In-Depth Analysis of Anaphora Resolution Requirements

Helene Schmolz
Chair of English Language and Culture
University of Passau
Passau, Germany
Email: Helene.Schmolz@uni-passau.de

David Coquil, Mario Döller
Chair of Distributed Information Systems
University of Passau
Passau, Germany
Email: David.Coquil, Mario.Doeller@uni-passau.de

***Abstract*—This paper aims to lay the foundations of an anaphora resolution framework able to process all types of hypertexts and treat all types of anaphors for the English language. To this end, we provide a linguistically unambiguous and extensive definition and categorization of the concept of anaphora. We introduce a new corpus, and use our proposed categorization to statistically analyze it. Finally, we describe a preliminary version of our framework and outline promising results of the first experimental evaluations.**

***Index Terms*—Entity resolution; text mining; corpus construction; anaphora**

I. Introduction

Anaphora can be defined as "the use of a word which refers to, or is a substitute for, a preceding word or group of words" (Simpson & Weiner [1]). Anaphora resolution then consists of identifying to which word or group of words (the *antedecent*) the anaphor refers. Consider, for example, the following sentences: *Tom plays the piano. He likes music*. Here, "he" is the anaphor, "Tom" the antecedent.

Automatic anaphora resolution is a problem in natural language processing. The potential applications in text processing are manifold: information extraction, information retrieval, text summarization etc. However, despite a long and rich history of research in computational linguistics, anaphora resolution has only seen limited deployment in real-life applications. This is mainly due to the shortcomings of existing corpora and of existing anaphora resolution methods.

Indeed, corpora compiled up to now are limited in the types of anaphors they consider (including only a small part of the large number of anaphor categories) and in the domains of their texts. This lack of a proper "gold standard" is particularly problematic with respect to a consistent evaluation of anaphora resolution methods. The methods themselves are usually also restricted to a small subset of anaphor types. Many of them focus on very specific text domains, which strongly limits their applicability. A more fundamental problem is that they frequently lack a sound linguistic definition and categorization of the concepts of anaphor and anaphora resolution.

In this paper, we lay the foundations of an anaphora resolution system addressing these issues. For this purpose, we first analyze the state of the art regarding existing corpora and anaphora resolution methods (Section II). Then, in Section III, we describe the main parts of our linguistic definition and categorization of anaphors, and introduce a new corpus. We present the results of a statistical analysis of this corpus with respect to the proposed categories of anaphors. In Section IV, we describe the structure of our anaphora resolution framework and show a proof-of-concept evaluation of its anaphor detection component. Finally, Section V concludes the paper and discusses future work.

II. Related Work

A. Existing Corpora

There are hardly any corpora that examine anaphors extensively and regarding several anaphor types. A few, however, merit attention. One of the most important corpus is part of the Syracuse study (cf. [2]). This corpus consists of 600 abstracts, half from the field of psychology, half from computer science. The following types of anaphors are analyzed: central pronouns, "nominal demonstratives" (i.e. demonstrative pronouns), relative pronouns, "definite article" (i.e. noun phrases with definite article), "pro-verb" (i.e. verb phrases with "do"), "nominal substitutes", "pro-adjectives", "pro-adverbials", and "subject references".

Other corpora are mainly concerned with central pronouns. To give examples, Mitkov & Hallett [3] investigated three corpora: technical manuals that were downloaded from the Internet, newswire texts that were taken from a part of the Penn Treebank corpus, and Jules Verne's *From the Earth to the Moon*. Besides, there are corpora consisting of technical manuals, e.g. from Mitkov, Evans & Orasan [4], who only analyzed anaphoric versus non-anaphoric items. Furthermore, Vicedo & Ferrández [5] examined central pronouns and "who", "whose", "whom" in a corpus consisting of news from the Time newspaper, medical journals, abstracts and extracts from information science and other computational and technical content.

The corpora mentioned above have several drawbacks. Most importantly, they focus on central pronouns and ignore the distribution of different types of anaphors in texts. The only corpus doing that - the one of the Syracuse study - shows, however, other insufficiencies. It seems doubtful, for instance, if some items that are treated as anaphors, e.g. "anybody", "everything", "someone", are in any context

anaphoric. Furthermore, this corpus ignores many types of anaphors: reciprocal pronouns, non-finite clauses, negations of verb phrases with "do" such as "don't", combinations of "do" such as "do this", and the adverbs "when", "while", "why", "whence", "whereby", "wherein", "whereupon", ellipses, cataphors. Finally, many existing corpora do not consider any hypertexts, e.g. Wikipedia texts, blogs or articles from online newspapers, and no corpus so far pays attention to the full range of hypertexts.

B. Anaphora Resolution Frameworks or Methods

One of the oldest algorithms that is still used as a reference is Hobbs' algorithm [6]. It solves personal and possessive pronouns with noun phrases as antecedents. Furthermore, Lappin and Leass' RAP [7] is frequently quoted when discussing anaphora resolution systems. In addition to Hobbs' algorithm, RAP solves reflexive pronouns and reciprocal pronouns. RAP also contains a procedure for identifying pleonastic pronouns i.e. non-anaphoric "it". Comparing Hobbs' and Lappin and Leass' algorithms, Lappin and Leass' RAP generally performs better (cf. [8], [3]).

Hobbs' and Lappin & Leass' algorithms are based on full parsing, but there are others which rely on partial parsing. Among commonly cited ones are Kennedy and Boguraev's algorithm [9], Mitkov's approach [10] and its newer version of MARS [4], and Baldwin's CogNIAC [11]. Kennedy & Boguraev's algorithm is based on Lappin and Leass' but only involves partial parsing. Baldwin's CogNIAC solves central pronouns. Finally, Mitkov's algorithm considers personal pronouns (cf. [8], [3]). Algorithms with partial parsing, however, generally perform worse than those with full parsing. Among the three algorithms (not considering MARS), Kennedy & Boguraev's algorithm scores best (cf. [3]). Apart from algorithms based on pronoun resolution, there are approaches that only focus on the detection of pleonastic "it". One important contribution comes from Boyd et al. [12].

Moreover, the above-mentioned algorithms all focus on pronouns. Research on other types of anaphors has started later, but has considerably increased in the last two decades of research. To give examples, Vieira & Poesio [13] focus on noun phrases with "the" ("definite descriptions" in their terminology). Meyer & Dale [14] describe an algorithm focusing on noun phrases with "the", especially cases where anaphora resolution is difficult to carry out, to which they refer as "associate anaphora".

Besides, there are algorithms that cover even more classes of anaphors. For example, Soon, Ng and Lim [15] consider noun phrases and proper nouns, next to central and demonstrative pronouns, however, only when they are coreferential. Another algorithm, SUPAR by Ferrández et al. [16] claims to solve anaphoric 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"). Apart from such algorithms for anaphora resolution, there is research focusing on the distinction between anaphoric and non-anaphoric noun phrases. Ng & Cardie [17], for example, consider coreferential noun phrases with "the" as well as proper names, pronouns and other noun phrases in their anaphoric and non-anaphoric use. To conclude, all algorithms up to now focus on one or a few anaphor types. No framework considers all types of anaphors.

III. General Framework

A. Anaphor Definition and Categorization

In this section, we define and categorize anaphors and illustrate our proposals using the following set of sentences:

1) **Tom** plays the piano. He likes music.
2) After she had come home, **Susan** answered her e-mails.
3) **Betty** repaired the lamp. Amazing! The girl is only twelve years old.
4) Tom bought a blue **shirt**. Simon bought a green one.
5) *There* are still plenty of cookies left.
6) **Simon** could not remember when to open the shop.
7) **The students** working on their theses attended a course about writing skills.
8) **The goals** scored by the team were impressive.

As current definitions of anaphors, such as the one given in the introduction, are too unspecific for our purpose, and anaphors show further properties, we outline six conditions that have to hold for items in order to be considered as anaphors here:

1) Anaphors refer back as well as forward, i.e. the antecedent either precedes or follows the anaphor. 2) Anaphors have to have an explicit antecedent, i.e. an antecedent which occurs in the same text. 3) Anaphors are interpreted in relation to their antecedents. 4) The relation between anaphor and antecedent is coreferential, substitutional, or shows other, miscellaneous features. 5) The use of anaphors leads to a reduction of the text and/or avoids excessive repetition. 6) Anaphors contribute to the cohesion of a text.

Condition one means that we also consider items taking a forward direction, e.g. (2), although these are infrequent. Such items are called "cataphors", but are here rather seen as special cases of anaphors [18]. Regarding condition three, it depends on the specific anaphor in how far their interpretation derives from the antecedent. Some anaphors have no particular meaning on their own but gain one from the context by referring back to antecedents, e.g. (1). Other anaphors contain further information such as noun phrases with "the", e.g. (3). As stated in condition four, we regard items where the relation between anaphor and antecedent is either coreferential i.e. both entities refer to the same thing in the real world, or substitutional i.e. the anaphor just replaces an expression without any intent to be coreferential (see Quirk et al. [19] for a more detailed definition). In rare cases, the relationship is both or neither and so falls into a third category with miscellaneous features. Examples of coreferential relationships are (1), (2), and (3); a substitutional relationship is shown in (4). Consequently, frameworks considering coreferential chains or only coreference have different objectives and fields of investigation and so are only partly comparable with our

Type of anaphor	Subcategories	Anaphoric items
CENP	personal pronouns	“he”, “she”, “it”, “they”, “him”, “her”, “them”
	possessive pronouns	“his”, “her”, “hers”, “its”, “their”, “theirs”, “mine”, “ours”, “yours”
	reflexive pronouns	“himself”, “ours”, “yours”
RECP		“each other”, “one another”
DEMP	dependent function	“this”, “that”, “these”, “those”
	independent function	“this”, “that”, “these”, “those”
RELP		“that”, zero “that”, “ who”, “whom”, “ whose”, “which”
ADV		“when”, “while”, “where”, “why”, “whence”, “wherein”, “whereupon”, “whereby”, “there”, “here”, “then”
INDP		“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”
OCS		“the same”, “such”, “so”
VPDO	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”, “so doing”, “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”, “did it”, “doing it”, “done it”, “do the same (thing)”, “does the same (thing)”, “did the same (thing)”, “done the same (thing)”
NFC		“to”, “-ing”, “-ed”

TABLE I
TYPES OF ANAPHORS

framework. As maintained in (6), anaphors contribute to the cohesion of a text because cohesion can be seen as “visualized semantic relation” between expressions in a text. As a result, anaphors are important means in disclosing the content of a text. Finally, all items that work anaphorically can - to a varying extent - also be used non-anaphorically (see (5)). It is therefore important for anaphora resolution systems to be able to distinguish between anaphoric and non-anaphoric items.

With that in mind, we now focus on the types of anaphors that can be distinguished. These types result from the six characteristics discussed above. The classification so is based on linguistic principles as well as practicability for computational tasks, considering mostly Quirk et al. [19] and Stirling & Huddleston [18]. This means that we take a broad definition of what is regarded as an anaphor in this paper, as Stirling & Huddleston [18], for instance, also did. Consequently, cataphors and ellipses, for example, are special cases of anaphors for us.

We distinguish twelve types of anaphors: central pronouns (CENP), reciprocal pronouns (RECP), demonstrative pronouns (DEMP), relative pronouns (RELP), adverbs (ADV), noun phrases with definite article (NPT), proper names (PROPN), indefinite pronouns (INDP), other forms of coreference and substitution (OCS), verb phrases with “do” and combinations with “so”, “this”, “that”, “it”, “the same (thing)” (VPDO), ellipses (ELL), and non-finite clauses (NFC). It is worth explaining that all DEMP can occur in dependent, i.e. when they are part of noun phrases, or independent function, i.e. when they are noun phrases themselves. NPT are phrases with “the”. With PROPN, we just distinguish between those for people (personal proper names) and all other instances (other proper names). The type OCS contains items which do not fall into one of the other categories. ELL fall into the subcategories nominal, verbal, and clausal ellipsis. All other anaphor types are presented in Table I, together with their subcategories, and the items for these types. Cataphoric instances do not form a separate category but rather are special cases within the corresponding anaphor types.

As described above, one type of anaphors are non-finite clauses. Up to now, this type has never been treated in the discussion of anaphora resolution. Even in linguistics, only Quirk et al. grant them a, however, meagre and unsatisfactory entry. Resolving non-finite clauses is of importance because English often prefers them, e.g. in (7), instead of finite clauses such as “The students who are working on their theses attended a course about writing skills.” (cf. [20]). Examples of anaphors in the form of “to”, “-ing”, and “-ed” are given in (6), (7), and (8). In the Appendix, we give an outline of how to identify anaphoric and non-anaphoric “to”-, “-ing”-, and “-ed”-items.

B. Sample Corpus

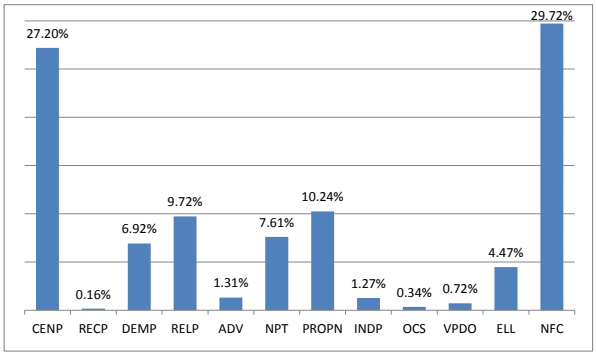

Fig. 1. Distribution of anaphors across the whole corpus (numbers relative to all anaphors)

As there is no corpus that examines the frequency of all types of anaphors in hypertexts, we have established a new corpus. We took Rehm’s [21] classification, which identifies all hypertext sorts on the Internet, and grouped these into three sorts for our purpose: Wikipedia texts as representation of online encyclopedias, texts from blogs, and texts from traditional websites (including homepages of companies, institutional homepages, personal homepages, and online newspapers). Wikipedia texts were chosen by using Wikipedia’s main topic classifications. Blog texts were selected by referring to Technorati’s blog directory (http://technorati.com/blogs/directory/) and Google’s blogsearch (http://www.google.com/blogsearch). Finally,

	CENP	RECP	DEMP	RELP	ADV	NPT	PROPN	INDP	OCS	VPDO	NFC	**Items in total**
Internet texts	1,666	10	424	596	80	466	627	78	21	44	1,822	6,127
in total (anaphoric)	(87.2)	(83.3)	(62.4)	(59.8)	(15.1)	(12.3)	(12.7)	(5.3)	(9.0)	(13.4)	(25.3)	(27.4)
Internet texts	244	2	256	401	450	3,314	4,295	1,394	212	285	5,376	16,229
in total (non-anaphoric)	(12.8)	(16.7)	(37.7)	(40.2)	(84.9)	(87.7)	(87.3)	(94.7)	(91.0)	(86.6)	(74.7)	(72.6)

TABLE II
RELATION OF ANAPHORIC AND NON-ANAPHORIC ITEMS IN ABSOLUTE NUMBERS (IN BRACKETS: RELATIVE DISTRIBUTION WITHIN EACH ANAPHOR TYPE IN PER CENT)

the search engine Google (http://www.google.com/), the open directory www.dmoz.org, and the website http://www.gksoft.com/govt/en/gb.html for institutional homepages helped to choose a representation of traditional website texts. In cases where one text was too long, only part of it was selected. In sum, each of the three sorts contains about 25,000 words, resulting in 75,974 words for the whole corpus. The corpus was then annotated with type of anaphor (regarding the twelve types introduced above), antecedent(s) of each anaphor, and relation between anaphor and antecedent (i.e. coreferential, substitutional, or other). More information about the annotated corpus can be found on the website http://www.dimis.fim.uni-passau.de/iris/index.php?view=anares

The analysis of the whole corpus revealed 6,127 instances of anaphors. Consequently, the corpus contains 80.7 anaphors in 1,000 words. How the twelve anaphor types are represented in the whole corpus is shown in Figure 1. Non-finite clauses with 1,822 cases are the most frequent, even slightly outnumbering central pronouns. This is a remarkable result because central pronouns are often considered to be the most widespread and important type (cf. [22], [8]). Apart from these two, proper names, relative pronouns, noun phrases with definite article, demonstrative pronouns, and ellipses are further important. The rest of the twelve anaphor types plays a minor role in the corpus.

As far as for the individual items, "to"-items are the most frequent anaphor items in the corpus, closely followed by "-ing"-items (both from NFC). They take 11.8% and 11.7% of all anaphor items respectively. Other frequent items are noun phrases with definite article (7.6%), personal proper names (from PROPN; 6.3%), and "it" (from CENP; 5.6%). Besides, most anaphor items show an anaphoric direction; only 1.2% of all anaphors are cataphoric. These cataphors are furthermore most common in non-finite clauses: 75.0% of all cataphors are found here.

Turning to each high-scoring anaphor type, the most frequent subcategories and items are as follows: CENP's most important subcategory are personal pronouns (64.0%). Possessive pronouns take 33.4%; the rest falls on reflexive pronouns. The most frequent item of CENP is "it" (20.6%). Furthermore, dependent demonstrative pronouns account for 55%, independent ones for 45%. The most common item here is "this". The subcategory "wh"-forms with 54.3% are the most frequent within RELP. The most common item within RELP is "that" (36.1%); within ADV it is "where" (56.3%). Besides, personal proper names lead within PROPN with 61.7%; and 32.1% of all INDP fall on "one" and "ones". Nominal ellipses make up 90.4% of all 274 anaphoric items of ELL. Finally, "to" (39.8%) and "-ing" (39.5%) are most common within NFC; "-ed"-items take 20.8%. Within "-ed"-forms 82.5% of all items are regular, 17.5% are irregular. Apart from that, we analyzed the ratio of anaphoric versus non-anaphoric items. The most anaphoric items are found with central pronouns; the most non-anaphoric with indefinite pronouns. The details are given in Table II. Yet, a high number of non-anaphoric items does not mean that anaphora resolution is impossible. Anaphora resolution then rather depends on an intelligent framework.

IV. THE ANARES FRAMEWORK

Based on the linguistic foundations introduced in the previous section, we have started to develop a general-purpose anaphora resolution framework called *Anares*. In this section, we present it briefly, with a focus on its rule-based filter, which uses rules defined by linguists to identify potential anaphors.

The global workflow of Anares is composed of a preprocessing step and a main loop. The preprocessing step includes the Stanford parser [23] for parsing and tagging, resulting in a syntax tree in which a number of properties have been identified for each node. These properties correspond to automatically acquired information about terms or group of terms that will be used to evaluate the rules for anaphor identification and execution later on; see the Appendix for examples of evaluated properties. The main loop processes the text sentence by sentence. A sentence is first analyzed in order to identify a set of potential anaphoric elements, and for each of these, a set of possible antecedents. This is done by evaluating rules for anaphora resolution established by linguists. Depending on the type of the rules that were triggered, more specialized (and computationally expensive) processes are executed to identify the "true" anaphors among the candidates as well as the most likely antecedent of each one. These specialized processes are also designed in such a way that they are based on the knowledge of the experts.

The linguistic rules are formalized using first-order logic. It is indeed easy to express our input data (syntax tree and properties) using this representation and to define rules as predicates that hold for certain variables. Multiple predicates can hold for an individual variable and complex systems can be constructed in this way. For the implementation of the rules as machine-readable objects, we used Prolog.

A proof-of-concept implementation of the rule-based filter has been developed in order to perform a first evaluation of our ideas. The set of implemented rules covers the full scope of anaphoric categories identified in Section III-A. The implemented rules include those derived from the examples

of information used for distinguishing anaphoric and non-anaphoric non-finite clause items presented in the Appendix. A small number of rules could not be properly applied because the acquisition of the necessary contextual information has not yet been implemented. Another restriction is that we have only considered rules for detecting anaphor candidates up to now, excluding rules for detecting antecedent candidates. This version of the rule-based filter was applied to a subset of the corpus described in Section III-B. The Prolog rules used by the filter were derived from such information.

Anaphor type	**Found**	**False-positives**	**False-negatives**	**Total**
Central pronouns	210 (94.2%)	13 (5.9%)	13 (5.8%)	223
Demonstrative pronouns	54 (96.4%)	1 (1.8%)	2 (3.6%)	56
Relative pronouns	82 (95.3%)	12 (14.0%)	4 (4.7%)	86
Adverbs	7 (100.0%)	0 (0.0%)	0 (0.0%)	7
Noun phrases with definite article	52 (96.3%)	45 (83.3%)	2 (3.7%)	54
Indefinite pronouns	8 (100%)	2 (25.0%)	0 (0.0%)	8
Non-finite clauses	282 (94%)	47 (15.7%)	17 (6%)	299
Total	695 (94.8%)	120 (16.4%)	38 (5.2%)	733

TABLE III
PRELIMINARY EVALUATION RESULTS

The results of this preliminary evaluation are detailed in Table III. These results obtained with a perfectible set of rules are encouraging. Indeed, the critical performance indicator at this stage is the amount of false negatives, as these actually anaphoric components will be ignored in the rest of the process. We can see that this indicator remains very low for all categories of anaphor. Moreover, the amount of false positives should ideally also be kept low because each of them will be subject to a more expensive identification process at the next stage before being (normally) discarded, resulting in useless additional computation time. We see here that the results are also rather satisfying for this indicator, even very good for the large "central pronouns" category. An exception is the "noun phrase with definite article" category, which probably requires more discriminating rules.

V. CONCLUSION

In this paper, we have described the fundamental components of a versatile anaphora resolution framework that can process all types of texts and anaphors. To achieve that, we have defined an extensive linguistic basis with a complete definition of anaphors and anaphoric categories, and introduced a generic corpus. We have statistically characterized this corpus based on our proposed categories. We have detailed a first version of an anaphora resolution system based on this work, and evaluated its rule-based anaphor identification component, which shows promising results even with a preliminary version of its rule set.

Regarding future work, one of our aims is to keep on expanding the corpus. Additionally, we plan to develop further processes for acquiring contextual knowledge about possible anaphors and antecedents (genus, animate/inanimate, etc.). This will enrich the information produced after the preprocessing step, and enable the integration of more complex rules into the framework.

REFERENCES

[1] J. A. Simpson and E. S. Weiner, eds., *The Oxford English Dictionary*, p. 436. Clarendon Press, 1989.

[2] E. Du Ross Liddy, "Anaphora in natural language processing and information retrieval," *Information Processing & Management*, vol. 26, no. 1, pp. 39–52, 1990.

[3] R. Mitkov and C. Hallett, "Comparing pronoun resolution algorithms," *Computational Intelligence*, vol. 23, no. 2, pp. 262–297, 2007.

[4] R. Mitkov, R. Evans, and C. Orasan, "A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method," in *Computational Linguistics and Intelligent Text Processing* (A. Gelbukh, ed.), LNCS 2276, pp. 168–186, Springer, 2002.

[5] J. Vicedo and A. Ferrández, "Applying anaphora resolution to question answering and information retrieval systems," in *Proceedings of the First International Conference on Web-Age Information Management* (H. Lu and A. Zhou, eds.), LNCS 2276, pp. 344–355, Springer, 2000.

[6] J. Hobbs, "Resolving pronoun references," in *Readings in Natural Language Processing* (B. Grosz, K. Sparck Jones, and B. Lynn Webber, eds.), Morgan Kaufmann, 1978.

[7] S. Lappin and H. Leass, "An algorithm for pronominal anaphora resolution," *Computational Linguistics*, vol. 20, no. 4, pp. 535–561, 1994.

[8] R. Mitkov, *Anaphora Resolution*. Longman, 2002.

[9] C. Kennedy and B. Boguraev, "Anaphora for everyone: Pronominal anaphora resolution without a parser," in *Proceedings of 16th Conference on Computational Linguistics*, pp. 113–118, 1996.

[10] R. Mitkov, "Robust pronoun resolution with limited knowledge," in *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference*, pp. 869–875, 1998.

[11] B. Baldwin, "CogNIAC. High precision coreference with limited knowledge and linguistic resources," in *Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pp. 38–45, 1997.

[12] A. Boyd, W. Gegg-Harrison, and D. Byron, "Identifying non-referential *it*. A machine learning approach incorporating linguistically motivated patterns," in *Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP*, pp. 40–47, 2005.

[13] R. Vieira and M. Poesio, "An empirically based system for processing definite descriptions," *Computational Linguistics*, vol. 26, no. 4, pp. 539–593, 2001.

[14] J. Meyer and R. Dale, "Mining a corpus to support associative anaphora resolution," in *Proceedings of the Fourth Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2002)*, 2002.

[15] N. Soon, Wee Meng, H. Tou, and D. Chung Yong Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, vol. 27, no. 4, pp. 521–544, 2001.

[16] A. Ferrández, M. Palomar, and L. Moreno, "An empirical approach to Spanish anaphora resolution," *Machine Translation*, vol. 14, no. 3–4, pp. 1–16, 1999.

[17] V. Ng and C. Cardie, "Identifying anaphoric and non-anaphoric noun phrases to improve coreference resolution," in *Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, pp. 1–7, 2002.

[18] L. Stirling and R. Huddleston, *The Cambridge Grammar of the English Language*, ch. Deixis and Anaphora, pp. 1449–1564. Cambridge University Press, 2010.

[19] R. Quirk *et al.*, *A Comprehensive Grammar of the English Language*. Longman, 2008. First published 1985.

[20] B. Kortmann, *English Linguistics: Essentials*. Cornelsen, 2005.

[21] G. Rehm, *Hypertextsorten. Definition – Struktur – Klassifikation*. Books on Demand, 2007.

[22] R. Carter and M. McCarthy, eds., *Cambridge Grammar of English. A Comprehensive Guide. Spoken and Written English Grammar and Usage*. Cambridge University Press, 2006.

[23] D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423–430, 2003.

APPENDIX

Examples of linguistic properties used to identify anaphoric and non-anaphoric items

Properties of anaphoric non-finite clauses:

	clause function					**phrase function**		
	adver-ver-bial	subject comple-ment	direct object	object com-plement	appositive use	prepositional complement	postmodifi-cation in noun phrases	postmodification in adjective phrases
"to"	✓[1]	✓	✓[2]	✓[3]	✓	×	✓	✓
"-ing"	✓[1]	✓	✓	✓[3]	✓	✓	✓	✓
"-ed"	✓[1]	×	✓[4]	✓[3]	✓	×	✓	×

[1] can be preceded by a conjunction
[2] wh-element can precede "to"
[3] rare
[4] always includes a subject

Non-anaphoric uses of non-finite clause items and other forms looking like non-finite clause items:

	nouns	ger-unds	adjec-tives	prepo-sitions	prepositional adverbs	as sub-ject	in extra-posi-tion	part of verbs and other fixed expressions	simple finite verb phrases		complex finite verb phrases		
									present forms	past forms	be	have	modal verbs
"to"	×	×	×	✓	✓	✓	×	✓	×	×	×	×	×
"-ing"	✓	✓	✓	✓	×	✓[1]	✓	×	✓	×	✓	✓	✓
"-ed"	✓	×	✓	×	×	✓	×	×	✓	✓	✓	✓	✓
irregular "-ed"	×	×	✓	×	×	✓	×	×	×	✓	✓	✓	✓

[1] anaphoric use is possible, but rare

Processes for identifying potential anaphoric items of non-finite clauses for anaphora resolution:

1. search for "to"-items, and all items ending in "-ing" and "-ed"
2. exclude non-anaphoric uses of these items as outlined above
3. the remaining items are potentially anaphoric

Example:

Text (from Wikipedia: Australia (continent), date of last access: 06/01/2012):

> When the last ice age *ended* in about 10,000 BC, *rising* sea levels *formed* Bass Strait, separating Tasmania from the mainland. Then between about 8,000 and 6,500 BC, the lowlands in the north were *flooded* by the sea, separating New Guinea and Australia.

Identifying potential anaphors:

1. the items "ended", "rising", "formed", "separating", "flooded", "separating" are found
2. the items "rising" (adjective), "ended", "formed" (past forms, simple finite verb phrases), "flooded" ("be", complex finite verb phrases) are non-anaphoric and consequently excluded
3. the two remaining items "separating" are anaphoric here