SciCorp: A Corpus of English Scientific Articles Annotated for Information Status Analysis

Ina Rösiger

Institute for Natural Language Processing
University of Stuttgart
Germany
ina.roesiger@ims.uni-stuttgart.de

Abstract

This paper presents SciCorp, a corpus of full-text English scientific papers of two disciplines, genetics and computational linguistics. The corpus comprises co-reference and bridging information as well as information status labels. Since SciCorp is annotated with both labels and the respective co-referent and bridging links, we believe it is a valuable resource for NLP researchers working on scientific articles or on applications such as co-reference resolution, bridging resolution or information status classification. The corpus has been reliably annotated by independent human coders with moderate inter-annotator agreement (average $\kappa = 0.71$). In total, we have annotated 14 full papers containing 61,045 tokens and marked 8,708 definite noun phrases. The paper describes in detail the annotation scheme as well as the resulting corpus. The corpus is available for download in two different formats: in an offset-based format and for the co-reference annotations in the widely-used, tabular CoNLL-2012 format.

Keywords: Information Status, Co-reference, Bridging, Corpus Resource, Scientific Domain, English

1. Introduction

Due to the rapid growth in publications, researchers are faced with the problem of having to extract relevant information out of an enormous number of journals, conference papers and workshop proceedings. Thus, the demand for a deeper, machine-driven understanding has gained interest in the research community, with applications such as summarisation, information extraction or citation analysis. As co-reference resolution and information status classification are important steps in the automatic processing of scientific papers, work on anaphora in scientific text is therefore beneficial for these applications as well as for a better understanding of referring phenomena in this domain.

Categorising the information status of noun phrases (NPs) or resolving anaphoric relations automatically requires annotated data for training and testing. However, for scientific text, annotated data are scarce: to the best of our knowledge, there is no full-text scientific corpus annotated with both information status and anaphoric relations.

We thus present SciCorp, a scientific corpus of two different disciplines, namely computational linguistics and genetics, for the purpose of information-structural analysis. Apart from standard predicted annotation layers the corpus features three types of manual annotation: co-reference clusters, bridging entities and their antecedents, and information status labels. We believe this corpus is useful for three tasks that are of interest to the NLP community: co-reference resolution, bridging resolution and information status classification. These tasks are briefly introduced in Section 2.

We addressed the research question of resolving co-referent and associated references in scientific literature in Rösiger and Teufel (2014). In this paper, an earlier version of the corpus was used as a basis for the experiments, but the corpus was not made publicly available as it was only annotated by one person.

This paper presents details of the annotation process and describes the new corpus that was annotated by three independent annotators and that can be downloaded from our website.[1]

2. Information Status Analysis

Noun phrase co-reference resolution is the task of determining which noun phrases (NPs) in a text or dialogue refer to the same discourse entities (Ng, 2010). Coreferent links exist between two NPs if the first NP refers back to a discourse entity that has already been introduced in the discourse and is thereby known to the reader. Coreferent entities include pronominal NPs (1), nominal NPs (2) and named entities (3).

(1) Pronominal:
     We present a paper that deals with the adaption to new domains. **It** starts with an overview of the biomedial domain.[2]

(2) Nominal:
     Sequences were mapped and deposited into a database. **The database** comprises 10000 entries.

(3) Named entities:
     In 2006, Andrew Fire and Craig Mello shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in the nematode worm C. elegans. **C. elegans** is unsegmented, vermiform, and bilaterally symmetrical.

Co-reference resolution in scientific articles is considered difficult due to the heavy use of abstract entities such as results or variables, while easy-to-resolve named entities are less frequently used. The more complex nature of the texts

---

[1] www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/scicorp.html

[2] Anaphors are typed in bold face, their antecedents are underlined.

is also reflected in the high proportion of definite descriptions (Watson et al., 2003). These typically require domain knowledge to be resolved. It has been shown in Rösiger and Teufel (2014) that in-domain training data help improve co-reference resolution in scientific text.

A similar task is the resolution of *bridging anaphora*, (Clark, 1975; Prince, 1981), a related phenomenon, that looks at the relation between non-identic, related concepts in a discourse. The interpretation of a bridging anaphor, often also called associative anaphor, is based on the associated antecedent, but the two are not co-referent. Examples 4 and 5 show two science-specific cases of associative anaphora from our data.

(4)   Xe-Ar was found to be in a layered structure with Ar on **the surface**.

(5)   We base our experiments on the Penn treebank. **The corpus size** is ...

Gasperin (2009) showed that biological texts differ considerably from other text genres, such as news text or dialogue, reporting that the proportion between non-referring and referring entities in scientific text differs from that reported for other genres. The same holds true for the type and relative number of linguistic expressions used for reference.

To address this issue, we decided to go beyond co-reference and bridging and also annotate the *information status* (Nissim et al., 2004) of noun phrases. Information status tells us whether a noun phrase refers to an already known entity or whether it can be treated as non-referring. This dataset will therefore also contribute to the task of *information status classification*.

3.   Corpus Creation

The computational linguistics (CL) papers were taken from the ACL anthology, the genetics (GEN) papers from PubMed. Papers were selected blindly, not singling out one topic, any specific year or the first language of the authors. The CL papers cover various topics ranging from dialogue systems to machine translation; the GEN papers deal mostly with the topic of short interfering RNAs, but focus on different aspects of it. The corpus contains a number of short papers as well as some long papers (see Table 4 for details). The manual annotations were performed on plain text versions of the papers.[3] After the annotation, we enriched the corpus with a number of automatic annotations.

4.   Manual Annotations

We manually annotated the corpus using the annotation tool Slate (Kaplan et al., 2012). Slate does not feature pre-defined mentions, so the identification of markables was part of the annotation task. The tool shows the whole text with a slide bar at the side and the annotator is asked to mark the markables with different colours depending on the information status category. Co-reference and bridging

---

[3]The papers were provided in the framework of the FUSE project (Foresight and Understanding from Scientific Exposition). The CL papers were converted from Latex source by Simone Teufel, the GEN papers by Dain Kaplan and Diarmuid Ó Séaghdha as well as other members of the FUSE project.

links are also highlighted in different colours. Three annotators, all graduate students of computational linguistics, independently annotated the documents according to the following annotation scheme. Detailed annotation guidelines were provided.[4] The annotators were given two papers (one from genetics, one from computational linguistics) to familiarise themselves with the task before starting the annotation work on the texts included in this corpus.

The remainder of this section describes the annotation scheme in detail. This fine-grained scheme is based on other schemes (Riester et al., 2010; Poesio and Vieira, 1998), but has been adapted to this special domain. More details about the related schemes can be found in Section 9.

4.1.   Markables

To limit the amount of markables, we restricted the annotation to definite NPs and allowed only nominal phrases as antecedents for both co-reference and bridging anaphors. Therefore, no event reference is covered in the corpus. We considered the following types of NPs as definite:

Definite descriptions or similar   Nominal phrases (NPs) starting with the definite determiner *the*, a demonstrative determiner such as *this*, a possessive pronoun like *my* or a universal quantifier such as *all*. Examples: *the most efficient siRNAs, the siRNAs, all the algorithms*

Named entities   such as *Noam Chomsky, siRNAs* but also variables like *x* and *y*.

Pronouns   including personal pronouns (*we, it, they*), possessive pronouns (*our, their, its*) and demonstrative pronouns like *this* or *these*.

4.2.   Non-Markables

Pronouns   We do not mark relative pronouns and expletive or pleonastic *it*. The *it* in cases like *since it was discovered that ...* is not considered a markable.

Indefinite NPs   including indefinite descriptions with the indefinite determiner *a*, such as *an important part*. It also comprises existential quantifier phrases like *some siRNAs, most siRNAs* or *15 siRNAs*. Bare plurals such as *proteins* are also considered indefinite and are thus not included in the annotation.

Bare singulars and the **existential *there*** are also not annotated.

4.3.   Overview: Annotated Categories and Links

We label information status and create reference links for a subset of the information status categories. Table 1 shows the categories in the annotation scheme and how they interact with the co-reference and bridging links: we create co-reference links for all entities of the category GIVEN and bridging links for all ASSOCIATIVE entities. The entity being marked is shown in bold face, referring expressions are marked by a box and their antecedents are underlined.

---

[4]www.ims.uni-stuttgart.de/institut/ mitarbeiter/roesigia/annotationguidelines. pdf

	Category	Example
Co-reference links	Given	We present the following experiment. $\boxed{\text{It}}$ deals with ...
Bridging links	Associative	Xe-Ar was found to be in a layered structure with Ar on $\boxed{\textbf{the surface}}$.
Categories without links	Associative (self-containing)	**The structure of the protein** ...
	Description	**The fact that the accuracy improves** ...
	Unused	**Noam Chomsky** introduced the notion of ...
	Deictic	**This experiment** deals with ... *(non-anaphoric use)*
	Predicative	Pepsin, **the enzyme, ...**
	Idiom	On **the one hand** ... on **the other hand** ...

Table 1: Categories and links in our classification scheme

4.4. Information Status Categories

We distinguish the following eight categories, as shown in Table 1.

Given We consider a definite noun phrase GIVEN when the entity refers back to a discourse entity that has already been introduced in the discourse and is thereby known to the reader. This includes lexically new material, pronouns and repetitions or short forms of entities that have been referred to before. GIVEN entities include synonyms and are not limited to entities that have the same head. See Examples (1),(2),(3) in Section 1.

Associative For ASSOCIATIVE anaphors, the text presents an antecedent NP which does not stand in the relation of identity, but in some other form of relation to the associative phrase. The antecedent may be an associate in a typical relation such as part-of, is-a, or any kind of associate as long as there is a clear relation between the two phrases. We do not limit associative references to any predefined relations. See Examples (4) and (5) in Section 1.

Associative (self-containing) In some constructions, e.g. genitives or PP modifiers, we identify an associative relation between the head noun phrase and the modifier. We consider them ASSOCIATIVE SELF-CONTAINING and do not create a link.

(6)   *The structure of the protein*

(7)   *the thoracic circuit stage in HK mutants*

(8)   *the giant fiber escape pathway of Drosophila*

Deictic All deictic expressions including all entities that refer to the current paper or aspects thereof are considered DEICTIC.

(9)   *This experiment* deals with ...

(10)   *This paper* talks about ...

Non-nominal deictic expressions (such as *here, now*) are not marked.

Unused If an entity is not mentioned before and is not related to some other entity in the text, but refers to something which is part of the common knowledge of the writer and the reader, it is called UNUSED. This means the entity can be interpreted due to world or domain knowledge. This is often the case for named entities upon first mention.

(11)   We note that *the accuracy* has improved.

(12)   *Noam Chomsky* introduced the notion of ...

(13)   You can look it up in *the ACL Anthology*.

Description The entity is either self-explanatory or given together with its own identification. This means the entity is not anaphoric, does not rely on information about the situation of utterance and is not associative of some trigger previously introduced in the discourse. The definite noun phrase refers to something new, but the syntactic construction makes the interpretation easier. DESCRIPTION entities are heavily related to (but do not exclusively appear with) the following syntactic patterns:

(14)   NP complements:
         the fact that the accuracy has improved

(15)   Relative clauses:
         the protein that is essential

Idiom Entities that include idiomatic expressions or metaphorical use are considered IDIOMS.

(16)   On *the one hand* [...] on *the other hand*

Predicative Any predicative expressions, including appositions, are annotated as PREDICATIVE.

(17)   Pepsin, *the enzyme, ...*

(18)   Pepsin is *the enzyme*

(19)   Short interferring RNAs (*RNAs*)

(Unmarked) This is a purely technical category that has nothing to do with information status itself.

Co-reference or associative links can refer to an entity that has not been marked before, e.g. because the entity is indefinite. As the annotation tool needs the entity to be classified to be able to create a link, we classify the entity as UNMARKED for practical reasons.

Attributes We additionally annotate two attributes that are only applied to entities in a co-reference chain (mostly GIVEN entities, but also to the first-mention entities).

- +/- **Generic**: Generic expressions include reference to a kind or a general quantification whereas a specific reading has a fixed referent, i.e. we know which referent we select out of the set of entities that fulfil the description.

(20) Generic:
In 2006, they shared the Nobel Prize in Physiology or Medicine for their work on RNA interference in <u>the nematode worm C. elegans</u>. **C. elegans** is unsegmented, vermiform, and bilaterally symmetrical.

(21) Specific:
We present <u>the following experiment</u>. **It** deals with ...

- +/-**Part of compound**:
It is controversial whether parts of compounds (heads and non-heads) should be markables (when they are definite, since we only mark definite NPs). On the one hand, one might want to include them in the set of co-referential entities, but one the other hand, they do not allow for anaphoric reference, cf. example (22).

(22) The <u>siRNA</u> activity. *It ...

We decided to include them in the list of mentions when they can be co-referenced to other mentions , but to mark them with the attribute PART-OF-COMPOUND so that they can be filtered out if required. Adjectives and common nouns are never marked.

(23) *The siRNA experiments*

⇒ two markables (= the siRNA experiments, siRNA)

4.5. Co-Reference Annotation

A pair of NPs is considered co-referent if the *anaphor*, refers back to a previous expression in the discourse, the *antecedent*. The anaphor must be a definite noun phrase. The criteria for definiteness are the same as the ones described in Section 4.1. and 4.2., but with one addition:

- **Bare singulars** ...

... if the insertion of a definite determiner is possible and more plausible than the insertion of an indefinite determiner.

(24) *The efficiency of RNAi is ... .*
RNAi efficiency can also be influenced by ...

Bare plurals are still excluded. The antecedent can be any type of nominal phrase (indefinite, definite, named entity, etc.). Abstract anaphora are not included in the corpus, i.e. verbal phrases or clauses are excluded as antecedents of a co-referent anaphor. The links follow the chain principle, so we always choose the closest occurrence of the entity as the antecedent.

4.6. Bridging Annotation

The anaphor must be a definite noun phrase as described in Section 4.1.; the antecedent can be any type of nominal phrase. In our guidelines, verbal phrases or clauses cannot be the antecedent of an associative anaphor. The links do not have to follow the chain principle, the annotators are told to choose the best fitting antecedent, not the last occurrence in the text. Associative antecedents can also have two antecedents (and two links), if this fits best. In our scheme, associative links are only annotated when there is a clear relation between the two entities. As we do not pre-define possible associative relations, this definition is vague, but it is necessary to keep the task as general as possible.

5. Agreement Study

After the annotators familiarised themselves with the annotation task and annotated two papers that are not part of the final corpus, we analysed the inter-annotator-agreement on two papers (one GEN, one CL) that are part of the corpus and computed Fleiss' κ (Fleiss, 1971). As can be seen in Table 2, for information status we achieve a κ between 0.68 (GEN) and 0.73 (CL), which is considered moderate agreement (Landis and Koch, 1977).[5] It is not surprising that the number for CL is a little higher given the fact that the annotators are students of computational linguistics.

	GEN	CL
Actual agreement	0.79	0.82
Agreement by chance	0.34	0.34
κ	0.68	0.73

Table 2: Overall inter-annotator-agreement (in κ)

Category	GEN	CL
κ given	0.72	0.77
κ associative	0.62	0.63
κ associative (sc)	0.68	0.74
κ description	0.67	0.69
κ unused	0.65	0.67
κ deictic	0.73	0.76
κ predicative	0.53	0.57
κ idiom	0.85	0.83

Table 3: Inter-annotator-agreement for the single categories (in κ)

Table 3 shows the inter-annotator agreement for the single categories.[6] It can be seen that given, deictic and idiom entities are easier to reliably annotate while associative, description, unused and predicative entities are more difficult.

For the co-reference links the agreement was 0.81 and for associative links it was 0.62. The agreement for the attribute GENERIC was 0.51 and for the attribute PART-OF-COMPOUND 0.85.

6. Annotation Challenges

This section presents a few observations concerning some of the difficulties that came up during the annotation. We

---

[5]Calculation based on markables. When there was disagreement about the markables, we resolved these cases via discussion between the three annotators.

[6]Calculation for category x based on those mentions where one of the annotators classified it as category x.

include this here because we think it might be helpful for further, similar annotation experiments.

One major obstacle was that not all the texts were written by native speakers. For example, sometimes the authors clearly had problems with definite articles. If the annotators are asked to mark only definite NPs and the authors leave out the definiteness marker, this is problematic. We resolved these cases by adding a rule to the guidelines that in cases were it was very clear that the author made a mistake, the entity should be marked. However, a few cases remained where it was less clear, and we did not mark these cases. Paying more attention to paper selection in the first place would have helped here. With this aspect in mind, while we originally intended to limit the annotation to definite NPs due to time constraints, in hindsight we think that it turned out to be more difficult and as a result also slower to identify definite markables than to just annotate every NP, disregarding their definiteness.

The annotation of the attribute GENERIC turned out to be difficult for the annotators, with an agreement of only 0.51. As the decision whether an entity is generic or not is not trivial (and probably needs much more detailed guidelines), the annotation of +/-GENERIC should be the focus of an annotation task, not a by-product. Nevertheless, we include this attribute in the distribution of the data. For PART-OF COMPOUND, this problem did not exist: deciding whether something is part of a compound or not is trivial enough to be annotated at the same time.

For the GEN texts it would have been nice to include an expert as it was difficult to understand what refers to what in a few cases.

7. Resulting Corpus

Our annotated corpus contains 14 full-text scientific papers, 7 papers for each of the two disciplines. As shown in Table 4 and 5, the annotated computational linguistics papers contain 968 sentences, 25,034 words and 3,564 annotated definite descriptions while the annotated genetics papers contain 1,320 sentences, 36,011 words and about 5,144 definite descriptions; the genetics subcorporus is thus a little bigger than the CL one.

The gold annotation was created by taking the majority vote of the three annotators. Disagreements with respect to the annotation or the markables were resolved via discussion between the annotators.

Table 5 and Table 6 show the distribution of categories in absolute numbers and in percent.

8. Automatic Annotations and Format

For the pre-processing of the texts, we used the Stanford Core NLP pipeline[7] to automatically do tokenisation, POS tagging, constituency parsing and named entity recognition.

The distribution of the data contains the source PDF and plain text versions of the papers, the annotated categories and links in an offset-based format as well as the co-reference annotations in the tabular CoNLL-12 format.

---

[7]nlp.stanford.edu/software/corenlp.html

	Total	CL	GEN
Markables (incl. Unmarked)	9,407	3,879	5,528
Markables (excl. Unmarked)	8,708	3,564	5,144
Given	4,730	1,851	2,879
Associative	1,366	561	805
Associative(sc)	321	113	208
Description	1,034	507	527
Unused	1026	424	602
Deictic	70	45	25
Predicative	147	58	89
Idiom	14	5	9
(Unmarked	699	315	384)
Links	6,201	2,436	3,765
Co-reference	4,712	1,837	2,875
Associative	1,489	599	890

Table 5: Distribution of information status categories, in absolute numbers

Category	CL	GEN
Given	51.9	56.0
Associative	15.7	15.6
Associative(sc)	3.2	4.0
Description	14.2	10.2
Unused	11.9	11.7
Deictic	1.3	0.5
Predicative	1.6	1.7
Idiom	0.1	0.2

Table 6: Distribution of information status categories, in percent

9. Related Work

The idea behind our corpus is to have information status labels as well as links for bridging and co-reference anaphors as we think that these annotations are clearly related and that our data can help research this interaction. There exists, to the best of our knowledge, no such corpus for the scientific domain.

For newspaper text, there is ISNotes, a corpus of 50 Wall Street Journal (WSJ) articles that are part of OntoNotes (Markert et al., 2012). The scheme is a little different from ours in that it follows Nissim et al. (2004)'s classification into the three main classes *old, mediated and new*, with a number of fine-grained subclasses for *mediated* and *old*. Nissim et al. (2004) have annotated 144 dialogues from the Switchboard corpus. The corpus comprises information status and co-reference annotations. However, bridging antecedents are not explicitly marked.

Poesio and Vieira (1998) were the first to do a corpus-based study of definite description use. They carried out two annotation experiments on a total of 34 WSJ articles. In the first experiment, they focused on the classification of definite NPs and did not annotate links. In the second experiment, co-reference and bridging links were also included. Our scheme is partially based on their categories, but with a number of differences. For example, they restricted their anaphoric category to entities having the same head and included those that do not have the same head in

CL (doc id)	words	sentences	GEN (doc id)	words	sentences
9704004	6,104	217	346034	3,040	116
9505025	5,085	222	135797	2,437	74
9812005	1,368	59	340083	4,030	154
9708001	4,416	160	149283	5,404	228
9607011	2,804	104	152674	5,711	223
9606028	1,981	68	148263	7,286	253
9606011	3,276	138	153544	8,103	275
Total	25,034	968	Total	36,011	1320

Table 4: Corpus statistics

the associative/bridging category.

An extensive overview of the different approaches and a terminological comparison is provided in Riester et al. (2008) as well as a DRT-inspired classification system based on the contexts that have to be consulted to bind the referent. The scheme has been extended in Riester et al. (2010) and Baumann and Riester (2012). Our scheme is a combination of Poesio and Vieira (1998) and Riester et al. (2010). It is different from Riester et al. (2010) in a few aspects, the most important one being that we do not annotate indefinite NPs.

As mentioned before, for our specialised domain there is no corpus incorporating all three types of information, but there has been extensive research on the single tasks.

Gasperin (2009) dealt with non-pronominal anaphora in biomedical text and seeked to find the antecedents for both co-referent and associative mentions, but restricted associative mentions to noun phrases that refer to biomedical entities, e.g. proteins or genes. The corpus consists of 5 full-text articles; two of them are annotated with associative references.

Similar to Gasperin (2009) is the approach of Batista-Navarro and Ananiadou (2011) which presents a corpus of full-text articles from biochemistry literature with co-reference annotations for a few more types of pre-defined entities, such as chemical compounds, organisms, drug targets or diseases. The HANAPIN corpus, annotated according to their own annotation scheme which does not include associative references, consists of 20 full-text articles.

Cohen et al. (2010) is one of the first to mark all co-referent relations between all noun phrases in full-text biomedical articles. The annotation scheme is based on the OntoNotes guidelines and does not include associative references. The CRAFT corpus comprises 97 articles.

Schäfer et al. (2012) present a large corpus of 266 annotated full-text computational linguistics papers from the ACL Anthology. The annotation comprises mainly noun phrase co-reference; associative references are not covered in this work.

If associative references are addressed at all in this previous work, they are limited to discipline-dependent, pre-defined relations of the biomedical domain. Our definition of associative references differs from these in that it is more general and attempts to cover more than only pre-defined types.

10. Conclusion

We have presented a corpus of 14 full-text scholarly papers annotated with information status and anaphoric links. It contains 8,708 definite noun phrases. The corpus has been marked by three independent human coders and the inter-annotator study has shown that we have achieved moderate agreement (κ = 0.68 for genetics, 0.73 for computational linguistics). The corpus is freely available for download in two formats: an off-set based format and in the tabular CoNLL-12 format.

Acknowledgements

11. Bibliographical References

Batista-Navarro, R. T. and Ananiadou, S. (2011). Building a coreference-annotated corpus from the domain of biochemistry. In *Proceedings of BioNLP 2011 Workshop*, pages 83–91. Association for Computational Linguistics.

Baumann, S. and Riester, A. (2012). Referential and Lexical Givenness: semantic, prosodic and cognitive aspects. In Gorka Elordieta et al., editors, *Prosody and Meaning*, number 25 in Interface Explorations. Mouton de Gruyter, Berlin.

Clark, H. H. (1975). Bridging. In *Proceedings of the 1975 workshop on Theoretical issues in natural language processing*, pages 169–174. Association for Computational Linguistics.

Cohen, K. B., Lanfranchi, A., Corvey, W., Baumgartner Jr, W. A., Roeder, C., Ogren, P. V., Palmer, M., and Hunter, L. (2010). Annotation of all coreference in biomedical text: Guideline selection and adaptation. In *Proceedings of BioTxtM 2010: 2nd workshop on building and evaluating resources for biomedical text mining*, pages 37–41.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gasperin, C. V. (2009). Statistical anaphora resolution in biomedical texts. Technical Report UCAM-CL-TR-764, University of Cambridge, Computer Laboratory, December.

Kaplan, D., Iida, R., Nishina, K., and Tokunaga, T. (2012). Slate – a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, pages 89–101.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Markert, K., Hou, Y., and Strube, M. (2012). Collective classification for fine-grained information status. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 795–804. Association for Computational Linguistics.

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1396–1411. Association for Computational Linguistics.

Nissim, M., Dingare, S., Carletta, J., and Steedman, M. (2004). An annotation scheme for information status in dialogue. *LREC 2004*.

Poesio, M. and Vieira, R. (1998). A corpus-based investigation of definite description use. *Computational linguistics*, 24(2):183–216.

Prince, E. F. (1981). Toward a taxonomy of given-new information. In *Radical Pragmatics*, pages 223–55. Academic Press.

Riester, A., Lorenz, D., and Seemann, N. (2008). A semantic explication of information status and the underspecification of the recipients knowledge. In *Proceedings of Sinn und Bedeutung*, volume 12, pages 508–522.

Riester, A., Lorenz, D., and Seemann, N. (2010). A recursive annotation scheme for referential information status. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, pages 717–722.

Rösiger, I. and Teufel, S. (2014). Resolving coreferent and associative noun phrases in scientific text. *Proceedings of the Student Research Workshop of EACL 2014*, page 45.

Schäfer, U., Spurk, C., and Steffen, J. (2012). A fully coreference-annotated corpus of scholarly papers from the acl anthology. In *Proceedings of the 24th International Conference on Computational Linguistics. International Conference on Computational Linguistics (COLING-2012), December 10-14, Mumbai, India*, pages 1059–1070.

Watson, R., Preiss, J., and Briscoe, T. (2003). The contribution of domain-independent robust pronominal anaphora resolution to open-domain question-answering. In *Proceedings of the Symposium on Reference Resolution and its Applications to Question Answering and Summarization. Venice, Italy June*, pages 23–25.