronoun, usually a central pronoun. It is also possible that an adverb occurs in between noun phrase and

[4] The Stanford parser also analyses dependencies in a clause and sentence. If a noun phrase precedes the *-ing* anaphor, the Stanford parser returns that the noun immediately preceding is the item that is postmodified by a “participial modifier” (Marneffe & Manning 2011: 8), which is, however, not always correct.

anaphor. If brackets are part of a sentence, these should be treated here as separate clauses and not considered in the search for a noun phrase.

If the noun phrase itself is preceded by *by*, the antecedent of the anaphor is this noun phrase for certain. Example (299) shows such a case. However, there must not be a comma in between *by*, the noun phrase or verb ending in *-ing*. This rule applies to 10 instances in the corpus.

Another case where the antecedent is the preceding noun phrase in every instance is if the anaphor takes the function of a postmodification of a noun phrase, as exemplified in (300). This noun phrase has to be the subject of a clause, but it need not necessarily occur at the beginning of a sentence. In this case, the noun phrase is preceded by a comma. Apart from that, the preceding noun phrase is the antecedent if the noun phrase is a prepositional complement at the beginning of a sentence, i.e. occurs after a preposition (example (301)). There are, however, only 2 cases in the corpus to which this rule applies. In sum, 14 items can be resolved by such information.

The noun phrase and the verb in *-ing* can also be interrupted by a comma (11 cases). Number (302) provides an example. Some other expressions can also occur in between anaphor and comma, if these are negative particles and/or adverbs. It is also possible to have a verb in between noun phrase and comma, i.e. a noun phrase is followed by a verb, a comma and finally an anaphor, if these occur at the beginning of a sentence (2 cases). In sum, these rules apply to 13 anaphors. Other instances are not considered with case 2, but are treated in case 5: these are instances where a comma immediately precedes an anaphor, or where an anaphor and a comma are only separated by negative particles/adverbs but no noun phrase (and verb) precedes this comma.

(299) There is general agreement that acupuncture is safe when administered by **well-trained practitioners** using sterile needles [...]. (W1)

(300) **Stone acupuncture needles** dating back to 3000 B.C. have been found by archeologists in Inner Mongolia. (W1)

(301) For **steps** taking a single beat the first half of the beat constitutes the foot movement and the second half is taken up by the hip movement. (W3)

(302) **Tevez**, outstanding last season, was unfairly shunted to one side as soon as Berbatov strolled through the door [...]. (B6)

There are cases where the preceding noun phrase is not always the antecedent: if a noun phrase is preceded by *of*, the antecedent is either only the noun phrase in the postmodification or the whole noun phrase including the postmodification with *of*. Example (303) illustrates the first case, example (304) the second. From the 36 instances in the corpus, 9 have the noun phrase including the *of*-postmodification as their antecedent. It remains debatable if such cases should be resolved because obviously only the context can determine how far the antecedent takes a noun phrase with postmodification. Nevertheless, the tendency for the anaphor is to have only the noun phrase in the postmodification as antecedent. Such a strategy would then resolve not incorrectly even those anaphors that have the whole noun phrase, including the postmodification, as antecedent, i.e. only in 3 cases the meaning would change entirely. For example, the anaphor *being* in example (304) refers to *the son of an Indian raja*. If only the postmodification in the preceding noun phrase is assumed to be the antecedent, i.e. *an Indian raja*, this is not completely wrong: the assumed antecedent still includes that the anaphor is about an Indian raja, only that it is not the Indian raja himself but his son who is meant. However, in cases such as *the Telecom act of 1996* (B8) it would be semantically totally false if an anaphor item just referred to the postmodification *1996*. It is unlikely that people automatically think of the Telecom act when only the year 1996 is mentioned, i.e. the year itself has nothing to do with the Telecom act.

(303) [...] American visitors to China brought back firsthand reports of **patients** <u>undergoing</u> major surgery using acupuncture as their sole form of anesthesia. (W1)

(304) Another version of the legend has **the son of an Indian raja** <u>being</u> rescued from a tiger by one of Grey's servants. (W6)

Finally, 133 items where a noun phrase precedes are then remaining from the previous steps. For these noun phrases it is important to pay attention to whether they are coordinated or not. If coordination occurs, the antecedent is the whole expression with all coordinated phrases. Items are also treated as case 2, if an adverb or a negative particle in between anaphor and noun phrase is present (9 cases). From 133 items, the rule that the antecedent is the preceding noun phrase applies to all instances, except for 17 cases. From these 17 instances, 7 items do have the preceding noun phrase as antecedent, but it is the whole noun phrase including the postmodification. The postmodification here is realised by a prepositional phrase other than an *of*-phrase that contains a noun phrase as prepositional complement. In 13 of 17 cases, the subject realised by a noun phrase is the antecedent. In 11 out of these 13 cases, the subject is human, but the preceding noun phrase is not human. Despite this discovery, it does not make sense to establish a rule that says that if the subject is human and the preceding noun phrase is not, the subject but not the preceding noun phrase is the antecedent: such a rule would lead to 17 items out of 133 being resolved incorrectly and therefore about the same number of false antecedents would be produced than without the rule. It is therefore better not to implement this rule, at least for this corpus.

In sum, 186 anaphoric items are resolved correctly with case 2 and 20 resolved with a false antecedent. However, also 16 false positives are resolved. Additionally, from 6 items needing information about antecedents – including two extrapositions – 5 are detected as non-anaphoric because of the preceding noun phrase, and 1 is not excluded but resolved.

3rd case: if a verb precedes the anaphor, the antecedent is the subject to this preceding verb.

If a verb precedes the anaphor, the antecedent is the subject to the verb (example (305)). With this case, adverbs can again occur in between anaphor and verb. In the corpus, this rule applies to 67 items, including one false antecedent. Case 3 also detects 9 items as non-anaphoric – including 1 extraposition – that need information about antecedents. Additionally, 1 false positive is excluded as there is no subject present. However, 6 false positives from the context category are resolved.

(305) **ManU** will start getting better results in the second half of the season [...]. (B6)

4th case: the antecedent is the preceding noun phrase, the subject of the preceding clause, or the subject or object of the matrix clause, depending on the preposition or conjunction that precedes the anaphor.

Items such as *after* or *than* can be either preposition or conjunction (cf. Quirk et al. 2012: 660-661). Consequently cases with prepositions and conjunctions are treated here together.[5] The prepositions and conjunctions that can precede an *-ing*-anaphor are given in Table 5. They were selected from Quirk et al.'s list (2012: 665-667, 920) regarding the likelihood to occur together with a non-finite clause.

[5] Cases where the preposition or conjunction occurs at the beginning of a sentence are not considered here as they were treated with case 1.

Table 5: Prepositions and conjunctions with *-ing*-anaphors

as	*about*	*into*	*and*
at	*above*	*notwithstanding*	*or*
but	*across*	*over*	*nor*
by	*after*	*throughout*	*both*
for	*along*	*toward(s)*	*either*
from	*among(st)*	*under(neath)*	*neither*
in	*around*	*unlike*	*so that*
like	*before*	*until*	
of	*behind*	*upon*	and all conjunctions
on	*below*	*versus*	mentioned in Table 28
since	*beneath*	*within*	of chapter 3
than	*beside*	*without*	
through	*between*		
till	*beyond*		
to	*despite*		
via	*during*		
with	*except*		

If a conjunction precedes the anaphor, the antecedent is usually the subject of the matrix clause, if not stated otherwise. However, with *and*, *or*, *but* and other coordinating conjunctions, the antecedent is the subject of the preceding clause.[6] With *unless*, the antecedent is not the subject, but rather the object of the matrix clause. If *by* occurs in the construction "prepositional phrase + *by* + verb in *-ing*" at the beginning of a sentence, the antecedent is also the subject of the matrix clause, which, however, follows the anaphor. This occurs twice in the corpus. Furthermore, if a preposition is not preceded by a noun phrase, or if it is only preceded by a noun phrase (as in example (306)), it is not resolved because it is likely to be non-anaphoric (except for cases where a preposition and a verb in *-ing* is sentence-initial; see case 1). However, if an *of*-phrase post-modifies a noun phrase within a prepositional phrase (example (307)), it is anaphoric. Such a structure is found once in the corpus.

A few further cases for prepositions are noteworthy: with *into*, the antecedent is not the subject, but rather the preceding object of the matrix clause (example (308)). As to *from*, the situation is slightly different. Such anaphors usually refer to the object of the matrix clause, if given. If there is no object, the antecedent is also the subject of the matrix clause. However, if *from* is later followed by *to* (example (309)), the antecedent is the subject (occurring once in

[6] If two anaphoric non-finite clause items are coordinated, this is not considered here but in the relevant sections.

the corpus). With *for*, it is presumably better not to resolve these items because they take the subject or the object as antecedent. Only 11 out of 22 have the subject as antecedent. Moreover, 8 non-anaphoric items would not be excluded if such anaphors were resolved.

(306) The usual method of *catching* lobsters has been to use baited, one-way traps [...]. (W10)

(307) After three years of <u>being</u> a housewife, <u>looking</u> after her daughter Maggie and not <u>working</u>, **Margot** decided it was time to let her emotions take control and get back into acting. (WS35)

(308) It is not helped by the fact that the local company with which we have a supply agreement employs a high-pressure salesman to answer the phone, quote the latest price per liter and coax **customers** into <u>buying</u> as many liters as possible. (B20)

(309) **Clarks** have always taken our role in the community seriously – from <u>providing</u> education and housing for our very first workers to sup-port-ing international initiatives initiatives to improve people's lives today. (WS2e)

An overview of how many anaphoric items occur with the individual prepositions and conjunctions and to what extent they are resolved correctly is given in Table 6:

Table 6: Anaphoric items of case 4

Preposition/conjunction	Anaphors in total	Correct antecedent identified	Wrong antecedent identified	Not resolved
about	6	6	0	0
after	14	11	3	0
and	10	9	1	0
as	5	3	2	0
at	5	5	0	0
before	7	7	0	0
but	2	2	0	0
by	37	37	0	0
despite	2	1	1	0
for	22	0	0	22
from	11	11	0	0
in	25	22	3	0
into	2	2	0	0
of	32	28	2	2

on	7	7	0	0
or	1	1	0	0
than	2	1	1	0
through	1	1	0	0
to	17	15	0	2
towards	1	1	0	0
when	1	1	0	0
while	8	6	2	0
with	1	1	0	0
without	5	5	0	0
In total	**224**	**183**	**15**	**26**

As for the non-anaphoric items, i.e. false positives and how these are treated, these are shown in Table 7. With the prepositions *at*, *in* and *of*, items are excluded due to the condition of only a noun phrase preceding. From the false positives of *to*, one item can be excluded if the fixed expression *due to* is explicitly excluded. In the case of *about* and *without*, the non-anaphoric items are eliminated because no subject is present in the clause.

The table contains both non-anaphoric items of the context category as well as of the category where information about antecedents is needed. For only the items needing information about antecedents, the individual items are as follows: 1 item occurs each with *about*, *like* and *to*, 2 items each with *by*, *for*, *on*, *in* (and 1 item is resolved as a false positive with *in*) and 9 items occur with *of*. Furthermore, 1 item each is found with *unless* and *or* and 3 items with *and*, the latter including 2 extrapositions.

Table 7: Non-anaphoric items with case 4

Preposition/conjunction	Items in total	Correctly excluded	Not excluded but resolved
about	3	2	1
and	5	4	1
as	5	0	5
at	1	1	0
by	3	2	1
for	10	10	0
from	1	1	0
in	5	3	2
like	4	1	3
of	19	14	5
on	3	2	1
or	1	1	0
than	1	0	1

to	4	2	2
unless	1	1	0
with	1	0	1
without	3	1	2
In total	**70**	**45**	**25**

5th case: if a comma or dash precedes the anaphor, the antecedent is the preceding noun phrase, the subject of the matrix clause or it is the full preceding matrix clause.

Generally, if the anaphor is preceded by a comma or a dash (Quirk et al. 2012: 1612, 1629, 1636) the antecedent is the subject of the preceding matrix clause (example (310)).[7] It is also possible that an adverb and/or negative particle occurs between anaphor and comma. These rules hold for 78 out of 135 cases. Furthermore, the antecedent of *including* is the preceding noun phrase (17 cases). If there is no immediately preceding noun phrase before the comma but a verb phrase, the preceding matrix clause is the antecedent (1 case). Noun and verb phrases must not be at the beginning of the sentence because then they belong to case 2. The remaining 39 items are hard to exclude or resolve correctly because here it depends on the context as to what the antecedent is. Furthermore, 12 false positives are resolved and 3 non-anaphoric items that need information about antecedents are excluded correctly.

(310) Since then **she** has sustained her career in film, television, and theater, recently <u>appearing</u> in a Canadian stage production [...]. (WS35)

6th case: if an anaphor occurs after colons, an opening mark for bracketing or after an adjective, the antecedent is the preceding matrix clause or the subject of the matrix clause.

Case 6 subsumes some minor important instances. First, if an anaphor occurs within a sentence, after an opening mark for bracketing, the antecedent is the preceding matrix clause, outside the brackets (example (311)). Between bracket and anaphor an adverb can occur. This case applies to 4 instances in the corpus at the expense of 2 false positives being resolved as well. Second, if the anaphor occurs after colons, the antecedent is the subject of the matrix clause before the colon (example (312)). By that rule, 2 items are resolved. Third, if the anaphor is

[7] Case 5 does not involve instances of syndectic or asyndetic coordinated phrases that are part of any of the cases above (cf. Quirk et al. 2012: 918).

preceded by an adjective, the antecedent is the subject of the matrix clause (2 items in the corpus). In sum, case 6 resolves 8 anaphors and 2 false positives.

(311) **We will sometimes ask you** (depending on the type of job you're applying for) to take part in a job specific selection exercise [...]. (WS2b)

(312) **They** act as a bridge between people and their rulers: representing people's interests to the government [...]. (WS72)

In total, the six cases detect the correct antecedent of 584 anaphors (see Table 8). Furthermore, 75 false antecedents are identified. There are 26 anaphoric items that are not resolved due to the likelihood of taking a false antecedent. Of the 131 false positives, 66 are excluded but 65 are resolved. Here, 45 of 66 and 2 of 65 are items that were categorised as needing information about antecedents. Consequently, precision for identifying anaphoric elements with *-ing* is in reality considerably higher and lies at 91.3 % and recall amounts to 99.0 %. The precision for resolving *-ing* anaphors is 88.6 %, recall is 85.3 %. The number for recall in anaphora resolution does not consider the seven anaphors that have not been detected in step one (cf. chapter 7.1.1), i.e. it is here calculated as 584 (anaphors resolved correctly) divided by 685 (anaphors detected in step one).

Table 8: Resolution of *-ing*-items[8]

Case	Anaphors resolved correctly	Anaphors resolved wrongly	Anaphors not resolved	False positives resolved	False positives excluded
1	45			3	3
2	10 (*by*)				
	14 (pm./pc.)				
	13 (NP, verb)				
	33 (*of*)	3 (*of*)			
	116 (NP)	17 (NP)		17 (NP)	5 (NP)
3	66	1		6	10
4	183	15	26	25	45
5	96	39		12	3
6	4 (brackets)			2 (brackets)	
	2 (colons)				
	2 (adjectives)				
In total	**584**	**75**	**26**	**65**	**66**

[8] The abbreviation *pm.* stands for "postmodification", *pc.* for "prepositional complement" and *NP* for "noun phrase".

7.2.2 Rules for *-ed* and their evaluation

The rules for anaphoric items ending in *-ed* are similar to those ending in *-ing*. As there are only 12 non-anaphoric items left from the previous step of anaphor detection, not so much attention needs to be paid to their exclusion as with *-ing*-items.

1st case: if an anaphor occurs at the beginning of a sentence, the antecedent is the subject of the following matrix clause.

With *-ed*-items, 9 anaphors (7 regular, 2 irregular) are resolved correctly in case 1. Additionally, two take a false antecedent because not the whole subject in the matrix clause but only the determinative of the phrase – here in the form of a possessive pronoun – is the antecedent. Additionally, 2 non-anaphoric items are resolved and 1 item that needs information about antecedents can be excluded.

2nd case: if a noun phrase precedes the anaphor, the antecedent is this noun phrase.

One type where the preceding noun phrase is the antecedent for sure is if it is preceded by the preposition *by*. This holds for only one instance in the corpus. There are 28 cases (19 regular, 9 irregular) where the anaphor functions as postmodification within a noun phrase in subject position. Of these, 2 regular anaphoric items take noun phrases as antecedents that are part of a prepositional complement at the beginning of a sentence. Furthermore, there are 8 regular and 1 irregular items where the noun phrase and the verb in *-ed* are interrupted by a comma.

Moreover, 24 items (17 regular, 7 irregular) occur where *of* precedes the noun phrase. Here, the tendency is again that the antecedent is only the immediately preceding noun phrase. This does not hold for 6 items. However, the meaning of the antecedent of these 6 items is in all cases not completely different and so these are not seen as false antecedents. Finally, the remainder of 167 items (136 regular, 31 irregular) take the preceding noun phrase as antecedent. There are 2 exceptions, i.e. producing 2 false antecedents. In sum, the subtypes in this case account for 227 anaphoric items and 2 false antecedents. Additionally, 2 false positives of irregular forms are resolved and 3 items that need information about antecedents are correctly excluded.

3rd case: if a verb precedes the anaphor, it is not resolved.

The third case of *-ing* does not occur with *-ed*-items. However, there is 1 anaphoric item in the corpus that occurs after a verb. This can be explained by a lack of a comma that should be present between the verb and *-ed*-item. In such cases, the anaphor is not resolved.

4th case: the antecedent is the subject of the preceding clause, or the subject or object of the matrix clause, depending on the preposition or conjunction that precedes the anaphor.

Generally, anaphoric items of this case have the subject of the preceding clause or matrix clause as antecedent, according to the conditions and exceptions stated with *-ing*. However, the exception with *of* does not apply for anaphoric -*ed*-items. Furthermore, *if* has the the object of the matrix clause as antecedent. In sum, 18 items from the corpus fall into the fourth case category. Only 5 types of prepositions and conjunctions occur: *as* (11 items, 2 of these are irregular), *but* (2 items), *if* (1 item), *though* (2 items), *when* (2 items). From the 18 items, 6 (with *as*) take a false antecedent. Finally, 3 non-anaphoric items are resolved (2 with *if*, 1 irregular item with *as*). There is no need – at least for this corpus – to exclude items whose prepositions or conjunctions are preceded by no noun phrase or a noun phrase only, due to the low number of non-anaphoric items.

5th case: if a comma or dash precedes the anaphor, the antecedent is the subject of the matrix clause, or it is the full preceding matrix clause.

According to the conditions stated with *-ing*-items, 31 (26 regular, 5 irregular) out of 42 anaphoric items have the subject of the preceding matrix clause as antecedent. One regular item has the whole preceding matrix clause as antecedent because a verb occurs before the comma. Furthermore, 1 regular anaphoric item is not resolved, as the antecedent is wrongly believed to be *there*. So the item is excluded, assumed to be non-anaphoric. The remainder of 9 items (6 regular, 3 irregular) take a false antecedent. Furthermore, 1 false positive of an irregular form is correctly excluded.

6th case: if an anaphor occurs after colons, an opening bracket or after an adjective, the antecedent is the preceding matrix clause or the subject of the matrix clause.

This category covers three minor cases again and follows the conditions outlined with *-ing*-items. There are 74 instances (73 regular, 1 irregular) where the anaphor appears at the beginning of an opening bracket. The anaphor here can be optionally preceded by an adverb, a conjunction and/or a negative particle. Of these 74 instances, 71 are correctly resolved, taking the preceding matrix clause as antecedent. No anaphors are found after colons or adjectives in the corpus, although these cases are possible with *-ed* as well.

In sum, 351 anaphoric items are resolved correctly, 22 take a false antecedent (see Table 9). Furthermore, two anaphoric items are not resolved. Of the 12 remaining false positives, 7 are resolved, 5 excluded. Consequently, the precision rate for detecting anaphors with information from anaphora resolution is 98.2 % with a recall of 100 %. The precision rate of anaphora resolution is 94.1 %, and recall is 93.6 %.

Table 9: Resolution of *-ed*-items

Case	**Anaphors resolved correctly**	**Anaphors resolved wrongly**	**Anaphors not resolved**	**False positives resolved**	**False positives excluded**
1	9	2		2	1
2	1 (*by*)				
	28 (pm./pc.)				
	9 (NP, verb)				
	24 (*of*)				
	165 (NP)	2 (NP)		2 (NP)	3 (NP)
3			1		
4	12	6		3	
5	32	9	1		1
6	71 (brackets)	3 (brackets)			
In total	**351**	**22**	**2**	**7**	**5**

7.2.3 Rules for *to* and their evaluation

How antecedents are found with *to* is similar to *-ing*- and *-ed*-items. Nevertheless, some rules can be left out, whereas some additional rules are partially necessary. Generally, the most frequent are the cases 2 and 3.

1st case: if an anaphor occurs at the beginning of a sentence, the antecedent is the subject of the following matrix clause.

This first case is not that important for *to*-items because only 4 anaphoric items have the subject of the following matrix clause as antecedent. Furthermore, 1 non-anaphoric item is resolved, 1 non-anaphoric item detected and excluded through anaphora resolution.

2nd case: if a noun phrase precedes the anaphor, the antecedent is this noun phrase, or it is the subject of the clause.

If *by* is followed by a noun phrase and then by an anaphor, this noun phrase is usually the antecedent, but not for certain with *to*-items. Only 5 such items can be found in the corpus. Of these, 4 items are anaphors, of which 2 are resolved correctly, 2 incorrectly. One non-anaphoric item preceded by *by* is also resolved. In addition, 5 anaphors that postmodify a noun phrase and 3 anaphors that are preceded by a noun phrase in a prepositional complement are found. The latter takes 1 false antecedent. Two non-anaphoric items postmodifying noun phrases and 4 items postmodifying noun phrases in prepositional complements are not excluded.

Furthermore, cases where a comma occurs between a noun phrase and *to* cannot be found, although the rules are also valid for *to*-items. Additionally, 22 items involving *of* occur: 6 have the antecedent as only the noun phrase before, 7 include the noun phrase with the *of*-postmodification, 1 item refers to the whole preceding clause and the antecedent of 8 items is the subject of the clause. Three non-anaphoric items occur with *of*. As the antecedent varies frequently, it makes sense not to resolve these anaphors.

The remainder of anaphors taking the preceding noun phrase as antecedent produces many wrong antecedents, i.e. it applies to 131 anaphoric items out of 234, which means that 103 would be resolved with a wrong antecedent. Furthermore, 28 of 52 false positives need anaphora resolution for exclusion, but only 5 of these 28 items could be correctly excluded.

Due to the high number of items taking a false antecedent, it makes sense to include a further rule that also considers the subject of the clause: if the subject is about a human denotation, e.g. names, personal pronouns referring to people, or nouns such as *speaker*, *players* and the immediately preceding noun phrase is not human, the antecedent is the subject. However, if both the subject and the preceding noun phrase are human, the subject is not preferred. This rule leads to the correct resolution of 67 out of the 103 items that would have taken a wrong antecedent. Additionally, only 10 items are resolved from those requiring information about antecedents, 18 are correctly excluded as their non-anaphoric status is clarified. This means that 34 out of 52 non-anaphoric items are resolved. However, 22 items of 131 correctly resolved anaphoric items are then assigned a false antecedent. Nevertheless, the rule is useful as the number of false antecedents is halved.

In sum, 61 anaphors take a false antecedent, and 185 anaphors are resolved correctly in case 2. However, 22 anaphors are not resolved. Furthermore, 41 non-anaphoric items are resolved and 21 can be detected as non-anaphoric.

3rd case: if a verb precedes the anaphor, the antecedent is the subject to this preceding verb.

This case is frequent with *to*-items: 357 instances occur in the corpus in sum.[9] Of these, 2 anaphoric items are resolved with the wrong antecedent. If there is no subject, the item should not be resolved, which also rules out 2 false positives. Apart from that, 8 false positives are resolved. Additionally, 57 non-anaphoric items whose status can be established by resolving the items are correctly excluded, 2 further items are wrongly resolved.

4th case: the antecedent is the subject of the preceding clause or of the matrix clause, depending on the preposition or conjunction that precedes the anaphor.

With *to*, all items have the subject of the preceding clause or of the matrix clause as antecedent. Nine items occur with *in order to*, 2 items with *but* and 1 item each with *and* and *or*, making up 13 items in sum. One item takes a false antecedent with *and*. Additionally, there are some prepositions that are part of verbs but immediately precede the anaphor. The antecedent here is the subject of the matrix clause. This occurs with *down* (1 item), *in* (1 item), *on* (3 items), *out* (2 items), *up* (2 items). Moreover, 2 false positives with *and*, 1 item each with *than* and *when* are resolved. One false positive is not resolved with *except* as there is no subject present. As to the detection of non-anaphoric items through anaphora resolution, 3 items can be excluded (2 with *and* and one with *out*, which is part of the verb), 2 items (both *and*) are not detected but resolved.

[9] Biber et al. (2007: 698, 722) also found that most *to*-infinitive clauses show this pattern.

In sum, 22 items are resolved with the correct antecedent, 1 item takes a false antecedent. Furthermore, 6 false positives are resolved and 4 non-anaphoric items are detected and so excluded.

5th case: if a comma or dash precedes the anaphor, the antecedent is the subject of the matrix clause.

Again, this rule is less complex with *to*. Ten items are resolved correctly, having the subject of the preceding matrix clause as antecedent and 2 of these have a preceding dash. One item takes the subject of the matrix clause that follows the anaphor as antecedent. One further item takes a false antecedent. From the non-anaphoric items whose status becomes clear through anaphora resolution, 3 items are identified and excluded, 1 is resolved. Additionally, 2 false positives are resolved, 1 is excluded. In sum, 4 non-anaphoric items can be detected, 3 are resolved.

6th case: if an anaphor occurs after colons, an opening bracket, after an adjective or after *how/what to*, the antecedent is the preceding matrix clause, the subject of the matrix clause, or the preceding noun phrase.

There are no cases involving colons or brackets in the corpus. However, there are some instances where the anaphor follows an adjective, in which instance the antecedent is the subject of the matrix clause. Negative particles and adverbs can occur in between anaphor and adjective. Here, 36 anaphors are resolved, 5 further items take a false antecedent. Four false positives can be excluded; 2 false positives are resolved. Furthermore, 19 non-anaphoric items are excluded due to the search for antecedents. In sum, 23 non-anaphoric items are detected and excluded, 2 are resolved.

Finally, there are cases where *to* following *how* or *what* are anaphoric. This is if a noun phrase or a noun phrase followed by a verb phrase precedes. The antecedent then is this preceding noun phrase. The preceding noun phrase must not be in the genitive because then the *to*-item is non-anaphoric. Two anaphoric items with *how* and one item with *what* occur in the corpus.

In sum, 618 anaphors can be assigned the correct antecedent, 22 are not resolved and 70 take a false antecedent (see Table 10). Consequently, a precision rate of 89.8 % and a recall of 87.0 % is achieved. The number for recall does not consider the one anaphor that has been discarded in the anaphor detection step. Of the 116 non-anaphoric items that need anaphora resolution, 101 can be excluded and 15 are resolved. Additionally, from the remaining 59 false positives, 48 are resolved, only 11 can be excluded. All this leads to a precision rate for identifying anaphoric items of 91.8 % and a recall of 99.9 %.

Table 10: Resolution of *to*-items

Case	**Anaphors resolved correctly**	**Anaphors resolved wrongly**	**Anaphors not resolved**	**False positives resolved**	**False positives excluded**
1	4			1	1
2	2 (*by*)	2 (*by*)		1 (*by*)	
	7 (pm./pc.)	1 (pc.)		6 (pm./pc.)	
			22 (*of*)		3 (*of*)
	176 (NP)	58 (NP)		34 (NP)	18 (NP)
3	357	2		10	59
4	22	1		6	4
5	11	1		3	4
6	36 (adjectives)	5 (adjectives)		2 (adjectives)	23 (adjectives)
	2 (*how to*)				
	1 (*what to*)				
In total	**618**	**70**	**22**	**63**	**112**

7.3 Conclusion

In this chapter, linguistic rules for resolving *non-finite clause* anaphors in computational systems have been proposed. They have been developed from information in grammar books (e.g. Quirk et al. 2012) and from a comprehensive analysis of the hypertext corpus. These rules have then been evaluated manually on the hypertext corpus in order to show their accuracy.

The task of anaphora resolution with *non-finite clause* anaphors has been split up into two parts. First, the detection of anaphors has been examined. Based on the categorisation proposed in chapter 3, anaphoric and non-anaphoric items have been distinguished. Besides, a few more rules have been established to account for non-anaphoric items of the context category. The context category subsumes non-anaphoric items that do not fall into any other category. With these rules of the anaphor detection stage, a precision rate of 84.8 % and a recall of 99.6 % can be achieved. This means that 318 false positives have been identified and 8 anaphors have not been detected.

In the second step, each anaphor has been assigned its antecedent. To do that, various rules have been designed that exploit the syntactic structure of sentences. With these, 1,553 out of 1,778 anaphors (or better, of 1,770 left from the first stage) have been assigned a correct antecedent. As a result, the precision rate for anaphora resolution lies at 90.3 % and the recall is 87.7 %. This means that 167 or 9.4 % of all anaphors have taken a false antecedent, 50 anaphors have not been resolved because of the high chance to be assigned a wrong antecedent. From the 318 false positives identified in the first stage, 135 have been resolved and 183 have been detected as non-anaphoric in stage two and have therefore been excluded. Consequently, the precision rate with anaphor detection of *non-finite clause* items has risen to 92.9 % with a recall of 99.6 %, due to the information available in stage two. The final results are summarised in Table 11.

Table 11: Final results of the anaphora resolution process with *non-finite clause* items

	-ing	*-ed*	*to*	***Non-finite clause* items in total**
Identification of anaphors				
Anaphors identified	685	375	710	**1,770**
Anaphors not identified	7	0	1	**8**
Non-anaphoric items identified	1,310	2,929	1,047	**5,286**
Non-anaphoric items not identified	65	7	63	**135**
Precision of anaphor identification	91.3 %	98.2 %	91.8 %	**92.9 %**
Recall of anaphor identification	99.0 %	100.0 %	99.9 %	**99.6 %**
Resolution of anaphors				
Anaphors resolved correctly	584	351	618	**1,553**
Anaphors resolved wrongly	75	22	70	**167**
Anaphors not resolved	26	2	22	**50**
Precision of anaphora resolution	88.6 %	94.1 %	89.8 %	**90.3 %**
Recall of anaphora resolution	85.3 %	93.6 %	87.0 %	**87.7 %**

The results that have been achieved by the established rules are pretty good, especially for anaphora resolution evaluations where values around 50 % are commonly achieved (compare with the benchmark results of central pronouns in chapter 6.5.2). Thus, in answering the second research question posed in chapter 4.5, it is to be expected that these promising linguistic rules produce far-reaching effects when used in computational anaphora resolution systems.

8 Conclusion

This book examined anaphora resolution from a linguistic point of view and with respect to its application in computational systems. The second chapter introduced important terms and concepts. It was shown that many current definitions of anaphors lack essential characteristics, with the consequence that the term “anaphor” is then only used for prototypical examples. Such definitions, for instance, usually name coreference as one criterion, i.e. anaphor and antecedent have to show a coreferential relation. But in that case, items with substitutional relations are not regarded as anaphors. Furthermore, computational approaches towards anaphora resolution do not usually discuss which conditions have to apply to items in order for them to be considered anaphors. Do anaphora resolution systems then also treat cataphoric items? And what about anaphors referring to entities that are not noun phrases? Chapter two proposed six conditions that have to hold for anaphoric items in this book. As a consequence, additionally to “prototypical” anaphors, anaphors that have not attracted much attention are investigated. Cataphors are included as special forms of anaphors and items are only analysed as anaphoric that have an explicit antecedent in the text. The interpretation of the anaphor is then derived more or less from the antecedent as anaphors lead to a reduction of the text and/or avoid excessive repetition. The relationship between anaphor and antecedent need not necessarily be coreferential, it can also be substitutional or it can belong to a third, minor miscellaneous category that comprises anaphors that are not specifically coreferential or substitutional. Finally, the resolution of anaphors contributes to the disclosure of the text content because anaphors are cohesive devices.

The third chapter proposed a new categorisation of anaphors, based on the conditions defined in chapter two. Current categorisations have not been adopted here as they are unsatisfactory and not complete from the point of view of the six conditions that have to hold for anaphors (cf. chapter 2.5). Standard grammar books (e.g. Huddleston & Pullum 2010; Quirk et al. 2012) lack important anaphor types, such as *non-finite clause* anaphors, or some items within these types. Similarly, categorisations in computational approaches (e.g. Mitkov 2002) show deficiencies because they are too unspecific for a linguistic discussion. The anaphor *it*, when referring to antecedents in the form of clauses or sentences, is not part of any of Mitkov’s categories. In consequence, a new categorisation has been established, with the aim of accounting for linguistic aspects as well as being of use in computational systems. The twelve categories comprise central pronouns, reciprocal pronouns, demonstrative pronouns,

DOI 10.1515/9783110416756

relative pronouns, adverbs, noun phrases with a definite article, proper names, indefinite pronouns, other forms of coreference and substitution, verb phrases with *do* and combinations with *so*, *this*, *that*, *it*, *the same (thing)*, ellipses and non-finite clauses. The characteristics of these anaphor types have then been specified, the individual items of each anaphor types presented and their anaphoric and non-anaphoric uses detailed.

In chapter four the frequency of anaphors in hypertexts was analysed. So far, the relative frequency of anaphor types as proposed in chapter three has not been investigated. Furthermore, there has been no corpus of hypertexts encompassing different hypertext types. Consequently, a corpus was established. Rehm's (2007) classification of hypertext types was adapted, which led to three categories for the purpose of this book: Wikipedia texts, blog texts and traditional website texts. Wikipedia texts were chosen as examples of an online encyclopedia. Half of all blog texts come from online newspapers, about half of the blogs are from companies and the remainder are some personal blogs and blogs from organisations. Traditional website texts consist of homepages from companies, personal homepages, institutional homepages and homepages from online newspapers. Thus, these categories represent important text forms of the Internet and are therefore highly relevant for text retrieval. The texts of each hypertext type contain about 25,000 words, amounting to approximately 76,000 words for the whole corpus.

Subsequently, this corpus has been analysed: *non-finite clause* anaphors turned out to be the most frequent type of anaphor, even outnumbering central pronouns. *Non-finite clause* anaphors amount to 29.1 % of all anaphors, followed by central pronouns with 27.4 % (cf. chapter 4.4.1). This result is particularly significant as no classification or study so far has treated non-finite clauses as one type of anaphor. As a consequence, it is of utmost importance that anaphora resolution systems consider *non-finite clause* anaphors and incorporate their resolution. Due to the substantial number of *non-finite clause* anaphors, profound effects can be expected in applications where an understanding of natural language is imperative (cf. chapter 6.1), such as in text retrieval systems. Other noteworthy anaphor types in the corpus are proper names, relative pronouns, noun phrases with a definite article, demonstrative pronouns and ellipses. Consequently, they are also vital for anaphora resolution. A comparison to previous studies is only limitedly possible because few surveys so far examine the distribution of different anaphor types. Most concentrate on just one type of anaphor, i.e. central pronouns (cf. chapter 4.1). Only the Syracuse study (e.g. Liddy 1990) carried out a study that included the largest number of different anaphor types. In her corpus of 600 abstracts, noun phrases with a definite article are the most frequent, followed by relative pronouns and central pronouns (cf. chapters 4.1 and 4.5). The comparability, however, is restricted, as this corpus, among other things, consists only of abstracts and as it does not consider *non-finite clause* anaphors. Finally, it is essential for anaphora resolution systems to distinguish between anaphoric and non-anaphoric uses of items. Thus, the ratio of anaphoric and non-anaphoric uses of each item and anaphor type has been calculated. The numbers vary with each anaphor type, but on average, an item is anaphoric in 27 % of all instances and non-anaphoric with a likelihood of 73 %. A high number of non-anaphoric uses, e.g. with indefinite pronouns where about 95 % of all items are non-anaphoric, means that an anaphora resolution system has to pay more attention to devising rules that exclude non-anaphoric items, in order not to elicit too many false positives.

The fifth chapter examined anaphors in the context of text retrieval. Hypertexts are frequently accessed by using text retrieval systems, i.e. search engines, so it is necessary to know how text retrieval works. As a rule, Web retrieval systems look at each hypertext and extract important terms. These terms are then stored in an index where they can be compared with the query of a user. Furthermore, natural language processing methods, such as sentence delimitation, tokenisation, stop word detection, stemming and lemmatisation, part-of-speech tagging and partial or full parsing, can help to find out about the content of a text and some methods can also condense more words to one term. However, current Web retrieval systems do not use many natural language processing methods, but rather restrict themselves to methods that are absolutely necessary and that are not cost-intensive, e.g. tokenisation. Anaphors and anaphoric relations are frequently ignored in Web retrieval. Central pronouns, for example, are normally considered as stop words and if stop word detection is applied, they are not represented in the index.

Chapter six looked at approaches to anaphora resolution from a computational point of view and discussed current uses of anaphora resolution systems. It was shown that only a few studies investigate the effect of anaphora resolution on text retrieval. However, anaphora resolution can be helpful for term representation in the index and for proximity searching. Chapter six also showed that anaphora resolution systems can be divided up into rule-based and data-based methods. Although rule-based methods consider some linguistic information, both rule-based and data-based methods only take into account as much linguistic knowledge as is absolutely necessary. This has fatal consequences: first, no current anaphora resolution system returns a satisfactory value for precision and recall. For instance, benchmark results for central pronouns revealed that both the precision and the recall is around 50 % (cf. chapter

6.5.2). A main reason for such a low performance is the lack of comprehensive linguistic knowledge incorporated in these systems. It demonstrates that natural language cannot be adequately analysed only by statistical or related means. They have to be complemented by methods that look at how language is structured and used: "it is widely agreed that more linguistic knowledge can indeed play a role in improving today's statistical systems, in all phases of the process" (Way 2010: 568). This insight has not yet arrived in anaphora resolution and text retrieval systems. Second, anaphora resolution systems do not encompass all anaphor types. Current systems rather focus on the resolution of central pronouns and on the resolution of anaphors that are noun phrases, i.e. pronouns, noun phrases with a definite article and proper names. Part of the reason why these anaphors are examined is that they are more striking on first sight and can be handled easier. However, the most frequent anaphor type – *non-finite clause* anaphors – has not been considered in any anaphora resolution system so far. Third, systems do not include anaphoric relations in their entirety. The majority of anaphora resolution systems are only concerned with coreferential relations. Thus, anaphora resolution systems are, not only from a linguistic point of view, pretty unsatisfactory. In addition, chapter six showed that no current corpus considers all anaphor types as defined here, which is why the hypertext corpus compiled in this book has been annotated with anaphoric relations. It can therefore be used for computational anaphora resolution purposes, whether in machine learning approaches or in evaluations of different anaphora resolution systems.

Chapter seven presents linguistic rules that can be implemented in anaphora resolution systems. Here, rules for detecting and resolving *non-finite clause* anaphors were established exemplarily because for that anaphor type, no rules are available so far. As a basis for these rules, the information about non-finite clauses given in chapter three as well as the findings of a thorough analysis of the hypertext corpus have been exploited. These rules for *non-finite clause* anaphors have then been evaluated with the precision and recall measures: anaphors in sum were detected with a precision rate of 92.9 % and a recall of 99.6 %. In other words, 135 non-anaphoric items cannot be identified as such and are wrongly assigned an antecedent in the end, 5,286 non-anaphoric items could be excluded and only 8 out of 1,778 anaphors are not detected. Generally, the ratio between anaphoric and non-anaphoric uses with *non-finite clause* items is average in the corpus, so excluding non-anaphoric items does not need proportionally more consideration than with other anaphor types (cf. chapter 4.4.2). The precision for resolving anaphors, i.e. assigning each anaphor the correct antecedent, amounts to 90.3 % and the recall rate is 87.7 %. This means that 1,553 anaphors out of the 1,770 anaphors left from the detection stage are assigned a correct antecedent, 167 take a false antecedent and 50 anaphors are not resolved due to the likelihood of taking a false antecedent.

This book thus contributed to research in anaphora resolution in several ways. Most importantly, anaphors and anaphora resolution so far have been restricted to either a linguistic or a computational discussion. Here, an interdisciplinary approach was taken that unifies the two perspectives. A pure linguistic debate is doomed to irrelevance from a computational point of view. A computational treatise is in danger of missing fundamental linguistic knowledge that is needed for an anaphora resolution system to work with a satisfying degree. Consequently, a classification of anaphors was developed based on six conditions, which were established with respect to both a linguistic and computational perspective. In this classification, the anaphoric cases were distinguished in detail from non-anaphoric uses. Such an exhaustive discussion of which items are considered and in which uses these are regarded as anaphors is not often found when anaphora resolution and anaphora resolution systems are presented and surely not systematically for all anaphor types.

In order to analyse the frequency of anaphor types, anaphors and non-anaphoric uses of items and as no current corpus so far represents various types of hypertexts, a hypertext corpus consisting of different hypertext types has been established. The corpus has also been annotated with anaphoric relations, so that it can be used in anaphora resolution systems and their evaluation. The intention is to make the corpus publically available, so that it can be of benefit for future projects on anaphora resolution as it includes all anaphor types. The corpus analysis revealed that *non-finite clause* anaphors are the most frequent anaphor types. They were therefore chosen for the establishment of rules for text retrieval systems. Good evaluation results in terms of precision and recall have been achieved and so show promising prospects for a future implementation of these rules.

Several implications for future work on anaphors and anaphora resolution can be inferred from the results of the present study. The detailed linguistic information given for the anaphor types about anaphoric and non-anaphoric uses is worth examination not only for *non-finite clause* anaphors but also for all other anaphor types. These rules have then to be compared to rules that are already available for anaphora resolution. Missing or additional aspects have to be added. Consequently, these rules as well as the rules described for *non-finite clause* anaphors should be implemented. This would also require that the items of each anaphor type – as was done for *non-finite clause* items – are analysed in depth, with regard to the exhaustive features described in chapter three and their frequency in the hypertext corpus. Additionally, it would be worthwhile to examine the social Web, e.g. Facebook, in terms of its anaphor distribution. Social networks are not hypertexts, which means that if regarding such forms of the Internet, a more general approach has to be taken. As the social Web is rather informal, different relative frequencies of anaphors could arise. Apart from that, a comparison of hypertexts and printed texts is desirable because it could reveal where the use of anaphors varies between the media. It could be expected that differences occur, due to the specific characteristics of printed texts and hypertexts respectively. Furthermore, it would be interesting if hypertext types in other classifications which pay more attention to linguistic issues differed in their use of anaphors. It would also be necessary to know if the rules established for *non-finite clause* anaphors hold with a similar performance on other text types and larger corpora, or if adaptions are necessary. Furthermore, the effect of anaphora resolution on text retrieval, with the defined anaphor types and with regard to hypertexts, needs to be explored in future research. Previous studies report that anaphora resolution has an impact on text retrieval and, with the comprehensive classification established in this book, more profound effects could occur. Finally, anaphora resolution can be valueable not only for computational-related areas but also for any linguistic field which involves texts, such as discourse analysis.

In conclusion, it is considered imperative after this study that the current active research on anaphora resolution does not forget about linguistics. Reiter (2007), for instance, points out:

> The ACL [i.e. Association for Computational Linguistics] community seems to be focusing more on specific niches of the "language research" space, such as low-level syntactic analyses based on statistical corpus-based techniques. [...] This inward focus also goes against the belief in the larger scientific community that we need more inter-disciplinary work, and more interaction between researchers coming from different backgrounds. (ibid.: 285)

Similar statements can be found with Spärck Jones (2007), Moore (2009) and Uszkoreit (2009). The trend of anaphora resolution to focus on minimal linguistic knowledge is also present in natural language processing and computational linguistics in general. This book hopes to contribute to a direction where linguistics takes a more prominent role in anaphora resolution and in other computational research fields involving natural language analysis. It is highly desirable and necessary that linguistics and information technology combine their resources in the endeavour to develop ever better text retrieval systems. This will be beneficial for both disciplines and – last but not least – Internet users.

Bibliography

Aarts, Flor & Aarts, Jan (1986), *English Syntactic Structures. Functions & Categories in Sentence Analysis*, Oxford et al.: Pergamon Press.

"About the Open Directory Project" (2011), http://www.dmoz.org/docs/en/about.html (date of last access: 14/10/2011).

Agarwal, Amit (06/02/2012), "A Google Search Operator That You May Not Know About!", *Digital Inspiration*, http://www.booleanblackbelt.com/2011/06/beyond-boolean-search-proximity-and-weighting/ (date of last access: 08/11/2012).

Agnes, Michael et al., eds. (2007, 4th ed.), *Webster's New World College Dictionary*, Cleveland, Ohio: Wiley.

Allen, James (1995), *Natural Language Understanding*, Redwood City, CA: Benjamin/Cummings.

"Anaphora" (2010), *in* Albert Sydney Hornby, ed., *Oxford Advanced Learner's Dictionary of Current English* (CD-ROM), Oxford – New York: Oxford University Press.

"Anaphoric Bank Data" (2009), http://anawiki.essex.ac.uk/anaphoricbank/data.php (date of last access: 17/11/2012).

Anzinger, Gunnar (n.d.), "Governments on the WWW", http://www.gksoft.com/govt/ (date of last access: 27/10/2011).

Aone, Chinatsu & Scott William Bennett (1995), "Evaluating Automated and Manual Acquisition of Anaphora Resolution Strategies", *Proceedings of the 33rd Annual Meeting of the ACL (Association for Computational Linguistics)*, pp. 122-129.

Asher, Nicholas, Daniel Hardt & Joan Busquets (2001), "Discourse Parallelism, Ellipsis, and Ambiguity", *Journal of Semantics*, 18 (1), pp. 1-25.

Baeza-Yates, Ricardo (2004), "Challenges in the Interaction of Information Retrieval and Natural Language Processing", *in* Alexander Gelbukh, ed., *CICLing 2004* (LNCS 2945), Berlin – Heidelberg: Springer, pp. 445-456.

Baeza-Yates, Ricardo & Carlos Castillo (2006), "Web Searching", *in* Keith Brown, ed., *Encyclopedia of Language & Linguistics* (2nd ed.), vol. 13, Oxford et al.: Elsevier, pp. 527-538.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto (2011), *Modern Information Retrieval. The Concepts and Technology Behind Search* (2nd ed.), Harlow et al.: Addison Wesley.

Baicchi, Annalisa (2004), "The Cataphoric Indexicality of Titles", *in* Karin Aijmer & Anna-Brita Stenström, eds., *Discourse Patterns in Spoken and Written Corpora*, Amsterdam – Philadelphia: Benjamins, pp. 17-38.

Baldwin, Breck (1997), "CogNIAC: High Precision Coreference with Limited Knowledge and Linguistic Resources", *in* Ruslan Mitkov & Branimir Boguraev, eds., *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pp. 38-45.

Barbu, Cătălina (2002), "Error Analysis in Anaphora Resolution", *Proceedings of LREC 2002*, pp. 275-280.

Bartel, Dietrich (2007), *Handbuch der musikalischen Figurenlehre* (5th ed.), Laaber: Laaber.

Bauer, Laurie (1983), *English Word-Formation*, Cambridge et al.: Cambridge University Press.

Bauer, Laurie (2004), *A Glossary of Morphology*, Edinburgh: Edinburgh University Press.

Baugh, Albert C. & Thomas Cable (2002), *A History of the English Language* (5th ed.), London: Routledge.

Bean, David & Ellen Riloff (1999), “Corpus-Based Identification of Non-Anaphoric Noun Phrases”, *Proceedings of the 37th Annual Meeting of the ACL (Association for Computational Linguistics)*, pp. 373-380.

Becker, Tilman (2010), “(Multimodale) Dialogsysteme”, *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 624-632.

Bell, David (2009), “On the Net: Navigating the World Wide Web”, *in* Glen Creeber & Royston Martin, eds., *Digital Cultures*, Maidenhead: Open University Press, pp. 30-38.

“Beyond Boolean Search: Proximity and Weighting” (27/06/2011), http://www.booleanblackbelt.com/2011/06/beyond-boolean-search-proximity-and-weighting/ (date of last access: 08/11/2012).

Berry, Michael & Murray Browne (2005), *Understanding Search Engines. Mathematical Modeling and Text Retrieval* (2nd ed.), Philadelphia: Society for Industrial and Applied Mathematics.

Biber, Douglas et al. (2007), *Longman Grammar of Spoken and Written English*, Harlow: Longman.

Bittner, Johannes (2003), *Digitalität, Sprache, Kommunikation. Eine Untersuchung zur Medialität von digitalen Kommunikationsformen und Textsorten und deren varietätenlinguistischer Modellierung*, Berlin: Schmidt.

Bloomfield, Leonard (1984, originally published in 1933), *Language*, Chicago – London: University of Chicago.

Bolter, Jay David (2001), *Writing Space. Computers, Hypertext, and the Remediation of Print* (2nd ed.), Mahwah – London: Lawrence Erlbaum Associates.

Bonzi, Susan & Elizabeth Liddy (1989), “The Use of Anaphoric Resolution for Document Description in Information Retrieval”, *Information Processing & Management*, 25 (4), pp. 429-441.

Born, Günter (2011), *HTML5. Das umfassende Praxis- und Referenzwerk*, München: Markt+Technik.

Bos, Johan & Jennifer Spenader (2011), “An Annotated Corpus for the Analysis of VP Ellipses”, *Language Resources and Evaluation*, 45 (4), pp. 463-494.

Botley, Simon & Anthony Mark McEnery (2000), “Discourse Anaphora: The Need for Synthesis”, *in* S.B. & A.M.McE., eds., *Corpus-based and Computational Approaches to Discourse Anaphora*, Amsterdam – Philadelphia: Benjamins, pp. 1-41.

Boyd, Adriane, Whitney Gegg-Harrison & Donna Byron (2005), “Identifying Non-Referential *It*: A Machine Learning Approach Incorporating Linguistically Motivated Patterns”, *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, pp. 40-47.

Brennan, Susan, Marilyn Friedman & Carl Pollard (1987), “A Centering Approach to Pronouns”, *Proceedings of the 25th Annual Meeting of the ACL*, pp. 155-162.

Brinker, Klaus (2010), *Linguistische Textanalyse. Eine Einführung in Grundbegriffe und Methoden* (7th ed.; editor: Sandra Ausborn-Brinker), Berlin: Schmidt.

Bublitz, Wolfram (1994), “In the Eye of the Beholder: ‘The Rather Mystical Notion of Coherence’”, *in* Keith Carlon, Kristin Davidse & Brygida Rudzka-Ostyn, eds., *Perspectives on English. Studies in Honour of Professor Emma Vorlat*, Leuven – Paris: Peeters, pp. 213-230.

Bublitz, Wolfram (1998), “Cohesion and Coherence”, *in* Jef Verschueren et al., eds., *Handbook of Pragmatics*, Amsterdam – Philadelphia: Benjamins.

Bublitz, Wolfram (2009), *Englische Pragmatik. Eine Einführung* (2nd ed.), Berlin: Schmidt.

Bühler, Karl (1982, originally published in 1934), *Sprachtheorie*, Stuttgart – New York: Fischer.

Bühler, Karl (1990), *Theory of Language. The Representational Function of Language* (translated by Donald Fraser Goodwin), Amsterdam – Philadelphia: Benjamins.

Bußmann, Hadumod, ed. (2008), *Lexikon der Sprachwissenschaft* (4th ed.), Stuttgart: Kröner.

Büttcher, Stefan, Charles L. A. Clarke & Gordon V. Cormack (2010), *Information Retrieval. Implementing and Evaluating Search Engines*, Cambridge, MA – London: MIT.

Byron, Donna (2002), "Resolving Pronominal Reference to Abstract Entities", *Proceedings of the 40th Annual Meeting of the ACL (Association for Computational Linguistics)*, pp. 80-87.

Cardie, Claire (1992), "Learning to Disambiguate Relative Pronouns", *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 38-43.

Carroll, John (2004), „Parsing", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 233-248.

Carter, Ronald & Michael McCarthy, eds. (2006), *Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English Grammar and Usage*, Cambridge et al.: Cambridge University Press.

Cheng, Winnie (2012), *Exploring Corpus Linguistics. Language in Action*, London – New York: Routledge.

Chomsky, Noam (1993), *Lectures on Government and Binding. The Pisa Lectures* (7th ed.), Berlin: De Gruyter.

Christiansen, Thomas (2011), *Cohesion: A Discourse Perspective*, Bern: Lang.

Chu Min Xian, Benjamin, Fadzly Zahari & Diskon Lukose (2011), "Benchmarking ARS: Anaphora Resolution System", *Proceedings of the 11th International Conference on Knowledge Management and Knowledge Technologies*.

Claridge, Claudia (2007), "Constructing a Corpus from the Web: Message Boards", *in* Marianne Hundt, Nadja Nesselhauf & Carolin Biewer, eds., *Corpus Linguistics and the Web*, Amsterdam – New York: Rodopi, pp. 87-108.

Clark, Alexander (2009), "Case Study: Natural Language Processing (NLP)", *in* Glen Creeber & Royston Martin, eds., *Digital Cultures*, Maidenhead: Open University Press, pp. 164-169.

Clark, Eve V. (2009), *First Language Acquisition* (2nd ed.), Cambridge et al.: Cambridge University Press.

Claus, Volker & Andreas Schwill (2006), *Duden Informatik A-Z. Fachlexikon für Studium, Ausbildung und Beruf* (4th ed.), Mannheim et al.: Dudenverlag.

Consten, Manfred (2004), *Anaphorisch oder deiktisch? Zu einem integrativen Modell domänengebundener Referenz*, Tübingen: Niemeyer.

Cornish, Francis (1996), "'Antecedentless' Anaphors: Deixis, Anaphora, or What? Some Evidence from English and French", *Linguistics*, 32, pp. 19-41.

Cornish, Francis (1999), *Anaphora, Discourse and Understanding. Evidence from English and French*, Oxford: Clarendon Press.

Cornish, Francis (2006), "Discourse Anaphora", *in* Keith Brown, ed., *Encyclopedia of Language & Linguistics* (2nd ed.), vol. 3, Oxford et al.: Elsevier, pp. 631-638.

Croft, Bruce, Donald Metzler & Trevor Strohman (2010), *Search Engines. Information Retrieval in Practice* (International ed.), Upper Saddle River, NJ et al.: Pearson.

Crystal, David (1994), *An Encyclopedic Dictionary of Language and Languages*, Oxford – Cambridge, MA: Blackwell.

Crystal, David (2008), *Language and the Internet* (2nd ed.), Cambridge: Cambridge University Press.

Crystal, David (2009), *A Dictionary of Linguistics and Phonetics* (6th ed.), Malden, MA et al.: Blackwell.

Crystal, David (2011), *Internet Linguistics: A Student Guide*, London – New York: Routledge.

Damascelli, Adriana Teresa & Aurelia Martelli (2003), *Corpus Linguistics and Computational Linguistics: An Overview with Special Reference to English*, Torino: Celid.

De Beaugrande, Robert-Alain & Wolfgang Ulrich Dressler (1981), *Introduction to Text Linguistics*, London – New York: Longman.

Diekmannshenke, Hajo (2000), "Die Spur des Internetflaneurs – Elektronische Gästebücher als neue Kommunikationsform", *in* Caja Thimm, ed., *Soziales im Netz. Sprache, Beziehungen und Kommunikationskulturen im Internet*, Opladen – Wiesbaden: Westdeutscher Verlag, pp. 131-155.

Do Carmo Pereira, Santiago, Hilário Seibel Júnior & Sérgio Antônio Andrade de Freitas (2009), "An Anaphora Based Information Retrieval Model Extension", *Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering*, vol. 4, pp. 330-334.

Durant, Alan & Marina Lambrou (2009), *Language and Media. A Resource Book for Students*, London – New York: Routledge.

Eberle, Kurt (2003), "Anaphernresolution in flach analysierten Texten für Recherche und Übersetzung", *in* Ute Seewald-Heeg, ed., *Sprachtechnologie für die multilinguale Kommunikation. Textproduktion, Recherche, Übersetzung, Lokalisierung*, Sankt Augustin: Gardez!, pp. 216-232.

Endres, Brigitte Odile (2004), "Ist Hypertext Text?", *in* Ulla Kleinberger Günther & Franc Wagner, eds., *Neue Medien – Neue Kompetenzen?*, Frankfurt am Main et al.: Lang, pp. 33-48.

Endres-Niggemeyer, Brigitte (2004), "Automatisches Textzusammenfassen", *in* Henning Lobin & Lothar Lemnitzer, eds., *Texttechnologie. Perspektiven und Anwendungen*, Tübingen: Stauffenburg, pp. 407-432.

"English Wikipedia" (2014), http://stats.wikimedia.org/EN/SummaryEN.htm (date of last access: 17/12/2014).

Esser, Jürgen (2009), *Introduction to English Text-linguistics*, Frankfurt am Main et al.: Lang.

Evans, Richard (2000), "A Comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal *It*", *in* Dimitris Christodoulakis, ed., *Natural Language Processing – NLP 2000. Proceedings of the Second International Conference*, Berlin et al.: Springer, pp. 233-240.

Evans, Richard (2001), "Applying Machine Learning Toward an Automatic Classification of *It*", *Literary and Linguistic Computing*, 16 (1), pp. 45-57.

"Exalead: Web Search Syntax" (2012), http://www.exalead.com/search/web/search-syntax/ (date of last access: 19/11/2012).

Ferrández, Antonio, Manuel Palomar & Lidia Moreno (1998), "Anaphora Resolution in Unrestricted Texts with Partial Parsing", *Proceedings of the 36th Annual Meeting of the ACL and 17th International Conference on Computational Linguistics*, pp. 385-391.

Ferrández, Antonio, Manuel Palomar & Lidia Moreno (1999), "An Empirical Approach to Spanish Anaphora Resolution", *Machine Translation*, 14 (3-4), pp. 1-16.

Ferrández, Antonio, Yenory Rojas & Jesús Peral (2007), "The Successful Application of Natural Language Processing for Information Retrieval", *Journal of Computer Science and Technology*, 7 (1), pp. 79-85.

Fillmore, Charles J. (1975), *Santa Cruz Lectures on Deixis 1971*, Bloomington: Indiana University Linguistics Club.

Finch, Geoffrey (2005), *Key Concepts in Language and Linguistics* (2nd ed.), Basingstoke, Hampshire: Palgrave Macmillan.

Fitschen, Arne & Piklu Gupta (2008), "Lemmatising and Morphological Tagging", *in* Anke Lüdeling & Merja Kytö, eds., *Corpus Linguistics. An International Handbook*, vol. 1, Berlin et al.: De Gruyter, pp. 552-564.

Fix, Ulla (2014), "Aktuelle Tendenzen des Textsortenwandels – Thesenpapier", *in* Stefan Hauser, Ulla Kleinberger & Kersten Sven Roth, eds., *Musterwandel – Sortenwandel. Aktuelle Tendenzen der diachronen Text(sorten)linguistik*, Bern et al.: Lang, pp. 15-48.

Fliedner, Gerhard (2010), "Korrektursysteme", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 555-565.

Fromkin, Victoria, Robert Rodman & Nina Hyams (2011), *An Introduction to Language* (9th, international ed.), Boston: Wadsworth, Cengage Learning.

Fuhr, Norbert (2011), "Einführung in Information Retrieval. Skriptum zur Vorlesung im WS 2011/12", http://www.is.informatik.uni-duisburg.de/courses/ir_ws11/folien/skript_1-6.pdf (date of last access: 28/02/2012).

Gansel, Christina & Frank Jürgens (2009), *Textlinguistik und Textgrammatik. Eine Einführung* (3rd ed.), Göttingen: Vandenhoeck & Ruprecht.

Garnham, Alan (2001), *Mental Models and the Interpretation of Anaphora*, Hove, East Sussex: Psychology Press.

Gauntlett, David (2009), "Case Study: Wikipedia", *in* Glen Creeber & Royston Martin, eds., *Digital Cultures*, Maidenhead: Open University Press, pp. 39-45.

Gibbon, Dafydd (2010), "Lexika für multimodale Systeme", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 515-523.

Giltrow, Janet & Dieter Stein (2009), *Genres in the Internet. Issues in the Theory of Genre*, Amsterdam – Philadelphia: Benjamins.

Glück, Helmut, ed. (2010), *Metzler Lexikon Sprache* (4th ed.), Stuttgart – Weimar: Metzler.

Gödert, Winfried, Klaus Lepsky & Matthias Nagelschmidt (2012), *Informationserschließung und Automatisches Indexieren. Ein Lehr- und Arbeitsbuch*, Berlin – Heidelberg: Springer.

Green, Keith (2006), "Deixis and Anaphora: Pragmatic Approaches", *in* Edward Brown, ed., *Encyclopedia of Language and Linguistics* (2nd ed.), vol. 3, pp. 415-417.

Greiffenstern, Sandra (2010), *The Influence of Computers, the Internet and Computer-Mediated Communication on Everyday English*, Berlin: Logos.

Grishman, Ralph (2004), "Information Extraction", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 545-559.

Grishman, Ralph (2010), "Information Extraction", *in* Alexander Clark, Chris Fox & Shalom Lappin, eds., *The Handbook of Computational Linguistics and Natural Language Processing*, Malden, MA et al.: Wiley-Blackwell, pp. 517-530.

Grishman, Ralph & Beth Sundheim (1996), "Message Understanding Conference – 6: A Brief History", *Proceedings of the 16th International Conference on Computational Linguistics (COLING)*, vol. 1, pp. 466-471.

Grosz, Barbara, Aravind Joshi & Scott Weinstein (1983), "Providing a Unified Account of Definite Noun Phrases in Discourse", *Proceedings of the 21st Annual Meeting of the ACL (Association for Computational Linguistics)*, pp. 44-50.

Gundel, Jeanette, Nancy Hedberg & Ron Zacharski (2005), "Pronouns Without NP Antecedents: How Do We Know When a Pronoun is Referential?", *in* António Branco, Anthony Mark McEnery & Ruslan Mitkov, eds., *Anaphora Processing. Linguistic, Cognitive, and Computational Modelling. Selected Papers from DAARC 2002*, Amsterdam – Philadelphia: Benjamins, pp. 351-364.

Hagenbruch, André (2010), "Flache Satzverarbeitung", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 264-279.

Haghighi, Aria & Dan Klein (2009), "Simple Coreference Resolution with Rich Syntactic and Semantic Features", *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 1152-1161.

Halliday, Michael A.K. & Ruqaiya Hasan (2008, first published 1976), *Cohesion in English*, Harlow: Pearson Education.

Harabagiu, Sanda & Dan Moldovan (2004), "Question Answering", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 560-582.

Hardt, Daniel (1997), "An Empirical Approach to VP Ellipsis", *Computational Linguistics*, 23 (4), pp. 525-541.

Hasler, Laura, Constantin Orasan & Karin Naumann (2006), "NP for Events: Experiments in Coreference Annotation", *Proceedings of the 5th Edition of the International Conference on Language Resources and Evaluation*, pp. 1167-1172.

Hasselgård, Hilde, Per Lysvåg & Stig Johansson (2012), *Glossary of Grammatical Terms Used in English Grammar: Theory and Use* (2nd ed.), http://folk.uio.no/hhasselg/terms.html#notional (date of last access: 25/06/2012).

Henrich, Andreas (2007a), *Information Retrieval 1. Kurs im Wintersemester 2007/2008* (script of the VHB-course), published in 2008 as *Information Retrieval 1. Grundlagen, Modelle und Anwendungen*, Bamberg: Otto-Friedrich-Universität Bamberg, http://www.uni-bamberg.de/minf/ir1-buch/ (date of last access: 16/02/2012).

Henrich, Andreas (2007b), *Information Retrieval 1. Einführung und Überblick* (PowerPoint slides to the VHB-course).

Herbst, Thomas (2010), *English Linguistics. A Coursebook for Students of English*, Berlin: De Gruyter.

Herbst, Thomas, Rita Stoll & Rudolf Westermayr (1991), *Terminologie der Sprachbeschreibung. Ein Lernwörterbuch für das Anglistikstudium*, Ismaning: Hueber.

"Here" (n.d.), *TheFreeDictionary.com*, http://www.thefreedictionary.com/here (date of last access: 09/07/2012).

Hering, Ekbert, Jürgen Gutekunst & Ulrich Dyllong (2000), *Handbuch der praktischen und technischen Informatik* (2nd ed.), Berlin et al.: Springer.

Hirschman, Lynette & Nancy Chinchor (1997), "MUC-7 Coreference Task Definition", http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/co_task.html (date of last access: 16/11/2012).

Hobbs, Jerry R. (1976), "Pronoun Resolution" (research report), New York: City University of New York, http://www.isi.edu/~hobbs/PronounResolution.pdf (date of last access: 14/01/2013).

Hobbs, Jerry R. (1986, first published 1978), "Resolving Pronoun References", *in* Barbara Grosz, Karen Spärck Jones & Bonnie Lynn Webber, eds. (1986), *Readings in Natural Language Processing*, Los Altos, CA: Morgan Kaufmann, pp. 339-352.

Hobbs, Jerry R. & Andrew Kehler (1997), “A Theory of Parallelism and the Case of VP Ellipsis”, *Proceedings of the 35th Conference of the ACL (Association for Computational Linguistics)*, pp. 394-401.

Hoffmann, Christian (2012), *Cohesive Profiling. Meaning and Interaction in Personal Weblogs*, Amsterdam – Philadelphia: Benjamins.

Hoffmann, Ludger (2000), “Anapher im Text”, *in* Klaus Brinker et al., eds., *Text- und Gesprächslinguistik. Ein internationales Handbuch zeitgenössischer Forschung* (vol. 1), Berlin – New York: De Gruyter, pp. 295-304.

Holzinger, Andreas (2002), *Basiswissen Multimedia. Band 1: Technik* (2nd ed.), Würzburg: Vogel.

Hornby, Albert Sydney, ed. (2010, 8th ed.), *Oxford Advanced Learner’s Dictionary of Current English*, Oxford – New York: Oxford University Press.

Huang, Yan (2000), *Anaphora. A Cross-linguistic Approach*, Oxford – New York: Oxford University Press.

Huber, Oliver (2002), *Hyper-Text-Linguistik. TAH: Ein textlinguistisches Analysemodell für Hypertexte. Theoretisch und praktisch exemplifiziert am Problemfeld der typisierten Links von Hypertexten im World Wide Web*, München: Utz.

Huddleston, Rodney (2010a), “Syntactic Overview”, *in* Rodney Huddleston & Geoffrey K. Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press, pp. 43-69.

Huddleston, Rodney (2010b), “The Clause: Complements”, *in* Rodney Huddleston & Geoffrey K. Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press, pp. 213-321.

Huddleston, Rodney (2010c), “Comparative Constructions”, *in* Rodney Huddleston & Geoffrey K. Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press, pp. 1097-1170.

Huddleston, Rodney (2010d), “Non-finite and Verbless Clauses”, *in* Rodney Huddleston & Geoffrey K. Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press, pp. 1171-1271.

Huddleston, Rodney & Geoffrey Pullum, eds. (2010), *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press.

Huddleston, Rodney, Geoffrey K. Pullum & Peter Peterson (2010), “Relative Constructions and Unbounded Dependencies”, *in* Rodney Huddleston & Geoffrey K. Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge et al.: Cambridge University Press, pp. 1031-1096.

Hutchins, John (2004), “Machine Translation: General Overview”, *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 501-511.

Ince, Darrel (2012), *A Dictionary of the Internet. Over 4450 entries* (2nd ed., online version), Oxford: Oxford University Press.

Jackson, Howard & Etienne Zé Amvela (2007), *Words, Meaning and Vocabulary. An Introduction to Modern English Lexicology*, London – New York: Continuum.

Jackson, Peter & Isabelle Moulinier (2002), *Natural Language Processing for Online Applications. Text Retrieval, Extraction and Categorization*, Amsterdam – Philadelphia: Benjamins.

Jacquemin, Christian & Didier Bourigault (2004), "Term Extraction and Automatic Indexing", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 599-615.

Jakobs, Eva-Maria (2003), "Hypertextsorten", *Zeitschrift für germanistische Linguistik*, 31, pp. 232-252.

Jakobs, Eva-Maria & Katrin Lehnen (2005), "Hypertext – Klassifikation und Evaluation", *in* Torsten Siever, Peter Schlobinski & Jens Runkehl, eds., *Websprache.net. Sprache und Kommunikation im Internet*, Berlin – New York: De Gruyter, pp. 159-184.

Jekat, Susanne & Martin Volk (2010), "Maschinelle und computergestützte Übersetzung", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 642-658.

Jucker, Andreas (2005), "Hypertext Research: Some Basic Concepts", *in* Lilo Moessner & Christa M. Schmidt, eds., *Anglistentag 2004 Aachen. Proceedings*, Trier: Wissenschaftlicher Verlag Trier, pp. 285-295.

Jurafsky, Daniel & James H. Martin (2009), *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.), Upper Saddle River, New Jersey: Pearson.

Katzer, Jeffrey, Susan Bonzi & Elizabeth Liddy (1986), *Impact of Anaphoric Resolution in Information Retrieval. Final Report*, Syracuse, NY: Syracuse University.

Kehler, Andrew (2002), *Coherence, Reference, and the Theory of Grammar*, Stanford, CA: Center for the Study of Language and Information.

Kennedy, Christopher & Branimir Boguraev (1996), "Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser", *Proceedings of the 16th Conference on Computational Linguistics*, pp. 113-118.

Klein, Dan & Christopher D. Manning (2003), "Accurate Unlexicalized Parsing", *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Kleinberg, Jon (1999), "Authoritative Sources in a Hyperlinked Environment", *Journal of the ACM*, 46 (5), pp. 604-632.

Knowles, Gerry & Zuraidah Mohd Don (2004), "The Notion of a 'Lemma'", *International Journal of Corpus Linguistics*, 9 (1), pp. 69-81.

Kolhatkar, Varada & Graeme Hirst (2012), "Resolving 'This-issue' Anaphora", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1255-1265.

Kortmann, Bernd (2005), *English Linguistics: Essentials*, Berlin: Cornelsen.

Kowalski, Gerald (2011), *Information Retrieval Architecture and Algorithms*, New York et al.: Springer.

Krüger-Thielmann, Karin & Hans Paijmans (2004), "Informationserschließung", *in* Henning Lobin & Lothar Lemnitzer, eds., *Texttechnologie. Perspektiven und Anwendungen*, Tübingen: Stauffenburg, pp. 353-378.

Kübler, Sandra (n.d.), "Linguistic Fundamentals", http://www.sfs.uni-tuebingen.de/files/Kursmaterialien/Kuebler/Anaphor/INTRO.doc (date of last access: 17/06/2012).

Lappin, Shalom (2005), "A Sequenced Model of Anaphora and Ellipsis Resolution", *in* António Branco, Anthony Mark McEnery & Ruslan Mitkov, eds., *Anaphora Processing. Linguistic, Cognitive, and Computational Modelling. Selected Papers from DAARC 2002*, Amsterdam – Philadelphia: Benjamins, pp. 3-16.

Lappin, Shalom & Herbert J. Leass (1994), "An Algorithm for Pronominal Anaphora Resolution", *Computational Linguistics*, 20 (4), pp. 535-561.

Lemnitzer, Lothar & Heike Zinsmeister (2010), *Korpuslinguistik. Eine Einführung* (2nd ed.), Tübingen: Narr.

Lenz, Friedrich (1997), *Diskursdeixis im Englischen. Sprachtheoretische Überlegungen und lexiko-grammatische Analysen*, Tübingen: Niemeyer.

Levene, Mark (2010), *An Introduction to Search Engines and Web Navigation* (2nd ed.), Hoboken, NJ: Wiley.

Lewandowski, Theodor (1994), *Linguistisches Wörterbuch* (6th ed.), Heidelberg – Wiesbaden: Quelle & Meyer.

Liddy, Elizabeth DuRoss (1990), "Anaphora in Natural Language Processing and Information Retrieval", *Information Processing & Management*, 26 (1), pp. 39-52.

Linsky, Leonard (1974), "Reference and Referents", *in* Danny Steinberg & Leon Jakobovits, eds., *Semantics. An Interdisciplinary Reader in Philosophy, Linguistics and Psychology*, Cambridge: Cambridge University Press, pp. 76-85.

Lyons, John (1977), *Semantics*, vol. 2, Cambridge et al.: Cambridge University Press.

Lyons, John (2005), *Linguistic Semantics. An Introduction*, Cambridge et al.: Cambridge University Press.

Malmkjær, Kirsten, ed. (2010), *The Routledge Linguistics Encyclopedia* (3rd ed.), London – New York: Routledge.

Manning, Christopher (2011), "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?", *in* Alexander Gelbukh, ed., *CICLing 2011* (LNCS 6608), part I, Berlin – Heidelberg: Springer, pp. 171-189.

Manning, Christopher, Prabhakar Raghavan & Hinrich Schütze (2008), *Introduction to Information Retrieval*, Cambridge et al.: Cambridge University Press.

Marcus, Mitchell, Mary Ann Marcinkiewicz & Beatrice Santorini (1993), "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19 (2), pp. 314-330.

Markert, Katja & Malvina Nissim (2005), "Comparing Knowledge Sources for Nominal Anaphora Resolution", *Computational Linguistics*, 31 (3), pp. 367-402.

Marneffe, Marie-Catherine de, Bill MacCartney & Christopher D. Manning (2006), "Generating Typed Dependency Parses from Phrase Structure Parses", *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, pp. 449-454.

Marneffe, Marie-Catherine de & Christopher D. Manning (2011), "Stanford Typed Dependencies Manual", http://nlp.stanford.edu/software/dependencies_manual.pdf (date of last access: 25/05/2012).

Matthews, Peter (2007), *The Concise Oxford Dictionary of Linguistics* (2nd ed.), Oxford – New York: Oxford University Press.

Mayor, Michael, ed. (2009), *Longman Dictionary of Contemporary English* (5th ed.), Harlow: Pearson.

McEnery, Tony & Andrew Hardie (2012), *Corpus Linguistics. Method, Theory and Practice*, Cambridge et al.: Cambridge University Press.

McEnery, Tony & Andrew Wilson (2001), *Corpus Linguistics. An Introduction* (2nd ed.), Edinburgh: Edinburgh University Press.

Meadow, Charles et al. (2007), *Text Information Retrieval Systems* (3rd ed.), Amsterdam et al.: Elsevier.

Menke, Peter (2012), "Evaluation of Technical Communication", *in* Alexander Mehler & Laurent Romary, eds., *Handbook of Technical Communication*, Berlin – Boston: De Gruyter, pp. 285-314.

Meyer, Josef & Robert Dale (2002), "Mining a Corpus to Support Associative Anaphora Resolution", *Proceedings of the Fourth International Conference on Discourse Anaphora and Anaphor Resolution (DAARC 2002)*.

Mikheev, Andrei (2004), "Text Segmentation", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 201-218.

Mitkov, Ruslan (1998), "Robust Pronoun Resolution with Limited Knowledge", *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869-875.

Mitkov, Ruslan (2001), "Outstanding Issues in Anaphora Resolution", *in* Alexander Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing, Second International Conference, CICLing 2001, Mexico City, Mexico, February 18-24, 2001, Proceedings*, Berlin et al.: Springer, pp. 110-125.

Mitkov, Ruslan (2002), *Anaphora Resolution*, London et al.: Longman.

Mitkov, Ruslan (2004a), "Anaphora Resolution", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 266-283.

Mitkov, Ruslan, ed. (2004b), *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 729-762.

Mitkov, Ruslan et al. (n.d.), "Annotation of Cross-Document Coreference: A Pilot Study", http://clg.wlv.ac.uk/projects/NP4E (date of last access: 27/03/2012).

Mitkov, Ruslan (2008), "Corpora for Anaphora and Coreference Resolution", *in* Anke Lüdeling & Merja Kytö, eds., *Corpus Linguistics. An International Handbook*, vol. 1, Berlin – New York: De Gruyter, pp. 579-598.

Mitkov, Ruslan & Catalina Hallett (2007), "Comparing Pronoun Resolution Algorithms", *Computational Intelligence*, 23 (2), pp. 262-297.

Mitkov, Ruslan, Richard Evans & Constantin Orasan (2002), "A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method", *in* Alexander Gelbukh, ed., *Computational Linguistics and Intelligent Text Processing, Third International Conference, CICLing 2002, Mexico City, Mexico, February 17-23, 2002, Proceedings*, Berlin et al.: Springer, pp. 168-186.

Mooney, Raymond (2004), "Machine Learning", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 376-394.

Moore, Robert (2009), "What do Computational Linguists Need to Know About Linguistics?", *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pp. 41-42.

Muñoz, Rafael, Manuel Palomar & Antonio Ferrández (2000), "Processing of Spanish Definite Descriptions", *in* Osvaldo Cairo, L. Enrique Sucar & Francisco J. Cantu, eds., *MICAI 2000 – Advances in Artificial Intelligence*, Berlin et al.: Springer, pp. 526-537.

Nelson, Mike (2010), "Building a Written Corpus. What are the Basics?", *in* Anne O'Keeffe & Michael McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, London – New York: Routledge, pp. 53-65.

Nelson, Ted (1972), "As We Will Think", *in* James Nyce & Paul Kahn, eds., *From Memex to Hypertext: Vannevar Bush and the Mind's Machine* (reprinted in 1991), Boston et al.: Academic Press, pp. 245-260.

Neumann, Günter (2010), "Text-basiertes Informationsmanagement", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 576-615.

Ng, Vincent et al. (2005), "A Machine Learning Approach to Identification and Resolution of *One*-Anaphora", *International Joint Conference on Artifical Intelligence (IJCAI)*, pp. 1105-1110.

Ng, Vincent (2010), "Supervised Noun Phrase Coreference Research: The First Fifteen Years", *Proceedings of the 48th Annual Meeting of the ACL*, pp. 1396-1411.

Ng, Vincent & Claire Cardie (2002a), "Identifying Anaphoric and Non-Anaphoric Noun Phrases to Improve Coreference Resolution", *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pp. 730-736.

Ng, Vincent & Clarie Cardie (2002b), "Improving Machine Learning Approaches to Coreference Resolution", *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 104-111.

Nielsen, Leif A. (2004), "Verb Phrase Ellipsis Detection Using Automatically Parsed Text", *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pp. 1093-1099.

O'Keeffe, Anne & Michael McCarthy, eds. (2010), *The Routledge Handbook of Corpus Linguistics*, London – New York: Routledge.

O'Keeffe, Anne, Michael McCarthy & Ronald Carter (2007), *From Corpus to Classroom: Language Use and Language Teaching*, Cambridge et al.: Cambridge University Press.

Olson, David & Dursun Delen (2008), *Advanced Data Mining Techniques*, Berlin – Heidelberg: Springer.

"OntoNotes: Coreference" (2012), http://www.bbn.com/ontonotes/components/coreference (date of last access: 16/12/2012).

Paroubek, Patrick (2007), "Evaluating Part-of-Speech Tagging and Parsing. On the Evaluation of Automatic Parsing of Natural Language", *in* Laila Dybkær, Holmer Hemsen & Wolfgang Minker, eds., *Evaluation of Text and Speech Systems*, Dordrecht: Springer, pp. 99-124.

Payne, John & Rodney Huddleston (2010), "Nouns and Noun Phrases", *in* Rodney Huddleston & Geoffrey Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press, pp. 323-523.

Pedersen, Sarah (2010), *Why Blog? Motivations for Blogging*, Oxford: Chandos.

"Personal pronoun" (n.d.), *TheFreeDictionary.com*, http://www.thefreedictionary.com/personal+pronoun (date of last access: 23/06/2012).

Pirkola, Ari (1999), "Studies on Linguistic Problems and Methods in Text Retrieval. The Effects of Anaphor and Ellipsis Resolution in Proximity Searching, and Translation and Query Structuring Methods in Cross-Language Retrieval" (PhD Dissertation), Tampere: University of Tampere, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.138.8052&rep=rep1&type=pdf (date of last access: 27/03/2014).

Pirkola, Ari & Kalervo Järvelin (1996a), "The Effect of Anaphor and Ellipsis Resolution in Proximity Searching in a Text Database", *Information Processing & Management*, 32 (2), pp. 199-216.

Pirkola, Ari & Kalervo Järvelin (1996b), "Recall and Precision Effects of Anaphor and Ellipsis Resolution in Proximity Searching in a Text Database", *in* Peter Ingwersen & Niels O. Pors, eds., *Proceedings CoLIS2, Second International Conference on Conceptions of Library and Information Science. Integration in Perspective*, Kopenhagen: The Royal School of Librarianship, pp. 459-475.

Poesio, Massimo & Mijail Kabadjov (2004), "A General-Purpose, Off-the-shelf Anaphora Resolution Module: Implementation and Preliminary Evaluation", *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pp. 663-666.

Poesio, Massimo & Renata Vieira (1998), "A Corpus-Based Investigation of Definite Description Use", *Computational Linguistics*, 24 (2), pp. 183-216.

Poitou, Jacques (1996), "Pro-Formen und verwandte Erscheinungen", *in* Marie-Hélène Pérennec, ed., *Pro-Formen des Deutschen*, Tübingen, Stauffenburg, pp. 123-134.

Porter, Martin F. (1980), "An Algorithm for Suffix Stripping", *Program*, 14 (3), pp. 130-137.

Pscheida, Daniela (2010), *Das Wikipedia-Universum. Wie das Internet unsere Wissenskultur verändert*, Bielefeld: transcript.

Puttenham, George (1589), *The Arte of English Poesie*, London: Richard Field.

Quirk, Randolph et al. (2012, first published 1985), *A Comprehensive Grammar of the English Language*, London – New York: Longman.

Rehm, Georg (2007), *Hypertextsorten. Definition – Struktur – Klassifikation*, Norderstedt: Books on Demand.

Rehm, Georg (2010), "Texttechnologische Grundlagen", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 159-168.

Reiter, Ehud (2007), "Last Words: The Shrinking Horizons of Computational Linguistics", *Computational Linguistics*, 33 (2), pp. 283-287.

Reppen, Randi (2010), "Building a Corpus. What are the Key Considerations?", *in* Anne O'Keeffe & Michael McCarthy, eds., *The Routledge Handbook of Corpus Linguistics*, London – New York: Routledge, pp. 31-37.

Rijsbergen, Cornelis J. van (1979), *Information Retrieval* (2nd ed.), London – Boston: Butterworths.

Rothlauf, Franz (2011), *Design of Modern Heuristics. Principles and Application*, Berlin – Heidelberg: Springer.

Sachs, Lothar (2006), *Einführung in die Stochastik und das stochastische Denken*, Frankfurt am Main: Harri Deutsch.

Sasse, Hans-Jürgen (1993), "Syntactic Categories and Subcategories", *in* Joachim Jacobs et al., eds., *Syntax. Ein internationales Handbuch zeitgenössischer Forschung*, vol. 1, Berlin – New York: De Gruyter, pp. 646-686.

Schmid, Helmut (2010), "Decision Trees", *in* Alexander Clark, Chris Fox & Shalom Lappin, eds., *The Handbook of Computational Linguistics and Natural Language Processing*, Malden, MA et al.: Wiley-Blackwell, pp. 180-196.

Schmitz, Ulrich (2004), *Sprache in modernen Medien. Einführung in Tatsachen und Theorien, Themen und Thesen*, Berlin: Schmidt.

Schubert, Christoph (2012), *Englische Textlinguistik. Eine Einführung* (2nd ed.), Berlin: Schmidt.

Schubert, Christoph (forthcoming), "Anapher", *in* Stefan Schierholz, ed., *Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK) online. Grammatik: Syntax*, Berlin: De Gruyter.

Schuler, Günter (2007), *Wikipedia inside. Die Online-Enzyklopädie und ihre Community*, Münster: UNRAST.

Schütte, Daniela (2004), *Homepages im World Wide Web. Eine interlinguale Untersuchung zur Textualität in einem globalen Medium*, Frankfurt am Main et al.: Lang.

"SemEval 2010: VP Ellipsis Processing" (2011), http://www.sigsem.org/wiki/SemEval_2010:_VP_Ellipsis_Processing (date of last access: 16/11/2012).

Siddiqui, Tanveer & Uma S. Tiwary (2008), *Natural Language Processing and Information Retrieval*, New Delhi: Oxford University Press.

Simpson, John A. & Edmund S.C. Weiner, eds. (1989, 2nd ed.), *The Oxford English Dictionary*, Oxford: Clarendon Press.

Sinclair, John (1991), *Corpus, Concordance, Collocation*, Oxford et al.: Oxford University Press.

Sladovníková, Šárka (2010*), Textverstehen. Analysen zu Kohäsion und Kohärenz am Beispiel journalistischer Texte*, Ostrava: Universität Ostrava.

"So How Many Blogs are There, Anyway?" (n.d.), http://www.hattrickassociates.com/2010/02/how_many_blogs_2011_web_content/ (date of last access: 27/12/2011).

Soanes, Catherine & Angus Stevenson, eds. (2005, 2nd rev. ed.), *Oxford Dictionary of English*, Oxford – New York: Oxford University Press.

Somers, Harold (2004), "Machine Translation: Latest Developments", *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 512-528.

Soon, Wee Meng, Hwee Tou Ng & Daniel Chung Yong Lim (2001), "A Machine Learning Approach to Coreference Resolution of Noun Phrases", *Computational Linguistics*, 27 (4), pp. 521-544.

Spärck Jones, Karen (2007), "Last Words: Computational Linguistics: What About the Linguistics?", *Computational Linguistics*, 33 (3), pp. 437-441.

Speake, Jennifer, ed. (2008), *Oxford Dictionary of Proverbs* (5th ed.), Oxford – New York: Oxford University Press.

Stede, Manfred (2007), *Korpusgestützte Textanalyse. Grundzüge der Ebenen-orientierten Textlinguistik*, Tübingen: Narr.

Stirling, Lesley & Rodney Huddleston (2010), "Deixis and Anaphora", *in* Rodney Huddleston & Geoffrey Pullum, eds., *The Cambridge Grammar of the English Language*, Cambridge: Cambridge University Press, pp. 1449-1564.

Stock, Wolfgang (2007), *Information Retrieval. Informationen suchen und finden*, München – Wien: Oldenbourg.

Storrer, Angelika (1999), "Was ist eigentlich eine Homepage? Neue Formen der Wissensorganisation im World Wide Web", *Sprachreport*, 15 (1), pp. 2-8.

Storrer, Angelika (2000), "Was ist „hyper" am Hypertext?", *in* Werner Kallmeyer, ed., *Sprache und neue Medien*, Berlin – New York: De Gruyter, pp. 222-249.

Storrer, Angelika (2003), "Kohärenz in Hypertexten", *Zeitschrift für germanistische Linguistik*, 31 (2), pp. 274-292.

Storrer, Angelika (2004), "Text und Hypertext", *in* Henning Lobin & Lothar Lemnitzer, eds., *Texttechnologie. Perspektiven und Anwendungen*, Tübingen: Stauffenburg, pp. 13-49.

Storrer, Angelika (2007), "Hypertext und Texttechnologie", *in* Karlfried Knapp, ed., *Angewandte Linguistik* (2nd ed.), Tübingen – Basel: Francke, pp. 207-228.

Storrer, Angelika (2008), "Hypertextlinguistik", *in* Nina Janich, ed., *Textlinguistik. 15 Einführungen*, Tübingen: Narr, pp. 315-331.

Storrer, Angelika (2012), "Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia", *in* Helmuth Feilke, Juliane Köster & Michael Steinmetz, eds., *Textkompetenzen in der Sekundarstufe II*, Stuttgart: Fillibach, pp. 277-304.

Stoyanov, Veselin et al. (2010), "Coreference Resolution with Reconcile", *Proceedings of the ACL 2010 Conference Short Papers*, pp. 156-161.

Strube, Michael (2010), "Anaphernresolution", *in* Kai-Uwe Carstensen et al., eds., *Computerlinguistik und Sprachtechnologie. Eine Einführung* (3rd ed.), Heidelberg: Spektrum, pp. 399-409.

Stuckhardt, Roland (2001), “Design and Enhanced Evaluation of a Robust Anaphor Resolution Algorithm”, *Computational Linguistics*, 27 (4), pp. 479-506.

Summers, Della (2006), *Longman Dictionary of English Language and Culture. Gets to the Heart of the Language* (3rd ed.), Harlow: Longman.

Swan, Michael (2005), *Practical English Usage* (3rd ed.), Oxford et al.: Oxford University Press.

Swan, Michael (n.d.), “Pronouns”, http://www.phon.ucl.ac.uk/home/dick/pronoun.htm (date of last access: 19/09/2014).

Technorati (2011a), “State of the Blogosphere 2011”, http://technorati.com/blogging/article/state-of-the-blogosphere-2011-introduction/ (date of last access: 27/12/2011).

Technorati (2011b), “Top 100 Blogs”, http://technorati.com/blogs/top100/ (date of last access: 27/12/2011).

Technorati (2011c), “Blog Directory”, http://technorati.com/blogs/directory/ (date of last access: 28/12/2011).

Tengi, Randee (1998), “Design and Implementation of the WordNet Lexical Database and Searching Software”, *in* Christiane Fellbaum, ed., *WordNet. An Electronic Lexical Database*, Cambridge – London: Massachusetts Institute of Technology, pp. 105-127.

Tetreault, Joel (2001), “A Corpus-Based Evaluation of Centering and Pronoun Resolution”, *Computational Linguistics*, 27 (4), pp. 507-520.

“The Free Dictionary” (2012), http://www.thefreedictionary.com/ (date of last access: 16/01/2012).

“The Stanford Parser: A Statistical Parser” (n.d.), http://nlp.stanford.edu/software/lex-parser.shtml (date of last access: 17/01/2013).

Trask, Robert L. (1993), *A Dictionary of Grammatical Terms in Linguistics*, London – New York: Routledge.

Trask, Robert L. (1997), *A Student’s Dictionary of Language and Linguistics*, London – New York: Arnold.

Trask, Robert L. & Stockwell, Peter (2007), *Language and Linguistics. The Key Concepts* (2nd ed.), Abingdon – New York: Routledge.

Tsujii, Jun’ichi (2011), “Computational Linguistics and Natural Language Processing”, *in* Alexander Gelbukh, ed., *CICLing 2011* (LNCS 6608), part I, Berlin – Heidelberg: Springer, pp. 52-67.

Tzoukerman, Evelyne, Judith Klavans & Tomek Strzalkowski (2004), “Information Retrieval”, *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 529-544.

Ule, Tylman & Erhard Hinrichs (2004), “Linguistische Annotation”, *in* Henning Lobin & Lothar Lemnitzer, eds., *Texttechnologie. Perspektiven und Anwendungen*, Tübingen: Stauffenburg, pp. 217-243.

Uryupina, Olga (2010), “Corry: A System for Coreference Resolution”, *Proceedings of 5th International Workshop on Semantic Evaluation*, pp. 100-103.

Uszkoreit, Hans (2009), “Linguistics in Computational Linguistics: Observations and Predictions”, *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics*, pp. 22-25.

Valentin, Paul (1996), “Anapher als kognitiver Prozeß”, *in* Marie-Hélène Pérennec, ed., *Pro-Formen des Deutschen*, Tübingen: Stauffenburg, pp. 179-189.

Van Dijk, Teun A. (2010), *Discourse and Context. A Sociocognitive Approach*, Cambridge et al.: Cambridge University Press.

Vater, Heinz (2001), *Einführung in die Textlinguistik. Struktur und Verstehen von Texten* (3rd ed.), München: Fink.

Vater, Heinz (2005), *Referenz-Linguistik*, München: Fink.

Versley, Yannick et al. (2008), “BART: A Modular Toolkit for Coreference Resolution”, *Proceedings of the ACL-08: HLT Demo Session*, pp. 9-12.

Vicedo, José L. & Antonio Ferrández (2000), “Applying Anaphora Resolution to Question Answering and Information Retrieval Systems”, *in* Hongjun Lu & Aoying Zhou, eds., *Proceedings of the First International Conference on Web-Age Information Management*, Berlin – Heidelberg: Springer, pp. 344-355.

Vieira, Renata & Massimo Poesio (2000), “An Empirically Based System for Processing Definite Descriptions”, *Computational Linguistics*, 26 (4), pp. 539-593.

Voorhees, Ellen (1998), “Using WordNet for Text Retrieval”, *in* Christiane Fellbaum, ed., *WordNet. An Electronic Lexical Database*, Cambridge – London: Massachusetts Institute of Technology, pp. 285-303.

Voutilainen, Atro (2004), “Part-of-Speech Tagging”, *in* Ruslan Mitkov, ed., *The Oxford Handbook of Computational Linguistics*, Oxford – New York: Oxford University Press, pp. 219-232.

Waltinger, Ulli & Alexa Breuing (2012), “Internet-based Communication”, *in* Alexander Mehler & Laurent Romary, eds., *Handbook of Technical Communication*, Berlin – Boston: De Gruyter, pp. 533-569.

Wawra, Daniela (2004), *Männer und Frauen im Job Interview. Eine evolutionspsychologische Studie zu ihrem Sprachgebrauch im Englischen*, Münster: LIT.

Wawra, Daniela (2008), *Public Relations im Kulturvergleich. Die Sprache der Geschäftsberichte US-amerikanischer und deutscher Unternehmen*, Frankfurt am Main et al.: Lang.

Wawra, Daniela (2011), “Wandel und Variation des Mediums Sprache. Ursachen – Tendenzen – Ergebnisse”, *in* Institut für interdisziplinäre Medienforschung, ed., *Medien und Wandel*, Berlin: Logos, pp. 91-109.

Way, Andy (2010), “Machine Translation”, *in* Alexander Clark, Chris Fox & Shalom Lappin, eds., *The Handbook of Computational Linguistics and Natural Language Processing*, Malden, MA et al.: Wiley-Blackwell, pp. 531-573.

Webber, Bonnie & Nick Webb (2010), “Question Answering”, *in* Alexander Clark, Chris Fox & Shalom Lappin, eds., *The Handbook of Computational Linguistics and Natural Language Processing*, Malden, MA et al.: Wiley-Blackwell, pp. 630-654.

Weber, Roger (2006), “Multimedia Retrieval” (script for the course in the winter term 2006/2007), Universität Basel, http://informatik.unibas.ch/lehre/ws06/cs342/slides/ (date of last access: 02/01/2012).

“Website” (n.d.), *TheFreeDictionary.com*, http://www.thefreedictionary.com/website (date of last access: 16/10/2012).

“Where” (n.d.), *TheFreeDictionary.com*, http://www.thefreedictionary.com/where (date of last access: 10/07/2012).

Wikimedia Deutschland (2011), *Alles über Wikipedia und die Menschen hinter der größten Enzyklopädie der Welt*, Hamburg: Hoffmann und Campe.

“Wikipedia: About” (2013), http://en.wikipedia.org/wiki/Wikipedia:About (date of last access: 17/11/2013).

“Wikipedia: Featured articles” (2014), http://en.wikipedia.org/wiki/Wikipedia: Featured_articles (date of last access: 17/12/2014).

"Wikipedia: Good articles" (2014), http://en.wikipedia.org/wiki/Wikipedia:Good_articles (date of last access: 17/12/2014).

Wilpert, Gero von (2001), *Sachwörterbuch der Literatur* (8th ed.), Stuttgart: Kröner.

Yule, George (2010), *The Study of Language* (4th ed.), Cambridge et al.: Cambridge University Press.

Yus, Francisco (2007), "Weblogs: Web Pages in Search of a Genre?", *in* Santiago Posteguillo, María José Esteve & M. Lluïsa Gea-Valor, eds., *The Texture of Internet. Netlinguistics in Progress*, Newcastle: Cambridge Scholars.

Żebrowska, Ewa (2013), *Text – Bild – Hypertext*, Frankfurt am Main et al.: Lang.

Zhai, ChengXiang (2009), *Statistical Language Models for Information Retrieval*, San Rafael, CA: Morgan & Claypool.

Index

DOI 10.1515/9783110416756