Influence of Text Type and Text Length on Anaphoric Annotation

Daniela Goecke[1], **Maik Stührenberg**[1], **Andreas Witt**[2]

Universität Bielefeld[1], Universität Tübingen[2]
Fakultät für Linguistik und Literaturwissenschaft, SFB441
Postfach 10 01 31, Nauklerstr. 35
33501 Bielefeld, 72074 Tübingen
Germany, Germany
{daniela.goecke, maik.stuehrenberg}@uni-bielefeld.de, andreas.witt@uni-tuebingen.de

Abstract

We report the results of a study that investigates the agreement of anaphoric annotations. The study focuses on the influence of the factors text length and text type on a corpus of scientific articles and newspaper texts. In order to measure inter-annotator agreement we compare existing approaches and we propose to measure each step of the annotation process separately instead of measuring the resulting anaphoric relations only. A total amount of 3642 anaphoric relations has been annotated for a corpus of 53038 tokens (12327 markables). The results of the study show that text type has more influence on inter-annotator agreement than text length. Furthermore, the definition of well-defined annotation instructions and coder training is a crucial point in order to receive good annotation results.

1. Introduction

The need for high quality corpus data is widely accepted for linguistic analyses and the development of linguistic applications. Data quality might be measured against reference corpora and – as reference corpora are not available when annotating data – by measuring inter-annotator agreement (IAA) that gives information on how consistent annotations of different coders are. The work presented is based on the development of an anaphora resolution system for both cospecification and bridging relations. In order to annotate a corpus for analysis, training and evaluation we investigated factors that influence coder agreement. The study focuses on two hypotheses:

1. Text type has an influence on the IAA of anaphoric annotation; agreement values are lower for more complex texts types.
2. Text length has an influence on the IAA of anaphoric annotation; long texts show less agreement than short texts e.g. due to the large number of antecedent candidates.

In order to check these hypotheses we have chosen two text types: German Newspaper texts and scientific articles. We assume scientific articles to have a more complex text structure and more complex topics than newspaper texts. Thus, we expect scientific articles to be less easy to understand and, therefore, to have lower agreement values than newspaper texts. In order to investigate the influence of text length independently from from text type, we have chosen texts of varying length for both text types. Influence of text length is thus investigated first for homogeneous sets of texts according to text type and checked afterwords for the heterogeneous group of two different text types.
The remainder of the paper is structured as follows: Section 2. describes the annotation of anaphoric relations and Section 3. reviews existing approaches to measure anaphoric agreement and presents our proposal to measure different annotation steps independently. Section 4. describes the setting of the study as well as the results. Finally, Section 5. derives a conclusion and gives clues for further development.

2. Annotating Anaphoric Relations

In order to annotate anaphoric relations, two types of information have to be specified. First, the markables, i. e. the elements that can be part of a relation, have to be identified. Second, the relation(s) between markables with their respective types and subtypes have to be chosen from a set of disjoint categories. In the domain of anaphoric relations, markables are those text units that evoke discourse entities. Discourse entities – or discourse referents – are constants within a discourse model that are evoked by NPs and which can be referred to in the subsequent discourse (Karttunen, 1976; Webber, 1988; Kamp and Reyle, 1993). These markables form the basis for the annotation process and are annotated in advance. By separating the tasks of markable detection and anaphora annotation proper we follow Hirschman et al. (1998) who describe for an annotation study that coder agreement can be increased if markables are detected in advance.
For our corpus, each text has been preprocessed using the dependency parser Machinese Syntax[1] which provides lemmatisation, POS information, dependency structure, morphological information and grammatical function. Based on this information, markables have been detected automatically afterwards by identifying nominal heads (i.e. nouns or pronouns) and their premodifiers.
For the annotation process we have investigated several annotation schemes for annotating anaphoric relations that have been developed in the last years, e.g. the text-based UCREL anaphora annotation scheme (Fligelstone, 1992; Garside et al., 1997), the SGML-based MUC annotation scheme (Hirschman, 1997), and the XML-based MATE/GNOME scheme for anaphoric annotation (Poesio, 2004), amongst others. The annotation scheme used for our ap-

[1] http://www.connexor.eu

proach is XML-based, too, and is based on the one presented by Holler et al. (2004) and has been adapted for the annotation of bridging relations (Clark, 1977). We introduce two primary relation types to distinguish cospecification (direct anaphora; *cospecLink*) and bridging relations (associative or indirect anaphora; *bridgingLink*). For each primary relation, a set of secondary relations has been defined.

For direct anaphora we annotate eight secondary relation types: *ident*, *namedEntity*, *propName*, *synonym*, *hyperonym*, *hyponym*, *addInfo*, *paraphrase*. The relation *ident* is chosen for pronominal anaphors or anaphor-antecedent pairs with identical head noun. Markables that are not of type *namedEntity* but refer to a markable of type *namedEntity* are annotated with the respective relation type. The value *propName* is chosen if the anaphoric element is a proper name that refers to an NP markable. Synonymy between the head nouns of anaphor and antecedent is annotated using the value *synonym*. Hyperonymy and Hyponymy between the head nouns of anaphor and antecedent are annotated by choosing the respective secondary relation types. The values *addInfo* and *paraphrase* are chosen if the anaphoric markable adds new information to the discourse or if the anaphor is a paraphrase of its antecedent.

For bridging relations six secondary relation types have been defined: *poss*, *meronym*, *holonym*, *hasMember*, *setMember*, *bridging*. The value *poss* is chosen if a possession relation between the anaphoric element and its antecedent is marked by a possessive pronoun or a NP_{gen}. The value *meronym* is chosen in case of a part-whole-relation between the head nouns of anaphora and antecedent; *holonym* is chosen accordingly. The value *hasMember* is chosen if the anaphor describes a set and the antecedent(s) are part of of that set and *setMember* is chosen if the anaphoric elements is part of a set described by its antecedent. If none of the previous relation types holds the relation *bridging* is used (e.g. wedding – bride).

Figure 1 shows a flowchart that serves as a guidance for the annotators. For each primary relation the coders can check each of the secondary relation types: the most general relation is checked last and is thus only chosen if none of the previous holds.

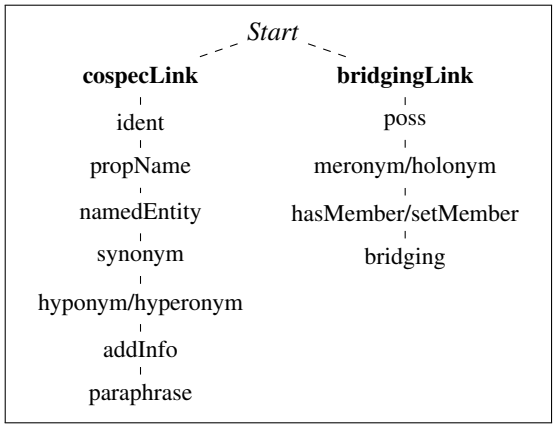

Figure 1: Flowchart for annotation process

The annotation schema is defined both as an annotator's manual as well as an XML DTD and XML Schema (XSD), the latter being the basis for the annotation tool *Serengeti*[2] that has been developed for the task under investigation and which is described in detail in Stührenberg et al. (2007). *Serengeti* is a web application thus its architecture is separated into a client and a server side. Documents and their annotations are managed centrally on the server side, all user interactions are rendered locally on the client side. The Graphical User Interface (GUI) of *Serengeti* is subdivided into several areas (cf. Figure 2). The main area renders the text to be annotated, roughly laid out in terms of paragraphs, lists, tables and non-text sections according to the input XML data. Additionally, the predefined markables are underlined and followed by boxes containing the markables' unique identifiers. These boxes serve as clickable buttons to choose markables during the annotation. A section at the bottom of the interface represents the annotation panel with a list of all annotated relations on the left and all editing tools – e.g. for choosing the relation type – on the right side. The annotators do not have to code relations as XML elements and thus the use of the annotation tool leads to good annotation results both in terms of quantity and quality.

On the basis of the annotated data we investigate the influence of text length and text type on the inter-annotator-agreement. During the annotation process four aspects of disagreement might occur: The coder has to decide if a markable is discourse-new or used anaphorically and for each anaphoric element its antecedent and the respective primary and secondary relation type has to be chosen. The annotation process can thus be subdivided according to these aspects that are analysed separately in Section 4.

1. For each discourse entity (DE), decide if the DE is used anaphorically;
2. For each anaphoric DE, identify the correct antecedent;
3. For each anaphora-antecedent-pair, choose the primary relation;
4. For each anaphora-antecedent-pair and identified primary relation, choose the secondary relation.

We investigate these aspects independently in order to derive solutions for best annotation practice: If one of the aspects is identified as a source for low IAA agreement this step has to be improved, e.g. by additional preprocessing or further coder training.

3. Agreement

When comparing annotation results, observed (percentage) agreement is not sufficient as it does not take chance into account: percentage figures do not show how much agreement one can expect simply due to chance. Nevertheless, percentage agreement shows trends that have to be checked against chance. Hirschman et al. (1998) and Mitkov et al. (2000) describe inter-annotator agreement on the basis of

[2] http://coli.lili.uni-bielefeld.de/serengeti/

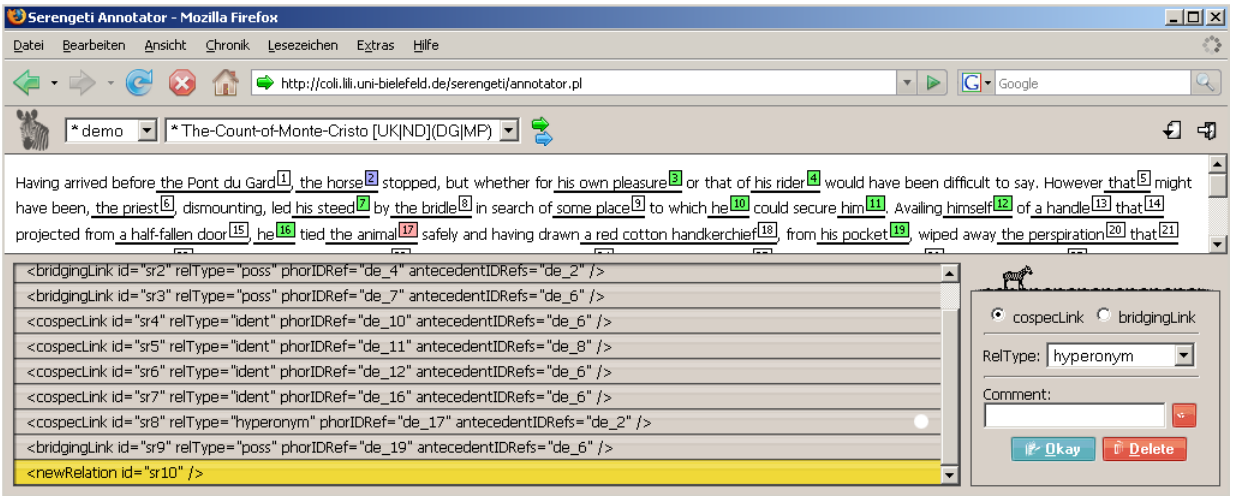

Figure 2: *Serengeti*'s User Interface.

precision and recall measures where one coder is compared against another. Several chance-corrected coefficients have been discussed for their application for linguistic tasks, especially anaphoric annotations (Carletta, 1996; Di Eugenio, 2000; Di Eugenio and Glass, 2004; Artstein and Poesio, 2005). Cohen's κ and Scott's π are of special interest when investigating coder agreement for nominal scales. κ and π both take chance into account but differ in the calculation of the expected agreement. Whereas expected agreement for Scott's π is computed as the sum of squared proportions over all categories, Cohen's κ computes expected agreement as joint probabilities of the coder's single proportions. κ (as well as π) is computed as

$$\kappa = \frac{P(A_o) - P(A_e)}{1 - P(A_e)}$$

where $P(A_o)$ and $P(A_e)$ are the probabilities of observed and expected agreement. An IAA value of 1 describes total agreement, whereas 0 describes agreement with no difference to chance. Negative values describe lower IAA than expected by chance. In the domain of anaphoric annotation, coders agree if they choose the same category (e.g. *cospecLink* or *bridgingLink*) for a given item (e. g. anaphora-antecedent-pair). Considering IAA values as an indicator for the reliability of data we follow Carletta (1996) and thus Krippendorff (1980) and assume $.67 < \kappa < .8$ to allow for tentative conclusions and $\kappa > .8$ to show good reliablilty.

An alternative to the investigation of single items (anaphora-antecedent-pairs) is the analysis of anaphoric chains (Passonneau, 2004; Poesio and Artstein, 2005). The use of anaphoric chains allows coders to choose different antecedents as long as these are part of the same anaphoric chain. As the categories (i. e. chains) need not be to be disjoint, total agreement as well as partial agreement can be considered in terms of chain identity, chain intersection, and chain subsets.

In our study we use Cohen's κ because of three reasons: (1) We are especially interested in the annotation of primary and secondary relation types. This information, however, is not accessible from anaphoric chains. (2) Our annotation guidelines define the correct antecedent to be the previous non-pronominal candidate. Therefore antecedent disagreement is tested for explicitly. (3) Our annotation scheme defines both cospecification and bridging relations. However, anaphoric chains are properly applicable for cospecification only. For bridging relations, we expect short chains only. In case of chains including both cospecification and bridging relations, chains are no longer equivalence classes of the respective discourse entities but rather describe topic chains or lexical chains. In fact, analysis of our data reflects the tendency that bridging relations combine separate cospecification chains into larger combined chains.[3] In the following Section we describe the setting of the study as well as the results.

4. Setting and Results

Data The data consists of two sets of German texts. The first consisting of three scientific articles (14485 tokens, 3805 markables) and twelve newspaper articles (2663 tokens, 763 markables), the second consisting of another three scientific articles (25138 tokens, 4757 markables) and seven newspaper texts (10752 tokens, 2782 markables). A total amount of 3642 different anaphoric relations has been annotated for the two sets. In order to investigate the influence of text length independently from from text type for the second data set, we have chosen texts of varying length for both text types. The average text length is 81 sentences for the newpaper articles and 470 sentences for the scientific articles. Figure 1 gives an overview on the texts of the two data sets.

Coders Two coders participated in the study. Both are students at the department of Linguistics and Literary Studies at the University of Bielefeld.

[3]For example, 163 cospecification chains (minimum length: 2, maximum length: 11) and 40 bridging chains (minimum length: 2, maximum length: 4) have been annotated for one of the scientific articles containing 1776 markables. The combination of cospecification and bridging leads to 179 chains with a minimum length of two and an maximum length of 13 markables.

	#tokens	#markables	#sentences
scientific-1	432	6467	1190
scientific-2	523	9286	1776
scientific-3	455	9385	1791
newspaper-1	40	682	163
newspaper-2	55	795	183
newspaper-3	37	722	188
newspaper-4	55	894	244
newpaper-5	124	2147	549
newpaper-6	106	2043	561
newspaper-7	149	3469	894

Table 1: Texts of the second data set

Annotation Annotation was done using the annotation tool *Serengeti*. After the annotation of the first set, the results were discussed and the annotation guidelines were extended by a flowchart defining the decision process to find the correct secondary relation. Afterwards the second set was annotated independently by the two coders. The coders were advised not to discuss their annotations with each other during the annotation process.

Evaluation of the annotation process Figure 3 shows higher overall agreement for the second set than for the first set. The set of anaphoric relations annotated by the first coder is compared with the set of relations from the second coder and overall agreement is given as percentage agreement of the total set of anaphoric relations. These results reflect the positive effects of coder training and well-defined annotation instructions on IAA.

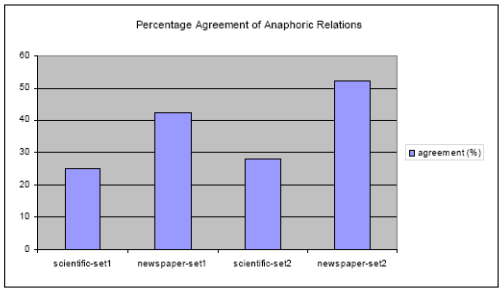

Figure 3: Overall agreement of anaphoric relations

We compute separate agreement values for the four annotation steps described in Section 2. The investigation of text type and text length is based on the annotation of the second data set. Tables 2 and 3 give an overview of the agreement values for scientific and newspaper articles of the second data set. The results for the separate annotation steps are given in the following.

4.1. Identification of anaphoric DEs

We use κ to measure agreement for the identification of anaphoric DEs. Two coders agree, if they choose the same value of the dichotomic category $+/-anaphoric$ for a given item (DE). They disagree if one coder chooses $+anaphoric$ and the second $-anaphoric$ and vice versa. Different subsets of markables (all markables, indefinite descriptions, definite descriptions) are investigated. The total number of items is the number of markables (and their respectives subsets) in the text.

Figure 4 shows κ values for the identification of anaphoric DEs. The results show that coders perform better on newspaper articles than on scientific articles where IAA values of $\kappa < .67$ are observed. For the newspaper articles, only two texts have $\kappa < .67$ and three of the texts have values of $\kappa > .8$ showing good reliability. Regarding agreement values within one set of texts, there is no evidence in the data that text length influences IAA values within one set of texts: Coders perform equally well on short and long newspaper texts. For the scientific articles, coders perform better on the longer texts than on the short one. We therefore assume that the different IAA values across the texts sets are due to text type and not due to text length. Nevertheless, the newspaper texts under investigation are generally shorter than the scientific articles and further investigation has to be done on even longer newspaper texts. We investigated different sets of markables in order to identify sources of disagreement: All markables, only definite description, and the set of markables without indefinite descriptions. In general, best performance values can be observed for the complete set and for the set without indefinite descriptions wich contains both pronouns and definite descriptions.

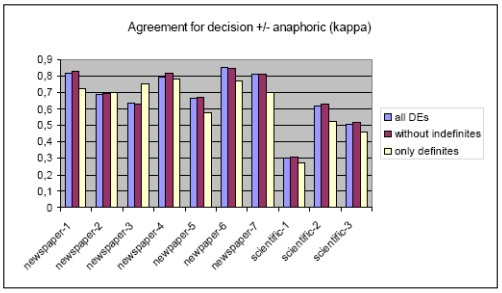

Figure 4: κ values for first annotation step

Good performance for these sets is due to the number of indefinite descriptions that are not anaphorically and due to the number on pronouns that are – except from expletiva – always used anaphorically. Regarding the set of definite description, agreement values tend to be lower, especially for the scientific articles. We assume that disagreement occurs because of missing domain knowledge that is necessary in order to distinguish anaphorical and non-anaphorical use of definite descriptions.

4.2. Identification of correct antecedent

Agreement of antecedent identification is measured in terms of percentage agreement. We use percentage agreement here, as κ is not applicable due to missing $P(A_e)$ for the choice of the antecedent. We do not compute κ for all anaphor-candidate-pairs and the category $+/-correct$ due to the fact that $P(A_e)$ is difficult to define for long texts because a fixed search window (e. g. window size in terms of sentences or markables) is not applicable. For the corpus under investigation only 51% of all anaphoric markables and only 39% of the non-pronominal anaphors find their antecedent within a search window of two sentences. The

	scientific-1			scientific-2			scientific-3		
	#items	#iai	κ	#items	#iai	κ	#items	#iai	κ
step 1	1190	845	0.3	1776	1565	0.62	1791	1482	0.51
step 2	128	98	–	236	231	–	250	211	–
step 3	98	97	.94	231	229	.96	211	206	.82
step 4 (cospec)	89	86	.66	200	190	0.8	194	194	1.0
step 4 (bridging)	8	6	.39	29	28	0.74	12	12	1.0

Table 2: IAA for scientific articles (#iai: items that have been annotated identically)

	newspaper-1			newspaper-2			newspaper-3			newspaper-4		
	#items	#iai	κ	#items	#iai	κ	#items	#iai	κ	#items	#iai	κ
step 1	163	151	.82	183	160	.69	188	157	.64	244	226	.79
step 2	39	34	–	39	36	–	49	43	–	47	43	–
step 3	34	34	1.0	36	36	1.0	43	43	1.0	43	42	.88
step 4 (cospec)	24	22	.84	28	26	.82	37	36	.96	38	35	.86
step 4 (bridging)	10	10	1.0	8	8	1.0	6	6	1.0	4	4	1.0

	newspaper-5			newspaper-6			newspaper-7		
	#items	#iai	κ	#items	#iai	κ	#items	#iai	κ
step 1	549	475	.66	561	526	.86	894	820	.81
step 2	114	96	–	157	139	–	245	210	–
step 3	96	94	.91	139	139	1.0	210	208	.97
step 4 (cospec)	83	83	1.0	126	125	.98	169	158	.87
step 4 (bridging)	11	9	.44	13	13	1.0	39	39	1.0

Table 3: IAA for newspaper articles (#iai: items that have been annotated identically)

total number of items is the number of markables that have been decided as anaphoric by both coders, i.e. the intersection of step 1.

Figure 5 shows percentage agreement for the identification of the antecedent of an anaphoric element. Coders perform well on all texts independently of the text type. Lowest percentage agreement can be observed for one of the scientific articles on which the coders performed least in step 1, too. There is no evidence for an influence of text length or text type on this task and we assume the annotation of antecedent to be solvable with rather high agreement.

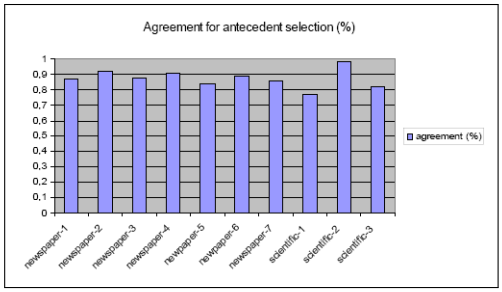

Figure 5: Percentage agreement for second annotation step

4.3. Identification of primary and secondary relation

Agreement values for primary (step 3) and secondary relation types (step 4) are measured using κ. Coders agree if they choose the same category from a set of categories for a given anaphor-antecedent-pair. The total number of items for the annotation of the primary relation type is the number of anaphoric markables for which both coders have chosen the same antecedent, i.e. the intersection of step 2. The total number of items for the annotation of the secondary relation types is the number of anaphor-antecedent-pairs for which both coders have chosen the same primary relation, i.e. the intersection of step 3. Figure 6 shows κ values for the annotation of primary and secondary relation types.

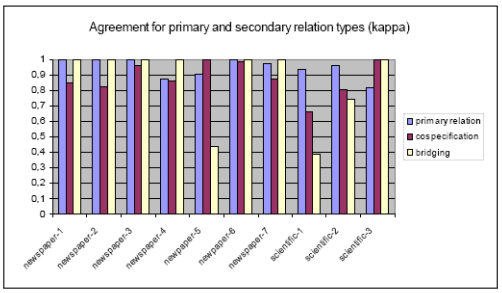

Figure 6: κ values for primary and secondary relation types (second data set)

Regarding the selection of the primary relation for one anaphor-antecedent pair, coders perform well for both newspaper articles and scientific articles with $.9 < \kappa \leq 1.0$ for eight texts and $.8 < \kappa < .9$ for two texts. Regarding the annotation of secondary relation types coders performed equally well for both cospecification and bridging relations and across text types. Coders agree with $.8 < \kappa \leq 1.0$ with one exception for cospecification and two exceptions for bridging relations. $\kappa < .67$ for both cospecification and briddging could be observed for the scientific article that led to low agreement values in the previous steps, too. One of the newspaper texts has $\kappa < .67$ for the bridging relations.

Apart from these exceptions κ values are remarkably high

for these tasks and there is no evidence for an influence of text length or text type on the tasks. Thus we consider the annotation of primary and secondary relation types as a solvable task leading to reliable data.

5. Discussion and Outlook

We have presented the results of a study investigating the agreement of anaphoric annotations on the basis of scientific articles and newspaper texts. The results support our first hypothesis (influence of text type) for the first annotation step: κ values for the classification of anaphoric and non-anaphoric markables are higher for newspaper texts than for scientific articles. However, there is no evidence for an influence of text type on agreement values for the annotation of primary and secondary relation types. The second hypothesis focuses on the effects of text length on agreement values. The results show high overall agreement for the choice of the correct antecedent and the selection of primary and secondary relation types that are independent from text length. However, further investigations should be done using even longer newspaper texts.

The results show that the annotation of anaphoric relations can be further improved by focusing on the detection of anaphoric and non-anaphoric markables either by preprocessing markables or by enhancing guidelines and training. The data shows an remarkably high agreement for the selection of primary and secondary relation types for anaphora-antecedent-pairs. Thus, coder training and the definition of well-defined annotation guidelines leads to high coder agreement even for tasks like anaphora annotation that are considered as rather difficult to annotate reliably.

6. Acknowledgements

The work presented in this paper is part of the project A2 *Sekimo* of the Research Group *Text-technological Modelling of Information* funded by the German Research Foundation. We thank the anonymous reviewers for their valuable comments.

7. References

R. Artstein and M. Poesio. 2005. Kappa^3=alpha (or beta). Technical Report NLE Technical Note 05-1, University of Essex, Natural Language Engeneering and Web Application Group, Department of Computer Science, September.

J. Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, 22(2):249–254.

H. H. Clark. 1977. Bridging. In P.N. Johnson-Laird & P.C. Wason, editor, *Thinking: Readings in Cognitive Science*, pages 411–420. Cambridge University Press, Cambridge.

B. Di Eugenio and M. Glass. 2004. The kappa statistic: a second look. *Computational Linguistics*, 30(1):95–101.

B. Di Eugenio. 2000. On the usage of kappa to evaluate agreement on coding tasks. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece.

S. Fligelstone. 1992. Developing a Scheme for Annotating Text to Show Anaphoric Relations. In G. Leitner, editor, *New Directions in English Language Corpora: Methodology, Results, Software Developments*, pages 153–170. Mouton de Gruyter, Berlin.

R. Garside, S. Fligelstone, and S. Botley. 1997. Discourse Annotation: Anaphoric Relations in Corpora. In R. Garside, G. Leech, and A. McEnery, editors, *Corpus Annotation: Linguistic Information from Computer Text Corpora*, pages 66–84. Addison-Wesley Longman, London.

L. Hirschman, P. Robinson, J. Burger, and M. Vilain. 1998. Automating Coreference: The Role of Annotated Training Data. In *Proceedings of the Workshop on Linguistic Coreference*, Granada, Spain.

L. Hirschman. 1997. MUC-7 Coreference Task Definition (version 3.0). In L. Hirschman and N. Chinchor, editors, *Proceedings of Message Understanding Conference (MUC-7)*.

A. Holler, J. F. Maas, and A. Storrer. 2004. Exploiting coreference annotations for text-to-hypertext conversion. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 651–654, Lisboa.

H. Kamp and U. Reyle. 1993. *From Discourse to Logic*. Kluwer: Dordrecht.

L. Karttunen. 1976. Discourse referents. *Syntax and Semantics: Notes from the Linguistic Underground*, 7:363–385.

K. Krippendorff. 1980. *Content analysis: An introduction to its Methodology*. Sage Publications.

R. Mitkov, R. Evans, C. Orasan, C. Barbu, L. Jones, and V. Sotirova. 2000. Coreference and anaphora: developing annotation tools, annotated resources ans annotation strategies. In *Proceedings of the Discourse, Anaphora and Reference Resolution Conference (DAARC 2000)*, pages 49–58, Lancester, UK.

R. J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, pages 1503–1506, Lisboa.

M. Poesio and R. Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proc. of ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83, Ann Arbor, Michigan, June.

M. Poesio. 2004. The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. In *Proceedings of SIGDIAL*, Boston, April.

M. Stührenberg, D. Goecke, N. Diewald, I. Cramer, and A. Mehler. 2007. Web-based annotation of anaphoric relations and lexical chains. In *Proceedings of The Linguistic Annotation Workshop (LAW)*, pages 140–147, Prague, Czech Republic, June. Association for Computational Linguistics.

B. Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistic (ACL–88)*, pages 113–122, State University of New York at Buffalo, June 27–30.