# Error analysis in anaphora resolution

**Cătălina Barbu**

School of Humanities, Languages and Social Sciences
University of Wolverhampton
Stafford Street, Wolverhampton, United Kingdom
C.Barbu@wlv.ac.uk

**Abstract**

This paper deals with error analysis and their influence in comparative and qualitative evaluation of systems performing anaphora resolution. It presents a corpus-based analysis of errors reported by four anaphora resolution systems, leading to an investigation of the type and source of errors; as a direct application of the investigation's results, a simple probabilistic hybrid method is described, that takes advantage of the strong points of each of the methods analysed, while trying to avoid their weak points.

## 1. Introduction

The problem of evaluation is an important issue in any natural language processing application; not only it allows for comparative assessment of different individual systems, but it gives important clues about the areas that can be improved to obtain higher success rates. A glance at the literature on automatic anaphora resolution shows that the resolution rate of pronouns lies around 80%, (Hobbs, 1978; Ge et al., 1998; Mitkov, 1998) but the results decrease significantly when fully automatic methods are employed. Despite many years of research in the field, and several original approaches tried, pronominal anaphora resolution is not significantly more successful now than it was twenty years ago.

The relatively low success rate of anaphora resolution methods can be intuitively attributed to a number of problems that the task arises: the impossibility of using world knowledge (due to the prohibitive expense needed for encoding this kind of knowledge), the lack of semantic information, errors in the pre-processing stage and, not in the last, the inherent ambiguity of the natural language.

Although this problem has been acknowledged by different authors (Mitkov, 2000), to our knowledge there has been no study of the way different factors induce a decrease in the performance of anaphora resolvers. Such a study could enable researchers working in the field to identify areas that can be improved and to deal with phenomena less studied and that could increase the performance. As we are not as yet able to make use of extensive semantic and world knowledge information, it is nonetheless possible to make better use of easily acquired knowledge sources, as morphological or syntactical information.

## 2. Goal of the paper

This paper describes a series of experiments targeting the analysis of errors reported by four pronominal anaphora resolvers. The main idea was to identify classes of errors that appear frequently and that are common to more than one system. Section 3 presents a first direction of investigation that consisted in identifying classes of pre-processing errors and in assessing their influence in the overall result. Section 4 describes a different thread of investigation, that followed the differences in resolution of certain types of anaphoric expressions. Statistics were generated with regard to the resolution potential of each of the methods on each of the classes of anaphors. This analysis aimed at discovering if there are statistically significant differences in the resolution of certain types of anaphors. The investigation led to a pilot implementation of a probabilistic system (presented in Section 5) that incorporates the best features of each of the methods analysed in order to take advantage of their ability to deal with certain classes of anaphors.

## 3. Error analysis

### 3.1. Methodology

The investigation is based on the analysis of the results provided by four anaphora resolvers: three rule-based approaches (Mitkov's knowledge-poor method (Mitkov, 1998), Kennedy&Boguraev's parser-free method (Kennedy and Boguraev, 1996), the robust version of Baldwin's Cogniac (Baldwin, 1997) and a machine learning method (Barbu, 2001). The choice of these particular approaches was based on the fact that they all make use of (more or less) the same set of indicators, although combined in different ways: Cogniac applies a set of rules in the order of their confidence, Kennedy&Boguraev compute scores associated to coreference classes and classify a pronoun in the highest ranked class, Mitkov associates scores to possible antecedents and link a pronoun to the noun phrase with the highest score, while Barbu builds decision trees using a set of features computed over a training corpus. Moreover, all methods have been implemented to run in a fully automatic mode, using the same pre-processing tools and the same testing data[1].

These common features make it possible to assess the influence of both the errors in the pre-processing stage for computing the indicators, and the errors due to malfunctioning of the methods themselves.

The performance of the systems has been computed in

[1]For a detailed description of the implementation and the way we used these implementations for evaluation of anaphora resolution systems, see (Barbu and Mitkov, 2001)

| | | | ACC | WIN | BEO | CDR | **Total** |
|---|---|---|---|---|---|---|---|
| **#words** | | | 9617 | 2773 | 6392 | 9490 | 28272 |
| **#pronouns** | | | 182 | 51 | 92 | 97 | 422 |
| **#anaphoric pronouns** | personal | it | 103 | 30 | 31 | 51 | 215 |
| | | he, she | 5 | 0 | 4 | 0 | 9 |
| | | they | 32 | 6 | 17 | 22 | 77 |
| | possessives | | 18 | 11 | 18 | 10 | 60 |
| | reflexives | | 3 | 0 | 0 | 2 | 5 |
| | **Total anaphoric** | | 161 | 47 | 70 | 85 | 366 |

Table 1: Distribution of pronouns in the training corpus

terms of four evaluation measures: precision and recall (as defined in (Aone and Bennett, 1995)), f-measure, that combines the two, and success rate (Mitkov, 1999):

- *Precision = number of correctly resolved anaphors / number of anaphors attempted to be resolved*
- *Recall = number of correctly resolved anaphors / number of all anaphors identified by the system*
- *F = 2*Precision*Recall / (Precision+Recall)*
- *Success rate = number of correctly resolved anaphors / number of all anaphors*

For the sake of the clarity, in some of the reports in the paper we have only expressed the performance of the resolution in terms of success rate.

### 3.2. Corpus

The corpus used for this investigation consisted of technical manuals (collected on the Internet) annotated with co-referential links. The corpus contains 28,272 words, with 19,305 noun phrases and 422 pronouns, out of which 363 are anaphoric. The files we used were: "Beowulf HOW TO" (further referred to as BEO), "Linux CD-Rom HOW TO" (CDR), "Linux Access HOW TO" (ACC) and "Windows Help file" (WIN). The full description of the corpus appears in Table 1, along with the distribution of different types of pronouns.

Technical manuals presented some characteristic features, like a well defined structure (sections, subsections), the presence of several types of text formatting (tables, lists) and an important number of "unknown" words (commands, product names, company names). As for the composition in pronouns, technical manuals contained an important number of non-anaphoric instances of *it* (amounting to almost a quarter of the total number of occurrences of *it*, as it can be seen in Table 1), and a reduced number of instances of singular masculine and feminine pronouns.

The texts were initially pre-edited in order to remove irrelevant data (tables, table of contents, bibliography, lists). These modifications did not however change the character of the texts.

### 3.3. Analysis of the influence of the pre-processing tools

Several works and experiments have demonstrated that fully automatic anaphora resolution is more difficult than previous work has suggested (Orăsan et al., 2000),(Barbu and Mitkov, 2001). Errors are inevitably introduced at each pre-processing step, and these errors are reflected in the overall success of the system. However, it has not yet been investigated what percentage of the errors in the resolution of pronouns are due to the pre-processing, and, more precisely, which aspects of pre-processing are responsible for failures in pronoun resolutions. Although this work tries to give an answer to some of these problems, it still has to be mentioned that the results presented here cannot be indiscriminately applied to other anaphora resolution systems, since they are based on a certain set of pre-processing tools and on a certain type of texts (technical manuals).

The influence of pre-processing increases exponentially with the number and complexity of the pre-processing tools used. All the aforementioned systems require a limited amount of easily computed morphological, syntactical and textual knowledge, acquired by using a shallow parser (we have employed the FDG shallow parser (Tapanainen and Järvinen, 1997)) and a noun phrase extractor (built on top of the output of the shallow parser).

The methodology used for assessing the influence of the pre-processing tools consisted in manually post-editing the results returned by the shallow parser and re-running the systems on the perfect input. Due to the expenses involved in this operation, it was not possible to fully post-edit the results, thus we have only decided to correct those errors that were intuitively considered more likely to influence the performance of anaphora resolvers. Accordingly, we have corrected the delimitation of sentences, the prepositional phrase attachment, the identification and attachment of articles, the composition of noun phrases, the attachment of noun phrases to verbs and the grammatical function of noun phrases; we have ignored the attachment of adverbials, the composition of verb-phrases, the mal-recognition of adverbials and of other non-anaphoric entities, the grammatical function and morphological features of non-referential entities, the features of anaphoric or non-anaphoric pronouns which are not tackled by the system (demonstratives, relatives, interrogatives, personal and possessive pronouns of first and second person).

The first step was to run the individual systems over the uncorrected input. The success rate ranged from 49.7% for Cogniac to 61.6% precision for Kennedy&Boguraev's method. The full results are displayed in Table 2.

| | | ACC | WIN | BEO | CDR | Average |
|---|---|---|---|---|---|---|
| Success rate | Mitkov | 52.7% | 55.3% | 48.5% | 71.7% | 56.2% |
| | Cogniac | 45.9% | 44.6% | 42.8% | 67.0% | 49.7% |
| | K&B | 55.0% | 63.8% | 55.7% | 74.1% | 61.6% |
| | ML | 55.9% | 63.8% | 52.8% | 72.9% | 59.8% |
| Precision | Mitkov | 42.8% | 50.9% | 36.9% | 62.8% | 48.8% |
| | Cogniac | 37.1% | 41.1% | 32.6% | 58.7% | 42.6% |
| | K&B | 48.3% | 58.8% | 42.3% | 64.9% | 52.8% |
| | ML | 49.4% | 58.8% | 40.21% | 63.9% | 51.2% |

Table 2: Initial evaluation results

The second step was to re-run the systems on the corrected input. As a result, the performance of all systems improved considerably, by up to 9% on one of the files, with an average improvement of 6.5% in precision (full results are shown in Table 3). Discovering that pre-processing influences significantly and consistently the performance of the anaphora resolvers, the second step was to break-down the cause of errors.

By analysing the common indicators that all anaphora resolution systems used, we made the assumption that three main types of pre-processing errors could account for failures: mal-identification of noun-phrases, errors in verb attachment and errors in the identification of the syntactic function of noun phrases. Several other types of errors (such as wrong delimitation of sentences) were considered important, but not frequent enough to allow space for investigation, therefore they were ignored.

The analysis of the individual influence of the selected types pre-processing errors has proved difficult, especially because they were strongly inter-connected: errors in verb attachment led to wrong identification of the syntactic function of noun phrases, just as the wrong identification of noun phrases did.

#### 3.3.1. Misidentification of noun phrases

The analysis of the errors introduced by misidentification of noun phrases was done by matching the noun-phrases in the un-corrected parser results with those in the post-edited results. This experiment showed that, although the noun phrase extractor did not eliminate any of the correct noun-phrases, it introduced additional ones that made the search space for antecedents on average 12% larger. This obviously has an influence not only on the final accuracy of the anaphora resolvers, but in their time efficiency as well.

The second type of misidentification of noun phrases that introduced errors in the anaphora resolvers was the wrong delimitation of noun phrases. The main consequence of this type of error was that some to noun phrases are wrongly identified as embedded. As all anaphora resolvers penalise embedded noun phrases, there have been cases where the correct antecedent was eliminated as a result. The example below shows this kind of misidentification, where the correct antecedent ”the Beowulf HOWTO” was rejected in the favour of the wrongly identified noun phrase ”a year”:

”Over <NP>a year <NP>*the Beowulf HOWTO*</NP> </NP> grew into a large document, and in August 1998 *it* was split into three.”

Wrong delimitation of noun phrases was the most common type of error produced by the noun phrase extractor; nevertheless its influence could not be fully assessed due to the evaluation method employed. All methods considered the resolution of a pronoun correct if the antecedent found spanned a substring of the correct antecedent, including the head noun; if the type of evaluation took into account perfect matchings only, the influence of errors in the noun phrase extractor could be far more extensive.

The assessment of each individual error in the output of the noun phrase extractor has proved too time consuming, for this reason only a global assessment has been performed. This was done by using the post-edited results of the NP extractor and the initial, uncorrected output of the shallow parser[2]. We’ve noticed that all methods have approximately equally improved (about 10%), with a slightly higher improvement for Cogniac and slightly lower for the machine learning method.

#### 3.3.2. Verb phrase attachment

The second step was to assess the importance of correct VP attachment. VP attachment can influence pronoun resolution due to several rules that make use of this information: collocation patterns (Mitkov), existential constructions (Kennedy&Boguraev), resolution of reflexives. By leaving the noun phrases as identified by the noun phrase extractor and with the un-corrected syntactical function and using the corrected results of the shallow parser, we have noticed that the method least sensitive to errors in the VP attachment was Cogniac (1.9% improvement), while the most sensitive was Mitkov’s (4.4% improvement).

#### 3.3.3. Syntactic function

All anaphora resolution methods make use of information about syntactic function, in rules such as: preference for a subject antecedent, syntactic parallelism, resolution of reflexives to the subject, Mitkov’s collocations pattern rule. An experiment involving the syntactic

[2]Of course, as the noun phrase extractor was built on top of the output of the shallow parser, this uncorrected input conflicted with the corrected output of the noun phrase extractor; all conflicts were ignored, meaning that the corrected output was preferred to the initial one. The same observation applies to all subsequent experiments.

function, similar to the ones described before showed that Cogniac was the most sensitive to pre-processing errors, while again the machine learning method was the least influenced. This experiment also showed that the identification of the syntactic function of noun phrases was relatively reliable, being the least important cause of errors in our implementation (approximately 1.2% improvement). It has to be mentioned that none of these experiments capture the (unlikely) situation where a pronoun is correctly resolved due to errors in the pre-processing stage. Although theoretically possible, intuitively we have considered this possibility too remote to benefit from a special treatment; we are nevertheless aware that such cases may occur and account for a small percent of the resolution performance.

Table 3 summarises the improvement in success rate obtained when using the input that was selectively corrected. The results are global, for all the evaluation files. These results do not fully reflect the influence of pre-processing on pronoun resolution, but rather give an estimate, due to the fact that the input was not entirely correct, and the errors are not independent.

| | Mitkov | K&B | Cogniac | ML |
|---|---|---|---|---|
| Initial results | 56.2% | 61.6% | 49.7% | 59.8% |
| NP identification | 67.7% | 71.0% | 61.7% | 67.7% |
| Syntactical function | 58.1% | 62.8% | 52.3% | 61.7% |
| VP attachment | 60.6% | 63.9% | 51.7% | 62.5% |
| Perfect input | 69.9% | 75.2% | 65.8% | 74.9% |

Table 3: Improvement of the success rate when using corrected input

#### 3.3.4. Other types of errors

During the analysis of the data, it became apparent that some of the errors appearing in the pre-processing stage were not due to the malfunctioning of the parser, but to the composition of the text itself. In this category enter spelling mistakes (one of the most repetitive was the employment of *it's* as a possessive determiner instead of *its*), wrong verb agreement (*The drivers is...*), inconsistencies in using references to gender underspecified individuals (*The user* sometimes referred to by *they*, and later by *he*), missing punctuation marks (e.g, full stop at the end of sentence). All these errors directly reflect on the performance of the parser and propagate towards the anaphora resolvers. We did not consider necessary to correct any of the spelling or style mistakes, in order to preserve the character of the file; although we deal with input of not the best quality, we have to take into account that this is the kind of texts usually found on the Internet, so any automatic natural language processing system should find ways of dealing with malformed input.

## 4. Reliability assessment

A drawback of existent anaphora resolution algorithms designed for English is that, to our knowledge, none of them applies a specialised treatment to different classes of pronouns. This is even more surprising considering the fact that it has been theoretically acknowledged the fact that different pronouns have characteristic anaphoric properties.

Subsequently, we attempted to analyse the reliability of each of the methods in the identification of the types of pronouns resolved. All three methods targeted the same types of pronouns (personal-third person only, possessives and reflexives), which made the comparison reliable. It has to be mentioned that none of the methods apply specific resolution rules according to the type of pronoun processed (apart from reflexives, which will be discussed later). However, differences in resolution rates may result from the application of other rules, apparently not related to the type of pronoun (for example, verb attachment and grammatical function can indirectly distinguish between a personal pronoun and a possessive determiner). We calculated the success rate of all methods in the resolution of three categories of pronouns.

Firstly, we constructed a category based on the morphological type of pronoun: neuter singular pronoun (*it*), masculine and feminine singular pronouns (*he* and *she*), plural pronouns (*they*), possessives (*his*, *her*, *their*, *its*) and reflexives (*himself*, *herself*, *itself*, *themselves*). Table 3 describes the resolution accuracy of each method for each type of pronoun. As it can be noticed, results are not included for the resolution of reflexives. This is due to the fact that there has been no significant difference between the methods with respect to the resolution of reflexives (only a very small number appeared in the testing corpus); this can be explained by the fact that all methods use the same constraints drawn from the Government and Binding Theory, which are never violated in the occurrences found in our corpus. Therefore, the reflexives have been omitted from all the subsequent results reporting.

| | Mitkov | Cogniac | K&B | ML |
|---|---|---|---|---|
| it | 63.2% | 43.2% | 70.2% | 59.5% |
| he, she | 30% | 22.2% | 66.6% | 55.5% |
| they | 50.6% | 38.9% | 54.5% | 45.4% |
| possessives | 40% | 30% | 73.3% | 78.3% |

Table 4: Success rate according to the morphological category of pronouns

The second classification was based on the syntactic function of the pronouns: subject, direct object, indirect object, attributive and others.

The third category was based on the distance (in number of intervening noun phrases and sentences) between the anaphor and the real antecedent: one or two noun phrases (same sentence), same sentence more than two intervening noun phrases, previous sentence, distance greater than one sentence. Statistics were collected from the corpus for the success rate of all the methods for each category of pronouns.

| | Mitkov | K&B | Cogniac | ML |
|---|---|---|---|---|
| Subj | 40.7% | 48.2% | 38.9% | 50.4% |
| Dir obj | 55.1% | 34.8% | 40.6% | 56.5% |
| Ind obj | 33.3% | 66.7% | 33.3% | 66.7% |
| Attributive | 76.5% | 86.8% | 63.2% | 51.5% |
| Other | 50.0% | 50.0% | 50.0% | 100.0% |

Table 5: Success rate according to the syntactic function of pronouns

| | Mitkov | Cogniac | K&B | ML |
|---|---|---|---|---|
| 1 or 2 NPs | 80.3% | 83.6% | 95.1% | 60.7% |
| more than 2 NPs | 36.9% | 35.1% | 66.7% | 58.6% |
| previous sentence | 59.7% | 66.0% | 23.3% | 61.6% |
| more than one sentence | 65.6% | 81.3% | 40.6% | 59.4% |

Table 6: Success rate according to the distance between anaphor and antecedent

## 5. A hybrid anaphora resolver

This section presents a hybrid system for anaphora resolution, currently under development, that uses a probability model to calculate the likelihood of a method to correctly solve a certain type of pronoun.
Unlike Ge&Charniak's statistical model (Ge et al., 1998), this is a much simpler probabilistic system, that only chooses between candidates already proposed by other anaphora resolvers. It is therefore not an independent system, does not perform a search for an antecedent in the space of possible antecedents, does not incorporate any new knowledge sources and does not aim at achieving ground-breaking accuracy rates. The only goal of this model was to show that it is still possible to improve the performance of anaphora resolvers by simply using the same knowledge sources in different ways, and by taking the best out of classical ideas.

### 5.1. Description

Intuitively, given a pronoun and the antecedents identified by the four systems, the method tries to estimate which is probability for a certain system to have found the correct antecedent, given that:

- the pronoun had the morphological function *m* (m in the set {*it/he/she, they/them, possessive*})
- the pronoun had the syntactic function *s* (s in the set {*subject, direct object, indirect object, attributive*})
- the distance between the pronoun and the antecedent was *d* (*d* in the set {*1/2 NPs in the same sentence, same sentence and more than 2 NPs, previous sentence, more than 1 sentence*})
- the antecedent had the syntactic function *as* (*as* in the same set as *s*)

On the basis of the statistics collected from the training corpus, we've calculated the probabilities for each pair <pronoun, system> and selected as correct the antecedent found by the system that maximised the probability.
As mentioned before, there was no significant difference between the systems in the resolution of reflexives, thus we cannot assume that a certain system is more likely to solve a reflexive than an other. In this case, we have always selected the antecedent returned by Kennedy&Boguraev's method. This was only done for the purpose of consistent comparative evaluation.

### 5.2. Evaluation

The qualitative and comparative evaluation envisaged comparison with the individual methods and with a combined method.

#### 5.2.1. Testing corpus

In order to assess the performance of the new system, we have evaluated it on unseen data, independent of the observation corpus described in section 3.2.. The testing corpus consisted of 3 technical manuals, containing 113 pronouns, out of which 86 anaphoric.

#### 5.2.2. A simple voting system

In order to evaluate the new system, we had at the same time in mind the time efficiency, so important in applications where anaphora resolution is only a component. For that reason, we have tried to show that the results obtained could not be surpassed or equaled by using a much less time-consuming system that combines the three methods using a simple voting procedure. Hence, we have implemented a voting system that considers as correct a result reported by the majority of the systems; in case of a tie, the correct result was the one returned by Kennedy&Boguraev's system, as the one that outperforms systematically the others. The voting system was used as a baseline. By evaluating the baseline against the three systems, we discovered an improvement of up to 4% in success rate compared to the best results on one testing file; the average improvement for all testing files was about 2%. However, this improvement was not consistent across all files used for testing. In some cases, the best results of the systems were better than the results of the baseline, therefore the voting system has decreased the performance of the best independent system.

#### 5.2.3. Results

Table 6 displays the results obtained when running the hybrid system, as compared to the results of the individual systems and of the combined baseline. It can be easily seen that the increase in performance is significant, up to 7.5% over the best individual system on one of the files; the average improvement for all testing files over the best system was 4.6%. More important, the improvement is consistent over all testing files and the hybrid system always outperforms the baseline.
We are aware of the fact that the small amount of training data does not allows us to draw a definite conclusion as to the resolution power of the individual systems, the

| | Success rate | | | | | |
|---|---|---|---|---|---|---|
| | Mitkov | Cogniac | K&B | ML | Baseline | Hybrid |
| Observation corpus | 56.2% | 49.7% | 61.6% | 59.8% | 63.3% | 73.0% |
| Unseen data | 58.1% | 52.3% | 62.8% | 62.7% | 63.9% | 67.3% |

Table 7: Final evaluation results

differences in resolution not being statistically significant. Therefore, the probabilities do not fully express the likelihood of a certain method to be preferred over an other when resolving a certain type of pronoun. Nevertheless, the results show that the improvement in performance is consistent and significant.

## 6. Conclusions and future work

We have tried to show in this paper that breaking down the errors reported by anaphora resolution systems can lead to interesting findings and open new areas of improvement. A simple hybrid system is described that takes advantage of the strong points of each of the three methods investigated and combines them probabilistically to obtain an improvement of up to 7% in success rate. We are at the same time aware that a more sophisticated probabilistic model, trained on more data and taking into account more aspects could lead to even better results, so our further work will concentrate in this direction. A further direction of research would be to separate the features that make a system more reliable than another in the interpretation of a certain class of pronouns. This would enable us to integrate in a single, independent system the best features and processing methods belonging to different existent anaphora resolvers.

## 7. References


Chinatsu Aone and Scot W. Bennett. 1995. Evaluating automated and manual acquisition of anaphora resolution rules. In *Proceedings of the 33th Annual Meeting of the Association for Computational Linguistics (ACL '95)*, pages 122–129.

Breck Baldwin. 1997. Cogniac: High precision coreference with limited knowledge and linguistic resources. In R. Mitkov and B. Boguraev, editors, *Operational factors in practical, robust anaphora resolution for unrestricted texts*, pages 38 – 45.

Catalina Barbu and Ruslan Mitkov. 2001. Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of ACL'01*, Toulouse, France.

Catalina Barbu. 2001. Automatic learning and resolution of anaphora. In *Proceedings of RANLP'01*, Tzigov Chark, Bulgaria.

Niyu Ge, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora, COLING-ACL '98*, pages 161 – 170, Montreal, Canada.

Jerry Hobbs. 1978. Pronoun resolution. *Lingua*, 44:339–352.

Christopher Kennedy and Branimir Boguraev. 1996. Anaphora for everyone: pronominal anaphora resolution without a parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96)*, pages 113–118, Copenhagen, Denmark.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98*, pages 867 – 875. Morgan Kaufmann.

Ruslan Mitkov. 1999. Pronoun resolution: the practical alternative. In Simon Botley and Antony Mark McEnery, editors, *Corpus-based and Computational Approaches to Discourse Anaphora*, Studies in Corpus Linguistics, chapter 7, pages 129 – 143. John Benjamins Publishing Company.

Ruslan Mitkov. 2000. Towards more comprehensive evaluation in anaphora resolution. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume III, pages 1309 – 1314, Athens, Greece.

Constantin Orăsan, Richard Evans, and Ruslan Mitkov. 2000. Enhancing preference-based anaphora resolution with genetic algorithms. In *Proceedings of Natural Language Processing - NLP2000*, pages 185 – 195. Springer.

P. Tapanainen and T. Järvinen. 1997. A non-projective dependency parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, pages 64 – 71, Washington D.C., USA.