António Branco (Ed.)

LNAI 4410

Anaphora: Analysis, Algorithms and Applications

6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007
Lagos, Portugal, March 2007
Selected Papers

Springer

Lecture Notes in Artificial Intelligence 4410

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

António Branco (Ed.)

Anaphora: Analysis, Algorithms and Applications

6th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2007
Lagos, Portugal, March 29-30, 2007
Selected Papers

Springer

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editor

António Branco
Universidade de Lisboa, Faculdade de Ciências
Departamento de Informática Campo Grande
1749-016 Lisboa, Portugal
E-mail: antonio.branco@di.fc.ul.pt

Library of Congress Control Number: 2007922631

CR Subject Classification (1998): I.2.7, I.2, I.7, F.4.3, H.5.2, H.3

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-71411-1 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-71411-8 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12035274 06/3142 5 4 3 2 1 0

Preface

Anaphora is a central topic in the study of natural language and has long been the object of research in a wide range of disciplines in the area of cognitive science, such as artificial intelligence and human language technology, theoretical, corpus and computational linguistics, philosophy of language, psycholinguistics and cognitive psychology. The correct interpretation of anaphora has played an increasingly vital role in real-world natural language processing applications including machine translation, automatic abstracting, information extraction and question answering. Given the challenges its complexity poses to scientific inquiry and technological progress, anaphora has been one of the most productive topics of multi- and inter-disciplinary research, and has enjoyed increased interest and attention in recent years.

The cutting-edge results reported in the papers collected in the present volume address all these aspects. They are a selection of the best papers presented at the sixth edition of DAARC.

The Discourse Anaphora and Anaphor Resolution Colloquia (DAARC) have emerged as the major regular forum for presentation and discussion of the best research results in this area. Initiated in 1996 at Lancaster University and taken over in 2002 by the University of Lisbon, the DAARC series established itself as a specialized and competitive forum for the presentation of the latest results on anaphora. The series is unique in that it covers this research subject from a variety of multidisciplinary perspectives, while keeping a strong focus on automatic anaphora resolution and its applications.

The program of the sixth DAARC was selected from 60 initial submissions. It included 24 oral presentations and 15 posters from over 70 authors coming from 18 countries: Australia, Belgium, Czech Republic, Denmark, France, Germany, Israel, Italy, Japan, Norway, Poland, Portugal, Russia, Spain, The Netherlands, Turkey, UK and the USA. The submissions were anonymized and submitted to a selection process by which each received three evaluation reports by experts from the Program Committee.

The program also included two invited presentations, by Ruslan Mitkov and his team, from the University of Wolverhampton, and by Jos van Berkum, from the University of Amsterdam.

The articles in the present volume grew out of one of the invited talks, by Ruslan Mitkov *et al.*, and of 12 regular papers presented at DAARC. They are fully fledged versions of the submissions that got the best reviewing reports from the Program Committee.

On behalf of the Organization Committee, I would like to thank all the authors who contributed with their papers to the present volume and all the colleagues in the Program Committee for their generous and kind help in the reviewing process of DAARC, and in particular, of the papers included in the

present volume. Without them neither this colloquium nor the present volume would have been possible.

Last but not least, my warm thanks also go to my colleagues in the Organization Committee of the colloquium, Tony McEnery, Ruslan Mitkov and Fátima Silva.

Lisbon, January 2007 António Branco

Organization

The sixth DAARC colloquium was organized by the University of Lisbon, Faculty of Sciences, Department of Informatics.

Organization Committee

António Branco, University of Lisbon, Portugal
Tony McEnery, Lancaster University, UK
Ruslan Mitkov, University of Wolverhampton, UK
Fátima Silva, University of Oporto, Portugal

Program Committee

Mijail Alexandrov-Kabadjov, University of Essex, UK
Mira Ariel, Tel Aviv University, Israel
Sergey Avrutin, OTS, The Netherlands
Amit Bagga, Ask.com, USA
Patricio Martinez Barco, University of Alicante, Spain
Peter Bosch, University of Osnabrück, Germany
António Branco, University of Lisbon, Portugal
Donna Byron, Ohio State University, USA
Francis Cornish, University of Toulouse-Le Mirail, France
Dan Cristea, University of Iasi, Romania
Robert Dale, Macquarie University, Australia
Richard Evans, University of Wolverhampton, UK
Martin Everaert, OTS, The Netherlands
Lyn Frazier, University of Massachusetts, Amherst, USA
Claire Gardent, CNRS/Loria, France
Rafael Muñoz Guillena, University of Alicante, Spain
Jeanette Gundel, University of Minnesota, USA
Sanda Harabagiu, University of Texas at Dallas, USA
Lars Hellan, Norwegian University of Science and Technology, Norway
Erhard Hinrichs, University of Tübingen, German
Graeme Hirst, University of Toronto, Canada
Yan Huang, University of Reading, UK
Andrew Kehler, University of California San Diego,USA
Andrej Kibrik, Russian Academy of Sciences, Russia
Emiel Krahmer, Tilburg University, The Netherlands
Shalom Lappin, King's College, UK
Tony McEnery, Lancaster University, UK
Ruslan Mitkov, University of Wolverhampton, UK

Jill Nickerson, Ab Initio Software Corp, USA
Constantin Orasan, University of Wolverhampton, UK
Maria Mercedes Piñango, Yale University, USA
Georgiana Puscasu, University of Wolverhampton, UK
Costanza Navarretta, CST, Denmark
Massimo Poesio, University of Essex, UK
Eric Reuland, OTS, The Netherlands
Jeffrey Runner, University of Rochester, USA
Antonio Fernandez Rodriguez, University of Alicante, Spain
Tony Sanford, Glasgow University, UK
Frédérique Segond, Xerox Research Centre Europe, France
Roland Stuckardt, University of Frankfurt am Main, Germany
Joel Tetreault, University of Rochester, USA
Renata Vieira, Unisinos, Brazil

Table of Contents

Human Processing and Performance

Language Analysis and Representation

Resolution Methodology and Algorithms

Computational Systems and Applications

Nuclear Accent Placement and Other Prosodic Parameters as Cues to Pronoun Resolution*

Ekaterina Jasinskaja, Ulrike Kölsch, and Jörg Mayer

University of Potsdam
Institut für Linguistik, SFB 632,
Karl-Liebknecht-Straße 24-25, 14476 Golm, Germany
{jasinsk,ukoelsch}@uni-potsdam.de, mayer@ling.uni-potsdam.de

This paper investigates the influence of prosody on the interpretation of anaphoric pronouns, concentrating especially on the effect of nuclear accent placement. It is well-known that accented and unaccented pronouns generally have different resolution preferences, but it is less obvious that pronoun interpretation can be affected by almost any manipulation of the accentual pattern of the sentence in which the pronoun occurs, even by a manipulation that does not involve the pronoun. However, the latter follows from theories of accentuation such as [1] and in this paper we present experimental support for this prediction. Our results corroborate the view that the influence of accent on pronoun resolution should be derived from a general theory of focus interpretation, rather than from rules defined specifically for accents occurring on pronouns.

We start in Section 1 by presenting some background on accentuation and its impact on pronoun resolution. Since accent is a way of signaling contrast, and contrast in turn can be viewed as a rhetorical relation, constraints on pronoun resolution that result from rhetorical structure should be taken into account as well, which is done in Section 2. Section 2 also introduces hypotheses related to other prosodic parameters (pitch range and pause duration) which are known to be able to convey aspects of rhetorical structure. Finally, Section 3 describes our experiment, and Section 4 discusses the results.

1 Accent Placement and Pronoun Resolution

It is well-known that accentuation affects the resolution preferences of anaphoric pronouns. In particular, the effect of accenting the pronoun itself has been studied quite extensively and is illustrated by the following example (coindexing indicates coreference relations):

* We are indebted to Elke Kasimir for making her implementation of Schwarzschild's OT constraints system [1] available for deriving our hypotheses; and to Robin Hörnig for advice on the statistical analysis. Many thanks also go to Martin Neumann, Norman Schenk and Marcus Thienert. This research was funded by the German Research Community (DFG) as part of the Collaborative Research Center *Information Structure* (SFB 632).

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 1–14, 2007.

(1) a. Paul$_i$ called Jim$_k$ a Republican.
Then he$_i$ insulted him$_k$.
b. Paul$_i$ called Jim$_k$ a Republican.
Then HE$_k$ insulted HIM$_i$.

Recent studies have argued that there is nothing special about the role of accent when placed on a pronoun; rather, the effects of accent on pronoun resolution should be derived from a general theory of accentuation and focus semantics [1,2]. In particular, it is proposed that certain accentual patterns, including those that involve accented pronouns, require the presence of a *contrasting alternative* in the context or a possibility to accommodate such an alternative [3,4,5]. A contrasting alternative in this case is a constituent (often, a clause) that differs from the clause in question only in the focused ($\approx$ accented) subconstituent(s). Thus in (1b), the accents on the pronouns and the absence of accent on the verb *insulted* are licensed only if the sentence is taken to contrast with the preceding sentence *Paul called Jim a Republican*; this implies that (a) HE must be contrasted with, i.e. distinct from, *Paul* (hence HE $\mapsto$ Jim), (b) HIM must be contrasted with, i.e. distinct from, *Jim* (hence HIM $\mapsto$ Paul), and (c) *insulted*, since it is deaccented, must be viewed as "parallel" to *called a Republican*, so that calling someone a Republican has to be accommodated as a kind of insult.

Most of the existing theoretical analyses of accented/stressed pronouns, notably [6] and [4], seem to treat accent as an independent property of the pronoun.[1] However, theories of accent placement such as [7] or [1] suggest that the decision to accent or deaccent a pronoun is not independent from the decision to accent or deaccent other constituents in the sentence. Thus, for instance, placing no accent on the pronouns in the second sentence of (1a) means almost automatically that the verb *insulted* must be accented. In this paper we present further support for the idea that the dependence of pronoun resolution on accentuation is a by-product of the general functioning of prosodic focus as a contrast-signaling device; however, unlike the previous studies, we would like to emphasise the importance of the overall accentual pattern of a sentence. That is, it does not only matter whether the pronoun is accented or not, but as predicted by Schwarzschild [1], any occurrence of an accent in the sentence, as well as any occurrence of deaccenting is potentially relevant for determining an antecedent. We present below the results of an experiment which show that this prediction is indeed borne out. There is a well-known asymmetry between nuclear and pre-nuclear accents in marking given information, and we will

[1] This approach makes it look as if the opposition of stressed and unstressed pronouns behaves like the opposition of strong and weak pronouns, e.g. *it* vs. *that* in English, *er* vs. *der* in German. The latter indeed applies specifically to pronouns, in that the strong/weak pairs often have to be defined in the lexicon, rather than following a productive pattern, whereas accentuation is completely productive in languages like English and German, and is not restricted to pronouns. Although the choice between a strong and a weak pronominal form might not be completely independent from stress, the oppositions are *a priori* distinct and a uniform analysis should be empirically justified.

therefore follow Venditti et al. [8] in restricting our attention to the placement of *nuclear* accents—the most prominent, and usually the last, accent in a prosodic phrase.

To illustrate the prediction in question, consider the German example in (2) as well as its English translation in (3):

(2) a. Johann hat die Möhren geschnitten.
Johann has the carrots cut
b. Marek hat indes die Kartoffeln geschält.
Marek has meanwhile the potatoes peeled
c. Außerdem hat er die Kartoffeln geschnitten.
besides has he the potatoes cut

(3) a. Johann cut the carrots.
b. Meanwhile, Marek peeled the potatoes.
c. Besides, he cut the potatoes.

The most natural interpretation of the pronoun *er* 'he' in (2c)/(3c) is *Marek*, the only male individual mentioned in the immediately preceding sentence, while the most natural pronunciation of (2c) is with a nuclear accent on the verb *geschnitten* 'cut', indicated by small caps in (4), whith the direct object *die Kartoffeln* 'the potatoes' deaccented.

(4) Außerdem hat er die Kartoffeln GESCHNITTEN
besides has he the potatoes cut

This pattern is explained straightforwardly if we assume that only the previous sentence is relevant for establishing the contrast relation. In that case, the transitive relation *He/Marek X-ed the potatoes* is given, while the verb *geschnitten* 'cut' is contrasted with *geschält* 'peeled,' so the verb is narrowly focused and receives the nuclear accent. However, if both context sentences (2a) and (2b) are taken into account, the question arises, with which of them (2c) should be contrasted. This choice is essential for determining the accentual pattern, and as it turns out, it interacts in a crucial way with the choice of antecedent for the pronoun.

Suppose, as before, that *er* 'he' in (2c) refers to *Marek*, but (2c) is contrasted with (2a). Then the verb *geschnitten* 'cut' is given, but its arguments are contrasted: *er/Marek* with *Johann* and the potatoes with the carrots. So a contrast/givenness-based theory predicts accents on *er* and *Kartoffeln*, cf. (5a). Now suppose that the pronoun refers to the antecedent farther away—*Johann*. If (2c) is contrasted with (2b), the sentences differ in who did what to the potatoes, so accents are expected on the pronoun *er/Johann*, contrasting with *Marek*, and *geschnitten* 'cut,' contrasting with *geschält* 'peeled,' cf. (5b). If, in turn, (2c) is contrasted with (2a), the open proposition *He/Johann cut X* is given and only the objects of cutting, the potatoes and the carrots, are contrasted. Therefore, we predict an accent on *Kartoffeln* 'potatoes,' cf. (5c), whereas the pronoun receives no accent.

(5) a. Außerdem hat ER [Marek] die KARTOFFELN geschnitten
besides has he the potatoes cut
b. Außerdem hat ER [Johann] die Kartoffeln GESCHNITTEN
besides has he the potatoes cut
c. Außerdem hat er [Johann] die KARTOFFELN geschnitten
besides has he the potatoes cut

This short sketch shows that theories of accent placement based on the notions of contrast and givenness such as [5] and [1] predict a rather complex interplay between pronoun resolution possibilities and accentuation patterns. Since coreference relations play a role in identifying the parallel part of the contrasting clauses, the choice of pronoun antecedent does not only determine whether the pronoun should be accented or not, but also imposes constraints on which other parts of the sentence may be accented. Conversely, one would expect that in discourses like (2) the shift of accent between the direct object and the verb in (2c) should correlate with a change of antecedent for the pronoun *er* 'he.' This study concentrates specifically on the contrast between (4) and (5c) where the pronoun remains unaccented in both versions. Here, the nuclear accent on *geschnitten* 'cut' is expected to correlate with the resolution of *er* 'he' to *Marek*. Resolution to *Johann* would, of course, be dispreferred due to distance considerations, however a nuclear accent on *Kartoffeln* should favour this suboptimal resolution. Testing this hypothesis is the main goal of the experiment presented below, but first a few words on some further corollaries of this hypothesis.

2 On the Role of Discourse Structure

If the above theory is correct, then placement of accent can influence which part of the context a current sentence is contrasted with. Thinking of contrast as a rhetorical relation, along the lines of Mann & Thompson [9] or Asher & Lascarides [10], accent placement can thus affect the *attachment site* of the current sentence in the discourse structure: with the accentual pattern in (4) the sentence is attached with a contrast relation to the immediately preceding sentence; with the accent on the direct object as in (5c), the sentence is attached higher up in the discourse structure, to a sentence that is farther away. In other words, the latter case is an instance of *discourse pop*.

The present work is part of a larger study on prosody as a cue to discourse structure. There is a substantial body of research on discourse prosody (see e.g. [11,12,13,14,15,16,17,18]) showing that pitch range—the fundamental frequency span between the realizations of high and low tones—is higher at the beginning of a discourse unit (e.g. a paragraph) and lower at its end. A switch from one discourse unit to another (topic shift, or discourse pop) is therefore associated with a perceptible reset of pitch range back to higher and larger F0 span. Similarly, discourse pops correlate with relatively longer pauses between utterances [19,20].

Furthermore, it is well-known that the hierarchical organisation of discourse (global topic structure) affects anaphora resolution. Although, in general, referents mentioned in more recent sentences tend to be more accessible for pronominal

reference, a discourse pop can change this. If a less recent antecedent is related to a more global discourse topic, it can become more salient when that topic is reactivated after the pop. A more recent antecedent, on the other hand, can become less salient, if it is only locally important in the discourse segment just closed off.[2] Consequently, prosodic features signaling a discourse pop are expected to facilitate resolution of pronouns to less recent antecedents, which is supported by our previous experimental studies [29,30].

Applying these findings to example (2) above, one would expect that a pitch reset in the last sentence, as well as a long pause before it, should favour high attachment to (2a) with corresponding resolution of *er* 'he' to *Johann*. Lack of pitch reset and normal pause length before (2c) should correlate with low attachment to (2b) and resolution of *er* 'he' to *Marek*. In our present experiment the effects of accentuation and global prosodic parameters were studied simultaneously, as we were interested in possible interactions between different prosodic devices signaling discourse attachment.

3 Experiment

3.1 Method

Discourses: For the purposes of the experiment we constructed 40 discourses, each of which consisted of a set of 3 sentences similar to (2), and in which the last sentence could be understood as contrasting with either the first or the second sentence, depending on the interpretation of the pronoun. The potential antecedents were proper names referring to male or female humans (either both male or both female), and always constituted the grammatical subject of the sentence, occurring in sentence-initial preverbal position. Sentence 2 was related to sentence 1 by a discourse adverbial that appeared immediately after the finite verb, cf. *indes* 'meanwhile' in (2b). The target sentences started with a discourse adverbial, cf. *außerdem* 'besides' in (2c), while the ambiguous pronoun *er* 'he' or *sie* 'she,' which was also the grammatical subject, immediately followed the finite verb. We wanted to avoid placing the target pronoun in the absolute sentence-initial position so that the first prenuclear pitch accent could precede it thus enabling the listener to appreciate the pitch range of the utterance before he or she interpreted the pronoun. The nuclear accent in turn always occurred after the target pronoun.

As with (2), all the experimental discourses were designed in such a way that shifting the nuclear accent from one constituent to another in the target sentence would indicate contrast with the first or the second sentence of the context. It should be noted, though, that there is an asymmetry between the accentual patterns in (4) and (5c). The nuclear accent on the transitive verb as in (4)

[2] This generalisation has been expressed in various forms as the Right Frontier Constraint [21,10], the stack model [22], the cache model [23], the veins theory [24], the rhetorical distance theory [25], among others, and has been empirically substantiated by e.g. [26,27,28].

indicates more or less unambiguously that the verb bears *narrow focus*; that is, the sentence could only be used felicitously as an answer to a question like *What did he do to the potatoes?* or be uttered in a context where the potatoes are given, e.g. if it is contrasted with a sentence that explicitly mentions the potatoes. In contrast, the accent on the direct object in (5c) is ambiguous between narrow focus on *Kartoffeln* 'potatoes' and broad focus on the VP or the whole sentence (cf. e.g. [7,31]). Thus (5c) can answer both a question like *What did he cut?* and questions like *What did he do?* or *What happened?* Similarly, it can be contrasted with a sentence that only differs from (5c) in the referent of the direct object, e.g. (2a), or with one where, for instance, the whole VP is different: *A: Johann hat die Pfanne gewaschen. B: Nein, er hat [die Kartoffeln geschnitten]*$_{\text{FOC}}$. 'A: John washed the frying pan. B: No, he cut the potatoes.' Finally, a transitive sentence with a nuclear accent on the direct object need not be involved in a contrast relation at all and can be uttered "out of the blue," hence this accentual pattern is often called the *default* or the *neutral* pattern.

In order to prevent a confound between the neutral vs. non-neutral accentual pattern distinction and the factor under investigation—contrast with sentence 1 vs. contrast with sentence 2—we made sure that our materials were balanced with respect to whether the "neutral" pattern supported attachment to sentence 1 or 2. To achieve this, half of the discourses were like (2), in that the neutral pattern appeared in the 'contrast with sentence 1' condition, whereas in the other half this was reversed, in that the neutral pattern appeared in the 'contrast with sentence 2' condition. An example of the latter is a discourse like (6) below, cf. the English translation in (7). Here the neutral pattern with the nuclear accent on the direct object *Garten* 'garden' in (6c) appeared in the 'contrast with sentence 2' condition, whereas the marked pattern with the nuclear accent on the verb *gemalt* 'painted' was expected to trigger contrast with sentence 1.

(6) a. Dirk hat den Garten fotografiert.
Dirk has the garden photographed
b. Franz hat solange den Teich gemalt.
Franz has in that time the pond painted
c. Dann hat er den Garten gemalt.
Then has he the garden painted

(7) a. Dirk took some photos of the garden.
b. During that Franz painted the pond.
c. Then he painted the garden.

The syntactic functions of the constituents involved in the accent shift manipulation were varied. There were 12 discourses like (2) and (6) where the nuclear accent was shifted between the (monotransitive) main verb and the direct object. In 8 discourses the accent was shifted between the first and the second object

of a ditransitive verb, e.g. *hat Benno ein* BUCH *geschenkt* 'gave Benno a BOOK' vs. *hat* BENNO *ein Buch geschenkt* 'gave BENNO a book,' in (8)/(9).

(8) a. Martha hat Niklas ein Buch überreicht.
Martha has Niklas a book presented
b. Leonie hat dann Benno eine DVD beschert.
Leonie has then Benno a DVD presented
c. Außerdem hat sie Benno ein Buch geschenkt.
apart from that has she Benno a book presented

(9) a. Martha gave Niklas a book (as a present).
b. Then Leonie gave Benno a DVD.
c. Apart from that she gave Benno a book.

There were 8 discourses in which the accent shift manipulation concerned the direct object of a (mono)transitive verb and a PP- or adverbial modifier of the verb, e.g. HEUTE *ein Seminar versäumt* 'missed a class TODAY' vs. *heute ein* SEMINAR *versäumt* 'missed a CLASS today.' In 2 cases the accent was shifted between a head noun and its PP argument: *ein* BUCH *über Napoleon* 'a BOOK about Napoleon' vs. *ein Buch über* NAPOLEON 'a book about NAPOLEON'; in 4 cases between an NP and its PP modifier: *ein* REGAL *aus Nussbaum* 'a SHELF of walnut wood' vs. *ein Regal aus* NUSSBAUM 'a shelf of WALNUT wood'. Finally, 6 discourses were like (10)/(11) in which the accent was shifted between an NP and its adjectival modifier: *mit einer blonden* AMERIKANERIN 'with a blond AMERICAN' vs. *mit einer* BLONDEN *Amerikanerin* 'with a BLOND American.'

(10) a. Björn tanzte mit einer rothaarigen Amerikaner-in.
*Björn danced with a.*FEM *red-haired American*-FEM
b. Maik tanzte übrigens mit einer blonden Schwed-in.
*Maik danced by the way with a.*FEM *blond Swede*-FEM
c. Vorher tanzte er mit einer blonden Amerikaner-in.
*Before that danced he with a.*FEM *blond American*-FEM

(11) a. Björn danced with a red-haired American.
b. By the way, Maik danced with a blond Swede.
c. Before that, he danced with a blond American.

In the 28 discourses in which the main verb was not involved in the accent shift manipulation, it was important that the verb be part of the background, i.e. that it constitute the parallel (non-contrasting) part of the contrasting sentences. As a result, the verb had to be repeated in all three sentences in a set, e.g. *tanzte* 'danced' in (10), which often led to rather unnatural discourses. To avoid this, in 21 of these 28 discourses, the second and third occurrences of the verb were replaced by synonyms or near-synonyms as in (8) above, where the

verbs *überreichen*, *bescheren* and *schenken* all describe an act of giving a present to someone.[3]

Finally, 42 distractor discourses were constructed. As with the experimental discourses, these consisted of a set of three sentences and mentioned multiple human referents, but varied as to whether or not they contained contrast relations, and as to whether or not the pronouns in sentence 2 or 3 resolved unambiguously on the basis of number and gender features.

Each discourse (experimental or distractor) was accompanied by a *who?*-question of the form in (12) *Who cut the potatoes?* In the experimental items the question was derived from sentence 3 in order to reveal the hearer's interpretation of the pronoun. In distractors, the question addressed any of the three sentences.

(12) Wer hat die Kartoffeln geschnitten?
Who has the potatoes cut

Audio Materials: All materials were recorded in an anechoic chamber. The sentences were read by one female speaker in randomised order (i.e. not in the context of the respective discourses), aiming at producing constant pitch range and intensity values. The third sentence of each experimental discourse was recorded in two versions corresponding to the two nuclear accent placements, cf. Figs. 1 and 2.

The sentences were resynthesised and the discourses put together using uniform pitch range and pause duration values following the methodology of Mayer et al. [30]. All signal processing was done using PRAAT [32].

Pitch range was defined as the range between the highest intonationally relevant high tone (HT) and the lowest relevant low tone (LT) within one phrase (sentence). Relevant tones were labeled manually in the original recordings and corresponded usually to high or low tonal targets of pitch accents. For pitch range manipulations, 3 different ranges were defined: normal, compressed and expanded. We determined the normal pitch range of the female speaker as ranging from 150 Hz (baseline) to 270 Hz (topline). Using standard expansion and compression ratios, the expanded pitch range of the speaker was set to 150 Hz baseline and 310 Hz topline and the compressed range to 140 Hz baseline and 250 Hz topline. The first and the second sentence of each discourse were assigned an expanded and a normal range, respectively. Each accentual realization of sentence 3 of the experimental discourses (cf. Figs. 1 and 2) was resynthesised in two versions: once with a compressed pitch range corresponding to the continuity condition and once with an expanded pitch range for the discourse pop condition. Third sentences of distractor discourses were assigned one of the pitch range values (expanded or compressed) on a random basis. Based on the original

[3] Either all the three verbs in a discourse were the same like in (10) or all three were distinct synonyms like in (8). Our intuition was that if the verb of the target sentence were synonymous with the verb of one of the context sentences, but literally repeated that of the other, this could have created a bias for attachment to the sentence that contained the literal repetition.

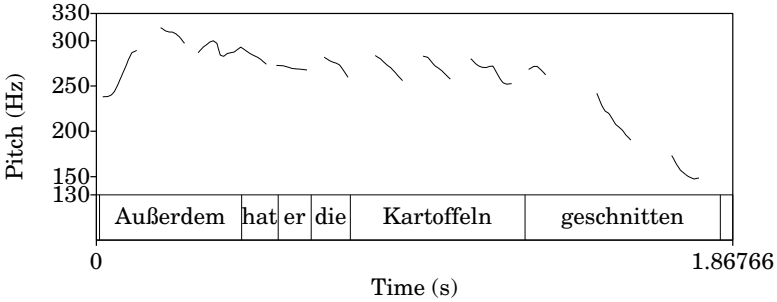


Fig. 1. Pitch track for (4). The falling nuclear accent occurs on *geschnitten* 'cut'.

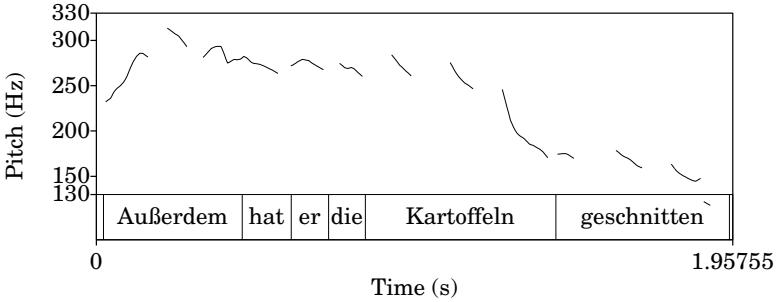


Fig. 2. Pitch track for (5c). The falling nuclear accent occurs on *Kartoffeln* 'potatoes'.

HT and LT and the target range values, the pitch contour of each sentence was shifted so that the LT was set to the target baseline and multiplied so that the HT reached the target topline.

The original discourses were re-created by concatenating the resynthesised sentences with intervening pauses (intervals of zero amplitude). The standard pause length was set to 400 ms. However, in the discourse pop condition an extra long pause of 800 ms was inserted before the last sentence. Figures 3 and 4 show the resynthesised and the reconcatenated realizations of (2) in the continuity and the discourse pop conditions, respectively (the accentual realization of sentence 3 is as in (4), cf. Fig. 1). The horizontal dashed lines indicate the top- and the baselines of the resynthesised sentences. Notice that in the continuity condition (Fig. 3) the pitch range "declines" from the beginning towards the end of the discourse, whereas in the discourse pop condition (Fig. 4) a pitch reset occurs in sentence 3.

In sum, each discourse appeared in four versions corresponding to the four experimental conditions resulting from a 2 by 2 design with accent placement and global prosody (GP) as factors: (1) accent placement in sentence 3 as contrasted with sentence 1 vs. sentence 2; and (2) pitch range of sentence 3 and pause duration before sentence 3 signaling discourse pop vs. discourse continuity.

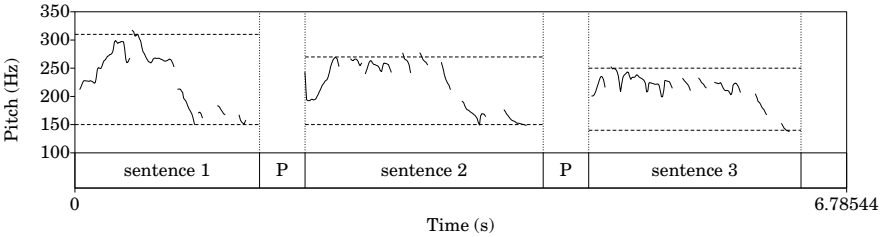


Fig. 3. Prosodic realization of (2) in the continuity condition: the pause between sentence 2 and 3 has standard length (400 ms); sentence 3 has a compressed pitch range

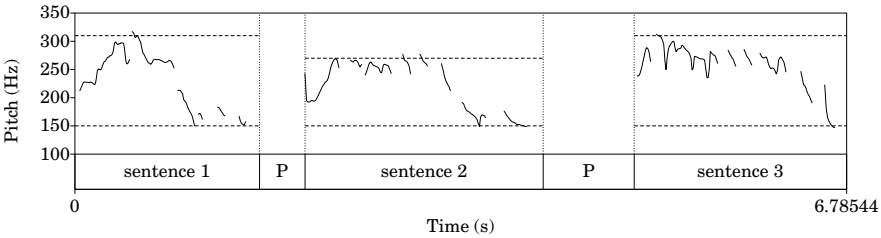


Fig. 4. Prosodic realization of (2) in the discourse pop condition: the pause between sentence 2 and 3 is long (800 ms); sentence 3 has an expanded pitch range

The final questions were spoken by a male speaker and were appended to the sequences after a silent interval of 1500 ms with the original unmanipulated question intonation.

Procedure: The experimental items were divided into four counterbalanced lists that contained only one version of each item, and mixed with the distractor items. The items were presented in a randomised order. After listening to each item only once, the participants had to answer the questions orally; no choice lists of possible answers were provided. The responses were recorded and classified as indicating resolution to the referent introduced in the first sentence (R1) or the second sentence (R2) or as "incorrect resolution". The response was classified as incorrect if it showed resolution to a referent other than R1 or R2, or if the subject refused to give a definite answer (e.g. by saying *I don't know*).

3.2 Subjects

53 subjects took part in the experiment, all of whom were undergraduate students of linguistics and native speakers of German, and were either paid or received credit for participation. The data of 13 subjects were excluded from the analysis since they failed to give an answer or gave an absurd answer to the test question three or more times. The data of the remaining 40 participants (10 per list) were subjected to statistical analysis.

3.3 Results

The data were aggregated by subjects and by items, the resulting relative frequencies of R1 resolution were square-root arcsine transformed and subjected to ANOVAs. The target pronoun was resolved more frequently (71.8% of times) to the most recent antecedent R2 than to R1 in all conditions, cf. Fig. 5, but there were more resolutions to R1 in the conditions where the accentual pattern corresponded to contrast of sentence 3 with sentence 1 (38.8%) as in (5c), than there were R1 resolutions in the conditions where the accentual pattern corresponded to contrast of sentence 3 with sentence 2 (17.4%) as in (4). The main effect of accentuation was significant both by subjects and by items [$F_1(1,39) = 32.65, p < .001$, and $F_2(1,39) = 59.75, p < .001$]. The main effect of global prosodic parameters was less strong (30.8% vs. 25.4%) and only significant by items [$F_1(1,39) = 1.66, p = .21$, and $F_2(1,39) = 7.15, p < .05$]. We found no interaction between the accentuation and global prosody factors [$F(1,39) < 1$].

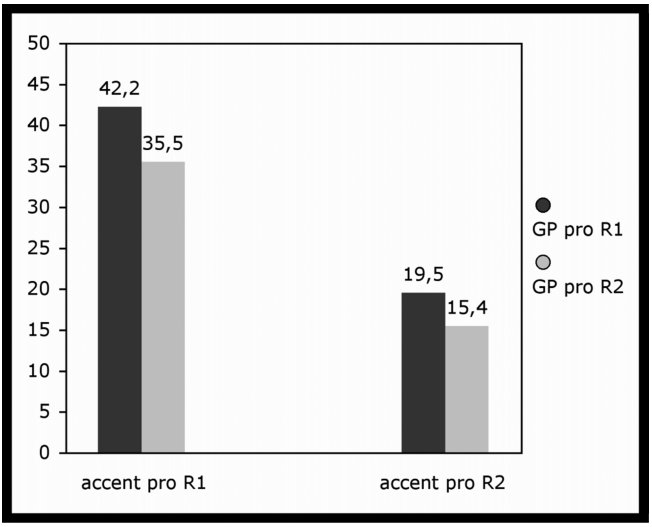


Fig. 5. Number of resolutions of the target pronoun to R1, the antecedent introduced in sentence 1, in %

4 Discussion and Conclusions

These results corroborate our hypothesis that the placement of nuclear accent can affect pronoun interpretation by determining with which sentence in the context the current sentence should be contrasted. Although, in general, pronoun resolution to the most recent antecedent is preferred, this preference is overridden more often if the accentual pattern of the sentence containing the pronoun indicates that it should be contrasted with an alternative realized earlier in the discourse, in which case the pronoun (if it is unaccented) is resolved

to an antecedent occurring in that alternative. This supports the predictions of theories that claim effects of accent on pronoun resolution to be a by-product of the general theory of accentuation as a contrast-signaling device. However, our results complement those of previous empirical studies in showing that the overall accentual pattern of the sentence also plays a role in determining the contrasting alternative, so that even the accentuation of constituents other than the pronoun can affect its resolution.

Since the effect of global prosody was only significant by items, this effect is more difficult to interpret, as are also the results regarding interaction between global prosody and accentuation as factors. It seems that accentuation and other prosodic parameters may work as independent factors. For this result to be conclusive, however, a stronger main effect of global prosody would need to have been measured. Our previous findings [30] show that the effect of pitch range and pause duration on pronoun resolution is rather subtle (affecting upto 10% of resolutions) and tends to disappear when the discourse pop is not signaled strongly enough, e.g. if different prosodic features do not "cooperate" in indicating a strong prosodic break. This could be one reason why the effect of global prosody in the present experiment was rather weak.Using more strongly expressed prosodic contrasts between the discourse pop and the continuity conditions could help increase the related effect. Another possible explanation for the weakness of the effect is that contrast is a coordinating, or multinuclear, discourse relation [9,10], and as such is generally thought to assign equal discourse-structural prominence to the sentences it connects. Under this view it is not clear whether the discourse segment that is closed off by the discourse pop in our experimental items (sentence 2) has a subordinated status with respect to sentence 1 or not (see discussion in Sect. 2). However, it is interesting that the global prosody effect that we found is nevertheless in the direction predicted by the Right Frontier Constraint and similar theories: if a pitch reset in the target sentence and a longer pause before it indicate a discourse pop, the pronoun is resolved to an earlier antecedent more frequently than in the continuity condition. This suggests that listeners can sometimes accommodate one of the segments connected by a contrast relation as being discourse-structurally subordinate to the other.

In conclusion, this work contributes to the study of prosody and its interpretation in discourse by demonstrating that pronoun resolution is only one of a whole range of semantic effects of discourse structure conveyed by prosody.

References

1. Schwarzschild, R.: GIVENness, AVOIDF and other constraints on the placement of accent. Natural Language Semantics (1999) 141–177
2. Rooth, M.E.: A theory of focus interpretation. Natural Language Semantics **1** (1992) 75–116
3. Wolters, M., Beaver, D.: What does *he* mean? In Moore, J., Stenning, K., eds.: Proceedings of the Twenty-Third Annual Meeting of the Cognitive Science Society, New Jersey, Lawrence Erlbaum Associates (2001) 1176–1180

4. de Hoop, H.: On the interpretation of stressed pronouns. In Blutner, R., Zeevat, H., eds.: Optimality Theory and Pragmatics. Palgrave Macmillan (2003) 25–41
5. Kehler, A.: Coherence-driven constraints on the placement of accent. In: Proceedings of the 15th Conference on Semantics and Linguistic Theory (SALT-15), Los Angeles, CA (2005)
6. Kameyama, M.: Stressed and unstressed pronouns: Complementary preferences. In Bosch, P., van der Sandt, R., eds.: Focus: Linguistic, Cognitive and Computational Perspectives. Cambridge University Press (1999) 306–321
7. Selkirk, E.: Sentence prosody: Intonation, stress and phrasing. In Goldsmith, J., ed.: Handbook of Phonological Theory. Blackwell, Oxford, UK (1995) 550–569
8. Venditti, J.J., Stone, M., Nanda, P., Tepper, P.: Discourse constraints on the interpretation of nuclear-accented pronouns. In: Proceedings of the 2002 International Conference on Speech Prosody, Aix-en-Provence, France (2002)
9. Mann, W.C., Thompson, S.: Rhetorical Structure Theory: Toward a functional theory of text organization. Text **8**(3) (1988) 243–281
10. Asher, N., Lascarides, A.: Logics of Conversation. Studies in Natural Language Processing. Cambridge University Press (2003)
11. Lehiste, I.: The phonetic structure of paragraphs. In Cohen, A., Nooteboom, S.G., eds.: Structure and Process in Speech Perception. Springer (1975) 195–203
12. Silverman, K.E.A.: The Structure and Processing of Fundamental Frequency Contours. PhD thesis, University of Cambridge (1987)
13. Swerts, M., Geluykens, R.: The prosody of information units in spontaneous monologue. Phonetica **50** (1993) 189–196
14. Sluijter, A.M.C., Terken, J.M.B.: Beyond sentence prosody: Paragraph intonation in Dutch. Phonetica **50** (1993) 180–188
15. Nakajima, S., Allen, J.F.: A study on prosody and discourse structure in cooperative dialogues. Phonetica **50** (1993) 197–210
16. Ayers, G.M.: Discourse functions of pitch range in spontaneous and read speech. Ohio State University Working Papers in Linguistics **44** (1994) 1–49
17. Möhler, G., Mayer, J.: A discourse model for pitch-range control. In: Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, Perthshire, Scotland (2001)
18. den Ouden, H.: Prosodic Realizations of Text Structure. PhD thesis, University of Tilburg (2004)
19. Grosz, B., Hirschberg, J.: Some intonational characteristics of discourse structure. In: Proceedings of the 2nd International Conference on Spoken Language Processing, Banff, Canada (1992) 429–432
20. Swerts, M.: Prosodic features at discourse boundaries of different strength. Journal of the Acoustical Society of America **101** (1997) 514–521
21. Polanyi, L.: A formal model of the structure of discourse. Journal of Pragmatics **12** (1988) 601–638
22. Grosz, B.J., Sidner, C.L.: Attention, intentions and the structure of discourse. Computational Linguistics **12**(3) (1986) 175–204
23. Walker, M.A.: Limited attention and discourse structure. Computational Linguistics **22**(2) (1996) 255–264
24. Cristea, D., Ide, N., Romary, L.: Veins theory: A model of global discourse cohesion and coherence. In: Proceedings of COLING-ACL 1998. (1998) 281–285
25. Kibrik, A.A.: Reference and working memory: Cognitive inferences from discourse observations. In van Hoek, K., Kibrik, A.A., Noordman, L.G.M., eds.: Discourse Studies in Cognitive Linguistics. Benjamins, Amsterdam (1999) 29–52

26. Anderson, A., Garrod, S.C., Sanford, A.J.: The accessibility of pronominal antecedents as a function of episode shifts in narrative text. Quarterly Journal of Experimental Psychology **35a** (1983) 427–440
27. Hitzeman, J., Poesio, M.: Long distance pronominalisation and global focus. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal, Canada (1998) 550–556
28. Cristea, D., Ide, N., Marcu, D., Tablan, V.: Discourse structure and co-reference: An empirical study. In: Proceedings of the ACL Workshop on The Relationship Between Discourse/Dialogue Structure and Reference, USA, University of Maryland (1999)
29. Jasinskaja, E., Kölsch, U., Mayer, J.: Global prosodic parameters and anaphora resolution. In Auran, C., Bertrand, R., Chanet, C., Colas, A., Cristo, A.D., Portes, C., Reynier, A., Vion, M., eds.: Proceedings of the International Symposium on Discourse and Prosody as a Complex Interface, Aix-en-Provence, France (2005)
30. Mayer, J., Jasinskaja, E., Kölsch, U.: Pitch range and pause duration as markers of discourse hierarchy: Perception experiments. In: Proceedings of Interspeech 2006, Pittsburgh PA, USA (2006)
31. Wagner, M.: Asymmetries in prosodic domain formation. MIT Working Papers in Linguistics **49** (2005) 329–367
32. Boersma, P., Weenink, D.: Praat: doing phonetics by computer. University of Amsterdam (2005)

Empirically Assessing Effects of the Right Frontier Constraint

Anke Holler[1] and Lisa Irmen[2]

[1] University of Heidelberg
Institute of General and Applied Linguistics
[2] Institute of Psychology
holler@cl.uni-heidelberg.de,
lisa.irmen@psychologie.uni-heidelberg.de

Abstract. In a questionnaire study the effects of discourse structural information on resolving inter-sentential anaphora were investigated. The Right Frontier Constraint, first proposed by Polanyi (1988), states that potential antecedents of an anaphor that are placed at the right frontier of a discourse unit can be accessed more easily than antecedents that are placed somewhere else. Participants (N=36) received written experimental passages of six lines each that contained a pronominal anaphor in the last line and two potential antecedents in the preceding text, one introduced in the first, one in the fourth line of a passage. Antecedents' relative position to the right frontier was manipulated through the discourse relation between the first and the second antecedent and through the filler information interposed between the second antecedent and the anaphor. The two potential antecedents either had the same or different grammatical gender. In the latter case only the first antecedent was gender congruent to the anaphor. Participants' task was to name the anaphor's antecedent. Results show that in case of unequal gender antecedents, participants almost always chose the gender congruent first antecedent, irrespective of its position relative to the Right Frontier. In case of equal gender antecedents choice patterns point to an influence of an antecedent's position relative to the Right Frontier. Alternative theoretical approaches such as centering theory or situational models cannot account for the found results. The findings in the same gender antecedent condition are therefore interpreted as an effect of the Right Frontier Constraint.

1 Introduction

It is a well-known fact that the resolution of anaphora is made up of many processes operating at different linguistic and non-linguistic levels. Even if we confine ourselves to looking at linguistic processes, there are various factors influencing the accessibility of information. Substantial empirical research has shown that phonologic and morpho-syntactic as well as semantic and pragmatic information guides the way an anaphor may find its antecedent. Even subtle changes in the grammatical form of sentences may influence anaphor resolution as Klin, Guzmán, Weingartner and Ralano (2006) have recently shown. Whereas one thread of research concentrates on aspects of the complexity of the anaphor (see for instance Ariel (2001)), another

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 15 – 27, 2007.

strand of research deals with the question of which properties of an anaphor's potential antecedents affect their salience. Our work addresses this latter aspect.

There is ample evidence that morpho-syntactic information such as gender or number congruency disambiguates the relation between an anaphor and its antecedent. In addition, it has been argued that anaphor resolution is subject to certain semantic inferences that also help to clearly determine the accessible antecedents of an anaphor. But in all cases where this kind of information is not available or does not vary with regard to a set of potential antecedents other linguistic factors come into play. Although a diversity of criteria for the selection of the antecedent of an anaphor has been provided depending on the respective theoretical background, there is wide agreement that the following factors are relevant for anaphor resolution: First, the syntagmatic distance between anaphor and antecedent, which is also known as the recency effect. Second, the grammatical function (or obliqueness), which on the one hand condensed in a subject assignment strategy, whereupon ambiguous pronouns will be assigned to antecedents which function as subjects. On the other hand, the parallel function assignment strategy refers to this factor as it proposes that pronouns will be assigned to antecedents with identical grammatical functions. The latter strategy is also semantically reformulated in terms of thematic role assignment. Third, further semantic aspects such as whether the potential antecedent is animate or not, and whether it functions as a topic or not. Fourth, the information-structural status of a potential antecedent in terms of providing new or familiar information seems to be a relevant factor. See Garnham (2001) for an overview of factors usually accepted as influencing anaphor resolution. These factors determining the salience of nominal antecedents of anaphora have been subject to a variety of psycholinguistic research. Several empirical studies aimed at investigating how an appropriate antecedent is assigned to potentially ambiguous pronouns during interpretation. In the course of this research it is generally agreed that the resolution of anaphora depends on what entities are currently in the focus of attention, e.g. Gordon, Grosz and Gilliom (1993), Hudson-d'Zmura and Tanenhaus (1998). Although it has been shown, that all of the aforementioned factors are certainly involved in establishing the preferred antecedent, it is still an open question inasmuch further factors have to be considered. To the best of our knowledge, the influence of discourse relations on the salience of potential antecedents of anaphora has not been in focus of psycholinguistics, yet.

The study presented here investigates inasmuch discourse structural information affects the way inter-sentential anaphora are resolved. It aims at an empirical verification of the so-called Right Frontier Constraint first proposed by Polanyi (1988). Following this constraint, it is hypothesized that potential antecedents of an anaphor that are placed at the right frontier of a discourse unit can be accessed more easily than antecedents being placed somewhere else. In other words, readers are more likely to resolve anaphora that are perceived as discourse-structurally salient. We examined this hypothesis in a questionnaire study.

2 The Structure of Discourse

A significant body of work on discourse structure has developed over the last twenty years. In the course of this research it has been shown that parts of texts can pertain to

previous parts in various ways. For instance, subsequent sentences may elaborate the content of preceding sentences, contrast to it or offer background information. Discourse-functional relations describe the variety of rhetorical roles utterances play in their discourse contexts. Accordingly, discourse-functional relations connect bits of the text and thereby organize texts. Discourse relations (sometimes also called rhetorical relations or coherence relations) can either be expressed explicitly using relation specific discourse markers, so-called cue words, or they are implicit and must be inferred by the readers. Much research has focused on the question of which discourse relations to distinguish and how to classify them (cf. among others Matthiessen and Thompson 1987, Mann and Thompson 1988, Sanders, Spooren, and Nordman, 1992). It is beyond dispute that certain discourse relations have something in common, and that the set of discourse relations can be partitioned with respect to various criteria. One relevant parameter in terms of which discourse relations can vary is their discourse-hierarchical status.

Many researchers have observed that discourse segmentation has a hierarchical structure (e.g. Grosz and Sidner (1986), Mann and Thompson (1987), Polanyi (1988, 1996), Asher (1993), Asher and Lascarides (2003)). Although the proposed discourse models differ in formal setting as well as in the set of stipulated discourse relations, they share the basic assumption that discourse portions can either be coordinated or subordinated to each other, and that accessibility of parts of discourse is determined by the coordination and subordination relation specified by the respective model of discourse. Intuitively, if a discourse relation subordinates a subsequent part of text to a preceding one, then this part provides more detailed information about the event or the proposition expressed in the prior part of text. However, if a discourse relation coordinates two parts of texts, then the level of detail does not change since the subsequent sentence somehow continues the information given in the preceding part of text. In other words, "coordination and subordination reflect the different effects these [discourse] relations have on the 'granularity' or the level of detail being given in the discourse" (Asher and Lascarides (2003:8)). Example (1) taken from Asher and Lascarides (2003) illustrates this claim.

(1) π_1: Max had a great evening last night.
π_2: He had a great meal.
π_3: He had salmon.
π_4: He devoured lots of cheese.
π_5: He then won the dancing competition.

In this example, π_2 to π_5 elaborate the information provided by π_1, which means that π_2 to π_5 are subordinate to π_1. The part from π_2 to π_5 is even further structured. Sentences π_3 and π_4 provide more detailed information about the event expressed by π_2. The sentence π_5 on the other hand continues the elaboration of Max's evening started by π_2.

The fundamental assumption of theories describing discourse structures is thus that discourse consists of a set of discourse units, which are connected by two sorts of discourse relations: subordinating relations "that push the discourse structure down" as Asher (2004) phrases it and coordinating relations "that push the structure from left to right." In other words, a relation is considered as subordinating discourse portions

in case one constituent discourse unit dominates another. On the other hand, a relation is considered as coordinating if no constituent discourse unit dominates another. Asher & Vieu (2005) provide several theory-internal tests evaluating discourse relations in terms of their discourse-hierarchical status. These tests substantiate the view that Elaboration is a primary subordinating relation, while Narration is a typical representative of the class of coordinating relations.

The hierarchical structure of discourse is relevant in that it has implications for the salience of information. It is a well-known fact that hierarchical discourse structure imposes restrictions on pronominal reference. In particular, antecedents to anaphora of the current clause must be introduced in the proposition expressed by the prior clause or in any proposition subordinate to the prior proposition. This observation has been generalized as a restriction saying that antecedents of anaphora must be introduced by proposition lying on the right edge (Polanyi 1988, 1996) or on the right frontier (e.g. Webber (1988), Asher (1993), Asher and Lascarides (2003)) of the discourse structure.

One of the recent linguistically influential frameworks that provide formal means for the analysis of discourse structure is Segmented Discourse Representation Theory (SDRT, Asher & Lascarides, 2003). SDRT offers a formal account of the hypothesis that discourse has a hierarchical structure upon which interpretation depends.

According to SDRT's formal setting, discourse structures consist of a set of labels for discourse constituents and a function that assigns formulas to these labels. To be reminiscent of DRT's discourse representation structures (DRS), discourse constituents are called Segmented DRS (SDRS). Asher & Lascarides (2003:138) define a SDRS in the following way:

A discourse structure or SDRS is a triple $<A, \mathcal{F}, LAST>$, where:

- A is a set of labels;
- $LAST$ is a label in A (intuitively, this is the label of the content of the last clause that was added to the logical form); and
- $\mathcal{F}$ is a function which assigns each member of A a member of Φ, which is the set of well-formed SDRS-formulae.

A SDRS can be converted into a graph whose nodes represent its labeled constituents and whose edges represent the discourse relations established between these constituents. Thereby, each subordinating relation creates a downward edge and each coordinating relation a horizontal one. An important restriction is that two nodes in a graphical representation cannot be connected using both a subordinating and a coordinating relation.

The central constraint on discourse update and anaphor resolution in SDRT is the so-called Right Frontier Constraint (RFC). Asher (1993) formulates a right-frontier rule for attachment saying that new information must either attach to the last entered constituent β in a discourse structure or to some constituent γ such that (β, γ) is in the transitive closure of the subordination relation. More formally, the right frontier is defined in SDRT as the set of available nodes for attachment falling under the following possibilities:

1. The label $\alpha = LAST$;
2. Any label $\gamma \geq_D^* \alpha$, where $\geq_D^*$ is defined recursively:

a. $R(\gamma,\alpha)$ is a conjunct in $\mathcal{F}(l)$ for some label l, where R is a subordinating discourse relation;
b. $R(\gamma,\delta)$ is a conjunct in $\mathcal{F}(l)$ for some label l, where R is a subordinating discourse relation and $\mathcal{F}(\delta)$ contains as conjunct $R'(\delta',\alpha)$ or $R'(\alpha,\delta')$, for some R' and δ'; or
c. $R(\gamma,\delta)$ is a conjunct in $\mathcal{F}(l)$ for some label l, where R is a subordinating discourse relation and $\delta \geq_D{}^* \alpha$.

The RFC affects anaphor resolution as the antecedent for an anaphoric expression is accessible only at the right frontier, i.e. at the right hand side of any level of a linearly ordered discourse parse tree.

That the right frontier has semantic effects can be illustrated by the aforementioned example (1). The sequence π_1 to π_5 can neither be continued by the sentence *It was a beautiful pink*, where the pronoun *it* is intended to refer to the *salmon*, nor by the sentence *It was delicious*, where the pronoun *it* is intended to refer to the *meal*. There is a discourse-structural explanation for this observation: In both cases the intended antecedent does not lie on the right frontier. The antecedents are, thus, not accessible. This is indicated by (2), which represents the discourse structure of example (1) according to the SDRT framework. Figure 1 depicts the corresponding graph.

(2) $<A, \mathcal{F}, LAST>$, where:

- $A = \{\pi_0, \pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6, \pi_7\}$
- $\mathcal{F}(\pi_1) = K_{\pi 1}$
- $\mathcal{F}(\pi_2) = K_{\pi 2}$
- $\mathcal{F}(\pi_3) = K_{\pi 3}$
- $\mathcal{F}(\pi_4) = K_{\pi 4}$
- $\mathcal{F}(\pi_5) = K_{\pi 5}$
- $\mathcal{F}(\pi_o) = \text{Elaboration}(\pi_1, \pi_6)$
- $\mathcal{F}(\pi_6) = \text{Narration}(\pi_2, \pi_5) \wedge \text{Elaboration}(\pi_2, \pi_7)$
- $\mathcal{F}(\pi_7) = \text{Narration}(\pi_3, \pi_4)$
- $LAST = \pi_5$

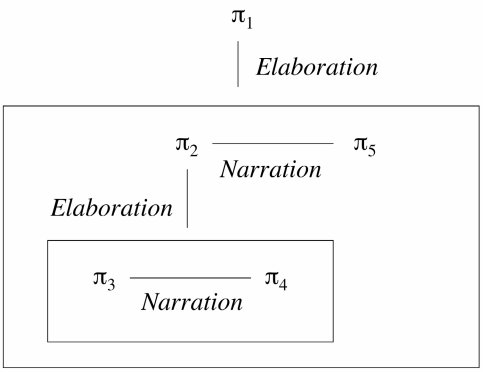


Fig. 1. Discourse structure of example (1)

This article is concerned with an empirical assessment of the Right Frontier Constraint. We ask whether the right frontier indeed constrains the accessibility of anaphora in discourse. To answer this question we conducted an experiment described in the next section.

3 An Empirical Study

In the following we will present a questionnaire-based experiment that investigated the effects of the discourse-relational structure on the interpretation of pronominal anaphora. We tested to which extent the discourse-structural position of antecedent candidates of anaphoric expressions is a salience-influencing factor.

3.1 Method

This section will provide details of the experimental materials used, the experimental procedure and the hypotheses tested.

3.1.1 Materials

Experimental materials consisted of short passages made up of six lines with a pronominal anaphor in the last line and two potential antecedents in the preceding text, one of which was introduced in the first, the other one in the fourth line (an example is provided in Table 1). The discourse relation between the first and the second antecedent and the filler information interposed between the second antecedent and the anaphor determined an antecedent's relative position to the right frontier.

All potential antecedents were comparable in terms of their information status as they were either definite noun phrases or proper names and hence hearer-old following Prince (1992). Thus, the speaker assumes that the addressee is already acquainted with the referent of the respective noun phrase. Because definite noun phrases and proper names differ as to their semantic properties, within items the first and the second antecedent always were identical in this respect.

The experimental passages occurred in three possible versions: a) In items of the first condition, only the first antecedent stood at the right frontier (Type A). In these passages the second antecedent appeared in a discourse unit that stood in subordinate relation to the discourse unit containing the first antecedent and was followed by coordinate filler information. b) In the second type, only the second antecedent stood at the right frontier (Type B). The second antecedent occurred in a discourse unit that was discourse-structurally coordinated with the discourse unit of the first antecedent and was followed by subordinate filler information. c) In the third type, both potential antecedents stood at the right frontier (Type C). The discourse unit containing the second antecedent was connected to the discourse unit containing the first antecedent by a subordinating relation and was followed by subordinate filler information.

Within each type the two potential antecedents either had the same or different grammatical gender. In the latter case only the first antecedent was gender congruent to the anaphor. Table 1 displays an example in the three outlined versions.

Table 1. Example of the three types of item structure

Type A Second antecedent in subordinate relation to first antecedent	Am Morgen ging **die Studentin** in die Universität (π_1), denn es war mal wieder an der Zeit, die Vorlesung über die Vor– und Nachteile von Kants Kategorischem Imperativ zu besuchen. (π_2) Im Hörsaal war es sehr voll. (π_3) **Die Kommilitonin/Der Kommilitone** war wie immer schlecht gelaunt (π_4), und es hörte niemand zu. (π_5) Nachmittags musste **sie** noch viel erledigen.
Only first antecedent at right frontier	In the morning **the student** went to the university because it was time to attend the lecture on advantages and disadvantages of Kant's categorical imperative. The lecture hall was busy. **The fellow student** was as always in a bad mood and nobody listened. In the afternoon **she** still had many things to do.
Type B Second antecedent in coordinate relation to first antecedent	Am Morgen ging **die Studentin** in die Universität (π_1), denn es war mal wieder an der Zeit, die Vorlesung über die Vor– und Nachteile von Kants Kategorischem Imperativ zu besuchen. (π_2) Im Hörsaal war es sehr voll. (π_3) **Die Kommilitonin/Der Kommilitone** war stattdessen in der Bibliothek (π_4), denn dort war es sehr ruhig. (π_5) Nachmittags musste **sie** noch viel erledigen.
Only second antecedent at right frontier	In the morning **the student** went to the university because it was time to attend the lecture on advantages and disadvantages of Kant's categorical imperative. The lecture hall was busy. **The fellow student** however was in the library because it was quiet there. In the afternoon **she** still had many things to do.
Type C Both antecedents at right frontier	Am Morgen ging **die Studentin** in die Universität (π_1), denn es war mal wieder an der Zeit, die Vorlesung über die Vor– und Nachteile von Kants Kategorischem Imperativ zu besuchen. (π_2) Im Hörsaal war es sehr voll. (π_3) **Die Kommilitonin/Der Kommilitone** war wie immer schlecht gelaunt (π_4), denn es hörte niemand zu. (π_5) Nachmittags musste **sie** noch viel erledigen. In the morning **the student** went to the university because it was time to attend the lecture on advantages and disadvantages of Kant's categorical imperative. The lecture hall was busy. **The fellow student** was as always in a bad mood because nobody listened. In the afternoon **she** still had many things to do.

Using the afore presented notions of SDRT, the material shows the graphical discourse structures given in figure 2, figure 3, and figure 4, resp.:

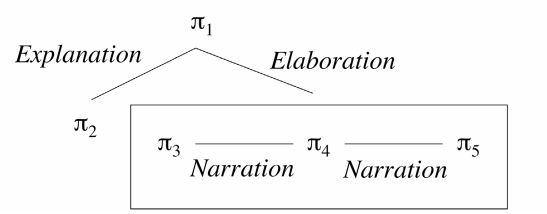


Fig. 2. Discourse structure of material of Type A

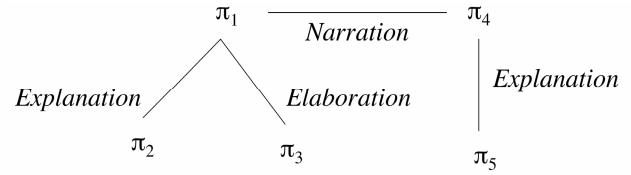


Fig. 3. Discourse structure of material of Type B

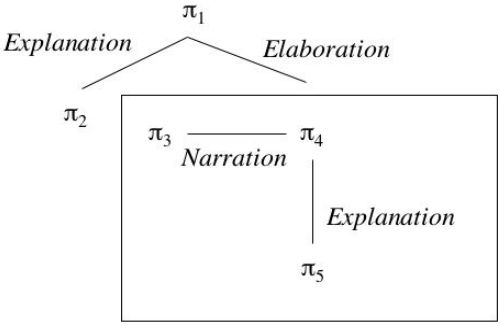


Fig. 4. Discourse structure of material of Type C

3.1.2 Procedure

Participants were presented with a questionnaire that contained six experimental passages of each type outlined above, resulting in 18 experimental passages. Three items of each type had equal gender antecedents, three items had unequal gender antecedents. After each passage participants were asked to name the antecedent of the pronominal anaphor by answering a question that rephrased the last line of a passage. If it read, for example, *In the afternoon she still had many things to do* the respective question was *Who was the one who had to do many things?*

3.1.3 Predictions

If anaphor resolution is constrained by the position of potential antecedents relative to the right frontier in the outlined way, participants should tend to choose the first antecedent in Type A and the second in Type B. No such difference is expected if other factors are more influential. Recency should, for example, favor the second antecedent across experimental conditions, gender congruency should always favor the first antecedent in case of unequal gender antecedents.

4 Results

In the following we will present the results first in the light of the RFC predictions, and then we will discuss possible alternative explanations of the data.

4.1 Right Frontier Constraint

The three different types of experimental passages were compared as to the frequencies of choices for the first or the second antecedent, separately for the two gender conditions (only first antecedent gender-congruent versus both antecedents gender congruent).

In case of equal gender antecedents, participants' choices differed between types of items and were in line with the idea of a right frontier constraint (see Table 2). Observed frequencies deviated reliably from the ones expected on the basis of marginal frequencies ($\chi^2_{(2)} = 8.323$, $p = .016$). While in Type A frequencies for Antecedent 1, the only right frontier antecedent, were higher than expected and lower for Antecedent 2, it was just the other way round for Type B (lower for Antecedent 1 and higher for Antecedent 2, the correct antecedent following RFC). Type C resembled results for Type A. As both antecedents conformed the RFC, this finding suggests that other factors also influenced participants' choices.

Table 2. Participants' choices in case of equal gender antecedents, separated by type of item (expected frequencies in parentheses)

Choices	1. Antecedent	2. Antecedent	Total
Type A (1. antecedent at right frontier)	72 (64.7)	38 (45.3)	110
Type B (2. antecedent at right frontier)	51 (63.0)	56 (44.0)	107
Type C (both antecedents at right frontier)	70 (65.3)	41 (45.7)	111
Total	193	135	328

The differing patterns of results between items of Types A and C on the one hand and Type B on the other indicate an influence of the passages' discourse structure on participants' choices. Even though there is no strong preference for the second antecedent in items of Type B, there clearly is a tendency against the dominance of Antecedent 1 that is present in Types A and C. As discourse structure was the only intended variation between types of items, we interpret the difference between items with Antecedent 1 as the only accessible (Type A) or as one possible antecedent (Type C) and items with Antecedent 2 as the only accessible antecedent (Type B) as an effect of the RFC.

In case of unequal gender antecedents, participants almost always chose the first antecedent, that is, the one that was gender congruent to the anaphor (see Table 3). Frequencies did not differ between the three types of passages, that is, in case of unequal gender antecedents there were no effects of discourse relation ($\chi^2_{(2)} < 1$, $p > .2$).

Table 3. Participants' choices in case of antecedents with unequal gender, separated by type of item

Choices	1. Antecedent	2. Antecedent	Total
Type A (1. antecedent at right frontier)	110	2	112
Type B (2. antecedent at right frontier)	111	2	113
Type C (both antecedents at right frontier)	110	3	113
Total	331	7	338

4.2 Alternative Accounts

In the following we address some alternative theoretical approaches that might also be considered influential for the findings reported above.

Syntagmatic distance. The experimental conditions of Type A and C revealed a clear preference for the antecedent with the greater distance to the anaphor. In Type B participants chose the closer and the further antecedent about equally often. These findings do not suggest a strategy of systematically going for the most recent matching antecedent. Therefore, the syntagmatic distance between the anaphor and the antecedents cannot account for the results of any experimental condition.

Local discourse coherence through centering. A central claim of Centering Theory (Grosz, Joshi and Weinstein, 1995) is that in the context of establishing discourse coherence certain entities are more central than others. According to Centering Theory, discourse is composed of three components: (i) a linguistic structure (i.e. the

structure of the sequence of the utterance), (ii) an intentional structure (i.e. the structure of purposes), and (iii) an attentional state (i.e. discourse participants' focus of attention). The latter records the objects (as well as properties and relations) that are salient at a given point in the discourse.

The centering approach predicts that an ambiguous pronoun would be interpreted as referring to the discourse entity that corresponds to the focus of attention, the so-called preferred center. In the absence of other discourse information, the subject noun is taken to be the default center of attention, and is, thus, continued by a pronoun in the following sentence. Centering theory does not make use of rhetorical relations. Nonetheless, it implements a hierarchical discourse structure by modeling dominance relationships on a so-called focus-stack, cf. Grosz and Sidner (1986).

Inspired by the centering framework, Chambers and Smyth (1998) pointed to the primary influence of parallel syntactic structures on resolving pronouns and its dominance over other effects on a discourse entities' salience. A non-subject pronoun will rather co-refer with a non-subject antecedent even if the subject is the most salient entity following Centering theory. We therefore checked the experimental materials as to their structural parallelism. In 16 out of 18 items the anaphoric pronoun was in the same grammatical role as both antecedents, namely the subject role. In one item the second antecedent was in the object position, in another item the first antecedent. That is, experimental materials were largely congruent in that both antecedents were structurally parallel to the anaphoric pronoun. An effect of different grammatical roles between anaphor and antecedents on participants' choices can therefore be excluded. Simultaneously, we can eliminate the possibility that pure subjecthood is responsible for our results.

Situation models. According to the concept of situation models (e.g. Anderson, Garrod and Sanford, 1983; Morrow, Greenspan & Bower, 1987) discourse representations are updated with incoming new information. When the situation described in a discourse changes, such as by shifts in space and/or time, the discourse model will then represent the new episode including new discourse entities as well as relevant background knowledge. Accessibility of antecedents depends on them being part of the current discourse model. Local characters' accessibility declines after a change in episodes while main characters are supposed to be accessible as antecedents throughout the discourse (Anderson, Garrod and Sanford, 1983).

In order to control for possible effects of constant versus changing episodes in the present experimental materials items were classified into those that in any version had a situational change between Antecedent 1 and 2 and those without a change (changes actually occurred only in item version B due to the coordinate relation between the discourse unit of Antecedent 1 and that of Antecedent 2). It turned out that in seven items situational contexts changed (e.g., Antecedent 1 went to a party while Antecedent 2 went to the cinema; Antecedent 1 rented a movie, Antecedent 2 went to see friends, etc.) while in seven they stayed constant. The remaining four items were not unambiguously classifiable in this respect (e.g., Antecedent 2 was just leaving the scene or did something else as Antecedent 1 without spatial or temporal specification).

Items that were clearly classifiable as to changing or constant contexts were compared concerning potential effects on participants' choices. Results are given in Table 4.

Table 4. Participants' choices in cases of constant or changing situations in Type B

Discourse Model	Type of Item	1. Antecedent	2. Antecedent
Constant	**Type A (1. antecedent at right frontier)**	57	26
Constant	**Type B (2. antecedent at right frontier)**	18	23
Changing		22	23
Constant	**Type C (both antecedents at right frontier)**	54	34
	Total	151	106

The basic finding that when the second antecedent is the only antecedent at the right frontier (Type B), it is chosen more often than when Antecedent 1 is the only antecedent at the right frontier (Type A) is apparently independent of changes in the situational model. In Type B Antecedent 1 and 2 were opted for in about equal numbers of cases, independent of a possible change in episodes between antecedents ($\chi^2_{(1)} < 1$, $p > .2$). Therefore, our initial interpretation of the data reflecting an effect of the right frontier constraint can be maintained.

5 Discussion

The present study represents a fruitful attempt to derive testable predictions from a linguistic theory on discourse hierarchy. Results indicate that discourse relations might be one factor to affect anaphor resolution. This is in accordance to the assumptions in theoretical linguistics. In particular, we were able to present effects that can plausibly be interpreted as reflecting the Right Frontier Constraint, which plays a prominent role in current theories on anaphor resolution. Not surprisingly, this constraint does not hold unrestrictively. The results regarding unequal gender information suggest that morpho-syntactic information overrides the RFC. This is in accordance with previous studies showing that explicit gender and number marking determines the reference of anaphoric pronouns.

Open questions remain as to the exact interplay of competing influences on antecedents' accessibility. Our study considered some factors in the experimental setup, such as morpho-syntactic cues and the RFC, and some post hoc in the data analysis, such as syntagmatic distance, structural parallelism and the effect of situation models.

Obviously, further research is needed to address how the RFC and other constrains interact in detail. In particular, future work should investigate the effects of the Right Frontier Constraint on other types of anaphoric expressions such as demonstrative pronouns and full nominal phrases.

References

Anderson, A., Garrod, S. C., Sanford, A. J. (1983) The accessibility of pronominal antecedents as a function of episode shifts in narrative text. Quarterly Journal of Experimental Psychology 35A: 427-440.

Ariel, M. (2001) Accessibility theory: An overview. In: Sanders, T, Schilperoord, J., Spooren, W. (eds.) Text representation: Linguistic and psycholinguistic aspects. John Benjamins, Amsterdam: 29-87.

Asher, N. (1993) Reference to abstract objects in discourse. Kluwer, Dordrecht.

Asher, N. (2004) Discourse topic. Theoretical Linguistics 30: 163-201.

Asher, N., Lascarides, A. (2003) Logics of discourse. Cambridge University Press, Cambridge.

Asher, N., Vieu., L. (2005) Subordinating and coordinating discourse relations. Lingua 115: 591-610.

Chambers, C. G., Smyth, R. (1998) Structural parallelism and discourse coherence: A test of centering theory. Journal of Memory and Language 39: 593-608.

Garnham, A. (2001) Mental models and the interpretation of anaphora. Psychology Press Ltd., Sussex.

Gordon, P., Grosz, B., Gilliom, G. (1993) Pronouns, names, and the centering of attention in discourse. Cognitive Science 3 (17): 311-347.

Grosz, B., Sidner, C. (1986) Attentions, intentions, and the structure of discourse. Computational Linguistics 12 (3): 175-204.

Grosz, B., Joshi, A., Weinstein, S. (1995) Centering: A framework for modeling the local coherence of discourse. Computational Linguistics 21 (2): 203-226.

Hudson-d'Zmura, S., Tanenhaus, M. (1998) Assigning antecedents to ambiguous pronouns. The role of the center of attention as the default assignment. In: M. Walker, Joshi, A., Prince, E. (eds.) Centering Theory in discourse. Clarendon Press, Oxford: 199-226.

Klin, C. M., Guzmán, A. E., Weingartner, K. M., Ralano, A. S. (2006) When anaphor resolution fails: Partial encoding of anaphoric inferences. Journal of Memory and Language 54: 131-143.

Mann, W., Thompson, S. A. (1988) Rhetorical structure theory toward a functional theory of text organization. Text 8(3): 243–281.

Matthiessen, C., Thompson, S. A. (1987) The structure of discourse and subordination.. In: Thompson, S. A., Haiman, J. (eds.) Clause combining in discourse and grammar. John Benjamins, Amsterdam.

Morrow, D. G., Greenspan, S. L., Bower, G.H. (1987) Accessibility and situation models in narrative comprehension. Journal of Memory and Language 26: 165-187.

Polanyi, L. (1988) A formal model of the structure of discourse. Journal of Pragmatics 12: 601-638.

Polanyi, L. (1996) The linguistic structure of discourse. Technical Report CSLI-96-200. CSLI, Stanford University.

Prince, E. F. (1992) The ZPG letter: Subjects, definiteness, and information status. In: Mann, W. C., Thompson, S. A. (eds.) Discourse description: Diverse linguistic analyses of a fund-raising text. John Benjamins, Amsterdam/Philadelphia: 295-326.

Sanders, T., Spooren, W., Noordman, L. (1992) Toward a taxonomy of coherence relations. Discourse Processes 15: 1–35.

Webber, B. L. (1988) Discourse deixis and discourse processing. Technical report MS-CIS-86-74. Linc Lab 42. Department of Computer and Information Science, University of Pennsylvania.

Pronoun Resolution and the Influence of Syntactic and Semantic Information on Discourse Prominence

Ralph Rose

Gunma Prefectural Women's University
Faculty of International Communication
Kaminote 1395-1, Tamamura-machi, Sawa-gun, Gunma Prefecture, Japan 370-1193
rose@gpwu.ac.jp
http:/www.gpwu.ac.jp/~rose

Abstract. Beginning with the observation that syntactic and semantic information often coincide (i.e., subjects are often agents, objects often patients), this study investigates the possibility that preference to resolve a sentence-initial pronoun to a syntactically prominent antecedent might actually be better explained in terms of preference for resolving to a semantically prominent antecedent. The study takes Discourse Prominence Theory (Gordon and Hendrick [11,12]) as an underlying framework. Results of three psycholinguistic experiments using a self-paced reading task show that *both* syntactic and semantic information guide readers' pronoun resolution preferences. This suggests a revised understanding of Discourse Prominence Theory in which the prominence of discourse referents is determined through a complex process depending on multiple linguistic factors. Results further show that the relative degree of prominence among competing candidates influences resolution processes.

Keywords: pronoun resolution, Discourse Prominence Theory, repeated-name penalty.

1 Introduction

Most pronoun resolution algorithms incorporate some method (explicitly or implicitly) for ranking candidate antecedents with higher-ranking candidates judged more likely to be the intended antecedent. One factor which practically all of these ranking schemata share is some measure of the syntactic prominence of candidate antecedents. In Lappin and Leass' Resolution of Anaphora Procedure [28], for instance, candidates are assigned a certain index value based on their grammatical role (subject, object, etc.). Hobbs' algorithm [19], on the other hand, employs a hierarchical search of the syntactic representation, effectively ranking candidates according to the syntactic structure. A simpler procedure is proposed by Gernsbacher and Hargreaves [7] using linear order-of-mention.

However, in English, syntactic and semantic information are often conflated: Syntactic subjects are often semantic agents while syntactic objects are often

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 28–43, 2007.

semantic patients. Thus, it is conceivable that the contribution of syntactic prominence to the ranking of candidates is better explained as the contribution of semantic prominence. The present study examines this possibility in a series of psycholinguistic experiments designed to disambiguate the effects of syntactic and semantic prominence in pronoun resolution preferences.

This paper is laid out as follows. In Section 2, I give an overview of some theoretical issues underlying the present research. Based on many existing models, I describe a general ranking schema for candidate antecedents in pronoun resolution. In Sections 3 and 4, the experiments are described. This is followed in Section 5 with some general discussion of the findings and their implications.

2 Background

2.1 Discourse Prominence Theory

In this paper, I assume as an underlying framework *Discourse Prominence Theory* (hereafter, DPT) introduced in Gordon and Hendrick [11,12]. While cast in terms of Discourse Representation Theory (DRT) [22,23], it is intended to be a general model of discourse processing which captures observations about how readers interpret reference and coreference in a discourse as well as the time-course of processing reference and coreference. Entities introduced in the discourse are referred to as *discourse referents* within the discourse representation, following Heim [16,17] and Karttunen [24]. The cumulative representation of the discourse thus far—that is, the context—is then seen to contain two things: a list of discourse referents and a list of semantic conditions on those referents. In this paper, I will be centrally concerned with the list of referents and how it is utilized during discourse processing.

In DPT, each new utterance is processed and incorporated into the representation with respect to the current context [25,27, inter alia]. As various linguistic objects or configurations are detected by the parser, corresponding operations are triggered which may access the context in order to be completed. Here I'll discuss three crucial operations in DPT. In DRT, these operations are called *constructions rules* (hereafter, CR) because they are operations that contribute to the construction of the discourse representation. First, when a proper name is encountered, a construction rule is triggered which introduces a new discourse referent into the representation (hereafter, CR.PN). Second, when a pronoun is encountered, a construction rule is triggered to search for a suitable referent in the context and then establish coreference with it (CR.PRO). Third, when it can be concluded from the semantic conditions that two independent discourse referents in the representation refer to the same real-world entity (i.e., corefer), then an operation is triggered to establish this equivalence in the representation (CR.EQ).[1]

For instance, consider the sentences in (1)-(2). Experimental evidence [11] shows that readers find it considerably easier to establish coreference between

[1] See [12] for a formal definition of their three constructions rules: CR.PN, CR.PRO, and CR.EQ.

the name and pronoun in (1) than between the two names in (2) and process the former sentence faster than the latter. In DPT, this is readily explained. The first occurrence of the name in both sentences triggers CR.PN which introduces a new discourse referent, say x, into the representation. Then the pronoun in (1) triggers CR.PRO which easily finds a suitable referent, x, and establishes coreference with it. However, the repeated name in (2) triggers CR.PN and introduction another new discourse referent, say y. Subsequently, the semantic information showing that x and y refer to the same real-world entity (i.e., *Jane(x)*, *Jane(y)*) trigger CR.EQ which then establishes equivalence between x and y. Thus, the additional operation accounts for the increased difficulty readers have with such a configuration.

(1) Jane$_i$ thinks she$_i$ is sick.
(2) Jane$_i$ thinks Jane$_i$ is sick.

In Gordon and Hendrick's description of DPT, they also imply the necessity of a further construction rule to handle cases where a pronoun has been incorrectly assigned. For instance, consider (3).

(3) a. John sent a package to Bill.
b. He received it two days later.

In DPT, the pronoun *he* in (3b) is initially interpreted as coreferent with the discourse referent introduce by John in the preceding sentence (because it is syntactically most prominent—discussed in further detail below). However, the more natural interpretation due to plausibility constraints is that the pronoun refers to Bill. Therefore, some sort of reanalysis operation must take place in order to 'repair' the discourse representation. I assume this operation is triggered by the recognition of some sort of inconsistency in the semantic conditions though I will leave an explicit description of this to future work. In this paper, I will refer to this construction rule as CR.RA.

In the present study, I am particularly interested in exploring how CR.PRO proceeds. Gordon and Hendrick do not discuss in great detail how the discourse processor determines what is a suitable referent, though they do seem to assume that referents introduced in syntactically more prominent positions are more suitable than those introduced in less prominent positions. In the following section I will discuss a general model of how the processor determines which referent is a suitable referent.

2.2 Pronoun Resolution

Most models of pronoun resolution incorporate two primary operations toward determining a suitable referent for a pronoun: a filtering operation and a ranking operation which take place over the referents in the context. The filtering operation removes from consideration referents which are morphosyntactically incompatible with the pronoun under consideration [2,4,28]. The ranking operation orders the referents with respect to some criteria. This ranking can then be

seen to reflect the degree of likelihood that each referent is the suitable referent for the current pronoun. In DPT, this ranking is referred to as the discourse prominence of a referent. In other theories and formalisms, such terms as 'focus' [35], 'givenness' [15], 'topichood' [8], 'salience' [3], and 'centering' [14,13] describe similar or overlapping conceptualizations.

The central question surrounding the ranking operation is the actual procedure and criteria for determining this ranking. In Gordon and Hendrick's description of DPT, the discourse prominence of referents is determined entirely by syntactic information. While this might be a convenient simplification, it is surprising they do not propose a more flexible approach which depends on numerous types of information because there is much evidence that many factors (e.g., recency, parallelism, coherence relations) influence pronoun resolution preferences. Mitkov [30] provides a useful overview of these factors while Lappin and Leass [28] evaluate the relative influence of a variety of these factors in their Resolution of Anaphora Procedure. Gordon and Hendrick [12] do discuss Lappin and Leass' work, yet still seem to reject other factors, excluding them from their idea of discourse prominence.

As an adaptation of DPT, therefore, I will view the procedure for determining the discourse prominence of referents as dependent on a number of different linguistic factors in some sort of combinatorial fashion and refer to this procedure as the *multiple prominence factor method* or MPFM, for short. Exactly how the various factors in the MPFM combine I will leave to future work, but one possibility might include a simple arithmetic summation across indices determined from each prominence factor. This is the approach taken in Lappin and Leass' procedure. Another possibility might be to determine discourse prominence in a constraint-based approach with constraints derived from the various factors.

While I acknowledge that a variety of factors may play a significant role in this procedure, in this study, I look at only two factors: syntactic prominence and semantic prominence. In the next two sections I discuss these two factors in greater detail.

Syntactic Prominence. Many researchers have observed preferences for an unbound pronoun to be interpreted as coreferent with a referent previously introduced in subject position [20,29] or in an utterance-initial position [7]. For example, the preferred interpretation of the (unaccented) pronoun in (4b) is to the referent introduced as the subject of the preceding utterance, Luke.

(4) a. Luke$_i$ hit Max$_j$.
b. He$_{i/\#j}$ ran home.

The typical account of these observations is to assume that the syntactic structure of an utterance imposes a prominence hierarchy on the referents introduced in that utterance. The exact way in which the syntactic information determines prominence varies from theory to theory—for example, grammatical role (e.g., subject, object, etc.) in the centering framework of Grosz and Sidner [13,14], relative height in the syntactic tree in Hobbs' algorithm [18], or linear

order-of-mention [7]—but crucially it is the structural configuration of an utterance which determines the relative prominence of referents.

Semantic Prominence. One problem with the syntactic prominence account is that in English, at least, syntactic information and semantic information are often conflated. That is, for example, referents introduced as sentential subjects are often semantic agents and bear more proto-agent entailments (e.g., sentience, volition) [5] while those introduced as objects are often semantic patients and bear more proto-patient entailments (e.g., affectedness). Thus, an alternative account of the observation in (4) above is to assume that the semantic information imposes a hierarchy on discourse referents such that those introduced as agents are more prominent than those introduced as patients. As such, there is a preference to interpret the pronoun in (4b) as coreferent with the more semantically prominent referent, Luke.

Some researchers have looked at the influence of semantic information in referential processing in somewhat different ways. Prat-Sala and Branigan [31] observed that animate entities were preferred over inanimate entities as antecedents in pronoun resolution. In other work, Stevenson and colleagues [33,34] and Arnold [1] suggest that in forward-looking discourse planning, referents introduced in certain roles (e.g., patients in agent-patient constructions, goals in source-goal constructions) are the default focal point for reference in an immediately following utterance. They suggest, however, that in pronoun resolution (a backward-looking process), only syntactic information is relevant—that the default referent of an utterance-initial pronoun is the subject of the preceding utterance.

For the present study, I will be taking a slightly different approach. I assume that the semantic prominence of discourse referents is determined by their semantic roles (e.g., agent, patient, etc.). Referents are ranked with respect to some hierarchy of semantic roles. Exactly what these roles are and how they are ranked I will leave unspecified. One possibility might include using role sets that have been proposed in the syntax-semantics literature (e.g., [6,21,32]). For the present study, I will assume the presence of both agent and patient semantic roles and that agent is ranked higher in the hierarchy than patient.

2.3 Summary

In this paper, I investigate semantic prominence by presenting data from a series of psycholinguistic experiments designed to evaluate and compare the effects of both syntactic and semantic prominence on pronoun resolution. Because, as noted above, syntactic and semantic information are often conflated, it is necessary to find a linguistic environment that allows the influence of each to be observed. I suggest that argument-reordering constructions are a good candidate for this. Consider the contrast between the non-*tough* and *tough* constructions in (5)-(6).

(5) John$_i$ could easily hit Matt$_j$ in the boxing match.
(6) Matt$_j$ was easy for John$_i$ to hit $\emptyset_j$ in the boxing match.

If syntactic information is what determines discourse prominence, then the prediction would be that an immediately following pronoun (i.e., *he*) should preferentially be interpreted as coreferent with the subject: John in (5) and Matt in (6). However, if semantic information determines discourse prominence, then the preference should be for the agent in both cases: John. Thus, the experiments described in the next section make use of this contrast in a self-paced reading task to compare the influence of syntactic and semantic information on discourse prominence.

3 Experiment I

3.1 Design

The goal in the first experiment was to compare the influence of syntactic and semantic information in pronoun resolution preferences during on-line discourse processing. The experiment takes advantage of the non-*tough*/*tough* alternation discussed above and extended in (7).

(7) a. John$_i$ could easily hit Matt$_j$ in the boxing match. CONTROL
a'. Matt$_j$ was easy for John$_i$ to hit $\emptyset_j$ in the boxing match. SPLIT
b. He$_i$ even managed to land a knockout punch. AGENT
b'. He$_j$ became bruised and bloodied all over. PATIENT

In the non-*tough* case, (7a), syntactic and semantic information converge to promote the same referent as more discourse prominent (i.e., John). I will refer to this case, therefore, as the CONTROL condition. In the non-*tough* case, on the other hand, syntactic and semantic information diverge and promote different referents. I will therefore refer to this case as the SPLIT condition. These two sentences, respectively for each condition, determine the context in which the continuation sentence (b/b') is processed. These continuation sentences begin with a pronoun and plausible under only one interpretation of the pronoun—coreferent with either the AGENT (John) or the PATIENT (Matt) of the preceding utterance. In terms of DPT, when the pronoun is encountered, it will be automatically assigned to the most prominent referent in the context in accordance with CR.PRO. However, if at some later point the reader realizes the assignment was incorrect, then CR.RA will be triggered costing time. Therefore, in a self-paced reading experimental task, shorter reading times will indicate which referent is perceived as more discourse-prominent. This approach was used in this experiment which was a 2×2 design pitting CONTEXT (CONTROL, SPLIT) against intended REFERENT of the pronoun (AGENT, PATIENT).

3.2 Method

Participants. The participants in this experiment were 32 undergraduate students from the Northwestern University Linguistics Department subject pool who were native speakers of North American English. The participants received course credit in return for their participation.

Materials. A total of 32 stimulus items were prepared using six adjectives (*tough*, *hard*, *fun*, *easy*, *difficult*, *a cinch*) which exhibit the non-*tough*/*tough* alternation and 32 agent-patient verbs (e.g., *hit*, *catch*, *capture*). Continuation sentences were prepared so that the initial pronoun referred to either the AGENT or PATIENT. Each pair of continuation sentences was also balanced for ASCII character length. These two-sentence test sequences were then embedded in longer discourses to make a five-sentence vignette as shown in (8).

(8) a. John and Matt took part in an important boxing match.
b. It was twelve rounds long.
c. John$_i$ could easily hit Matt$_j$ in the final round. CONTROL
c'. Matt$_j$ was easy for John$_i$ to hit $\emptyset_j$ in the final round. SPLIT
d. He$_i$ even managed to land a knockout punch. AGENT
d'. He$_j$ became bruised and bloodied all over. PATIENT
e. The judges had no trouble deciding the winner.

Each vignette was followed by one comprehension question. These questions focused on different parts of the vignette in order to encourage participants to read and process the entire discourse.

The 32 items were combined with 48 items from a different experiment. The items were randomized into blocks and presentation of the items in the different conditions was balanced across participants so that adjacent stimuli were not from the same experimental condition.

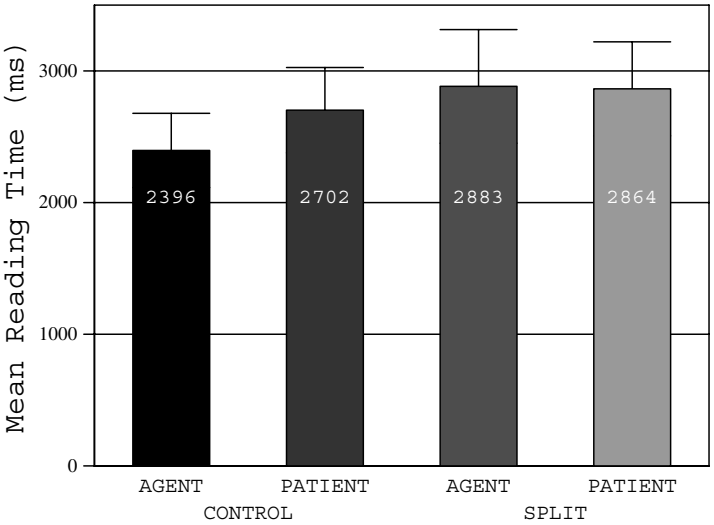
Procedures. The stimuli were presented one sentence at a time on a computer screen using Superlab by Cedrus Corporation. Participants were instructed to read each sentence and then press a button to continue to the next sentence. Participants were asked to read each vignette as quickly as possible, but also to concentrate on comprehension. The time between button presses was recorded as their reading time. In this study, only the reading times of the fourth sentences, (8d/d') are analyzed.

3.3 Results

The results of Experiment 1 are shown in Figure 1. There was a main effect of CONTEXT, no effect of REFERENT, but a marginally significant interaction between the two. In the CONTROL condition, participants read the AGENT continuation sentence faster suggesting they preferred the pronoun in the continuation sentence to be coreferent with John—the syntactically and semantically prominent entity in the context sentence. However, in the SPLIT condition, participants did not show any preference for either continuation sentence.

3.4 Discussion

The experimental results show that in the CONTROL condition, participants prefer to interpret the pronoun as coreferent with the syntactically and semantically prominent entity. This is consistent with previous experimental work described in Section 2.2 where preference is shown for a syntactically prominent



	by participants		*by items*	
CONTEXT	$F(1,31) = 12.5$	$p < 0.005$	$F(1,31) = 7.8$	$p < 0.01$
REFERENT	$F(1,31) = 2.6$	*n.s.*	$F(1,31) = 1.1$	*n.s.*
CONTEXT*REFERENT	$F(1,31) = 2.2$	$p = 0.15$	$F(1,31) = 4.3$	$p < 0.05$

Fig. 1. Mean reading times with 95% confidence intervals for the continuation sentences (8d/d') for participants ($n = 32$) in Experiment I. Two main factors were tested: CONTEXT (CONTROL, SPLIT) and intended REFERENT of pronoun (AGENT, PATIENT).

entity. The current experiment thus replicates those results. However, the results in the SPLIT condition are quite interesting: Participants showed no preference for either referent.

One explanation for these results is that *both* syntactic and semantic prominence influence the ranking of candidate antecedents in an independent fashion. When syntactic and semantic prominence coincide to promote one antecedent (as in the CONTROL condition), then pronoun resolution processes can select one candidate over others. However, when syntactic and semantic prominence diverge, promoting different entities, then pronoun resolution processes do not show any preference. This could be accounted for in the MPFM in different ways: if the method uses a simple summation across prominence factors to calculate the discourse prominence of referents, then in the CONTROL condition, the syntactically and semantically prominent referent is doubly boosted and has a clearly higher total discourse prominence index than other referents. Then, in terms of DPT, the search for a suitable referent is concluded successfully and the pronoun is subsequently resolved to this referent, the AGENT. With the AGENT continuation, then, nothing more happens and the correct discourse representation is achieved. However with the PATIENT continuation, semantic

information introduced later in the sentence results in an inconsistency which triggers CR.RA, leading to the increased reading times as observed.

In the SPLIT condition, however, the two different referents receive comparably-sized boosts from the different factors, respectively, such that their net discourse prominence values are essentially equal. In terms of the DPT, this would seem to be a case in which the search for a suitable referent might be unsuccessful because there is more than one such referent. DPT allows that when a search is unsuccessful, a new discourse referent is introduced. Later information, though, shows that the pronoun is coreferent with an existing referent, so CR.RA is triggered to establish equivalence between the new referent and the intended referent. In the SPLIT condition, this sequence of operations appears to have happened for both the AGENT and PATIENT continuations yielding comparable reading times in both.

Thus, the experimental results can be captured in DPT, but only with a richer conception of how suitable referents are determined—one that is based on multiple prominence factors.

4 Experiments IIa-b

One criticism that may be made of the first experiment is that reading time measurements are being compared across different sentences. While the length of the continuation sentences was controlled, the lexical items and syntactic structure and complexity were not. This could be one source of variation. One way to overcome this problem is to take advantage of the *repeated-name penalty* experimental technique described in Gordon, et al. [10]. They observed that readers take longer to read sentences containing reference to a currently focused entity when the reference is by name (e.g., *John* as in (9b)) rather than by pronoun (e.g., *he* as in (9b')).

(9) a. John walked to the supermarket.
b. John bought two fish.
b'. He bought two fish.

In DPT, this is explained in the same way as the c-commanding case discussed in Section 2.1. After the context sentence in (9a), John is the most discourse-prominent referent. Thus, when the pronoun in the continuation sentence in (9b') triggers CR.PRO, John will be judged the most suitable referent and coreference will be readily established. However, the proper name in (9b) will merely trigger CR.PN and then the introduction of a new discourse referent *different from* the existing referent of John in the context. Subsequent information indicating that these two referents point to the same entity in the real world will then trigger CR.EQ to establish equivalence between these referents. The additional operations necessary to establish coreference are presumed to lead to increased reading times and hence, the repeated-name penalty.

4.1 Design

In the present study, the repeated-name penalty experimental paradigm is a useful way to look more closely at how participants perceive the relative discourse prominence of referents in the context by comparing the repeated-name penalties across the various experimental conditions. Thus, the difference in reading times between the pronoun and name versions of (10b) can be compared to that of (10b') in both the CONTROL and SPLIT conditions. Based on the results of Experiment I, the prediction is that in the CONTROL condition, there should be a larger repeated-name penalty for the AGENT than for the PATIENT continuation sentence, but in the SPLIT condition, the repeated-name penalties should be approximately the same.

(10) a. John$_i$ could easily hit Matt$_j$ in the boxing match. CONTROL
a'. Matt$_j$ was easy for John$_i$ to hit $\emptyset_j$ in the boxing match. SPLIT
b. [John$_i$ / He$_i$] even managed to land a knockout punch. AGENT
b'. [Matt$_j$ / He$_j$] became bruised and bloodied all over. PATIENT

In order to test these predictions, two further experiments were thus performed, one looking at the CONTROL condition and the other looking at the SPLIT condition. Both experiments were a 2×2 design pitting intended REFERENT (AGENT, PATIENT) against FORM of reference (PRONOUN, NAME).

4.2 Method

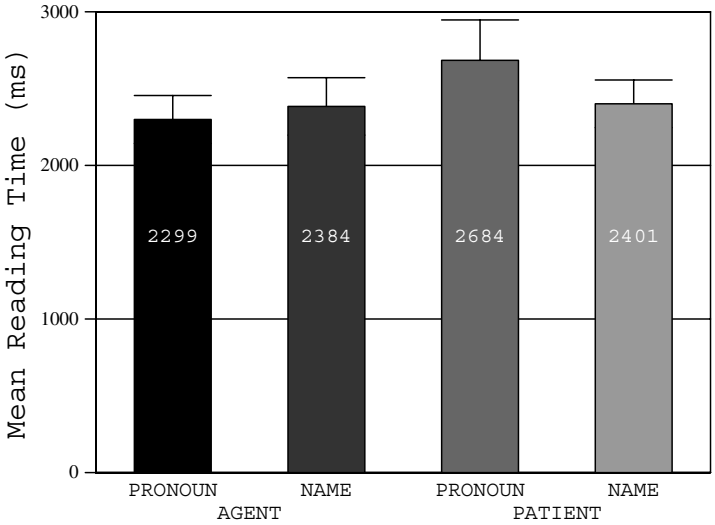
Participants. 32 undergraduate students from the Northwestern University Linguistics Department subject pool who were native speakers of North American English participated in each of the two experiments reported here. None of these students had participated in Experiment I. The participants received course credit in return for their participation.

Materials. The materials for this experiment were the same as those used in Experiment I except that two versions of the continuation sentences (i.e., (8d/d'))—one beginning with a pronoun and one with a repeated name—were used. Experiment IIa used stimuli only in the CONTROL condition while Experiment IIb used stimuli only in the SPLIT condition.

Procedures. The procedures for these two experiments were exactly the same as those reported above for Experiment I.

4.3 Results

The results of Experiment IIa are shown in Figure 2. In this experiment—the CONTROL condition from Experiment 1—there was a marginal main effect of REFERENT, no effect of FORM, but a significant interaction between the factors. These results appear to be driven by an 83ms repeated-name penalty with the AGENT continuation and a 270ms repeated-name *advantage* (i.e., a



	by participants	*by items*
REFERENT	$F(1,31) = 7.9$ $p < 0.01$	$F(1,31) = 3.6$ $p = 0.067$
FORM	$F(1,31) = 1.6$ *n.s.*	$F(1,31) = 1.5$ *n.s.*
REFERENT*FORM	$F(1,31) = 6.3$ $p < 0.05$	$F(1,31) = 8.5$ $p < 0.01$

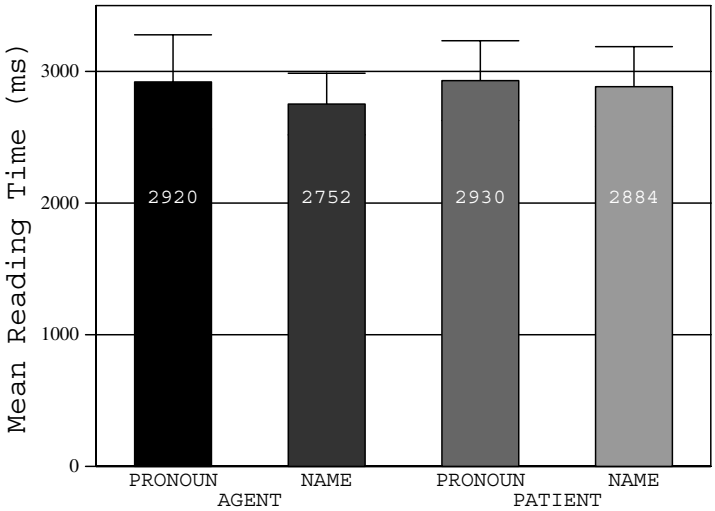
Fig. 2. Mean reading times with 95% confidence intervals for the continuation sentences (8d/d') for participants ($n = 32$) in Experiment IIa—the CONTROL condition from Experiment I. Two main factors were tested: intended REFERENT (AGENT, PATIENT) and referential FORM (PRONOUN, NAME).

Table 1. Experiment IIa: CONTROL Condition Repeated-Name Penalties

	penalty	*by participants*	*by items*
AGENT	83ms	$t(31) < 1.0$ *n.s.*	$t(31) < 1.0$ *n.s.*
PATIENT	-270ms	$t(31) = 2.5$ *n.s.*	$t(31) = 2.7$ $p = 0.07$

negative penalty) with the PATIENT continuation as shown in Table 1. Post-hoc t-tests, however, do not show that either of these penalties is significantly different from a null hypothesis of 0ms.

The main effect of REFERENT suggests that on the whole, participants prefer that the continuation contain reference (regardless of form: name or pronoun) to the most discourse-prominent entity. This is consistent with many theories of forward-looking discourse construction [33,34,13]. The significant interaction between REFERENT and FORM indicates that the AGENT continuation exhibited a significantly larger repeated-name penalty than the PATIENT



	by participants	*by items*
REFERENT	$F(1,31) < 1.0$ *n.s.*	$F(1,31) < 1.0$ *n.s.*
FORM	$F(1,31) < 1.0$ *n.s.*	$F(1,31) = 1.9$ *n.s.*
REFERENT*FORM	$F(1,31) < 1.0$ *n.s.*	$F(1,31) < 1.0$ *n.s.*

Fig. 3. Mean reading times with 95% confidence intervals for the continuation sentences (8d/d') for participants ($n = 32$) in Experiment IIb—the SPLIT condition from Experiment I. Two main factors were tested: intended REFERENT (AGENT, PATIENT) and referential FORM (PRONOUN, NAME).

Table 2. Experiment IIb: SPLIT Condition Repeated-Name Penalties

	penalty	*by participants*	*by items*
AGENT	-168ms	$t(31) < 1.0$ *n.s.*	$t(31) = 1.4$ *n.s.*
PATIENT	-46ms	$t(31) < 1.0$ *n.s.*	$t(31) < 1.0$ *n.s.*

continuation. What is interesting, though, is that—although these numbers are not statistically strong—it seems that the AGENT continuation incurs no repeated-name penalty, while the PATIENT continuation incurs a repeated-name advantage. Some implications of this will be discussed in the discussion section below.

The results of Experiment IIb using the SPLIT condition stimuli are shown in Figure 3. In contrast to Experiment IIa, there were no significant main effects and no significant interaction. There was a 168ms repeated-name advantage in the AGENT condition and a 46ms repeated-name advantage in the PATIENT condition as shown in Table 2. However, neither of these was significant.

In short, the results of Experiment IIb are basically flat-lined with participants showing no apparent preferences for any continuation across the board.

4.4 Discussion

Taken alone, the results of Experiment IIb are probably unremarkable, but taken together with the results of Experiment IIa, they reinforce the conclusion that both syntactic and semantic prominence influence the ranking of candidates for pronoun resolution: When syntactic and semantic prominence converge, then pronoun resolution prefers the promoted candidate, but when syntactic and semantic diverge, then pronoun resolution shows no preference. This can be captured in the DPT as a part of the process of determining a suitable referent for a pronoun: This process takes advantage of a ranking method which depends on a number of different factors such as the MPFM described above.

A secondary implication of the results of Experiments IIa-b is that the repeated-name penalty must been seen in a new light. Ultimately, this comes down to how the search for a suitable referent proceeds. In the original experiments which established the repeated-name penalty concept [9,10], most of the stimuli had contexts in which there was little or no chance of ambiguity because of parallelism effects, topicalization, or gender-disambiguation. If there is only one compatible referent, then the search for a suitable referent will be relatively straightforward and resolution should be quick. Similarly, if we assume that topicalizing a referent makes it very highly discourse prominent, then the search for a suitable referent may still be very easy because any competing candidates will be so low in the prominence hierarchy. Thus, in both of these cases it is not surprising that the contrasting case with repeated-name reference would take much longer because of the subsequent triggering of CR.EQ.

The present experiments indicate that the search for a suitable referent may actually be more costly when there is more than one compatible referent in the context. In the CONTROL condition of the experiment, the AGENT referent is more discourse prominent, but apparently not so much so that it is immediately deemed the most suitable referent (as it might if it were topicalized). Therefore, with the AGENT continuation, the processor must take roughly as much time in the PRONOUN condition as it must take in the NAME condition to establish equivalence among the discourse referents: In the latter condition, CR.PN and CR.EQ are triggered while in the former condition, only CR.PRO is triggered. Yet the same net time is taken in each case. This seems to be best explained by seeing the search for a suitable referent as being a more costly procedure when there are other compatible referents. With the PATIENT continuation, the same difficulty is faced by the processor except that in the PRONOUN condition, CR.RA is also triggered because the pronoun had been assigned by default to the more discourse-prominent AGENT. This leads to a large delay in this condition—thus an apparent advantage for the repeated name continuation.

In the SPLIT condition however, the search for a suitable referent is immediately concluded because no single suitable referent can be found—the two potential referents are equally ranked. Thus CR.PRO introduces a new discourse referent and later, CR.RA is triggered to establish coreference. In the NAME conditions CR.PN introduces a new discourse referent and later, CR.EQ is triggered. Thus, in all the conditions, the same net costs are incurred: those caused by

introducing a new referent and subsequently by establishing equivalence among referents.

In short, while DPT accounts nicely for the results of these experiments, it is necessary to bring in a more sophisticated conceptualization of how the process of finding a suitable referent proceeds. It is not—as originally suggested by Gordon and Hendrick [11,12]—as simple as selecting the referent realized in the most syntactically prominent position. Rather, there is at least one other factor (perhaps many) that determine discourse prominence; namely, semantic prominence. Furthermore, the relative discourse prominence of referents influences the speed with which the search process may be concluded.

5 General Discussion

In short, the experimental evidence here, combined with evidence from the Stevenson, et al. [34] and Arnold [1] studies described briefly in Section 2.2, suggest that semantic information affects both forward-looking and backward-looking referential processes in discourse. The results of those studies, however, show some interesting contrasts with the present study. For instance, one contrast is that while the present experiment shows agent-preference for pronoun reference, [34] shows a default patient-preference for topic continuation. This is not necessarily a contradiction. While it would be theoretically convenient if the same ranking scheme affected both forward-looking and backward-looking referential processes, this does not have to be the case. Further work is clearly necessary to to understand just how these processes are related to one another.

The results of this study point toward two other areas for further study. First, while there is much work looking at ambiguous pronoun resolution, much of this work seems to be limited to cases where one candidate outranks other candidates. The present study suggests that there are cases where ranking produces ties. This is not a new notion, however. There are many models which suggest that discourse entities are only partially-ordered in their prominence (e.g., the list of forward-looking centers in classical centering theory; Grosz, et al. [13]). Yet how pronoun resolution processes actually deal with cases of equally-ranked candidates seems to be much less studied.

A second area for further research concerns the ranking scheme. The evidence here strongly suggests, as noted previously, that ranking is based on a number of factors as in the MPFM. This is not a new concept, of course, and many pronoun resolution algorithms have achieved a fairly high degree of success with such methods (e.g., [26,28]). However, there is more work to be done on the way the ranking is actually utilized by the processor. The experimental evidence in this study suggests that the relative ranking of referents on the discourse prominence hierarchy affects how those referents are accessed during pronoun resolution processes. A referent which is ranked much higher than any other referent seems to block, in a sense, consideration of those other referents. While on the other hand, as referents are more closely ranked in the hierarchy, more

time is required to consider them. Yet when they become too closely ranked, then the search for a suitable referent fails.

6 Conclusion

In conclusion, the series of experiments presented here suggest that both syntactic and semantic prominence contribute to the ranking of candidates for pronoun resolution in a way that may result in a partially-ordered ranking. Furthermore, *tough*-constructions seem to be a useful construction for generating such-partially ordered rankings and therefore may prove a useful means for studying how pronoun resolution processes deal with equally-ranked candidates. DPT provides a useful framework in which to capture the time-course of discourse comprehension and pronoun resolution, but only with a more complex conceptualization of how the discourse prominence of referents is determined and how the processor makes use of the ranked list of referents.

References

1. Arnold, J.: The effects of thematic roles on pronoun use and frequency of reference. Discourse Processes **31** (2001) 137-162
2. Arnold, J., Eisenband, J., Brown-Schmidt, S., Trueswell, J.: The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. Cognition **76** (2000) B13-B26
3. Arnold, J.: Marking salience: The similarity of topic and focus. Unpublished Manuscript (1998)
4. Boland, J., Acker, M., Wagner, L.: The use of gender features in the resolution of pronominal anaphora. Cognitive Science Technical Report 17, The Ohio State University Center for Cognitive Science (1998)
5. Dowty, D.: Thematic proto-roles and argument selection. Language **67** (1991) 547-619
6. Fillmore, C.: The Case for Case. In Bach, E., Harms, R., ed.: Universals in Linguistic Theory. Holt, Rhinehart, and Winston, New York (1968) 1-90
7. Gernsbacher, M.A., Hargreaves, D.: Accessing sentence participants: The advantage of first mention. Journal of Memory and Language **27** (1988) 699-717
8. Givón, T.: Topic continuity in discourse: An introduction. In Givón, T., ed.: Topic Continuity in Discourse: A Quantitative Cross-Language Study. John Benjamins, Amsterdam (1983) 1-42
9. Gordon, P., Chan, D.: Pronouns, Passives, and Discourse Coherence. Journal of Memory and Language **34** (1995) 216-231
10. Gordon, P., Grosz, B., Gilliom, L.: Pronouns, names, and the centering of attention in discourse. Cognitive Science **17** (1993) 311-347
11. Gordon, P., Hendrick, R.: Intuitive knowledge of linguistic co-reference. Cognition **62** (1997) 325-370
12. Gordon, P., Hendrick, R.: The representation and processing of coreference in discourse. Cognitive Science **22** (1998) 389-424
13. Grosz, B., Joshi, A., Weinstein, S.: Centering: A framework for modeling the local coherence of discourse. Computational Linguistics **21** (1995) 203-225

14. Grosz, B., Sidner, C.: Attention, intentions, and the structure of discourse. Computational Linguistics **12** (1986) 175-204
15. Gundel, J., Hedberg, N., Zacharski, R.: Cognitive status and the form of referring expressions. Language **69** (1993) 274-307
16. Heim, I.: The Semantics of Definite and Indefinite Noun Phrases. PhD thesis, University of Massachusetts, Amherst (1982)
17. Heim, I.: File change semantics and the familiarity theory of definiteness. In Bauerle, R., Schwarze, C., Von Stechow, A., eds.: Meaning, Use, and Interpretation. DeGruyter, Berlin (1983) 164-189
18. Hobbs, J.: Resolving pronoun references. Lingua **44** (1978) 311-338
19. Hobbs, J.: Coherence and coreference. Cognitive Science **3** (1979) 67-90
20. Hudson-D'Zmura, S., Tanenhaus, M.: Assigning antecedents to ambiguous pronouns: The role of the center of attention as the default assignment. In Walker, M., Joshi, A., Prince, E., eds.: Centering Theory in Discourse. Clarendon Press, Oxford (1997) 199-226
21. Jackendoff, R.: Semantic Interpretation in Generative Grammar. MIT Press, Cambridge, MA (1972)
22. Kamp, H.: A theory of truth and semantic representation. In Groenendijk, J., Janssen, T., Stokhof, M., eds.: Formal Methods in the Study of Language. Mathematisch Centrum, Amsterdam (1981) 277-322
23. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic, Dordrecht (1993)
24. Karttunen, L.: Discourse referents. In McCawley, J., ed.: Syntax and Semantics, Vol. 7: Notes from the Linguistic Underground. Academic Press, New York (1976) 363-385
25. Kehler, A.: Coherence, Reference, and the Theory of Grammar. CSLI Publications, Stanford University, CA (2002)
26. Kennedy, C., Boguraev, B.: Anaphora for everyone: Pronominal anaphora resolution without a parser. Proceedings of the 16th International Conference on Computational Linguistics (COLING '96) (1996) 113-118
27. Kintsch, W., van Dijk, T.: Toward a model of text comprehension and production. Psychological Review **85** (1978) 363-394
28. Lappin, S., Leass, H.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20** (1994) 535-561
29. Mathews, A., Chodorow, M.: Pronoun resolution in two-clause sentences: Effects of ambiguity, antecedent location, and depth of embedding. Journal of Memory and Language **27** (1988) 245-260
30. Mitkov, R.: Anaphora Resolution. Longman, London (2002)
31. Prat-Sala, M., Branigan, H.: Discourse constraints on syntactic processing in language production: A cross-linguistic study in English and Spanish. Journal of Memory and Language **42** (1999) 168-182
32. Speas, M.: Phrase Structure in Natural Language. Kluwer Academic Publishers, Dordrecht (1990)
33. Stevenson, R., Crawley, R., Kleinman, D.: Thematic roles, focus and the representation of events. Language and Cognitive Processes **9** (1994) 519-548
34. Stevenson, R., Knott, A., Oberlander, J., McDonald, S.: Interpreting pronouns and connectives: Interactions among focusing, thematic roles and coherence relations. Language and Cognitive Processes **15**(3) (2000) 225-262
35. Ward, G.: The Semantics and Pragmatics of Preposing. PhD thesis, University of Pennsylvania (1985)

Anaphora Resolution as Equality by Default

Ariel Cohen

Ben-Gurion University
Beer Sheva 84105, Israel
arikc@bgu.ac.il
http://www.bgu.ac.il/~arikc

Abstract. The resolution of anaphora is dependent on a number of factors discussed in the literature: syntactic parallelism, topicality, etc. A system that attempts to resolve anaphora will have to represent many of these factors, and deal with their interaction. In addition, there must be a principle that simply says that the system needs to look for an antecedent. Without such a principle, if none of the factors recommend a clear winner, the system will be left without an antecedent. This principle should work in such a way that, if there is exactly one good candidate antecedent, the system will choose it; if there are more than one, the system will still attempt to identify one, or, at least, draw some inferences about the likely antecedent; and, in case there is no candidate, the system will produce an accommodated or deictic reading.

Many systems embody some version of this principle procedurally, as part of the workings of their algorithm. However, because it is not explicitly formalized, it is hard to draw firm conclusions about what the system would do in any given case. In this paper I define a general principle of Equality by Default, formalize it in Default Logic, and demonstrate that it produces the desired behavior. Since all other factors can also be formalized in Default Logic, the principle does not need to be left implicit in the algorithm, and can be integrated seamlessly into the rest of the explicit rules affecting anaphora resolution.

1 The Search for an Antecedent

Identifying the antecedent of an anaphoric trigger (a pronoun, definite DP, etc.) depends on the interaction of many factors: syntactic (e.g. Binding Theory), semantic (e.g. selectional restrictions), and pragmatic (e.g. Centering Theory). Some of these factors, such as selectional restrictions and syntactic binding requirements rule out certain antecedents, while other factors, e.g. topicality, suggest that a certain antecedent should be chosen.

Most, perhaps all of these factors are defeasible. Consider, for example, the following discourse, from [1]:

(1) The Vice-President entered the President's office. *He* was nervous and clutching his briefcase. After all, he couldn't fire the Vice-President without making trouble for himself with the chairman of the board.

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 44–58, 2007.

The pronoun in the second sentence has two potential antecedents: *the Vice-President* or *the President*. Clearly *the Vice-President* is preferred: it has the same syntactic position (subject) as the pronoun, and it is more salient. However, by the time the third sentence is processed, it is clear that this choice is wrong, and the intended antecedent is, in fact, *the President*.

Even what appear to be inviolable constraints, such as number agreement, can sometimes be overruled. For example, [2] note that

> numeric agreement in this corpus of Wall Street Journal articles is a defeasible constraint, because it includes so many mentions of organizations. An organization, such as "Wellington Industries" appears syntactically to be plural, but can be re-mentioned with the pronoun *it*.

Such examples abound; and they indicate that all anaphora resolution factors, or almost all of them, are best thought of as defaults, which may be overridden. It is therefore attractive to model anaphora resolution as a system of prioritized defaults (e.g. [2,3,4,5]).

Most such systems do not encode the constraints explicitly, but rather procedurally, as part of the algorithm. There are, however, strong arguments for having a declarative, explicit definition of the constraints, as argued by [2]. They implement a system of constraints for anaphora resolution proposed by [6], formulated in Optimality Theory [7]. They point out that a program that uses an explicit definition of constraints is easy to test, debug, and revise. It is also much easier to modify, say in order to apply it to another genre or another language. If constraints need to be added, removed, or the priorities between them changed, this can be done quickly, reliably, and transparently.

In this paper I am not going to consider the question of identifying these factors or their relative strengths. What I do wish to argue is that formalizing all these factors is not enough, and an additional rule is necessary; I will propose a formalization of this rule in Default Logic [8].

The rest of the paper is organized as follows. The next section contains a discussion of the additional rule: Don't Overlook Anaphoric Possibilities—DOAP [9]. The following two sections provide necessary background for the formalization of DOAP: section 3 discusses DRT as an underspecified representation for anaphora, and the significance of treating anaphoric relations as equality. Section 4 presents a brief overview of Default Logic, to be applied to equality in section 5, where DOAP is formalized as Equality by Default. Section 6 discusses the inferences that can be drawn using this relation, and contains examples demonstrating that they obey the desired patterns. The final section concludes the paper and points out potential additional applications of the theory.

2 Don't Overlook Anaphoric Possibilities

Consider the discourse in (1) again.

The antecedent that is eventually chosen, *the President*, is not suggested by any of the well known factors discussed in the literature: it is neither topical, nor

a subject, nor does it have the same syntactic position as the pronoun, etc. This antecedent is simply chosen as a last resort, since the other potential candidate is ruled out. This "last resort" rule must be defined somehow, for, without it, no antecedent would be chosen. Indeed, in the linguistics literature, such a rule has been proposed [9, p. 603]:

(2) *Don't Overlook Anaphoric Possibilities* (DOAP)
Opportunities to anaphorize text must be seized.

Essentially, this rule says simply that, when we encounter a trigger , we must try to find an antecedent. If we find an antecedent that is suggested by some rule, so much the better; but even a dispreferred antecedent is better than no antecedent at all. DOAP has been used by [10], who propose an Optimality Theoretic system of prioritized defaults for anaphora resolution.

However, while factors such as syntactic parallelism or selectional restrictions are, at least conceptually, easy to implement, it is not clear how to formalize DOAP in such a way that it could be implemented. This paper is an attempt to provide such a formalization, which, in combination with other factors, has the potential to bring about a fully explicit system of anaphora resolution.

Of course, in practice almost all anaphora resolution algorithm obey DOAP, in the sense that they always attempt to find (at least) one antecedent, even if the anaphora is ambiguous. However, if DOAP is not defined explicitly in the object level of the logic, but is left to a metalevel description, it is hard to be clear on, let alone prove, what the system will do when there is no clear choice of antecedent: which, if any, antecedent it will choose, and which inferences it will draw. Hence, formalization of DOAP on a par with all other factors is a desirable goal.

Take, for example, systems that use model building techniques. Such systems typically generate minimal models. Minimality could be seen as an implementation of DOAP: A model in which the antecedent of a referring expression is not identified is not minimal (since it has an additional entity, namely the reference of the trigger); it is therefore dispreferred, and the anaphoric reading is chosen, if possible.

However, for many of these systems, the model cannot always be relied upon to be minimal [11]. Even where it can, minimality of the model is not sufficient to ensure that an antecedent is chosen.

Consider, for example, the following discourse:

(3) John met Mary. *He* didn't talk to *her*.

A model builder would generate a model whose universe consists of John and Mary, and where the denotation of the predicate *talk to* is the empty set. This model satisfies the discourse in (3), and is clearly minimal, yet it says nothing about which antecedents the pronouns refer to.

An explicit formalization of DOAP should be able to deal with cases where there is one clear antecedent, as well as with cases where there isn't. In general, when an anaphoric trigger is encountered, there are three possibilities.

One possibility is that there is exactly one appropriate antecedent:

(4) John was eating ice cream. *He* was upset.

In this case, John is the only appropriate antecedent, and we would want to resolve the anaphora by equating the pronoun with John.

The second possibility is that there is no appropriate antecedent in the text:

(5) John was eating ice cream. *The waitress* brought him the check.

The text provides no appropriate antecedent for the definite description, so one must be accommodated. If the anaphoric trigger is a pronoun, whose informational content is minimal, accommodation may be impossible [12]. In this case, the pronoun will be interpreted deictically:

(6) John was eating ice cream. *She* brought him the check.

In (6) we interpret the pronoun as referring to some individual that is not introduced in the discourse, and is, perhaps, identified by pointing.

The third possibility is that there is more than one good candidate antecedent:

(7) John and Bill met at the ice cream parlor. *He* was upset.

There are few reasons to prefer either John or Bill as the antecedent of the pronoun. In this case, we have two choices: we can decide on some antecedent, perhaps at random, perhaps using some criterion such as recency; alternatively, we can acknowledge that the anaphora is genuinely ambiguous. Even if we take the latter course of action, all is not lost: although we do not know who the pronoun refers to, we can still draw some conclusions about him. For example, we know that, whoever *he* is, he was at the ice cream parlor.

3 An Underspecified Representation for Anaphora

Before formalizing DOAP, we need to say something about how the relation between trigger and antecedent is represented. Consider a simple case of ambiguous anaphoric reference:

(8) John shook hands with Bill and Mary. *He* hung out with *her* the whole evening.

What can we say about the resolution of the anaphora? The pronoun *her* probably refers to Mary, and the pronoun *he* is ambiguous between John and Bill, but probably refers to John. And, in the right context and/or intonation, either pronoun (or both) may be used deictically, referring to some other individual that is not denoted by a linguistic antecedent. What we would like is a system that allows us to represent all these options, pick those we consider plausible, and draw some inferences even in the absence of a clear resolution.

As the discourse in (8) exemplifies, anaphora is often ambiguous. Moreover, it is always possible, in principle, that what we had identified as the antecedent of a trigger actually is not, and we need to get an accommodated or deictic reading. In the case of (8), since we have two pronouns, one with three possible interpretations (John, Bill, or the deictic use) and the other with two (Mary or deictic), we will have six potential interpretations. We need to be able to represent the ambiguity, but still draw inferences as best we can on the basis of what we know. This calls for some sort of underspecified representation, and some inference mechanism to derive conclusions from it.

Many special formalisms have been proposed, whose sole purpose is to allow efficient representation of and reasoning with underspecification. I will not, however, go down this road, for several reasons. A formalism that is not independently motivated on linguistic grounds, and whose sole justification is to represent underspecification, may work in a practical system, but its explanatory adequacy from a linguistic point of view would be dubious.

To give one example, recall that deictic readings of a pronoun are always (given the right intonation and/or context) possible, and this is the case across languages. Why is this? Why don't we have languages where pronouns are restricted to linguistic antecedents only, and deictic readings are indicated only by, say, demonstratives? A formalism that is only geared toward underspecification would be quite adequate even if pronouns could only refer to linguistic antecedents, and it is hard to see why it would necessitate the availability of deictic readings. It is, of course, preferable to have the possibility of deictic readings follow directly from the representation, thus explaining the puzzle.

Furthermore, a nonstandard representation will typically require nonstandard inference methods, especially tailored for the representation.[1] Again, these inference methods would not be independently justified, unlike rules of common-sense inference that must, in one way or another, be used in order to understand natural language.

An additional reason for keeping the representation as simple and as close to standard linguistic representations as possible is the fact that it is not likely to be replaced by a fully specified representation during the interpretation process. Normally, one uses an underspecified representation in the hope that, in the fullness of time, or as the need arises, it will be fully specified. In this sense, an underspecified representation is only a "temporary measure." Unlike a fully specified representation, it is not really a description of the world (which has a truth value), but rather a description of readings. However, as examples like (1) demonstrate, we may choose some antecedent, only to find later on that it is inappropriate. Even if there is only one candidate antecedent, it is possible that it will later be ruled out, leaving us with an accommodated or deictic reading. Hence, the representation of anaphora cannot be thought of as a temporary measure, to be discarded once the ambiguity is resolved. The underspecified representation cannot therefore be *ad hoc*, and must be fully motivated.

[1] Though see [5], who uses a nonstandard representation of anaphora, but applies Default Logic to generate its perceived readings.

I suggest that we don't need to look far for a representation and its associated inference method. A standard, linguistically motivated representation, without special machinery for underspecification, will do: Discourse Representation Theory [13].[2]

Using this theory, the discourse in (8) will be represented by the following (simplified) DRS:

(9)

$x\ y\ z\ u\ v$
John(x)
Bill(y)
Mary(z)
shake-hands(x,y)
shake-hands(x,z)
male(u)
female(v)
hang-out(u,v)

Note that this DRS does not resolve the anaphora. In this representation, u and v are subject to existential closure, and all we know is that *some* antecedents exist. So, in effect, the DRS (9) is an underspecified representation, containing all the possible ways of resolving the anaphora.

The relation between anaphoric trigger and antecedent is represented in DRT as an equality relation. Thus, any specific resolution of the anaphora results in the addition of equalities identifying the referents of the pronouns. For example, if we identify *he* with John and *her* with Mary, we get the following DRS:

(10)

$x\ y\ z\ u\ v$
John(x)
Bill(y)
Mary(z)
shake-hands(x,y)
shake-hands(x,z)
male(u)
female(v)
hang-out(u,v)
u=x
v=z

While equalities such as the ones above are often treated as a mere notational convenience, it is clear from the formal definitions of [13] that they are *real* equalities, in the strictest logical sense. This means that we can apply the full power of the equality axioms, and get various desirable results for free. For

[2] Of course, it may be the case that some sort of special underspecified representation is needed for other reasons, e.g., to represent scope ambiguities. All I claim is that such special representations are not necessitated by the need to represent anaphora.

example, if the antecedent has a certain property, then it immediately follows that the trigger has this property too.

In this paper I propose a simple formalization of DOAP using Default Logic [8]. The idea is that a trigger and a potential antecedent are equated by default, unless this is prevented by some rule. This default rule is assigned low priority, so that other factors affecting anaphora resolution can rule out inadmissible antecedents, or suggest an antecedent before DOAP applies. The result it that this principle would apply only if there is no strong preference for any antecedent; but when it does apply, the behavior of the resulting system complies with the desiderata described above.

4 Default Logic

The relation between trigger and antecedent is equality, so the problem of anaphora resolution becomes the problem of inferring the necessary equalities from the representation. As discussed above, this inference must be defeasible, so some form of nonmonotonic reasoning is necessary to formalize it.

One could, following [10], use Optimality Theory to state DOAP, but this would be problematic. While Optimality Theory is suitable for expressing defeasible, prioritized constraints, it does not employ a formal language; constraints in Optimality Theory are typically expressed in natural language, and may consequently be underspecified or vague—indeed, [10] use nothing more precise than the natural language definition of DOAP in (2). Since the goal of the current paper is a formal system, which could be implemented, and about which statements could actually be proved, this is not good enough. I will, instead, use a formal system with well defined syntax and semantics—Default Logic [8].[3]

Default Logic is one of the most widely used nonmonotonic formalisms. A substantial body of theoretical work has been devoted to it, and a number of theorem provers have been implemented.

A *default theory* is a pair (D, A), where D is a set of defaults and A is a set of first-order sentences. Defaults are expressions of the form

$$\text{(11)} \qquad \frac{\alpha(x) : \beta_1(x), \ldots, \beta_m(x)}{\gamma(x)},$$

where $\alpha(x)$, $\beta_1(x), \ldots, \beta_m(x)$,and $\gamma(x)$ are formulas of first-order logic whose free variables are among $x = x_1, \ldots, x_n$. Note that the presence of $\alpha(x)$ is optional.

The intuitive meaning of a default is as follows. For every n-tuple of objects $t = t_1, \ldots, t_n$, if $\alpha(t)$ is believed, and the $\beta_i(t)$s are consistent with one's beliefs, then one is permitted to deduce $\gamma(t)$.

For example, the following rule says that if something is a bird, and you don't know anything to the contrary, you may believe that it flies:

$$\text{(12)} \qquad \frac{\mathbf{bird}(x) : \mathbf{fly}(x)}{\mathbf{fly}(x)}$$

[3] See [14] on implementing Optimality Theory in Default Logic.

Crucial to the interpretation of Default Logic is the notion of an *extension*. Roughly speaking, an extension of a default theory is a set of statements containing all the logical entailments of the theory, plus as many of the default inferences as can be consistently believed. A default theory may have more than one extension, as in the well known *Nixon diamond*. Suppose we have the following two defaults:

1. $\dfrac{\mathbf{Quaker}(x) : \mathbf{pacifist}(x)}{\mathbf{pacifist}(x)}$
2. $\dfrac{\mathbf{Republican}(x) : \neg\mathbf{pacifist}(x)}{\neg\mathbf{pacifist}(x)}$

The first rule says that Quakers are pacifist by default, and the second rule says that, by default, Republicans are not pacifist. If Nixon is both a Quaker and a Republican, in one extension he will be a pacifist, and in another he won't be. So, is Nixon a pacifist or isn't he?

When faced with multiple extensions, there are two general strategies we can use to decide which conclusions to accept: skeptical or credulous reasoning. Skeptical reasoning means accepting only what is true in all extensions. So, we will believe neither that Nixon is a pacifist, nor that he is not a pacifist. Credulous reasoning means picking one extension, based on whatever principles one deems appropriate, and accepting its conclusions. This means we will pick one extension, perhaps using our knowledge of Nixon's statements and actions, and based on this extension, conclude whether he is a pacifist or not.

A useful feature of some formalizations of Default Logic (e.g [15]) is the possibility of assigning priorities to defaults. Intuitively, this means that if default d_1 outranks default d_2, then it applies first, in the sense that there is no extension of the default theory that contains the conclusion of d_2 but not the conclusion of d_1, if both are applicable. While ranking is a very useful device, and we will use it too, it is important to emphasize that it doesn't add to the formal power of the system: for every ranked default theory, an equivalent unranked default theory can be constructed [16].

The semantics of Default Logic can be provided by *Herbrand models* [17,18]. Suppose we have a first order language $\mathcal{L}$, and we augment it with a set of new constants, b, calling the resulting language $\mathcal{L}_b$. The set of all closed terms of the language $\mathcal{L}_b$ is called the *Herbrand universe* of $\mathcal{L}_b$ and is denoted $\boldsymbol{T}_{\mathcal{L}_b}$. A *Herbrand b-model* is a set of closed atomic formulas of $\mathcal{L}_b$.

5 Equality by Default

Resolving anaphora means generating an equality between two discourse referents. I suggest generating such an equality by default: we assume that two elements are equal if they cannot be proved to be different. The idea underlying this notion has been proposed, though not formalized, in [19]. Charniak's approach is further explored in [20], and formalized more fully in [21,22], in which its potential for anaphora resolution is noted.

Equality by Default can be implemented in Default Logic very simply, with the following default:

(13) $$\frac{: x = y}{x = y}.$$

This rule means that whenever it is consistent to assume that two elements are equal, we conclude that they are. It would, of course, be *in*consistent to assume $x = y$ if we know that $x \neq y$. By the axioms of equality, then, (13) is equivalent to saying that we assume $x = y$ unless there is some property ϕ s.t. we know $\phi(x)$ but we also know $\neg\phi(y)$.

It might be objected that this is rather too liberal an assumption of equality, and that we allow two many elements to be equal by default. This, however, is not the case. Equality by Default does not apply in isolation; any reasonable system drawing inferences from natural language will require many more defaults, some of which deal specifically with anaphora, while others don't. If we assign low priority to Equality by Default, so that, if other defaults can apply, they will, inappropriate equalities will be ruled out, and rather few equalities will remain.

For example:

(14) a. John saw Bill. *He* greeted *him*.
b. John hates *him*.
c. John doesn't have a car. *It* is red.
d. A man came into the bar. *She* was upset.

The most likely interpretation of (14.a) is that the first pronoun refers to John, and the second one to Bill, hence they are not equal. This interpretation is brought about by a default rule that prefers antecedents that share the grammatical position of the pronoun (parallelism). In general, Equality by Default is a principle of *last resort*: it will not be invoked if other rules suggest some antecedent. Since in this case the parallelism rule applies, Equality by Default will not apply, and we are in no danger of concluding erroneously that the referents of the two pronouns are equal.

Sentence (14.b) does not have an interpretation where *him* is equated with John, for syntactic reasons. In (14.c), the pronoun *it* should not be equated with the discourse referent representing the indefinite *a car*, because, according to the rules of DRT, the indefinite is not accessible to the pronoun. The discourse in (14.d) is an example where the pronoun cannot be associated with the antecedent because of a gender mismatch. If all such constraints are formalized—as indeed they must be for any anaphora resolution system—and given a higher priority than Equality be Default, inadmissible antecedents will be ruled out.

We could restrict the definition of Equality by Default to apply only to anaphoric triggers and potential antecedents. However, this is not really necessary. Spurious equalities between arbitrary discourse referents will not be generated, because of independently motivated principles. Consider the following examples:

(15) a. John talked to Bill.
b. An officer talked to a gentleman.
c. John is meeting a woman tonight. His mother told me so.
d. John went to the clinic. The doctor had a busy day.

Sentence (15.a) involves two different names. Usually, it is assumed that two different names denote two different individuals; this is known as the Unique Names Assumption [23]. It might appear that our system cannot have the Unique Names Assumption, because different terms are assumed to be equal, rather than different, by default. However, this is not the case because, in DRT, names get their reference by *anchoring* them to individuals in the model, rather than by equality [13]. If the names *John* and *Bill* are anchored to different individuals, with different properties, then they must be different and cannot be equal by default.

Sentence (15.b) involves two indefinites. Standardly, indefinites are assumed to be novel [24]. This means that an indefinite must be different from any previously introduced discourse referent; hence, the referent of *a gentleman* must be different from the referent of *an officer*.

In sentence (15.c), *his mother* does not refer back to *a woman*. The reason is due to conversational implicature [25]: a speaker who knows that John is meeting his mother should say so, hence we conclude that the woman is someone else.

Sentence (15.d) is an example of bridging: *the doctor* is identified with the doctor associated with the clinic. Could it be equal to John by default? The answer is, in fact, yes, and the sentence does have this reading. But (15.d) also has another, perhaps more plausible reading, where John is a patient rather than a doctor. This reading is obtained because the notion of a clinic also introduces the notion of patients, together with the restriction that the patients are different from the doctor. According to one default conclusion, John is equated with the doctor, but according to another, he is equated with one of the patients, and is different from the doctor. Clearly, the two default conclusions are incompatible, hence we will have two extensions, one for each reading. We can then apply credulous reasoning to choose one of the readings, or skeptical reasoning, in which case the ambiguity remains unresolved.[4]

Thus, although the assumption of Equality by Default appears very permissive, in fact it allows rather few elements to be equal by default. These are intended to be anaphoric triggers and their potential antecedents, when no antecedent is suggested by an anaphora resolution factor.

Like other default theories, Herbrand models can provide a semantics for Equality by Default [22]. A clarification, however, is in order. Since the Herbrand universe of a language $\mathcal{L}_b$ is the set of all closed terms of $\mathcal{L}_b$, then, by definition, in a Herbrand model no two terms are identical. But in our default theory, two terms may be equal by default. Is this a contradiction? The answer is no. Equality is *any* relation that satisfies the equality axioms, and is not necessarily

[4] See section 6 for more on how the proposed system deals with ambiguity.

identity.[5] Hence, there is no problem about two terms being *equal*, even though they are not *identical*.

6 Unresolved Anaphora

As mentioned above, there are two cases where anaphora may remain unresolved: when there is no appropriate antecedent, or when there is more than one. In the first case, the trigger needs to be interpreted as referring to an entity not provided by the linguistic content (i.e. an accommodated or deictic interpretation). In the second case, the anaphora is truly ambiguous, and this ambiguity needs to be either resolved arbitrarily, or left unresolved, drawing as many inferences as possible.

6.1 No Potential Antecedent

It turns out that using Herbrand models has a consequence that is particularly important for our purposes. Note that the new elements introduced in b, by being new, are equal by default to any term. In particular, they are equal by default to any anaphoric trigger; this is how accommodated and deictic readings are possible.

This theory allows accommodated and deictic readings, but only as a last resort, when no other readings are possible. More precisely, it has been shown [22] that if E is an extension for Equality by Default, and w is a Herbrand b-model of E, then w is *minimal*. That is to say, there is no Herbrand b-model w' of E such that

(16) $\{\langle t_1, t_2\rangle : w \models t_1 = t_2\} \subset \{\langle t_1, t_2\rangle : w' \models t_1 = t_2\}$.

Now, consider a model w of extension E where trigger u is accommodated or interpreted deictically. This means that, in w, for every x_i, a potential antecedent of u, $u \neq x_i$; and for some new element $n \in b$, $u = n$. Since w is minimal, there is no Herbrand b-model w' of E that contains all the equalities in w and adds to them. Therefore, it is not only in w, but in all models of E, that u is different from all its potential antecedents.

What this means is that there is at least one extension, i.e. at least one plausible way of reasoning from the known facts, that is inconsistent with an anaphoric reading of u. Hence, accommodated or deictic readings are only available when the anaphoric reading is implausible (or impossible).

6.2 Multiple Potential Antecedents

Suppose we have two acceptable antecedents for some trigger. For example, in (7), repeated below, the pronoun may be equated with *John* or with *Bill*.

(17) John and Bill met at the ice cream parlor. *He* was upset.

[5] Of course, we can a have a non-Herbrand model where equality *is* identity—such models are called *normal*, see [26, p. 100] for details.

If we make the standard assumption, as described above, that different names are anchored to different individuals, we know that *John* is different from *Bill*, so it is impossible to believe that the pronoun is equal to both. We will therefore have two extensions: in one of them, the pronoun is equated with *John*, and in the other—with *Bill*.

How do we deal with these extensions? We may decide to force a decision for one or the other; for example, we can decide that the most recent antecedent (Bill) is appropriate, or that the first one mentioned (John) is more prominent, hence preferred. So, in effect, we would apply credulous reasoning and pick one extension.

Note that, by the axioms of equality, once such a choice is made, any property of the antecedent becomes also a property of the trigger. For example, if we choose the extension where *he* is equated with Bill, and if Bill is bald, it will immediately follow that *he* is bald.

There are cases, however, where the anaphora is genuinely ambiguous, and we may have no reason to prefer one reading over the other. But even if we decide not to resolve the anaphora, there are still inferences we can make. Recall that, given (17), we want to conclude that whoever the pronoun refers to was at the ice cream parlor.

In this case, it makes sense to apply skeptical reasoning, and accept only what is true in all extensions. This will generate the desired inference, since in both extensions, the pronoun has the properties that its antecedent has.

Of course, there are extensions where *he* is equated with one or more of the new terms introduced in b. However, this makes no difference to the inference pattern described above, for the following reason. Even if *he* is equated with one or more new elements, it must also be equated with John or Bill. This is because so long as it is possible to find at least one antecedent for the pronoun, a model for the deictic reading, i.e. a model where the pronoun is equated with a new element but with no other element, will not be minimal, hence it will not be the model of any extension. In every extension, then, the pronoun will be equated with some old discourse referent x (which may be equated with any number of new elements in b). Since x is either John or Bill, and both are at the ice cream parlor, x is at the ice cream parlor. Since this is the case in every extension, skeptical reasoning will still conclude that the *he* was at the ice cream parlor.

Now let us consider cases where one possible antecedent has a property than the other one lacks, or is not known to have:

(18) a. John walked along the sidewalk and saw that Bill was inside the ice cream parlor. *He* was upset.
b. John saw that Bill was eating ice cream. *He* was upset.

In (18.a), Bill is inside the ice cream parlor, but John is outside. Thus, in one extension, *he* will have the property of being inside the ice cream parlor, and in the other—its negation. If we apply skeptical reasoning, we will be able to conclude nothing—this appears intuitively correct.

In (18.b), we know that Bill was eating ice cream, but we do not know whether John was. Intuitively, we cannot conclude that *he* was eating ice cream, although

this is consistent with the pronoun being equated with either John or Bill. Skeptical reasoning predicts this result: while in one extension the property of eating ice cream is predicated of the antecedent of the pronoun, in the other extension, neither this property nor its negation will be so predicated. Therefore, it is not true in all extensions that *he* was eating ice cream.

7 Conclusions and Further Applications

I have proposed a formalization of DOAP, a rule that tells us to exhaust all anaphoric possibilities before accommodating or interpreting a pronoun deictically. This rule is formalized using a standard linguistic representation (DRT) and a standard default reasoning system (Default Logic); no special mechanisms for representation or inference are required. Yet this conceptually simple theory appears to produce exactly the sort of inferences regarding anaphora that are intuitively desirable. It is ensured that if it is possible to find an antecedent, we do so; if more than one is a good candidate, we can use Default Logic techniques for dealing with multiple extensions; and if there is none, we accommodate or interpret the pronoun deictically. Thus, the resolutions such a system would make, and the inferences it would draw, can be proved explicitly, rather than be left implicit in the workings of the algorithm.

In this paper I have concentrated on anaphora. Yet I believe the theory can be extended to additional, related phenomena. The phenomena of presupposition and bridging immediately come to mind. Intuitively, these phenomena share with anaphora the notion of some trigger that is looking for an antecedent. In all three cases, a DOAP-like principle applies: it is preferred to choose an antecedent that is, in some sense, already given, than introduce a new one.

The theory presented here can be applied to these other phenomena as well, provided that the association of trigger with antecedent be represented as an equality relation, so that the rule of Equality by Default may be applied. Fortunately, this requirement can, indeed, be easily satisfied in standard theories of bridging and presupposition, in the following manner.

Regarding presupposition, we have already considered example (5) above, where the presuppositions of a definite description is accommodated if no antecedent can be found. This phenomenon is not restricted to definite descriptions, however, and appears to be a general fact about presupposition [12]: binding is preferred to accommodation.

For example, consider the following sentence:

(19) If Jack comes in, then Mary will realize that a dangerous criminal came in.

Despite the factive verb in the consequent of the conditional, (19) does not presuppose that a dangerous criminal came in. Rather, the presupposition is bound by the antecedent of the conditional; in order for this to be possible, the discourse referent corresponding to *a dangerous criminal* must be equated with Jack.

In general, resolving presuppositions involves adding, for each discourse referent x from the universe of the presuppositional DRS, a condition of the form $x = y$, where y is some discourse referent of the antecedent DRS [12].

The same holds for cases of bridging, even when the existence of the entity in question is not entailed by the antecedent, but is merely associated with it. The following example is from [27]:

(20) John entered the room. He saw the chandelier sparkling brightly.

We tend to associate the chandelier with the room John entered, rather than with some other room, not mentioned in the discourse (which John can look into, say through a window).

What is the relation between the chandelier and the room mentioned in the discourse? Since normally rooms have some sort of light in them, we can assume that mentioning the room introduces a discourse referent representing this light source. Then, the relation between the chandelier and the light is that of equality, and Equality by Default can apply to produce the desired result.

It appears, then, that the proposed formalization of DOAP may be fruitfully applied to other phenomena besides pure anaphora. The precise details, however, must await another occasion.

References

1. Asher, N.: Linguistic understanding and non-monotonic reasoning. In: Proceedings of the 1st International Workshop on Nonmonotonic Reasoning, New Paltz (1984)
2. Byron, D., Gegg-Harrison, W.: Evaluating optimality theory for pronoun resolution algorithm specification. In: Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2004). (2004) 27–32
3. Lascarides, A., Asher, N.: Temporal interpretation, discourse relations and common sense entailments. Linguistics and Philosophy **16** (1993) 437–493
4. Mitkov, R.: An uncertainty reasoning approach for anaphora resolution. In: Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95), Seoul, Korea (1995) 149–154
5. Poesio, M.: Semantic ambiguity and perceived ambiguity. In van Deemter, K., Peters, S., eds.: Semantic Ambiguity and Underspecification. CSLI, Stanford (1996) 159–201
6. Beaver, D.: The optimization of discourse anaphora. Linguistics and Philosophy **27**(1) (2004) 3–56
7. Prince, A., Smolensky, P.: Optimality theory: Constraint interaction in generative grammar. Technical report, Rutgers University, New Brunswick, NJ and University of Colorado at Boulder (1993)
8. Reiter, R.: A logic for default reasoning. Artificial Intelligence **13** (1980) 81–132
9. Williams, E.: Blocking and anaphora. Linguistic Inquiry **28** (1997) 577–628
10. Hendriks, P., de Hoop, H.: Optimality theoretic semantics. Linguistics and Philosophy **24** (2001) 1–32
11. Bry, F., Yahya, A.: Positive unit hyperresolution tableaux and their application to minimal model generation. Journal of Automated Reasoning **25** (2000) 35–82
12. van der Sandt, R.: Presupposition projection as anaphora resolution. Journal of Semantics **9** (1992) 333–377

13. Kamp, H., Reyle, U.: From Discourse to Logic. Kluwer Academic Publishers, Dordrecht (1993)
14. Besnard, P., Mercer, R., Schaub, T.: Optimality theory through default logic. In Günther, A., Kruse, R., Neumann, B., eds.: KI'03: Advances in Artificial Intelligence, Proceedings of the Twenty-sixth Annual German Conference on Artificial Intelligence. Volume 2821 of Lecture Notes in Artificial Intelligence., Springer-Verlag (2003) 93–104
15. Brewka, G.: Adding priorities and specificity to Default Logic. In Pereira, L., Pearce, D., eds.: Proceedings of the 4th European Workshop on Logics in Articial Intelligence (JELIA-94). (1994) 247–260
16. Delgrande, J.P., Schaub, T.: Expressing preferences in Default Logic. Artificial Intelligence **123** (2000) 41–87
17. Lifschitz, V.: On open defaults. In Lloyd, J., ed.: Computational Logic: Symposium Proceedings, Berlin, Springer–Verlag (1990) 80–95
18. Kaminski, M.: A comparative study of open default theories. Artificial Intelligence **77** (1995) 285–319
19. Charniak, E.: Motivation analysis, abductive unification and nonmonotonic equality. Artificial Intelligence **34** (1988) 275–295
20. Cohen, A., Makowsky, J.A.: Two approaches to nonmonotonic equality. Technical Report CIS-9317, Technion—Israel Institute of Technology (1993)
21. Cohen, A., Kaminski, M., Makowsky, J.A.: Indistinguishability by default. In Artemov, S., Barringer, H., d'Avila Garcez, A.S., Lamb, L.C., Woods, J., eds.: We Will Show Them: Essays in Honour of Dov Gabbay. College Publications (2005) 415–428
22. Cohen, A., Kaminski, M., Makowsky, J.A.: Notions of sameness by default and their application to anaphora, vagueness, and uncertain reasoning. ms., Ben-Gurion University and The Technion (2006)
23. Reiter, R.: Equality and domain closure in first order databases. Journal of the ACM **27** (1980) 235–249
24. Heim, I.: The Semantics of Definite and Indefinite NPs. PhD thesis, University of Massachusetts at Amherst (1982)
25. Grice, H.P.: Logic and conversation. In Cole, P., Morgan, J.L., eds.: Syntax and Semantics 3: Speech Acts. Academic Press (1975)
26. Mendelson, E.: Introduction to mathematical logic. Chapman and Hall, London (1997)
27. Clark, H.: Bridging. In Johnson-Laird, P., Wason, P., eds.: Thinking. Readings in Cognitive Science. Cambridge University Press, Cambridge (1977) 411–420

Null Subjects Are Reflexives, Not Pronouns

António Branco

University of Lisbon
Faculdade de Ciências, Departamento de Informática
NLX - Natural Language and Speech Group
Campo Grande, 1749-016 Lisboa, Portugal
Antonio.Branco@di.fc.ul.pt
http://www.di.fc.ul.pt/~ahb

Abstract. It has been a prevalent assumption in the literature that phonetically null Subjects of finite clauses are pronouns. This paper examines in detail this empirical generalization and argues that null Subjects are reflexives rather than pronouns.

The critical point at stake here, which has obscured appropriate classification, is that null Subjects are reflexives in top-command positions, i.e. reflexives that have no immediate local commanders. The key issue is thus the observation that, for at least some languages, a top-command reflexive obeys Principle A but with respect to a reshuffled local domain, which is the local domain of the upstairs predicator immediately subcategorizing the predicational domain where the top-command reflexive occurs. Given that the anaphoric binding discipline of reflexives in such positions partially overlaps the binding discipline of pronouns, this gave rise to the mistaken classification.

1 Introduction

In large enough contexts, an anaphoric expression has more than one admissible antecedent. And when occurring in a given syntactic position, different anaphoric expressions may have different sets of admissible antecedents. This is illustrated in the examples below, with three anaphors — *herself*, *her*, and *the little girl* — occurring in the same position, each with different sets of admissible antecedents.

(1) Mary's brother told Paula's sister that the nurse described Joan to herself/her/the little girl.

For *the little girl*, its set of admissible antecedents contains *Mary* and *Paula*. For *her*, in turn, its set of admissible antecedents also includes *Paula's sister*, while *herself* only has *the nurse* and *Joan* as admissible antecedents.

Such differences in terms of sets of admissible antecedents is the basis for the partition of anaphoric expressions into different groups according to their anaphoric capacity. It has therefore been crucial for anaphor resolution to determine how many such types or groups of anaphoric expressions there are, what expressions belong to which type, and what exactly are the sets of admissible

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 59–76, 2007.

antecedents for each type. The results of this inquiry have been collected in what is known, in some linguistic frameworks, as the binding theory.[1]

For the purposes of the research reported in the present paper, it is sufficient to focus our attention in two types of anaphors, viz. reflexives and pronouns, such as *herself* and *her* above, respectively.

Like for other types of anaphors, their sets of admissible antecedents have been characterized intensionally. These definitions have been termed, respectively, Principle A and Principle B, and rely on a few auxiliary notions, such as the notions of command and locality, to be presented below.

1.1 Command

A first difference between the anaphoric capacity of reflexives and of pronouns, and *a fortiori* between their admissible sets of antecedents, is that reflexives cannot have, as antecedents, expressions that occur in "recesses" of the grammatical structure.

(2) [The doctor who called the nurse$_i$]$_j$ described [Joan's$_k$ sister]$_l$ to herself$_{*i/j/*k/l}$.

As represented by the starred indexes, the expressions *the nurse* and *Joan* are not admissible antecedents of the reflexive *herself*. This is in contrast with the larger expressions where they are included, respectively, *the doctor who called the nurse* and *Joan's sister*, which turn out to be admissible antecedents of *herself*.

This contrast results from the circumstance that, in terms of grammatical structure, the latter hold a certain relative position with respect to the reflexive that the former do not. Such relation has been termed in the binding theory literature as a *command relation* and its definition has evolved toward successive versions of enhanced empirical adequacy. We assume here the definition of command according to which A commands B iff A has a grammatical function that is less oblique than the grammatical function of B, if they are selected by the same predicator, or A commands some X that subcategorizes for B or is a projection of B — where, for instance, Subject is less oblique than Object or Indirect Object, Object is less oblique than Indirect Object, etc.[2]

It is worth noting that the notion of command integrates two distinct constraints that are relevant for the correct definition of the set of admissible antecedents of reflexives. By requiring that an antecedent of a reflexive be a commander of it, on the one hand, the antecedent cannot be in a grammatical "recess" with respect to the reflexive; on the other hand, the antecedent cannot be preceded by the reflexive in the obliqueness hierarchy of grammatical functions. The first constraint is exemplified in the data above, the second is illustrated in the contrast below:

[1] For a recent overview, see (Büring, 2005).

[2] This definition is proposed by Pollard and Sag (1994: Chap.6), who term it as o-command in order to differentiate it from earlier, empirically less accurate versions, such as c-command or theta-command.

(3) a. The nurse$_i$ didn't describe Joan$_j$ to herself$_{i/j}$.
b. The nurse$_i$ didn't describe herself$_{i/*j}$ to Joan$_j$.

As in b. the grammatical function of *Joan* — Indirect Object — is not less oblique than the grammatical function of the reflexive — Direct Object —, *Joan* cannot act as its antecedent.

The absence of contrast to topicalized constructions, as in (4)a., confirms that precedence in the grammatical obliqueness hierarchy rather than mere linear precedence or constituency-based precedence is actually at stake here. And the possibility of a reflexive in *by*-phrases, as in (4)b., — with the less oblique possible semantic role of *Agent* – confirms that the hierarchy of semantic roles is not at stake here either:

(4) a. To herself$_{i/j}$, the nurse$_i$ didn't describe Joan$_j$.
b. John$_i$ was shaved by himself$_i$.

The requirement that their antecedents can only be their commanders is a key difference of reflexives with regards to pronouns, for which such requirement does not hold:

(5) [The doctor who called the nurse$_i$]$_j$ told [Joan's$_k$ sister]$_l$ that Mary needs her$_{i/j/k/m}$.

This is illustrated in the example above, with the possibility that *the nurse* or *Joan*, which are in grammatical "recesses" and therefore do not command *her*, be antecedents for this pronoun.

1.2 Locality

A second key difference between pronouns and reflexives is that the antecedents for the first cannot be "too close" to them.

(6) The doctor$_i$ said the nurse$_j$ thinks [Mary$_k$ talked to her$_{i/j/*k}$].

As represented by the starred index, the expression *Mary* is not an admissible antecedent for the pronoun *her*. This is in contrast with the expressions more far apart, namely *the doctor* and *the nurse*, which turn out to be admissible antecedents.

Contrasts like these result from the circumstance that the admissible antecedents of a pronoun occur outside the predicational domain of the predicator selecting it as argument. Such relevant span of a sentence has been termed in the literature as the *local domain* and includes the arguments of the predicator directly selecting the anaphoric expression at stake.

The requirement that their antecedents cannot be in its local domain is a key difference of pronouns with regards to (short-distance) reflexives, for which such requirement is reversed:

(7) The doctor$_i$ said that the nurse$_j$ thinks [Mary$_k$ talked to herself$_{*i/*j/k}$].

Also for long-distance reflexives, such as the Portuguese *ele próprio*, their antecedents can occur in their local domain (as marked by index k in the example below). The difference from short-distance reflexives to long-distance ones is that the latter, but not the former, can also have antecedents outside the local domain (as marked by indexes i and j):

(8) A médica$_i$ disse que a enfermeira$_j$ acha [que a Maria$_k$
the doctor said that the nurse thinks that the Maria
conversou com ela própria$_{i/j/k}$].
talked with ELA PRÓPRIA.

But note that also for the long-distance ones, the requirement that their antecedents be their commanders is in force, as illustrated below:[3]

(9) [A doente que chamou a enfermeira$_i$]$_j$ acha que [a irmã da
the patient who called the nurse thinks that the sister of-the
Maria$_k$]$_l$ conversou com ela própria$_{*i/j/*k/l}$.
Maria talked with ELA PRÓPRIA.
'[The patient who called the nurse]$_j$ thinks [Maria's sister]$_l$ talked with her$_j$/herself$_l$.'

1.3 Top-Command Reflexives

On a par with the notions of command and locality, a key difference between pronouns and reflexives has to do with a different behavior with respect to extra-sentential antecedents. Given that the set of admissible antecedents of pronouns includes all the expressions that are not their local commanders, extra-sentential expressions can always be included in such set, as captured in its intensional definition:[4]

Principle B: A pronoun must be locally o-free.

As for reflexives, it is only in specific circumstances that this can happen. The admissible antecedents of short-distance reflexives are their local commanders; and for long-distance ones, their admissible antecedents are their (local and non local) commanders. Naturally, these restrictions cannot apply when reflexives have no relevant commanders, that is when they occur as the top-commanders in their relevant grammatical obliqueness hierarchies.

In such cases, two types of anaphoric behavior have been observed. As discussed in the subsections below, in some languages, the locally top-commanding short-distance reflexive follows no anaphoric discipline, in which case it is said to be a reflexive *exempt* from its binding Principle. In some other languages, the top-commanding reflexive keeps following its usual binding discipline but in the scope of a reshuffled local domain.

[3] For further details on long-distance reflexives, see (Branco and Marrafa, 2000).

[4] The notion of *o-binding of A by B* is an abbreviation for the conjunction of the requirements that B commands A and is its antecedent. It has a dual in the notion of *o-freeness*.

Exemption from locality or command. The following example provides an illustration of the behavior of short-distance reflexives in a top-command position and the associated exemption effect:[5]

(10) Whom he$_i$ was supposed to be fooling, he$_i$ couldn't imagine. Not the twins, surely, because Désirée, in the terrifying way of progressive American parents, believed in treating children like adults and had undoubtedly explained to them the precise nature of her relationship with himself$_i$.

Here, *himself* is the only argument of *relationship*, the (nominal) predicator selecting it, and therefore in a top-command position. The reflexive does not display its typical anaphoric binding discipline: Instead, it takes an antecedent from a previous sentence, that clearly is not in its local domain neither a commander of it.

A rationale for this can be found in the fact that besides the specific anaphoric binding discipline a reflexive complies with when it is not a top-commanding item, an overarching interpretability condition is admittedly in force in natural languages requiring the "meaningful" anchoring of anaphoric expressions, and *a fortiori* of reflexives, to antecedents. When a reflexive is in a top-command position, no local commander is available to function as its antecedent and anchor its interpretation. Hence, in such cases, reflexives appear to escape their specific binding regime to comply simply with such general interpretability condition and their interpretability be rescued.

Command in a reshuffled locality. For other languages, in turn, data involving reflexives in top-command positions indicate that the reshuffling of the domain may be induced. In these cases, what counts as the local domain for the reflexive is the local domain delimited by the predicator immediately selecting the predication domain where the reflexive is in the top command position.

The German *sich* seems to provide an example of a reflexive which induces local domain reshuffling when in top-command positions. First, when in such a position, admissible antecedents for the reflexive can be found only in the immediately upstairs local domain:[6]

(11) Gernot$_i$ dachte, [dass Hans$_j$ dem Ulrich [ein Bild von sich$_{*i/j}$]
Gernot thought, that Hans the Ulrich a picture of himself
überreichte].
gave.
'Gernot thought that Hans$_j$ gave Ulrich a picture of him$_j$.'

Second, also in a reshuffled local domain, directionality of anaphoric binding for reflexives is complied with, as a non commander in the domain immediately upstairs is not an admissible antecedent (Kiss, 2001:(8)a):

[5] From (Zribi-Hertz, 1989). Pollard and Sag (1994:Ch.4,ftn.18) note that this example, and similar ones, taken from quotes of various writers, "are uniformly judged ungrammatical by American speakers".

[6] Tibor Kiss, p.c.

(12) Ich überreichte dem Ulrich$_i$ ein Buch über sich$_{*i}$.
I gave the Ulrich a book about himself.
'I gave Ulrich$_i$ a book about himself$_{*i}$.'

Third, even in a reshuffled local domain, recesses in grammatical geometry are opaque to the anaphoric capacity of *sich*, as illustrated by a nominal inside of a commanding nominal that cannot be an antecedent for it:[7]

(13) Jan dachte, dass [die Mutter von Hans$_i$] dem Carl ein Bild von
Jan thought, that the mother of Hans the Carl a picture of
sich$_{*i}$ überreichte.
himself gave.
'Jan thought that Hans$_i$' mother gave a picture of himself$_{*i}$ to Carl.'

In order to take into account the anaphoric behavior of reflexives in top-command positions, the intensional definition of their admissible set of antecedents is such that the locality and command requirements are stated to be in force in case the reflexive is not in a top-command position:

Principle A: A locally commanded short-distance reflexive must be locally o-bound.

Therefore, the difference in terms of anaphoric behavior of different reflexives when in top-command positions is to be captured by the appropriate setting of the parameterized construct of locality. While for reflexives like *sich*, the local domain happens to undergo reshuffling, no such reshuffling is associated with reflexives like *himself*.[8]

* * *

In the present Section, the key grammatical constraints on the admissible antecedents of reflexives and pronouns were introduced. In the remainder of this paper, I will proceed by checking out these constraints with respect to null Subjects. Supported by the data to be discussed, the conclusion that will emerge is that null Subjects are reflexives in top-command positions.

In the next sections, the data taken into account are from Portuguese. In the next Sections 3 and 2, I will discuss data showing that antecedents of null Subjects occur in a reshuffled local domain and command them. In Section 4, the behavior of null Subjects in terms of split antecedents and ellipsis is examined. Finally, in Sections 5 and 6, open issues for further research are discussed and conclusions are presented.

[7] Manfred Sailer, p.c.

[8] For the purposes of the present paper, it is enough to focus on short-distance reflexives. For the intensional definition of the admissible set of antecedents of long-distance reflexives, the binding Principle Z, see (Branco and Marrafa, 2000). For a recent overview of the binding Principles, their auxiliary notions and corresponding parameterization see (Branco, 2005a). For the parameterization of the notion of locality in terms of reshuffling, see (Branco, 2005b).

2 Null Subjects and Locality

In this section, the anaphoric behavior of null Subjects is examined with respect to the locality requirements impinging on their admissible antecedents.

2.1 Reshuffled Locality

A null Subject occurs in the top-command position of the predicational domain supported by the predicator that immediately selects it. When this immediate predicational domain is subcategorized by another, upstairs predicator, there are conditions for the null Subject to find its antecedent in a reshuffled local domain, as illustrated in the following example:[9]

(14) O director$_i$ informou o médico$_j$ de [que $\emptyset_{i/j}$ vai receber novo
the director informed the doctor of that goes receive new
equipamento].
equipment.
'The director$_i$ informed the doctor$_j$ that he$_{i/j}$ is going to receive new equipment.'

There is robust evidence that the admissible antecedents of null Subjects can be found only in the immediately upstairs domain, as can be observed in different constructions such as completive, adverbial or relative clauses (in the next examples below, null Subjects will be also contrasted with overt pronouns in the same position):[10]

(15) O médico$_i$ acha [que o director$_j$ não percebeu [que $\emptyset_{*i/j}$ / ele$_{i/j}$
the doctor thinks that the director not noticed that / he
cometeu um erro]].
made a mistake.
'The doctor thinks the director$_j$ didn't notice that he$_j$ made a mistake.'

[9] It is worth noting that, given the semantics of some verbs, the null Subject may be restricted to pick as its antecedent only the Indirect Object or the Direct Object of the upstairs clause. That is the case of verbs like *ordenar* (to order) or *impedir* (to hamper):

(i) A Maria$_i$ ordenou à Ana$_j$ [que $\emptyset_{*i/j}$ / ela$_{*i/j}$ levasse o vestido
the Maria ordered to-the Ana that / she brought the dress
amarelo].
yellow.
'Maria ordered Ana to bring the yellow dress.'

(ii) A Maria$_i$ impediu a Ana$_j$ [de que $\emptyset_{*i/j}$ / ela$_{*i/j}$ levasse o vestido
the Maria hampered the Ana of that / she brought the yellow
amarelo].
dress.
'Maria hampered Ana from bringing the yellow dress.'

[10] When applicable, the English translations of the examples will indicate only the admissible anaphoric links for null Subjects.

(16) O médico$_i$ nunca atende o telefone [quando o director$_j$ decide
the doctor never answers the phone when the director decides
[que $\emptyset_{*i/j}$ / ele$_{i/j}$ vai operar o próximo doente]].
that / he goes operate the next patient.
'The doctor never answers the phone when the director$_j$ decides he$_j$ is going to perform an operation on the next patient.'

(17) O Pedro$_i$ cumprimentou o médico [a quem o director$_j$ disse [que
the Pedro greeted the doctor to whom the director said that
$\emptyset_{*i/j}$ / ele$_{i/j}$ precisava de ser operado]].
/ he needed of be operated.
'Pedro greeted the doctor to whom the director$_j$ said he$_j$ needed to undergo an operation'.

In (16), the null Subject appears in a completive clause selected by the main verb *decide* ("decides") of the adverbial clause, while in (15) it appears in a completive clause that is embedded in another completive clause. As expected from what is observed in (14), *o director*, the NP in the immediately upstairs domain, can act as antecedent. However, the other NP, *o médico*, which is outside this reshuffled local domain, cannot serve as an antecedent for the null Subject.

In (17), in turn, the null Subject of the relative clause cannot have *o Pedro* as antecedent because it lies outside the predicational domain immediately upstairs with regards to the null Subject, which is structured around the verb *disse* ("said", simple past).

Such an impossibility of reaching beyond the immediately upstairs domain holds also in constructions where there is no admissible antecedent intervening between the null Subject and the expressions outside that domain:

(18) O médico$_i$ espera [que nenhum aparelho de raios x revele [que $\emptyset_{*i}$ /
the doctor hopes that no device of rays x reveals that /
ele$_i$ deixou um bisturi dentro do doente]].
he left a scalpel inside of$_t$he patient]]
'The doctor$_i$ hopes that no X-ray machine reveals he$_{*i}$ left a scalpel inside the patient.'

This indicates that the anaphoric capacity of null Subjects is not sensitive to eventual blocking effects by intervenors that are admissible antecedent candidates.[11]

In this vein, it is also worth noting that the anaphoric capacity of a null Subject is not sensitive to the mood of the predicator selecting it in its original, non reshuffled local domain.[12] As the examples above and the one below

[11] For examples of blocking effects induced by intervenors on the anaphoric capacity of Chinese long-distance reflexive and on English exempt short-distance reflexive see (Tang, 1989) and (Pollard and Sag, 1994), respectively, and the references therein.

[12] For examples of sensitivity to mood effects on the anaphoric capacity of Icelandic reflexives, see (Manzini and Wexler, 1987).

illustrate, null Subjects in both indicative and subjunctive completives can only reach admissible antecedents in the immediately upstairs clause:[13]

(19) O médico$_i$ disse-me [que o director$_j$ ainda não aceitou [que $\emptyset_{*i/j}$
the director told-me that the doctor yet not recognized that
/ ele$_{i/j}$ tenha cometido um erro]].
/ he had-SUBJUNCTIVE made a mistake.
'The doctor told me that the director$_j$ didn't acknowledge yet that he$_j$ made a mistake.'

All this evidence that the null Subject is following the anaphoric discipline of top-command reflexives (taking admissible antecedents in a reshuffled domain) is further stressed by the systematic contrast with the different anaphoric behavior of the pronoun that occurs in the same positions. As can been seen in every one of the examples above, none of the restrictions observed for the reflexive null Subject in terms of reshuffled locality is complied with by the pronoun: though the latter can always pick antecedents in such upstairs domain, it can do it also further away.[14]

[13] Also examples with verbs in subjunctive mood from other semantic classes are uniformly judged possible by Portuguese native speakers:

(i) A Maria$_i$ não acha [que $\emptyset_i$ / ela$_i$ consiga emagrecer dessa forma].
the Maria not thinks that / she is-able lose-weight that way.
'Maria$_i$ doesn't think she$_i$ is able to lose weight that way.'

(ii) A Maria$_i$ detesta [que $\emptyset_i$ / ela$_i$ seja obrigada a esperar pelo médico].
the Maria hates that / she be forced to wait by-the doctor.
'Maria hates to be forced to wait for the doctor.'

(ii) O director$_i$ ordenou [que $\emptyset_i$ / ele$_i$ fosse operado de imediato].
the director ordered that / he was operated of now.
'The director$_i$ ordered that he$_i$ was subjected to an operation right away.'

For a discussion of the specific behavior of null Subjects with volitive verbs see Section 5 below.

[14] It is likely that this partial similarity, together with a possible lower frequency in the usage of overt pronouns in some contexts (Barbosa *et al.*, 2005), might have been taken as a more disjunctive contrast than it really happens to be, thus leading some authors to suggest that the overt pronoun cannot take the antecedent that is taken by null Subject in the immediately upstairs domain. This appears, however, not to be the case: Irrespective of differences in frequency of usage, such cases are uniformly judged as possible by Portuguese native speakers, and a quick web search offers examples of such anaphoric links even in carefully written style, as the following sentence in a Portuguese newspaper online (Antunes, 2003):

(i) A culpa vai morrer solteira visto que o ministro$_i$ até já disse que ele$_i$ tinha feito tudo.
guilt will die unmarried since the minister$_i$ had even already said that he$_i$ had done everything.

For the sake of the main claim of the present paper, however, it is worth noting that if overt pronouns could not have the upstairs Subjects as antecedents, this would be a drawback for the empirical adequacy of Principle B, not for the claim that null Subjects are reflexives.

Moreover, while the third person pronoun can always entertain extra-sentential anaphoric links (be they deictic or not), this is not the case with null Subjects. As we are going to check in the subsection just below, only in very specific conditions an extra-sentential anaphoric link can be established for a third person null Subject.

2.2 Exemption from Locality or Command

Notice that though a null Subject in a top-command position induces domain reshuffling, such reshuffling, however, is not an option when the null Subject occurs in an absolute top-command position. As suggested by the discussion in the Section 1.3 above, in that case a null Subject may be exempt from its typical binding discipline. This is illustrated in the following example:

(20) O médico$_i$ falou com a Maria e $\emptyset_i$ / ele$_i$ vai operá-la de seguida.
'The doctor$_i$ talked with Maria and he$_i$ is going to perform an operation on her right away.'

Here, the null Subject appears in the absolute top-command position, as the Subject of a conjunct clause, and can take an antecedent that is not a local commander of it, i.e. it can entertain an anaphoric link that is exempt from the constraint captured in Principle A.

Besides, given that they turn out to be exempt from anaphoric binding principles, a null Subject in a top-command position accepts admissible antecedents in extra-sentential anaphoric links. This is illustrated in constructions with a null Subject of a matrix clause:

(21) A: Como é que o médico$_i$ resolveu o problema?
B: $\emptyset_i$ Foi falar com o director.
'A: How did the doctor$_i$ solve the problem?'
'B: He$_i$ went to talk with the director.'

The example below illustrates also the exempt behavior of the long-distance reflexive *ele próprio*: In an absolute top-command position, it can also entertain cross-sentential anaphoric links.[15]

(22) A: Como é que o médico$_i$ resolveu o problema?
B: Ele próprio$_i$ foi falar com o director.

As underlined by the examples above, there continues to be a parallelism between reflexives and null Subjects, thus indicating that null Subjects display the behavior of reflexives also in absolute top-command positions.

[15] Note that the Portuguese phonetically overt short-distance reflexive *si próprio* bears a residual non nominative case: Given that it cannot occur in Subject positions, it is not possible to design examples like (21) with it.

3 Null Subjects and Command

Having checked that null Subjects behave like reflexives in top-command positions with respect to the locality requirement, in the present Section, we will examine now the anaphoric behavior of null Subjects with respect to the two dimensions of the command relation, recess and directionality.

3.1 Recess

The two examples below present relevant contrasts concerning the constraint according to which null Subjects cannot entertain anaphoric links to antecedents in grammatical "recesses":

(23) a. [O médico do Pedro$_i$]$_j$ disse-me [que $\emptyset_{*i/j}$ / ele$_{i/j}$ tem de ser
the doctor of-the Pedro told-me that / he has to be
operado].
operated.
'Pedro's doctor$_j$ told me that he$_j$ has to undergo an operation.'

b. [O exame do Pedro$_i$] mostra [que $\emptyset_{*i}$ / ele$_i$ tem de ser
the test of-the Pedro shows that / he has to be
operado].
operated.
'Pedro$_i$'s medical test reveals that he$_{*i}$ has to undergo an operation.'

In a., *o Pedro* occurs in the predicational domain of a commander of the null Subject, viz. *o médico do Pedro*, but it is not itself a commander of it, and the anaphoric link between *o Pedro* and the null Subject turns out not to be admissible.

In example b., the anaphoric link is not possible either though *o Pedro* is now the only NP in the sentence that could act as the antecedent of the null Subject. This illustrates that even when there is no alternative antecedent available which may serve as a blocking intervenor, non commanding NPs are not admissible antecedents of null Subjects.

3.2 Directionality

Besides "recess", the other dimension of the command relation is directionality: As a (commanded) reflexive has to be commanded by its antecedent, it has to be more oblique than the latter.

The two examples below present key data to test this constraint with respect to null Subjects:

(24) a. O médico informou a Ana$_i$ [de que $\emptyset_i$ / ela$_i$ vai ser
the doctor informed the Ana of that / she goes to-be
operada].
operated]
'The doctor informed Ana$_i$ that she$_i$ will undergo an operation.'

b. O médico disse à Ana$_j$ [que $\emptyset_{?i}$ / ela$_i$ vai ser operada].
the doctor said to-the Ana that / she goes to-be operated.
'The doctor said to Ana$_i$ that she$_{?i}$ will undergo an operation.'

In example a., *a Ana* is a commander of the null Subject — given that *a Ana* is the Direct Object and the null Subject occurs in an embedded clause that is the Oblique Complement clause —, and *a Ana* is an admissible antecedent for the null Subject, as expected.

In example b., in turn, *a Ana* is not a commander of the null Subject — given that it is the Indirect Object and the null Subject occurs in the Direct Object clause. Here appears to be only a slight contrast, if any, with respect to example a. Such contrast is however more sharp in the following example, with the topicalization of the Indirect Object in order to avoid possible garden-path effects shadowing grammaticality judgments:

(25) À Ana$_i$, o médico disse [que $\emptyset_{??i}$ / ela$_i$ vai ser operada].
To-the Ana, the doctor said that / she goes to-be operated.
'To Ana$_i$, the doctor said she$_{??i}$ will undergo an operation.'

Nevertheless, contrasts are not so sharp here, specially with respect to (24), as they tend to be in all the other examples above. This may interpreted as indicating that there might be some difference between reflexives in top-command positions in nominal and verbal domains. In example (12), we saw that a top-command reflexive in a nominal domain induces a reshuffled local domain that preserves the command relation of the upstairs domain. The example (24) above, however, seems to indicate that this may not be completely the case for reflexives in the top-command position of a verbal domain, and that all the elements of the upstairs domain can act, at least weakly, as their commanders.[16]

4 Plurals and Ellipsis

In the previous two Sections, the data presented provide key evidence that a null Subject cannot be a pronoun and support the plausibility that it is a reflexive. In this respect, it is worth noticing the systematic contrast between the anaphoric behavior of null Subjects and that of pronouns: Anaphoric links that

[16] In this connection and in connection with the observations in footnote 9, it is worth noting that given their specific semantic value, some verbs may superimpose the constraint that the null Subject has an antecedent in the upstairs clause that is less oblique than the embedded clause where the null Subject occurs. This is illustrated by example (i) in that footnote 9, with the verb *ordenar* (to order), and by the following example, with the verb *permitir* (to allow):

(i) Esse dinheiro permitiu à Ana$_i$ [que $\emptyset_i$ / ela$_i$ fosse operada de imediato].
that money permited to-the Ana that / she was operated of now.
'That amount of money allowed Ana$_i$ to undergo an operation right away.'

are blocked for a null Subject are always admissible for pronouns throughout the constructions illustrated by the examples above.

In this section, the anaphoric behavior of null Subjects is examined in further contexts where they also exhibit an anaphoric behavior that is specific of reflexives.

4.1 Null Subjects with Split Antecedents

Besides command and locality, another dimension along which pronouns differ from reflexives concerns the possibility of accepting so called split antecedents. While plural pronouns may have more than one antecedent, as in (26)c., that is not the case with plural short-distance reflexives, as in (26)a., and long-distance reflexives show an anaphoric behavior whose acceptability somehow lies between those two classes of anaphors, as illustrated in (26)b.:

(26) a. O médico$_i$ descreveu o Pedro$_j$ a si próprios$_{*(i+j)}$.
the doctor$_i$ described the Pedro$_j$ to themselves$_{*(i+j)}$.

b. O médico$_i$ descreveu o Pedro$_j$ a eles próprios$_{??(i+j)}$.
the doctor$_i$ described the Pedro$_j$ to ELES PRÓPRIOS$_{??(i+j)}$.

c. O director$_i$ informou o médico$_j$ de que a Maria os$_{i+j}$
the director$_i$ informed the doctor$_j$ of that the Maria them$_{i+j}$
ouviu.
heard.

Interestingly, null Subjects seem to go along more with long-distance than with short-distance reflexives:[17]

(27) A enfermeira$_i$ informou o médico$_j$ [de que $\emptyset_{??(i+j)}$ / eles$_{i+j}$ serão
the nurse informed the doctor of that / they will-be
avaliados em breve].
evaluated in brief.
'The nurse$_i$ informed the doctor$_j$ that they$_{??(i+j)}$ will be evaluated soon.'

In any case, even with split antecedents, null Subjects keep patterning not like pronouns but like reflexives with respect to locality or command for each of their antecedents.

In the examples below, the null Subjects are tested in contexts of split antecedents. In a., one of the two antecedents does not command the null Subject, and in b., one of the two antecedents is not in the (reshuffled) local domain. In both examples, the anaphoric links to split antecedents are not admissible:[18]

[17] One should not exclude the possibility that this is a side effect of the top-command position of null Subjects: As there is no overt nominative reflexive in Portuguese to design key contrasts here, the verification of this hypothesis has to be left open.

[18] Data similar to example a. was pinpointed by Figueiredo Silva (2000). Contrary to what is reported in (Barbosa *et al.*, 2005), we do not find any difference from European to Brazilian Portuguese here.

(28) a. [O director que chamou a enfermeira$_i$] informou o médico$_j$
the director who called the nurse informed the doctor
[de que $\emptyset_{*(i+j)}$ / eles$_{i+j}$ serão avaliados em breve].
of that / they will-be evaluated in brief.
'[The director who called the nurse$_i$] informed the doctor$_j$ that they$_{*(i+j)}$ will be evaluated soon.'

b. A enfermeira$_i$ disse que o médico$_j$ acha [que $\emptyset_{??(i+j)}$ /
the nurse said that the doctor thinks [that /
eles$_{i+j}$ serão avaliados em breve].
they will-be evaluated in brief.
'The nurse$_i$ said the doctor$_j$ thinks they$_{??(i+j)}$ will be evaluated soon.'

4.2 Null Subjects in Antecedents of VP Ellipsis

Another key difference between reflexives and pronouns can be found in their anaphoric behavior in ellipsis contexts. Constructions of VP ellipsis whose antecedent contains a pronoun allow for two readings, the so-called sloppy and strict readings. When the antecedent of VP ellipsis constructions contains a reflexive, in turn, only a sloppy reading is available.

Also in this respect, null Subjects behave like reflexives.

In examples like a. above, where null Subjects are in the antecedent of a VP ellipsis, only the sloppy reading is available, while in b., with the same construction but with a pronoun, both sloppy and strict readings obtain:[19]

(29) a. A Ana$_i$ acha que $\emptyset_i$ será operada em breve e a Maria$_j$
the Ana thinks that will-be operated in brief and the Maria
também.
too.
'Ana$_i$ thinks she$_i$ will undergo an operation soon and Maria$_j$ thinks she$_j$ will too.' (SLOPPY)

b. A Ana$_i$ acha que ela$_i$ será operado em breve e a
the Ana thinks that she will-be operated in brief and the
Maria$_j$ também.
Maria too.
'Ana$_i$ thinks she$_i$ will undergo an operation soon and Maria$_j$ thinks she$_j$ will too.' (SLOPPY)
'Ana$_i$ thinks she$_i$ will undergo an operation soon and Maria also thinks she$_i$ will.' (STRICT)

5 Further Issues

In the discussion above, there is broad and cogent empirical evidence supporting the generalization that null Subjects are reflexives. A possible twin viewpoint

[19] Data like these are noticed in (Figueiredo Silva, 2000).

could have been that the classification of null Subjects as reflexives leads to correctly predicting the set of their admissible antecedents.

Taken in this latter perspective, the classification of null Subjects as reflexives deliver a *prima facie* prediction that happens not to hold in a very specific set of constructions. As illustrated below, in completive clauses with volitive verbs, the Subject of the matrix clause cannot be picked as an antecedent by the null Subject:

(30) a. O médico$_i$ quer que $\emptyset_{*i}$ / ele$_{*i}$ seja operado amanhã.
'The doctor$_i$ wants that he$_{*i}$ will be subject to an operation tomorrow.'

b. O médico$_i$ quer [ser operado amanhã].
'The doctor wants to be subjected to an operation tomorrow.'

The example in a. illustrates this point. Interestingly, that construction is replicated in b. with a variant where the completive includes not a finite, like in a., but a non finite verb. As this example b. shows, there is no deep semantic incompatibility for a volitive verb to select a completive clause whose Subject is anaphorically dependent on the Subject of the volitive verb, a circumstance which cannot thus be invoked to explain away the data in example a.

Interestingly, however, this impossibility of anaphorically linking the Subject of the finite completive with the Subject of the volitive verb is not limited to null Subjects. As it is also shown in a., it extends also to (phonetically overt) pronouns, thus clearly suggesting that rather than a predictive failure of the claim that null Subjects are reflexives what may be at stake here is a very specific grammatical construction that calls to be appropriately taken into account.

In fact, the options in terms of the tense of the completive clause of a volitive verb appear to be strongly correlated to the options for the tense of the matrix, volitive verb itself:

(31) a. O médico quis que a Ana fosse / *seja
the doctor wanted-PAST that the Ana be-PAST / *be-PRES
operada.
operated.
'The doctor wanted that Ana was subjected to an operation.'

b. O médico quer que a Ana *fosse / seja
the doctor wants-PRES that the Ana *be-PAST / be-PRES
operada.
operated.
'The doctor wants that Ana is subjected to an operation.'

While some kind of correlation between tenses may be apparent between matrix and completive clauses in general, the strong constraining effect illustrated above is not found with verbs from other classes:

(32) a. O médico informou / informa que a Ana foi
the doctor informed-PAST / informs-PRES that the Ana was-PAST
/ é operada Terça-feira.
/ is-PRES operated Tuesday.
'The doctor informed/informs that Ana is/was subjected to an operation Tuesday.'

b. A Maria lamentou / lamenta que a Ana tivesse / tenha
the Maria$_i$ was-sorry / is-sorry that the Ana had-PAST / has-PRES
de esperar pelo médico.
of wait by-the doctor.
'Maria was/is sorry that Ana had/has to wait for the doctor.'

These data suggest that the underlying grammatical structure of the constructions with finite completives induced by volitive verbs may be quite specific and distinct from the general case. Of particular relevance here is the fact that a phonetically overt pronoun in the Subject position of the completive cannot take the Subject of the matrix clause as its antecedent. This is a behavior that is in contradiction with the typical anaphoric behavior of overt pronouns in similar constructions, in general. This seems thus an important indication that, in spite of the apparent embedding of the finite completive clause into the matrix clause, the Subject of the completive and the Subject of the matrix are in the same underlying predicational domain, which counts as a local domain for the sake of the anaphoric discipline of the pronoun.

In this connection it is worth noting that, under this hypothesis, what surfaces as the Subject of the completive turns out not to be the least oblique item of its underlying grammatical obliqueness hierarchy. However, this is very likely to be a key feature for the licensing of null Subjects. Accordingly, this hypothesis may also help to understand the other odd aspect at stake here, namely why null Subjects are not admissible in finite completives of volitive verbs.

While this hypothesis is very compelling for its plausibility, it calls naturally to be further worked on. It is important to research whether it can receive further empirical validation. It is also important to discuss how it could be accommodated in formal grammars, and check what implications it may bring for current assumptions on the grammatical structure of sentences of Portuguese. Given the central aims of the present article, this lies, however, outside of its scope and has to be reported in future papers.

6 Discussion and Conclusions

Much of the interest around null Subjects was triggered by initial proposals about the specifics of (i) their anaphoric type and (ii) the conditions licensing their occurrence: In a nutshell, a null Subject was assumed (i) to be a phonetically null pronoun (thus complying with Principle B, and abbreviated as "little *pro*") and (ii) to be licensed in contexts bearing discernible inflectional features (sometimes abbreviated as "strong Φ features"), namely the contexts of

inflectional agreement between a Subject and its verb. The appealing functional rationale was thus that the phonetically null anaphoric expression had to occur in an agreement context where the other, perceptible term of the agreement relation could somehow supplement its null phonetics and reveal its occurence.[20]

Subsequent research on a wider range of languages, focusing mainly on claim (ii) above, brought to light data showing that this functionalist rationale was not empirically supported: Some of the languages that have no inflectional morphology, e.g. Chinese, Japanese, Korean, etc. — but not all[21] — have null anaphoric expressions not only occurring in Subject positions but also in positions with other grammatical functions.[22]

The research reported in the present paper, in turn, focus mainly on issue (i), with the similar outcome that this claim is not also empirically grounded: Not only is the data discussed above incorrectly predicted if null Subjects are classified as pronouns, but also these data provide overwhelming evidence that null Subjects are reflexives. The correct account of the anaphoric behavior of null Subjects was thus shown to simply fall out from:

- their classification as **reflexives** — with their set of admissible antecedents captured by Principle A; together with
- the observation that, given that they are Subjects, these reflexives occur **in a top-command position** — with the corresponding effects:
  - the inducing of a reshuffled local domain, which is the local domain of the upstairs predicator immediately subcategorizing the predicational domain where the top-command reflexive occurs, in case such upstairs domain exist;
  - or else the exemption from the grammatical constraint on their anaphoric discipline, captured by Principle A, in case they occur in an absolute top-command position.

It would not be fair, however, not to mention that in previous works, a few aspects of the anaphoric behavior of null Subjects were brought to light that were noticed as problematic for the empirical adequacy of claim (i).[23] Nevertheless, these problems tend typically to be detected or handled in the frame of linguistic inquiries whose major concern is to relate claim (i), about anaphoric type, with claim (ii), about licensing contexts, mainly in view of improving the empirical adequacy of the latter.

Hence, such drawbacks were calling to be systematically aligned together and discussed under a fresh perspective, decisively focused on the anaphoric behavior of null Subjects and illuminated by advanced results on binding theory. As reported here, this permits to obtain an important progress with respect to issue (i). In our view, this progress, with a more accurate classification of null Subjects

[20] Vd. (Chomsky, 1981) and (Rizzi, 1982).
[21] E.g. Scandinavian languages (Platzack, 1987).
[22] Vd. (Huang, 1989).
[23] For recent discussion about Portuguese, see (Kato *et al.*, 2000), (Barbosa, 2004) and (Barbosa *et al.*, 2005).

as reflexives, may now have a serendipitous effect on the inquiry about issue (ii) as well: It may well foster progress on the research about the licensing conditions of null Subjects when crossed with the key data concerning this research issue, eventually helping to reinterpret such data under a new perspective or eliciting new relevant data that has remained unnoticed or undervalued so far.

References

Antunes, F., 2003. "2004 vai arder menos que 2003". *Fábrica de Conteúdos*, www.fabricadeconteudos.com.

Barbosa, P., 2004. "Two kinds of Subject *pro*". paper presented at *14th Colloquium of Generative Linguistics*, University of Oporto.

Barbosa, P., M. E. L. Duarte and M. Kato, 2005. "Null Subjects in European and Brasilian Portuguese". *Journal of Portuguese Linguistics* 4, 11–52.

Branco, A., 2005a. "Anaphoric Constraints and Dualities in the Semantics of Nominals". *Journal of Logic, Language and Information* 14, 149–171.

Branco, A., 2005b. "Reflexives: Escaping Exemption via Domain Reshuffling". In Stefan Mller (ed.) *Proceedings of the HPSG05 Conference*. Stanford: CSLI Publications, 467–481.

Branco, A. and P. Marrafa, 1999. "Long-distance Reflexives and the Binding Square of Opposition". In Webelhuth, Koenig and Kathol (eds.) *Lexical and Constructional Aspects of Linguistic Explanation*. Stanford: CSLI Publications, 163–177.

Büring, D., 2005. *Binding Theory*. Cambridge: CUP.

Chomsky, N., 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Figueiredo Silva, M. C., 2000. "Main and Embedded Null Subjects in Brazilian Portuguese". In Kato and Negrão (eds.), 2000.

Huang, J., 1989. "Pro-drop in Chinese: a generalized control theory". In Jaeggli and Safir (eds.) *The Null Subject Parameter*. Dordrecht: Kluwer, 185–214.

Kato, M. and E. V. Negrão (eds.), 2000. *The Null Subject Parameter in Brazilian Portuguese*. Frankfurt-Madrid: Vervuert-Ibero Americana.

Manzini, M. R. and K. Wexler, 1987. "Parameters, binding theory and learnability". *Linguistic Inquiry* 18, 413–444.

Platzack, C., 1987. "The Scandinavian languages and the null-subject parameter". *Natural Language and Linguistic Theory* 5, 377–402.

Pollard, C. and I. Sag, 1994. *Head-driven Phrase Structure Grammar*. Stanford: CSLI Publications.

Rizzi, L., 1982. *Issues in Italian Syntax*. Dordrecht: Foris.

Tang, C.-C. Jane, 1989. "Chinese Reflexives". *Natural Language and Linguistic Theory* 7, 93–121.

Zribi-Hertz, A., 1989. "Anaphor Binding and Narrative Point of View: English Reflexive Pronouns in Sentence and Discourse". *Language* 65, 695–727.

Using Very Large Parsed Corpora and Judgment Data to Classify Verb Reflexivity

Erik-Jan Smits[1], Petra Hendriks[1], and Jennifer Spenader[2]

[1] Center for Language and Cognition
[2] Artificial Intelligence
University of Groningen
Dutch Department, Faculty of Arts, P.O. Box 716 - 9700 AS Groningen, The Netherlands
E.J.Smits@rug.nl

Abstract. Dutch has two reflexive pronouns, *zich* and *zichzelf*. When is each one used? This question has been debated in the literature on binding theory, reflexives and anaphora resolution. Partial solutions have attempted to use syntactic binding domains, semantic features and pragmatic concepts such as focus to predict reflexive choice, but until now no experimental data either in favor of or against one of these theories is available. In this paper we look at reflexive choice on the basis of empirical data: a large scale corpus study and an online questionnaire. On the basis of the results of both experiments, we are able to predict the choice between the two reflexive items in Dutch without assuming a distinction between verbs that occur with *zich* or *zichzelf* a priori (cf. a distinction in terms like 'inherent reflexivity' (Reinhart and Reuland, 1993)). Instead, we examine the distribution of *zich* and *zichzelf* using the Clef corpus, a 70 million word Very Large Corpus of Dutch. The corpus is tagged and parsed. This allows us to identify the typical action the verbs are used to describe: reflexive or non-reflexive actions. Regression analysis shows that, by doing so, 21% of the distribution of the two reflexive items in Dutch can be predicted. Using the verb reflexivity found in the corpus study even allows us to explain 83% of the participants' choices in the online study between *zich* and *zichzelf*. As such, both the corpus study and the online questionnaire confirm the group of verbs called 'inherent reflexive verbs' without postulating the group beforehand. We further discovered that even inherently reflexive verbs, which are argued to never co-occur with *zichzelf*, sometimes had *zichzelf* chosen as the preferred argument in the questionnaire, and to a lesser degree, in the corpus suggesting that the verb classes are tendential and not categorical.

1 Two Reflexives, One Meaning?

Dutch, like German, French, Swedish and Danish, but unlike English, has two reflexive pronouns: *zich* and *zichzelf*, both unspecified for gender, number and case:

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 77–93, 2007.

(1) Jan wast zich/zichzelf.
Jan washes SE/SELF
'Jan washes himself'

(2) Jan schaamt zich/*zichzelf.
Jan schames SE/*SELF
'Jan is ashamed of himself'

(1) can be used with both *zich* and *zichzelf*, while (2) seems only to be possible with *zich*. There has been much theoretical debate about what features predict the choice of *zich* or *zichzelf*. The choice has been argued to be the result of syntactic constraints (Broekhuis 2004, Reuland and Koster 1991), to be strongly affected by semantic properties of the verb (Haeseryn et al. 2002 (Algemene Nederlandse Spraakkunst, ANS), Reinhart and Reuland 1993, Lidz 2001) by the degree of affectiveness of the situation (Everaert 1986, Geurts 2004), or by the placement of focus (Everaert, 1986). However, as far as we know there are no large-scale corpus studies or questionnaire studies documenting the use of *zich* and *zichzelf*. Such data, however, is important for several reasons: first, heuristics for the types of objects a given verb tends to co-occur with can improve parsing. Second, the choice of reflexive *zich* with a non-reflexive verb is suggested to be related to the habitualness of the event in the context. Confirming this empirically would mean we have a new surface clue to habitual events, an interesting result for natural language understanding. Third, the acquisition of reflexives and pronouns is a major topic in child language. To correctly make materials and interpret results for Dutch and other language with two reflexives we need to know what their uses are. Finally, the results should be relevant to the choice of the reflexive in natural language generation.

The purpose of this study is to see to what degree a large-scale corpus study and an online questionnaire can help predict the choice between *zich* and *zichzelf*. Through an analysis of the distribution of *zich* and *zichzelf* among predicate types, we also address the existence of a number of different classes of reflexivity which can be found in the literature (among other terms *inherent reflexive verbs*, *necessarily reflexive verbs*, *accidental reflexive verbs*). We do this by examining the use of each predicate and looking at how often the action denoted by the verb is reflexively performed in the corpus compared to how often it is performed to some other party. The experimental data points out that it is only possible to do so if both reflexive and non-reflexive transitive uses are taken into account, considering both corpus and questionnaire data.

2 *Zich* vs. *Zichzelf*

In Binding Theory in Generative Grammar approaches to syntax, Principle A governs the use of reflexive forms:

Principle A: A reflexive must be bound in its local domain.[1]

Because the distinction between the use of pronouns and reflexives can largely be explained purely on the basis of syntactic criteria (i.e. their binding relations), a similar syntactic based approach has been suggested for explaining the distribution of the two reflexive forms in Dutch (*zich* and *zichzelf*) (Broekhuis 2004, Reuland and Koster 1991). Reinhart and Reuland (1993) however argue against standard Binding Theory and its characterization in terms of the syntactic characteristics of the NP, assuming instead a much closer relation between anaphora and argument structure. Put differently, they claim that reflexivity is a syntactic property of predicates. Most important for the current paper, they make a syntactic distinction between *zich* and *zichzelf*, respectively called SE and SELF anaphora.

The debate in the literature seems to have focused on two questions: 1) Are there different classes of verbs that differ in their choice of a SE or SELF reflexive argument? and 2) Is there a difference in meaning between a SE and a SELF reflexive?[2] A third question that has yet to be consistently addressed is 3) What effect does context have on the felicity of a SE or SELF anaphor?

Looking first at the question of verb classes, most theorists claim that there are at least two classes: inherently reflexive verbs and regular transitive verbs. The Dutch Grammar ANS (Haeseryn et al, 2002) identifies a group of verbs as "noodzakelijk reflexieve werkwoorden", or necessarily reflexive verbs, including such verbs as *zich vergissen* (to err), or *zich zorgen maken* (to worry). These verbs are claimed to only occur with *zich* and never with *zichzelf*, and further can never occur with a non-reflexive object. They also recognize "toevallige reflexieve werkwoorden", or accidental reflexive verbs, which can occur with *zich*, *zichzelf*

[1] For current purposes, it suffices to take 'local domain' as the sentence containing the reflexive and, taking a standard Chomskyan perspective, define 'binding' as a relation between two elements A and B for which holds that A and B are co-indexed and A c-commands B (i.e. there is a structural relation between A and B or, more precisely, A c-commands B or A is in a higher structural position than B). Crucially, pronouns must be free in their local domain, i.e. are not co-indexed with an element A in the same sentence and are not c-commanded by this A.

[2] We know that there are several differences between the two forms related to information structure. Only *zichzelf* can occur in coordination and focus positions such as questions answers, in clefts or in topicalization, e.g. (i) and (ii) based on examples from Geurts (2004):

(i) De trainer heeft *zich/zichzelf en zijn hond aangemeld.
The trainer has *SE/SELF and his dog registered
'The trainer registered himself and his dog'

(ii) De kok heeft alleen *ZICH/ZICHZELF gesneden.
The cook has only *SE/SELF cut
'The cook only cut himself'

Zich can never be phonetically focused, and has been argued to not be possible in a context in which there are no salient alternatives (see Reinhart 2003 and Geurts 2004).

and with a third person object. These two groups should be distinguished from a third group of verbs that can occur with a third person object or *zichzelf* but never with *zich*. The groups themselves are defined according to the types of arguments that are felicitous and as such cannot be used to predict the use of *zich* and *zichzelf*. ANS has nothing to say about a possible semantic difference between SE and SELF reflexives.

In summary, based solely on their distribution with different arguments ANS distinguish between three types of verbs.

(3) Necessarily reflexive verbs (only with *zich*)

a. Jan vergist zich
Jan mistake SE
'Jan makes a mistake'

b. *Jan vergist zichzelf
Jan mistake SELF
'Jan makes a mistake'

c. *Jan vergist de hond
Jan mistake the dog
'Jan makes the dog a mistake'

(4) Non-reflexive verbs (never with *zich*)

a. *Zij begrijpen zich niet
They understand SE not
'They don't understand themselves'

b. Zij begrijpen zichzelf niet
They understand SELF not
'They don't understand themselves'

c. Zij begrijpen de melkboer niet
They understand the dairy farmer not
'They don't understand the dairy farmer'

(5) Accidental reflexive verbs (with both *zich* and *zichzelf*)

a. Jan wast zich
Jan washes SE
'Jan washes himself'

b. Jan wast zichzelf
Jan washes SELF
'Jan washes himself'

c. Jan wast de melkboer
Jan washes the dairy farmer
'Jan washes the dairy farmer'

Unlike ANS, Reinhart and Reuland (1993) recognize only two classes of predicates: inherently reflexive predicates[3], that can occur with *zich*, and

[3] As Reinhart and Reuland 1993) point out, the observation that some verbs express an intrinsic reflexive relation between its arguments actually goes back to Jesperson (1933), Gleason (1965) and Partee and Bach (1981) (where it is attributed to Montague).

transitive predicates. Similar to ANS they offer no independent criteria to determine whether or not a given verb belongs to one or the other class. This is only determinable by looking at which arguments the verb can co-occur with. Reinhard and Reuland derive the third class, accidental reflexive verbs, by arguing that *zichzelf* is an operator that can reflexivize a transitive predicate,[4] imposing an identity relation on the two arguments of a predicate:

> "a transitive predicate that is not inherently reflexive may turn into a reflexive predicate if reflexivity is marked on one of its arguments with a SELF-anaphora" (1993: 662)

Zich is not an operator since it can only occur with a predicate that is already inherently reflexive. Thus when the same surface verb form occurs with *zich* it is the necessarily reflexive predicate form, and when it occurs with *zichzelf* it is the transitive predicate that has been turned into a reflexive predicate through the operator *zichzelf*. According to Reinhart and Reuland, this explains why *zich* is not allowed in (6), but is in (7); *wassen* (to wash) has an inherently reflexive and a transitive lexical entry; *haten* (to hate) has no inherently reflexive predicate counterpart, and its transitive entry can only be turned into a reflexive predicate by using a SELF anaphor. Since this is not the case in (6), *zich* is ungrammatical.

(6) Jan haat *zich/zichzelf
Jan hates self
'Jan hates himself'

(7) Jan schaamt zich/*zichzelf
Jan is ashamed of self
'Jan is ashamed of himself'

In sum, Reinhart and Reuland state that it is the type of predicate that determines the distribution of *zich* and *zichzelf* since inherently reflexive predicates will be able to just use *zich*. For accidental reflexive verbs that allow both *zich* and *zichzelf* like (1), they must postulate an ambiguity in the lexicon: the same surface verb *wassen* (to wash) has both an inherently reflexive and a transitive form.

Lidz (2001), like Reinhard and Reuland, believes that there is both a transitive and an inherently reflexive lexical entry for accidental reflexive verbs like, e.g. *scheren* (to shave). But Lidz considers *zich* and *zichzelf* to have different meanings. Reinhart and Reuland's account predicts that all reflexively marked predicates correspond to the same type of semantic reflexivity, regardless of how the reflexivity was achieved, i.e. lexically on the verb or with a SELF operator. Lidz (2001) argues against this conclusion, by convincingly showing that there is in fact a difference. Consider this example: Ringo Starr goes into Madame Tussaud's wax museum. Once he sees his own statue, he notes that they have portrayed him with a beard. But he does not have a beard in real life. Displeased,

[4] See for a slightly different view Keenan (1988), who argues that the SELF anaphor turns a transitive verb into an intransitive one.

Ringo decides to shave the beard off the statue. According to Lidz, it is felicitous in Dutch to utter (9), but not (8) (from Lidz, 2001:128).

(8) Ringo scheert zich
Ringo shaves SE
'Ringo shaves himself'

(9) Ringo scheert zichzelf
Ringo shaves SELF
'Ringo shaves himzelf'

Conversely, if Ringo does have a beard in real life and he decides that he looks better without one given the way they portrayed him in the Madam Tussaud's wax museum (i.e. in this case without a beard) and begins to shave his own face, (8) is felicitous and (9) is not. Both sentences are marked for reflexivity, in terms of Reinhart and Reuland either lexically on the verb or syntactically with *zichzelf*, yet they differ in the situations in which they are true or false. The operation of changing the transitive *scheren* (to shave) into a reflexive action by applying a *zichzelf* operator results in a reflexive shaving action that differs in meaning from the inherently reflexive lexical entry *scheren*.

In order to capture this observation, Lidz (2001) replaces the distinction between SE and SELF anaphora of Reinhart and Reuland with what he considers a more semantic one: he calls SE reflexive modified predicates *pure-reflexive predicates* and SELF reflexive modified predicates *near-reflexive predicates*. In near reflexive predicates the reflexive *zichzelf* object is a function of the subject, but not identical with it, unlike in the true reflexive predicates:

(10) Semantic/pure reflexive predicates:
$\lambda x\,[P(x,x)]$

(11) Near-reflexive predicates:
$\lambda x\,[P(x,f(x))]$

Lidz's (2001) account gives different semantic representations for (8) and (9) cases with reference only to the sentence itself. However it's clear that in his example the context plays an important role in determining which form is felicitous, and he gives no account of how to distinguish contexts appropriate for *zich* from those appropriate for *zichzelf*. Since the inherent reflexive form and the transitive form of the verbs are homophones, and there is currently no way to determine when we are dealing with the inherently reflexive form or the transitive form besides looking at the argument, this account also cannot help us predict the choice of *zich* vs. *zichzelf*.

We need to find independently motivated features that correlate with the choice between *zich* and *zichzelf*. Work in this direction is found in Zubizarreta (1987), who looks at the *semantic affectiveness* of the predicate. The idea of affectiveness was originally discussed in Anderson (1979). A verb is +affective if its action results in a change in the abstract or physical state of its object. Because +affective verbs cause a change, the events they refer to are also

necessarily delimited, though the converse is not true: not all verbs that refer to delimited events are +affective. We can call a verb +affective if it denotes a delimited event favoring a coreferential interpretation between the subject and object, as a result of which actions such as *brushing*, *washing* or *shaving* are +affective while *admiring* or *promising* are -affective.

By using affectiveness Zubizarreta tries to distinguish between inherent reflexive verbs and transitive verbs without defining them in terms of the potential to use *zich* or *zichzelf*.[5] Zubizarreta (1987) begins by arguing that +affective verbs have an internal argument. She gives *eat* as a typical example. When used intransitively *eat* is argued to have a hidden argument of food. She presents *out*-prefixation as a distinguishing test; verbs with affected internal arguments can take *out*-prefixation, e.g. John outate Bill, while standard transitive verbs like *confuse* can't, e.g. *John outconfused Bill. Zubizarreta (1987) claims that inherently reflexive predicates are a subset of verbs that have affected internal arguments, and supports the claim with data from Dutch based on an observation in Everaert (1986). Everaert argues that *zich* can't be a co-argument with the subject in its binding domain because it behaves as a clitic and is a bound anaphor, illustrating this with examples like the following, where (12) illustrates that the *zich* cannot be a co-argument with *zij*, and (13) where it can function as an argument in an adjectival small clause. (examples from Everaert 1986:126):

(12) *Zij begrijpen zich niet.
They understand themselves not.
'They do not understand themselves'

(13) Marie maakt zich niet druk
Marie makes herself not stressed
'Marie is not stressed'

In (13) *zich* is not considered by Everaert to be a co-argument of *druk maken* (to stress out), but in (12) it is considered to be a co-argument of *begrijpen* (to understand). But there are a number of clear counterexamples to this claim. Actually for accidental reflexive verbs *zich* can be a co-argument with the subject:

(14) Jan wast zich.
Jan washes himself
'Jan washes himself'

(15) Jan verbergt zich
Jan hides himself
'Jan hides himself'

Zubizarreta (1987) explains the binding of *zich* with the subject in (14) and (15) by stating that these verbs are 'inherently reflexive'. She then argues that the verbs are syntactically *intransitive*, despite their misleading appearance of transitivity; this is because *zich* can be considered to be an internal affected argument. Further, Zubizarreta (1987) claims these verbs are semantically *transitive*. The

[5] Note that Zubizarreta is concerned with the difference between transitive verbs and inherent reflexive verbs and actually her paper never mentions *zichzelf*.

fact that they realize *zich*, which mistakenly appears to be a subject-coargument, is merely because they wear their semantics on their sleeve. (14) and (15) are thus not counterexamples to the generalization that *zich* cannot be syntactically bound by a subject because *zich* isn't a syntactic argument. *Zich is* a semantic argument because the verbs are +affective and their meaning requires that they act on some sort of object.

Zubizarreta's account means verbs like (14) and (15) should pattern with verbs like (13) and not like (12). A potential problem is that within the group of accidentally reflexive verbs like (14) and (15), some verbs seem to require *zich*, e.g. *verbergen* (to hide), while others like *wassen* (to wash) realize *zich* optionally, e.g. intransitively. Further, Zubizarreta (1987) claims that the realization of *zich* with the last group is a lexical idiosyncracy.

Zubizarreta classifies verbs differently than ANS, Reinhart and Reuland and Lidz: the accidental reflexive verbs with their reflexive uses are classified together with the inherently reflexive verbs.

By not taking *zichzelf* into consideration Zubizarreta misses an important characteristic that distinguishes the class of verbs like (13) from those like (14) and (15). The first group never co-occurs with *zichzelf*, while the latter group can. Also, the first group cannot take a third person argument, while the second group clearly can and does. There is also clearly a tangible difference in meaning between 'Jan wast' (Jan washes) and 'Jan wast zich' (Jan washes himself); in the former Jan can be washing any object but in the latter he must be washing himself. Finally, it is hard to think of what evidence could confirm Zubizarreta's assumption that there is a hidden semantic argument in certain inherently reflexive verbs like *wassen* (to wash) when they are used with *zich*.

Affectiveness has also been appealed to by Jakubowicz (1992) to explain the binding possibilities of the Danish SE reflexive *sig*, quite similar to Dutch *zich*. Jakubowicz argues that verbs that allow local binding with the Danish SE reflexive *sig* are only those that are +affective. Because these verbs also co-occur with the Danish SELF reflexive *sig selv* and with non-reflexive objects, the class of +affective verbs seems to coincide with the class of accidentally reflexive verbs. The local binding ability in *sig* is attributed to an argument present in the +affective verbs, again because the action predicated by the verb must act on something or someone, and thus is concrete enough to be bound locally. In contrast to Zubizarreta (1987) Jakubowicz considers the +affective verbs to be syntactically, as well as semantically, transitive.

Zubizarreta's and Jakubowicz's work is interesting in that they try to ground the idea of reflexivity in terms unrelated to the features they are trying to predict. However, because the definition of +affectiveness is quite vague, it doesn't help us that much with predicting the choice between *zich* and *zichzelf*; we lack a method for objectively determining affectiveness.[6] Because the above explanations are either circular or incomplete we will work with the surface characteristics considering there to be three classes of verbs.

[6] How general the process of out-prefixation is, isn't clear; further, it isn't applicable to Dutch.

2.1 Flexible Class Membership?

A question that has not been addressed in the earlier work is whether membership in one of the above three classes is categorical or whether membership is flexible. Geurts (2004) brings up an interesting example of a case where explicitly emphasizing the reflexivity of an event can make *zich* possible with a verb that most informants would in a neutral context immediately classify as non-reflexive (Geurts, 2004: 4).

(16) De zuster dient *zich/zichzelf opium toe
The sister injected *SE/SELF opium in
'The nurse injected herself with opium'

(17) Betty dient zich/zichzelf weer eens opium toe
Betty injected SE/SELF once again opium in
'Betty injected herself once again with opium'

A nurse normally gives medicine to patients, i.e. others. However, if we know that Betty is a drug addict, and habitually injects herself with opium, when we refer to a token event of this type, *zich* becomes possible. It seems then that the class which the verb falls into is changed by purely pragmatic characteristics, e.g. pragmatic coercion. This example is problematic for the classification of Reinhart and Reuland (1993) and Lidz (2001) because a verb that is generally *not* consider to have a inherently reflexive version seems to acquire just such a lexical entry when the context is appropriate. This suggests that the choice of argument is a regular alteration more than the existence of two lexical entries.

Inherent reflexivity as a semantic feature is perhaps evaluated against the sum of all events in our experience, in which case normal injections are not reflexive. But it seems that in a delimited context, e.g. Betty's life, the sum of all events can be the realm of evaluation, in which case injecting is typically a reflexive event, and *zich* becomes possible.

The example given by Lidz (2001) could also be analyzed along these lines. The use of either *zich* or *zichzelf* to express verb reflexivity results in a difference in meaning because of the habitualness of the situation; in the Madame Tussaud examples, *zich* is possible when Ringo shaves his own face (e.g. not the statue's) because that is a normal reflexive shaving action. Because Ringo's shaving his statue is not standard shaving, *zichzelf* is preferred.

2.2 Towards Objective Predictors

The problem of the current classification of verbs seems to be that they are all based on the feature we want to predict: inherent reflexive verbs are defined as those verbs that only occur with *zich*. A verb is non-reflexive if it cannot occur with *zich*. Also, the divisions that exist of what verbs fall into each category have been determined entirely introspectively. It is therefore not clear to what degree they are correct, and to what degree they have been subjectively determined by the analyst. Further, for the accidental reflexive verbs where both *zich* and *zichzelf* are possible it would be advantageous to know if one was more frequent

than the other, and under what conditions each occurs. There seems to be a relationship between the frequency with which an action is performed reflexively and the 'class' to which the verb belongs.

To gain more objective information about the use of reflexive arguments we decided to do two empirical studies, a corpus study and an online questionnaire. We predict that for verbs that frequently occur with a third person object, and therefore are referring to a non-reflexive event, the use of *zichzelf* will be more frequent than the use of *zich* among the accidental reflexive verbs. Further, we predict that verbs that are seldom used to refer to non-reflexive events will have a higher frequency of co-occurrence with *zich* than with *zichzelf*. Further we also predict that because argument co-occurrence has to do with the ratio of the frequency of reflexive or non-reflexive actions in the world, the classes are not lexically determined.

Our aims are, first, to experimentally verify the difference between necessarily and accidental reflexive verbs, and, second, to experimentally test the hypothesis that the choice between *zich* and *zichzelf* correlates with the typical relation a predicates denotes with respect to its argument(s). The theoretical literature mentioned above predicts that we will not find any necessarily reflexive predicates with *zichzelf*. This follows from the definition of *zichzelf* as an operator which turns a non-reflexive predicate into a reflexive predicate; *zichzelf* can only be applied to a non-reflexive predicate and not to a necessarily reflexive predicate. Conversely, the theories predict that non-reflexive predicates typically occur with *zichzelf* and not with *zich*. In order to answer these questions we did two empirical studies.

3 Empirical Data

3.1 Corpus Study: Method and Results

For the corpus study we used the CLEF (Cross-Language Evaluation Forum) corpus for Dutch made up of 72 million words and 4,150,858 sentences taken from the full content of the 1994 and 1995 Dutch daily newspapers of Algemeen Dagblad and NRC Handelsblad (Jijkoun, Mishne, and de Rijke 2003). The corpus was parsed with the LFG-based Alpino parser (Bouma, van Noord, and Malouf 2001).

We focused on 60 verbs, where 28 of the verbs were defined as inherently reflexive by ANS (Haeseryn et al. 2002). Third person subjects with objects were searched for in the corpus for these 60 verbs. We counted how often each verb occurred with a reflexive *zich*, *zichzelf* or with a non-reflexive object.

First a comparison of *zich* and *zichzelf* was made. The results are displayed in the boxplots in figure 1 in which necessarily reflexive verbs and accidental reflexive verbs are plotted on the x-axis and the use of *zich* on the y-axis (in percentages of the total number of transitive usages). Statistical analysis shows that the distribution of *zich* and *zichzelf* in corpus data to a great extent confirms ANS's classification. *Zich* is significantly more often found to occur with the

verbs that are labeled necessarily reflexive verbs in ANS than with the accidental reflexive verbs. A t-test shows that *zich* is significantly more often used with necessarily reflexive verbs (mean = 82.4%, sd. = 25.5, std. error mean = 4.6) than with accidental reflexive verbs (mean = 99.3%, sd. = 2.7, std. error mean = 4,5) ($t(58)$ = -3,5, p = .001). Most members of the class of necessarily reflexive verbs never occur with *zichzelf*. One of a few exceptions is *ontpop* (turn into), that was used 638 times with *zich*; however, it was used once with *zichzelf*:

(18) Aan het slot van zijn tweede informateurschap heeft Tjeenk Willink zichzelf ontpopt als het activistische type.
'At the end of his second informator-ship Tjeenk Willink turned (himself) into the activist type.'

Because *zich* is also possible with (18), and because this is not a typical focus position, it is not clear how to distinguish this usage. A verb like *straffen* (to punish) is, in line with our predictions, seldom found to occur with *zich* (cf. fig. 1 in which *straffen* is marked as an outlier with an asterix). Below is one of the two examples *straffen* did occur with a SE-anaphor, which interestingly enough, is also an example with *straffen* and a SELF anaphor, where the SELF anaphor is probably chosen for contrast:

(19) Straft de tragische held Oedipus *zich* lijfelijk, deze Eddie
Punish the tragic hero Oedipus SE physically, this Eddie
straft *zichzelf* door het onmogelijke te willen ...
punishes SELF through the impossible to want ...
'While the tragic hero Oedipus punishes himself physically, this Eddie punishes himself by wanting the impossible ...'

For current purposes, an even more important question for the corpus study is: can we predict the distribution of *zich* and *zichzelf* without *a priori* assuming a distinction à la ANS? Or, put differently, can we find a relationship between the frequency with which a verb occurs with a reflexive object versus a non-reflexive object, and the frequency with which the same verb in only reflexive events occurs with *zich* versus *zichzelf*. For this reason we looked at all transitive uses of each verb, including uses with a non-reflexive object. We made a simple linear regression analysis using the use of *zich* and the frequency of reflexive usages as regressors. The regression analysis shows that 21% of the use of *zich* can be predicted by the frequency of reflexive events with the same verb (R^2 = 0.21%, $t(63)$ = 3.9, p < .001).

We can explain 21% of the data by knowing how frequently a verb occurs with a reflexive action from the corpus data alone. However, people might use *zich* or *zichzelf* for other reasons. To see how closely the corpus data reflects the intuitions of naïve speakers, we also did an online questionnaire. Further, because many of the verbs occurred infrequently even in our 70 million word corpus, we felt it was important to supplement results based on a handful of examples with intuitions.

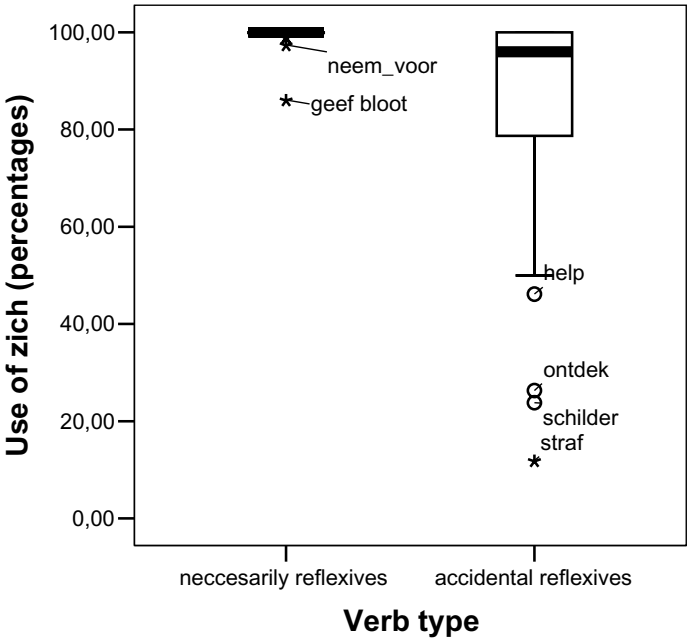


Fig. 1. The use of *zich* (expressed on the y-axis in terms of the percentage of the total number of arguments a predicate is found to occur with) for both necessarily reflexive verbs and accidental reflexive verbs (following the definitions of ANS (Haeseryn et al. 2002)). Translations of the displayed verbs are as follows: *neem voor*,'have intentions' *geef bloot*, 'reveal', *help*, 'help', *ontdek*, 'discover', *schilder*, 'paint', *straf*, 'punish'.

3.2 Online Questionnaire: Method and Results

Twenty-nine adult native speakers of Dutch took part in an online test where they were asked to make a forced choice between *zich* and *zichzelf* as the best argument for 78[7] potentially reflexive verbs. The stimuli were presented in short sentences like (20):

(20) Maria schaamde _____
'Maria was ashamed of _____'

Similar to the corpus data, the data from the online questionnaire reveals a significant difference between the distribution of *zich* and *zichzelf* for necessarily reflexive and accidental reflexive verbs. *Zich* is used in 21.9% of the cases (sd. = 8.0, std.error mean = 1.5) with accidental reflexive verbs and in 93.7% of the

[7] In the corpus study we excluded a number of otherwise interesting verbs because of the existence of a homonym with a very different sense that would have required checking examples by hand. For example, the verb *wegscheren* can mean 'to shave away' but also, with *zich* in the combination *zich wegscheren* 'get out of here', where only the former is truly transitive.

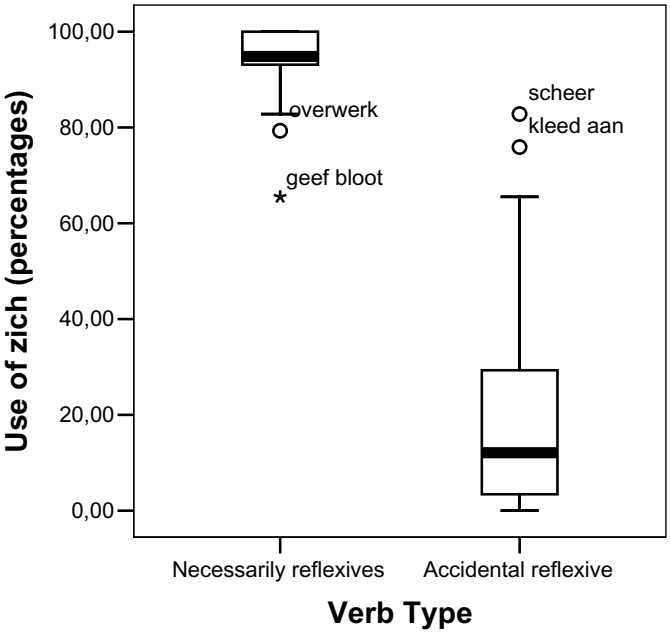


Fig. 2. The use of *zich* in the questionnaire (expressed on the y-axis in terms of the percentage of the total number of arguments a predicate is found to occur with) for both necessarily reflexive verbs and accidental reflexive verbs (following the definitions of ANS (Haeseryn et al. 2002)). Translations of the displayed verbs are as follows: *overwerk*,'overwork' *geef bloot*, 'reveal', *scheer*, 'shave' and *kleed aan*, 'dress'.

cases (sd. = 26.5, std.error mean = 4.7) with necessarily reflexive verbs, $t(58)$ = -13.8, p < .001). This again experimentally confirms Haeseryn et al.'s distinction between necessarily and accidental reflexive verbs.

3.3 Comparing the Data

Comparing the results from the questionnaire study with the results from the corpus study, statistical analysis reveals a significant difference between the use of *zich* for Haeseryn et al.'s necessarily and accidental reflexive verbs (respectively 93.7% versus 99.3% for neccessary reflexive verbs, $t(27)$ = -4.5, p < .001, and 21.9% versus 82.4% for accidental reflexive verbs, $t(31)$ = -11.2, p < .001). We suspect that the difference has to do with the sparse data problem for some of the reflexive verbs in the corpus, which was the motivation for doing the questionnaire study in the first place. Since the difference between the two classes is still the same we see both types of data as complementary confirmation of Haeseryn's classification.

To test the hypothesis that the distribution of *zich* and *zichzelf* in the questionnaire also correlates with verb reflexivity, we compared the online choices for *zich* or *zichzelf* in the questionnaire for the 60 verbs that occurred in both

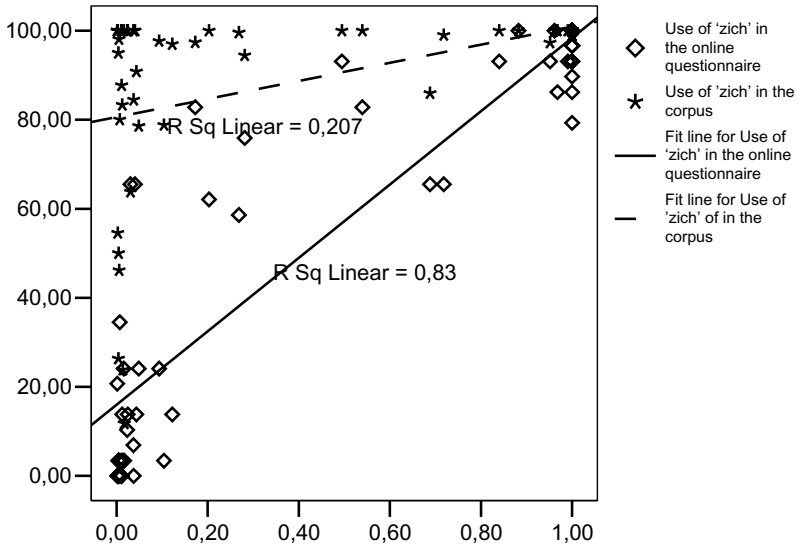


Fig. 3. Typical reflexive usage of each verb expressed on the x-axis in terms of the percentage of reflexive uses among all its uses (i.e. reflexive and non-reflexive transitive uses) versus the use of *zich* per verb on the y-axis. The dotted line represents the correlation between the use of *zich* in the corpus study and the typical reflexive usage of each verb (found in the corpus study). The solid line represents the correlation between the use of *zich* found in the online questionnaire and the typical reflexive character of each verb as found in the corpus study.

experiments (see Appendix A). Simple linear regression shows that 83% of the distribution is predicted along these lines (R^2=0.83, $t(61) = 16.9$, $p < .001$). This shows that the inherent reflexive nature of the verb, defined as the frequency with which a verb is used to refer to a reflexive action or a non-reflexive transitive action in the corpus, is a correct predictor of the distribution of *zich* and *zichzelf* in the online questionnaire.

As a next step, we used Fisher's r to z-test to statistically analyze the difference between the correlation coefficients found for the distribution of *zich* and *zichzelf* in the corpus data and the questionnaire with respect to the typical predicate structure. This revealed a significant difference between the correlation in the corpus study and the online questionnaire regarding the reflexive nature of the verb and the distribution of *zich* and *zichzelf* (z = -5.5, $p < .001$). Put differently, the data from the corpus study gives us a better picture how often a verb is used with a reflexive or a non-reflexive action. This in turn significantly improves our ability to model the use of *zich* and *zichzelf* in the questionnaire. This leads us to conclude that the distribution of *zich* and *zichzelf* can be predicted solely on the basis of corpus and judgment data. No distinction has to be made a priori between necessarily and accidental reflexive predicates along the lines of ANS.

4 Discussion

The results show that it is possible to predict to a large degree what class a verb belongs to (either to the class of necessarily reflexive verbs or to the class of accidental reflexive verbs). Moreover, we have shown that a combination of a corpus study and an online questionnaire allows us to do so. Several reasons motivate the decision to supplement the corpus work with judgment data. Even with a corpus of over 70 million words, it is not possible to find reflexive uses of all the verbs that can possibly occur with a reflexive meaning. For example, *ruiken* (to smell) only occurred with reflexive objects twice in the corpus (however 451 times with a non-reflexive object). This is a very small number to draw conclusions from. Moreover, the corpus data alone was not a perfect predictor for the distribution of *zich* and *zichzelf*, we also needed to looked at judgment data. For example, the verb *schamen* (to be ashamed, reflexive in Dutch) was only used once with *zichzelf* in the corpus, and the case was a direct translation from English to Dutch which may have influenced the choice. However, in the online test 6 respondents preferred *zichzelf* instead of *zich*. The context sentence was extremely short and neutral, so the preference for *zich* might be explained by the tendency for *zich* to avoid focus positions. The end of the sentences is a typical focus position in Dutch. However, this explanation would make it hard to explain why the other 23 respondents preferred *zich*.

Because we did the corpus work for the most part automatically, some of the results might be incorrect. Incorrect parses of imperative and topicalized sentences were found when the data was hand-checked. This has certainly introduced some noise in the data, but it is just this type of parsing error that empirically founded reflexive classes might be able to help avoid.

Do corpus results and judgment results give us a way to predict the choice of reflexive? Yes and no. We can derive the main classes without assuming a priori classes, but we cannot predict individual choices for accidental reflexive verbs. We can use the corpus results to confirm what verbs belong to the class of inherent reflexive verbs (preference for *zich*), and to confirm what verbs are typically 'non-reflexive' (preference for *zichzelf*). But because the online study shows that subjects can deviate from the predictions, other factors such as focus and the habitualness of the action need to be considered. Various participants in the online questionnaire pointed in a similar way at the habitualness that seems to play a role in the meaning of *Het meisje snijdt zich* 'The girl cuts herself' for unintentional cutting, versus *Het meisje snijdt zichzelf* for intentional cutting.

Zubizarreta's work brought up a possible additional factor. She suggests that the intransitive uses of the accidental reflexives somehow play a role in the frequency of the use of *zich* vs. *zichzelf*. This could be tested. If the choice to omit *zich* is totally idiosyncratic, then we should be able to count purely syntactically intransitive uses of e.g. *wassen* (to wash) as being reflexive uses. We can then consider whether ratios of reflexive uses to non-reflexive uses are more predictive if we count syntactically intransitive uses as being among the reflexive transitive uses. We leave this for future work.

Revisiting the questions brought up at the beginning of the paper, we found evidence confirming the existence of at least three classes of reflexives, though the membership is not completely categorical as many researchers have thought, with the categories being just strong tendencies. We were not able to evaluate whether or not there is a difference in meaning between SE and SELF reflexives because we did not look at individual examples, but we think the fact that we found exceptions among the class of necessarily reflexive verbs that took *zichzelf* as an argument seems to suggest that there is some difference: how otherwise can we explain this deviation from the strong trend for these verbs only to occur with *zich*? Finally we did not address the question of how context effects the choice in our experimental studies because this also involves manual checking of examples, but note that this is an obvious future endeavor.

In sum, we have found evidence that verbs do roughly belong to classes of necessarily reflexive verbs and accidental reflexive verbs. We conclude that the corpus data alone does not completely predict the choice between *zich* and *zichzelf*. Because judgment data reveals significantly different patterns we conclude that both sets of data are necessarily to make a good model. By doing so, we are able to, unlike previous research, predict class membership to a high degree based on the frequency with which the verb is used to refer to a reflexive action or a non-reflexive action. In doing so we come to the conclusion that the transitive uses of the verb and the reflexive uses are actually related. In fact, it strongly calls into question the underlying assumption in the work of Reinhart and Reuland (1993), Lidz (2001) and Zubizarreta (1987) that there are two identical surface forms mapping to two different underlying verb forms, the inherently reflexive predicate form and the transitive predicate form. Remember, the motivation for this distinction was to be able to account for the difference between verbs that have no transitive form and allow only *zich* and those that allow both. This analysis requires postulating two distinct lexical entries for each accidental reflexive verb surface form. Since these authors offer no independently motivated way to prove two distinct forms exist, and we can distinguish them on the basis of the frequency of all the arguments they co-occur with (e.g. non-reflexive as well as reflexive), it seems unnecessary to maintain this view.

References

Anderson M. (1979). Noun phrase structure. Unpublished doctoral dissertation, MIT, Cambridge, MA.

Bouma, G., G. van Noord, and R. Malouf. (2001). Alpino: Wide-coverage computational analysis of Dutch. In Computational Linguistics in The Netherlands 2000. Rodopi

Broekhuis, H. (2004). The referential properties of noun phrases I (2nd edition). Modern grammar of Dutch occasional papers 1, University of Tilburg.

Everaert, M. (1987). *The syntax of reflexivization.* Foris Publications, Dordrecht, The Netherlands / Riverton, USA

Geurts, B. (2004). Weak and Strong Reflexives in Dutch, In: Philippe Schlenker and Ed Keenan (eds.), Proceedings of the ESSLLI workshop on semantic approaches to binding theory.

Gleason, H. (1965). Linguistics and English Grammar. New York: Holt, Rinehart and Winston.
Haeseryn, W., K. Romijn, G. Geerts, J. de Rooij & M.C. van den Toorn.(2002). Algemene Nederlandse Spraakkunst. Second, totally revised version of 2002. Groningen/Deurne, Martinus Nijhoff /Wolters Plantyn
Jakubowicz, C. (1992). Sig en danois: syntaxe et acquisition, in H.-G. Obenauer and A. Zribi-Hertz (eds.), Structure de la phrase et thêorie du liage, Presses Universitaires de Vincennes, Saint Denis, pp. 121-149.
Jesperson, P. (1933). Essentials of English grammar. London: Allen and Unwin (1983).
Jijkoun, V., Mishne, G. & M. de Rijke. (2003). Preprocessing Documents to Answer Dutch Questions. In: Proceedings 15th Belgian-Dutch Conference on Artificial Intelligence (BNAIC'03), 2003.
Keenan, E. (1988). On semantics and the binding theory. In J. Edwards (ed) *Explain language universals.* John Hawkings. Oxford: Blackwell.
Lekakou, Marika. (2005). Reflexives in contexts of reduced valency: German vs. Dutch. In The Function of Function Words and Functional Categories, Dikken, M. den and C. M. Tortora (eds.). John Benjamins.
Lidz, J. (2001). Condition R. Linguistic Inquiry **32** (1), 123-140
Partee, B. and Bach, E. (1981). Quantification, pronouns, and VP-anaphora. In Formal methods in the study of language. Mathematisch Centrum, Amsterdam University.
Reinhart, T. & Reuland, E. (1993). Reflexivity. Linguistic Inquiry **24**, 656 - 720
Reuland, E. and J. Koster. (1991). Long-distance anaphora: an overview. In Long-distance anaphora, ed. J. Koster and E. Reuland, 1-25. Cambridge:Cambridge University Press
Zubizarreta, M.L. (1987) *Levels of Representation in the Lexicon and in the Syntax.* Foris Publications: Dordrecht, the Netherlands / Providence RI, USA

Appendix

Necessarily reflexive verbs tested (28), following ANS (Haeseryn et al. 2002):

abonneren (to subscribe to), *bedrinken* (to get drunk), *inbeelden* (to imagine), *behelpen* (to make do), *beijveren* (to), *bemoeien* (to interfere with), *beraden* (to think over), *beroemen* (to boast), *indenken* (to image)), *distantiëren* (to dissociate), *gedragen* (to behave), *bloot geven*, *generen* (to feel embaressed), *schuilhouden* (to hide), *inleven* (to imagine), *misdragen* (to misbehave), *voornemen* (to resolve), *ontfermen* (to take pity on), *ontpoppen* (to turn out to be), *ontspinnen* (to lead to), *overwerken* (to overwork), *schamen* (to be ashamed of), *vergrijpen* (to attack)), *verhouden* (to be in proportion), *verkneukelen* (to chuckle) , *verloven* (to engage), *verslikken* (to choke) , *verspreken* (to make a slip of the tongue).

Accidental reflexive verbs tested (32):

aaien (to pet), *achtervolgen* (to follow), *bedekken* (to cover), *beschermen* (to protect), *bewonderen* (to admire), *bijten* (to bite), *binden* (to bind), *geven* (to give), *ingraven* (to bury), *helpen* (to help), *horen* (to hear), *kietelen* (to tickle), *aankleden* (to dress), *knippen* (to cut), *kussen* (to kiss), *lachen* (to laugh), *opmaken* (to make up), *omhelzen* (to hug), *ontdekken* (to discover), *prikken* (to prick), *ruiken* (to smell), *scheren* (to shave), *schilderen* (to paint), *schoppen* (to kick), *slaan* (to hit), *snijden* (to cut), *straffen* (to punish), *tekenen* (to draw), *tillen* (to lift), *verstoppen* (to hide), *vertellen* (to tell), *zien* (to see).

An Empirical Investigation of the Relation Between Coreference and Quotations: Can a Pronoun Located in Quotations Find Its Referent?

Shana Watters[1] and Jeanette Gundel[2]

[1] Department of Computer Science and Engineering,
[2] Department of Linguistics
University of Minnesota
Minneapolis, MN, 55455, USA
watters@cs.umn.edu

Abstract. Many reference resolution studies omit anaphoric forms found in quotations, assuming that they may need special handling as there is insufficient discourse context to determine the referent. This paper reports on an empirical study performed to evaluate this assumption. Specifically, the study addresses the following questions: Are anaphoric expressions found in quotations sufficiently different to justify ignoring them, and is there enough context available for a system to determine the referents of anaphoric expressions found within quoted text? The current study focuses on the pronoun *it* within the Givenness Hierarchy framework of Gundel, Hedberg, and Zacharski [13]. We find that this framework can be used in most cases to locate the antecedent for referential *it* found in quoted text.

1 Introduction

Research on anaphora resolution within computational linguistics and natural language processing has primarily focused on pronominal reference and has produced accuracy rates in the vicinity of 80% (see [25], inter alia). This accuracy rate can be somewhat misleading since one problem researchers must face when testing their algorithms is what to choose as their test data and the framework to use for evaluation. This problem arises because there is currently no gold standard corpus or standard framework for evaluating (co-)reference resolution systems. As Salmon-Alt & Romary note, "there is an opportunity to stabilize the corresponding knowledge as an international standard in the context of the recently created ISO committee TC37/SC4 on language resource management. Indeed this committee aims at providing generic standards for the representation of linguistic data at various levels" [31]. Lack of a gold standard corpus allows for researchers to use a variety of corpora. In the past, researchers have used such corpora as MUC-6 and MUC-7 ([5], [33], [29], [39]), ACE ([8]), Penn Treebank ([4], [34], [35], [11]), and researcher created corpora

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 94–106, 2007.

([27], [6]) to test their reference resolution algorithms and frameworks. Because of the diversity of data used for testing and the different forms of evaluation, the studies are not always comparable. As Mitkov states "The studies carried out so far have not distinguished between the evaluation of an anaphora resolution algorithm and the evaluation of an anaphora resolution system. As a result, the findings reported often vary significantly and fail to provide common ground for comparison" [26]. One example of this lack of comparability, is that many studies omit occurrences of pronouns found in quotations since they may need special handling or there is not enough discourse context to determine the referent (e.g., [12], [7]); other studies, on the other hand, include material in quotations (e.g., [20], [29], [8]); and in most studies, there is no indication as to whether or not quoted text is included in the corpus. In [20] and [21], quoted material is listed as one of the reasons why the algorithm improperly chose an antecedent. Kennedy & Boguraev note that "Ensuring proper interpretation of anaphors both within and outside of quoted text requires, in effect, a method of evaluating quoted speech separately from its surrounding context" [20].

This paper discusses the results of an empirical study performed to answer the following question: Are anaphoric expressions found in quotations sufficiently different to justify ignoring them, and is there enough context available for a reference resolution system to determine the correct antecedent for a pronoun found within quoted text? The current study focuses on the pronoun *it*. It is conducted within the Givenness Hierarchy framework of Gundel, Hedberg, and Zacharski [13]. Within this framework, reference resolution is constrained by the cognitive (memory and attention) status conventionally signaled by different forms. The pronoun *it*, like other unstressed pronouns in English, is assumed to signal the status 'in focus', thus constraining possible interpretation to ones that can be assumed to be in the addressee's focus of attention because the referent has been introduced by a sufficiently salient linguistic form or extralinguistic object, or has been recently mentioned more than once. The study supports the hypothesis that this framework can be used in most cases to locate the antecedent for referential *it* found in quoted text.

2 The Givenness Hierarchy

Gundel, Hedberg, and Zacharski [13] propose that determiners and pronouns restrict their possible referents by conventionally signaling different cognitive statuses (memory and attention states) that the intended referent is assumed to have in the mind of the addressee. There are six statuses that form a hierarchy of memory and attention states. This hierarchy, known as the Givenness Hierarchy, is provided below along with the English forms that conventionally signal each status as part of their meaning.

Givenness Hierarchy

In Focus	>	Activated	>	Familiar	>	Uniquely Identifiable	>	Referential	>	Type Identifiable
{*it*}		{*that, this, this* N}		{*that* N}		{*the* N}		{indefinite *this* N}		{*a* N}

The Givenness Hierarchy is arranged from left to right in order of the most restrictive status ('in focus') to the least restrictive status ('type identifiable'). Each status entails all of the statuses that are less restrictive than itself. So anything that is 'in focus' is also 'activated', 'familiar', 'uniquely identifiable', 'referential', and 'type identifiable'. The entailment is unidirectional, in that a status only entails all lower statuses, but not vice-versa. Gundel, Hedberg, and Zacharski [14] note that the statuses are part of the conventional meaning of individual lexical items. By conveying information about the addressee's memory and attention states with respect to the referent, the different forms serve as processing signals that assist the addressee in restricting possible referents. The conveyed information for each status is as follows [15]:

- **In Focus:** Associate a representation that your attention is currently focused on.
- **Activated:** Associate a representation from working memory.
- **Familiar:** Associate a representation already in memory.
- **Uniquely Identifiable:** Associate a unique representation by the time the nominal is processed.
- **Referential:** Associate a unique representation (by the time the sentence is processed).
- **Type Identifiable:** Identify what kind of thing this is.

3 The Study

3.1 Methodology

Data for the study consists of 64,725 words from 59 newspaper articles[1], 1 Discover web exclusive article, and 11 Discover magazine articles; it does not include text from picture captions or graphics.

All instances of the 3rd person personal pronoun *it* were located in the corpus and each was coded as being either "not found within quotations" or "found within quotations". Each occurrence of *it* found within quotations was then coded as being either referential or expletive. To be classified as referential, the pronoun had to have an intended referent. In the majority of cases, the referent was an individual object denoted by a previous NP/DP, an event described by a previous clause, or an activity described by a previous VP.

[1] Due to the large number of magazine and newspaper articles reviewed for this study, only those articles referred to within this paper are included in the `bibliography`. A full listing of all articles used for the study is located at http://www.cs.umn.edu/~watters/daarc06_bib.pdf

Each occurrence of referential *it* was then analyzed to determine whether it satisfied one of the coding protocol criteria for the cognitive status 'in focus'[2] [16] and, if so, which one. If the form had a linguistic antecedent, but did not satisfy one of the coding protocol criteria, it was marked as being indeterminable by the Givenness Hierarchy. When the antecedent could not be located in the corpus, it was marked as 'antecedent not found'. The classification of the data was performed by the authors and an additional naive coder.

3.2 The Givenness Hierarchy Coding Protocol Criteria

According to the protocol in [16], a referent is considered to be 'in focus' if it meets one or more of the following criteria:

Criterion 1: It is mentioned in main clause subject position in the immediately preceding sentence/clause.

(1) **<u>The Ghan</u>**, which makes the 1,850-mile run from Adelaide to Darwin twice weekly, derailed about 18 miles south of the Adelaide River in Australia's Northern Territory. "We don't know why **it** came off the tracks. It's too early to tell," said Alan Stuart, a spokesman for the Great Southern Railway. [1]

In (1), the referent of *it* in *we don't know why it came off the tracks* was determined to be the Ghan. Since the Ghan is mentioned in main clause subject position in the immediately preceding sentence, it meets Criterion 1 for in focus status of the referent. The antecedent of *it* (*the Ghan*) is an NP/DP located outside the quotation where *it* is located.

(2) "**<u>Dodder</u>** is a very difficult pest to control," De Moraes says. "**It** attaches to the host plant, and it makes it very hard to kill the weed plant without killing the host." [23]

In (2), the referent of *it* in *It attaches to the host plant* was determined to be dodder. Since dodder was mentioned in main clause subject position in the immediately preceding clause, it meets Criterion 1 for in focus status of the referent. The antecedent *Dodder* is an NP/DP located in a separate quotation.

(3) **<u>The methane</u>** there "would probably take some decades or centuries to come out," he says. "But once **it** started, it would be essentially unstoppable." [22]

In (3), the referent of *it* in *once it started* was determined to be the methane which is mentioned in the main clause subject position in

[2] The study is concerned with the 'in focus' cognitive status only since determiners or pronouns that signal lower statuses are not reviewed.

the immediately preceding sentence and is, therefore, in focus by Criterion 1. The antecedent *The methane* is an NP/DP that is located in the non-quoted portion of a sentence that also contains quoted material.

(4) "**Laying of mines or fencing the border** will only separate people, families from each other," he told a news conference. "Rather than helping, **it** will cause people difficulty in movement, in trade." [32]

In (4), the referent of *it* was determined to be the activity/event of laying of mines or fencing the border, which is mentioned in subject position in the immediately preceding main clause and is, therefore, in focus by Criterion 1. This event/activity is a nominalized VP that occurs in a separate quotation than the one that contains the pronoun *it*.

(5) "But **the strange breathing** happened when he was lying down. After the X-ray, when they sat him up again, he stopped doing **it**.

In (5), the referent of *it* was determined to be the strange breathing which was mentioned in the main clause subject position of the immediately preceding sentence and is, therefore, in focus by Criterion 1. The antecedent *the strange breathing* is an NP/DP located in the same quotation as the pronoun *it*.

Criterion 2: It is mentioned earlier in the same sentence.

(6) **The decision** was made a year ago, "but nobody got around to executing **it**," he said Wednesday. [18]

In (6), the referent of *it* in *but nobody got around to executing it* was determined to be the decision. Since the decision is mentioned earlier in the same sentence, it meets Criterion 2 for in focus status of the referent. The antecedent of *it*, *the decision*, is an NP/DP located outside the quotation where the *it* is located.

(7) "We believe **our offer** is more than fair and don't feel any need to amend **it** at this point," Parker said. [38]

In (7), the referent of *it* in *and don't feel any need to amend it at this point* was determined to be the offer. Since the offer was mentioned earlier in the same sentence, it meets Criterion 2 for in focus status of the referent. The antecedent for the pronoun, *the offer*, is an NP/DP located in the same quotation.

(8) "If **we were to win**, **it** could have a favorable effect on other institutions that have residency programs, but whether or not it does would depend on the basis of the decision," said John W. Windhorst Jr., who, with Thomas Tinkham, filed the suits. [10]

In (8), the referent of *it* was determined to be the activity/event of winning the suit, which is mentioned earlier in the same sentence and, therefore, is in focus by Criterion 2. This event is described by the clause, *we were to win*, and occurs in the same quotation as the one that contains the pronoun *it*.

Criterion 3: It is mentioned in syntactic focus position of the immediately preceding clause (i.e. postcopular position of a cleft or existential sentence).

(9) "If it was a test environment, they said, 'Let's play dead,"' says Ofria. "There's **this thing** coming to kill them, and so they avoid **it** and go on with their lives." [40]

In (9), the referent of *it* in *and so they avoid it and go on with their lives* was determined to be the thing. The thing was mentioned in the syntactic focus position of the immediately preceding clause, in postcopular position of an existential sentence, and is, therefore, in focus by Criterion 3. The antecedent *this thing* is an NP/DP that is located in the same quotation as the pronoun *it*.

Criterion 4: It is a higher level topic that is part of the interpretation of the preceding clause (whether it is overly mentioned there or not).

(10) For pilgrims streaming in from all continents, **the hajj** is a crowning moment of faith, a duty for all able-bodied Muslims to carry out at least once. On Thursday morning, as they have for the past few days, hundreds of thousands **circled the Kaaba**, the black cubic stone in Mecca, Islam's holiest site, which Muslims face when they perform their daily prayers. "For us **it** is a vacation away from work and daily life to renew yourself spiritually," said Ahmed Karkoutly, an American doctor from Brownsville, Texas. "You feel you are part of a universe fulfilling God's will. It's a cosmic motion, orbiting the Kaaba. [19]

In (10), the referent of *it* in *For us it is a vacation away from work and daily life to renew yourself spiritually* was determined to be the hajj. The hajj is a higher level topic that is implicit in the immediately preceding sentence, as the latter describes what takes place during the hajj, which is explicitly mentioned in the sentence before that. Since Muslims circle the Kaaba during the hajj, the hajj is, therefore, in focus since it is what the previous sentence is discussing. The antecedent *the hajj* is an NP/DP and is located two sentences away from where the *it* is located.

Criterion 5: It was mentioned in the two immediately preceding clauses.

(11) Analysts said they didn't anticipate a rash of **iPod** returns because of the delays. "What you're seeing is the tremendous success of **the iPod**," said Michael Gartenberg, vice president and research director with Jupiter-Research. "No doubt **it** was a very, very popular gift, and no matter how well you plan on the server side of the equation, there are always times when you get caught short." [30]

In (11), the referent of *it* in *No doubt it was a very, very popular gift* was determined to be the iPod. Since iPod was mentioned in the two immediately preceding clauses, it meets Criterion 5 for in focus status of the referent. The antecedent *the iPod* is an NP/DP and is located outside of the quotation where *it* is located.

(12) The British demolished the building in an effort to disband **the unit**. "We identified **the serious crimes unit** as frankly, too far gone", Burbridge said. "We just had to get rid of **it**. [36]

In (12), the referent of *it* in *We just had to get rid of it* was determined to be the serious crimes unit which was mentioned in the two immediately preceding clauses by *the serious crimes unit* and *the unit*. Therefore, the referent is in focus by Criterion 5. The antecedent for *it*, *the serious crimes unit*, is an NP/DP and is located outside of the quotation where *it* is located.

(13) "Every single Mosasaurus shows **evidence of the bends**, but every single Clidastes lacks **it**. Every single Tylosaurus has **it**, but every single Halisaurus lacks it." [17]

In (13), the referent of *it* in *Every single Tylosaurus has it* was determined to be evidence of the bends. Since evidence of the bends is mentioned in the two immediately preceding clauses, it meets Criterion 5 for in focus status of the referent. The antecedent, *the evidence of the bends*, is an NP/DP and is located in the same quotation as the pronoun *it*.

Criterion 6: It is the event or activity denoted by the immediately preceding sentence.

(14) In July, Mayer John F. Street gave a televised address in which he pleaded with young people. "**Lay down your weapons**. Do **it** now. Choose education over violence." [24]

In (14), the referent of *it* in *Do it now* was determined to be the activity of laying down weapons, which is an activity denoted by the immediately

preceding sentence and is, therefore, in focus by Criterion 6. The activity is described by the sentence *Lay down your weapons* and occurs in the same quotation as the pronoun *it*.

(15) "We need **to deal as a nation with America's No. 1 health problem**," Ramstad said. "**It**'s not only the right thing to do, but the cost-effective thing to do." [9]

In (15), the referent of *it* in *It's not only the right thing to do* was determined to be the event of dealing as a nation with America's No. 1 health problem. Since the event is denoted by part of the immediately preceding sentence, it meets Criterion 6 for in focus status of the referent. This event/activity is described by the clause *to deal as a nation with America's No. 1 health problem* and is located in a separate quotation.

Not all referents *of it* were classifiable as 'in focus' by the coding protocol. In (16) and (17), the referents of *it* are marked as being indeterminable since they could not be classified as 'in focus' according to the current coding criteria.

(16) Just like when Brown was alive, the raucous throng of thousands cheered and applauded as pallbearers lifted his gold casket and carried it inside, for Brown, who died of heart failure Christmas morning, to lie in repose on the stage where he made his 1956 debut. As New Yorker Norman Brand waited for the procession to arrive, the 55-year-old recalled hearing **Brown's anthem** for the first time in his native Alabama. "**It** really changed the attitude of most black people. It was like a wake up call. Before that, if you were called black, it was like an insult," Brand said. [28]

In (16), the referent of *it* in *It really changed the attitude of most black people* was mentioned in the previous sentence but was marked as indeterminable since it was not possible to classify the referent as in focus according to the coding criteria. The antecedent, *Brown's anthem*, is mentioned in the previous sentence but is not mentioned in a syntactically prominent position.

(17) But they were struck by what else was in the sample: "**Nearly one ton of biological material**," says Walter Michaelis, a biogeochemist at the University of Hamburg in Germany, who led the expedition. "No sediment. No carbonates. **It** was a cubic meter of bacteria!." [22]

In (17), the referent of *it* in *It was a cubic meter of bacteria* was mentioned in a previous sentence but is marked as indeterminable. This referent could not be classified as in focus according to the coding protocol. The antecedent *nearly one ton of biological material* is mentioned in main clause subject position but is not located in the immediately preceding sentence.

4 Results

There was a total of 310 instances of *it* found within quotations in this study. As can be seen in Table 1, 263 instances were referential (84.84%) and 47 were nonreferential (15.16%).

Table 1. Classification of *it* in Quotations

Classification	**Number**	**Percentage**
Referential *It*	263	84.84%
Non-Referential *It*	47	15.16%
Total	310	100%

The antecedent was found somewhere in the corpus in 256 of the 263 instances (97.34%) of referential *it*; 3 out of the 263 (1.14%) instances could not be agreed upon by the coders so they were marked as ambiguous, and 4 out of 263 (1.52%) were marked as not found.

In (18), the referent of *it* could not be agreed upon by the coders. One coder believed the referent to be the event of the two-year legal battle, another the human error in the testing, and the third believed the referent to be the state of affairs described by the test results being incorrect.

(18) But last April, after a two-year legal battle that cost Chreky $800,000, the Fairfax, Virginia, circuit court found that human error in the testing was probable and that the DNA results were incorrect. "**It** hurt my family; my business," Chreky says. "My life will never be the same."

In (19), the referent of *it*, the outcome, could not be located in the corpus.

(19) Bringing diversity into Avida has brought more bad news for those who think complexity cannot evolve. Ofria decided to run the complexity experiment over again, this time with a limit on the supply of numbers. "**It** just floored me," he says. "I went back and checked this so many ways." In the original experiment, the organisms evolved the equals routine in 23 out of 50 trials. But when the experiment was run with a limited supply of numbers, all the trials produced organisms that could carry out the equals routine. [40]

As can be seen in Table 2, the antecedent was marked determinable for each referential *it* in 208 of the 256 instances. This means that using the Givenness Hierarchy framework, an appropriate antecedent could be found for 81.25% of all instances of referential *it* found in quotations using the current coding protocol for cognitive status, specifically in this case for the status 'in focus'. The

Table 2. Classification of Antecedents for Referential *it* in Quotations Using the Givenness Hierarchy

Classification of Antecedent	Number	Percentage
Determinable	208	81.25%
Indeterminable	48	18.75%
Total	256	100%

percentage is comparable to that obtained for referential *it* that is not in quotations. For example, Watters [37] found that 86.58% of all instances of referential *it* not found in quotations could find an appropriate antecedent using the current coding protocol for the cognitive status 'in focus'.[3]

The Coding Protocol only provides sufficient (not necessary) conditions for assigning cognitive status as the status of a referent is not fully determined by formal linguistic properties. It is, therefore, highly likely that the 48 instances of referential *it* whose antecedents were found to be indeterminable cannot be attributated solely (if at all) to the fact that these forms were found in quotations.

5 Conclusion

The results of this study support the conclusion that, using the Givenness Hierarchy Framework, there is sufficient context, either within or outside the quotation in which referential *it* is found, for determining the antecedent of this form; algorithms for resolving the referent of anaphoric *it* thus do not have to ignore, or otherwise include special restrictions on, occurrences of this form found in quotations. Future work is needed to determine whether this also applies to other anaphoric expressions such as demonstrative pronouns and determiners and phrases headed by a definite article. Since the Givenness Hierarchy assumes different constraints (hence different algorithms) for different forms, such studies will have to be conducted separately for each pronominal and determiner form in a given language.

[3] The empirical study performed by Watters [37] looked at 82 instances of referential *it* not found in quotations and 23 instances of referential *it* found in quotations. The criteria were used as defined by the Coding Protocol of Gundel [16] with minor clarifications.

For example, Criterion 2, it was mentioned in the same sentence, was used when the coordinating conjunction 'and' or 'but' was used to begin a new sentence.

Example: There was already a layer of weak sediments on the Norwegian continental slope, and it is probable that an earthquake was what triggered **the slide**. But it's equally probable that the gradual melting of the hydrates made **it** possible–and made it worse.

In the above example, Criterion 2 would find the referent of *it* in *that the gradual melting of the hydrates made it possible* to be the slide. The antecedent *the slide* is an NP/DP that is located in the same sentence as the pronoun *it*.

References

1. Associated Press (2006). Train Derails in Australia. December 12, 2006.
2. Bagga, Amit(1998). 'Evaluation of Coreferences and Coreference Resolution Systems'. In: *Proceedings of the First International Conference on Language Resources and Evaluation (LREC '98)*, May 1998, pp. 563-566.
3. Banbury, Jen (2006). "How to Build a T.Rex". *Discover*, Vol. 27, No. 05, May 2006.
4. Byron, Donna K. and Joel R. Tetreault (1999). 'A Flexible Architecture for Reference Resolution'. In: *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL '99)*, pp. 229-232.
5. Cardie, Claire and Kiri Wagstaff (1999). 'Noun Phrase Coreference as Clustering'. In: *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Association for Computational Linguistics, pp. 82-89.
6. Castaño, José, Jason Zhang, and James Pustejovsky (2002). 'Anaphora Resolution in Biomedical Literature'. In: *Proceedings of the International Symposium on Reference Resolution*, Alicante, Spain.
7. Cherry, Colin and Shane Bergsma (2005). 'An Expectation Maximization Approach to Pronoun Resolution'. In: *Proceedings of the 9th Conference on Computational Natural Language Learning (CoNLL)*, Ann Arbor, June 2005, pp. 88-95.
8. Dimitrov, Marin, Kalina Bontcheva, Hamish Cunningham, and Diana Maynard (2002). 'A Light-weight Approach to Coreference Resolution for Named Entities in Text'. In: *Proceedings of the 4th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, Lisbon, September 18-20.
9. Frommer, Frederic J. (2006). "Mental Health Bill to Face House Vote". Associated Press, December 28, 2006.
10. Furst, Randy (2006). "U, May Sue U.S. to Recover Residents' Social Security Taxes". *Star Tribune*, December, 27, 2006.
11. Ge, Niyu, John Hale, and Eugene Charniak (1998). 'A Statistical Approach to Anaphora Resolution'. In: *Proceedings of the 2nd Workshop on Very Large Corpora*, Montreal, Canada, pp. 161-170.
12. Gundel, Jeanette K. , Nancy Hedberg, and Ron Zacharski (2004). 'Demonstrative Pronouns in Natural Discourse'. In: A. Branco, T. McEnery and R. Mitkov (eds.), *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2004)*, Lisbon: Edições Colibri.
13. Gundel, Jeanette, Nancy Hedberg, and Ron Zacharski (1993). "Cognitive Status and the Form of Referring Expressions in Discourse". *Language*, Vol. 69, No. 2, pp. 274-307.
14. Gundel, Jeanette K., Nancy Hedberg, and Ron Zacharski (2001). "Definite Descriptions and Cognitive Status in English: Why Accommodation is Unnecessary." *Journal of English Language and Linguistics*, 5.2, pp. 273-295.
15. Gundel, Jeanette K. (2003). 'Information Structure and Referential Givenness/Newness: 'How Much Belongs in the Grammar?." In: Stefan Müller (ed) *Proceedings of the HPSG03 Conference*, CSLI Publications.
16. Gundel, Jeanette K. (2004). 'Coding Protocol for Status on the Givenness Hierarchy'. Language and Cognition Class Handout, 2004.
17. Hitt, Jack (2006). "CSI: Jurassic". *Discover*, Vol 27, No. 09, September 2006.
18. Jesdanun, Anick (2006). "Anti-Spam Tool Ceases As Spammers Evolve". Associated Press, December 27, 2006.

19. Keath, Lee (2006). "3 Million Muslims Begin Annual Hajj". Associated Press, December 28, 2006.
20. Kennedy, Christopher and Branimir Boguraev (1996). 'Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser'. In: *Proceedings of the 16th Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 113-118.
21. Kupść, Anna, Teruko Mitamura, Benjamin Van Durme, Eric Nyberg (2004). 'Pronominal Anaphora Resolution for Unrestricted Text'. In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2004).*
22. Kunzig, Robert (2004). "20,000 Microbes Under the Sea". *Discover*, Vol 25, No. 3, March 2004.
23. Limjoco, Victor (2006). "Vampire Weed". *Discover*, Web Exclusives, October 26, 2006.
24. Matthews, Karen (2006). "Murders Up in New York, Other Big Cities". Associated Press, December 27, 2006.
25. Mitkov, Ruslan (2002). *Anaphora Resolution.* Pearson Longman, Studies in Language and Linguistics, 2002.
26. Mitkov, Ruslan (2000). "Towards a More Consistent and Comprehensive Evaluation of Anaphora Resolution Algorithms and Systems". In: *Proceedings of the Discourse Anaphora and Anaphora Resolution Colloquium (DAARC-2000)*, Lancaster, UK., pp. 96-107.
27. Müller, Christoph, Stefan Rapp, and Michael Strube (2002). 'Applying Co-Training to Reference Resolution'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July (2002), pp. 352-359.
28. Neumeister, Larry (2006). "Fans Honor 'Godfather of Soul' at Apollo". Associated Press, December 28, 2006.
29. Ng, Vincent and Claire Cardie (2002). 'Improving Machine Learning Approaches to Coreference Resolution'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, July 2002, pp. 104-111.
30. Robertson, Jordan (2006). "Online Shoppers Overwhelm iTunes Store". Associated Press, December 28, 2006.
31. Salmon-Alt, Susanne and Laurent Romary (2004). 'Data Categories for a Normalized Reference Annotation Scheme'. In: *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC)*, S. Miguel, Azores, Portugal, September 23-24, 2004, pp. 145-150.
32. Shah, Amir (2006). "Karzai: Border Fence Won't Stop Terrorists." Associated Press, December 28, 2006.
33. Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim (2001). "A Machine Learning Approach to Coreference Resolution of Noun Phrases". *Computational Linguistics* 27(4):521-544.
34. Tetreault, Joel R. (1999). 'Analysis of Syntax-Based Pronoun Resolution Methods'. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 602-605.
35. Tetreault, Joel R. (2001). "A Corpus-Based Evaluation of Centering and Pronoun Resolution". *Computational Linguistics*, Volume 27, Issue 4, December 2001, pp. 507-520
36. Torchia, Christopher (2006). "U.S. Military Death Toll in Iraq Exceeds Number of Deaths in 9/11 Attacks." Associated Press, December 25, 2006.

37. Watters, Shana (2006). 'The Givenness Hierarchy and the Pronoun 'It': An Empirical Study Investigating the Cognitive Status of Being 'In Focus' ". Unpublished manuscript, M.A. in Linguistics project paper, University of Minnesota, MN.
38. Weber, Harry R. (2006). "US Airways CEO: Not Upping Delta Offer". Associated Press, December 28, 2006.
39. Yang, Xiaofeng, Guodong Zhou, Jian Su, and Chew Lim Tan (2003). 'Coreference Resolution Using Competition Learning Approach'. In: *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, Sapporo, Japan, pp. 176-183
40. Zimmer, Carl (2005). "Testing Darwin". *Discover*, Vol. 26, No. 02, February 2005.

Applying Backpropagation Networks to Anaphor Resolution

Roland Stuckardt

Johann Wolfgang Goethe University Frankfurt
Im Mellsig 25, D-60433 Frankfurt am Main, Germany
roland@stuckardt.de
http://www.stuckardt.de/

Abstract. Despite some promising early approaches, neural networks have by now received comparatively little attention as a machine learning model for robust, corpus-based anaphor resolution. The work presented in this paper is intended to fill the apparent gap in research. Based on a hybrid algorithm that combines manually knowledge-engineered antecedent filtering rules with machine-learned preference criteria, it is investigated what can be achieved by employing backpropagation networks for the corpus-based acquisition of preference strategies for pronoun resolution. Thorough evaluation will be carried out, thus systematically addressing the numerous experimental degrees of freedom, among which are sources of evidence (features, feature vector signatures), training data generation settings, number of hidden layer nodes, and number of training epochs. According to the evaluation results, the neural network approach performs at least similar to a decision-tree-based ancestor system that employs the same general hybrid strategy.

1 Introduction

Triggered by pioneering work in the nineties, the research on robust, operational anaphor resolution has seen a rapid progress in the last decade. Among the knowledge-poor approaches that operate on noisy data are rule-based as well as machine-learning-based systems.[1] A closer analysis reveals that the majority of the corpus-based approaches employs decision trees[2] or Naïve Bayes classifiers[3]. According to the recent survey by Olsson ([13]), there is only the early work of Connolly et al. ([6]) that investigates neural networks as a device for coreference resolution.

Notably, the research of Connolly et al. ([6]) gave evidence that neural networks, employed as classifiers making coreference predictions for instances of object (NP) anaphora, yield better results than other, less complex ML techniques,

[1] Among important recent work are the manually designed approaches of Lappin and Leass ([1]), Kennedy and Boguraev ([2]), Baldwin ([3]), Mitkov ([4]), Stuckardt ([5]) and the machine-learning-based approaches Connolly et al. ([6]), Aone and Bennett ([7]), Ge et al. ([8]), Soon et al. ([9]), Stuckardt ([10]).

[2] e. g., Ng and Cardie ([11]), Soon et al. ([9].), Aone and Bennett ([7])

[3] e. g., Ng and Cardie ([12]), Ge et al. ([8]).

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 107–124, 2007.

among which are Naïve Bayes and Posterior classifiers; regarding pronominal anaphors, formal evaluation indicated that neural networks might even outperform decision trees. In light of these promising early results, the question arises why neural networks have been largely neglected by subsequent research and, in particular, why the majority of approaches to ML-based anaphor and coreference resolution focused on decision trees or Naïve Bayes techniques. Moreover, there have been recent successful applications of neural networks to the problem of modeling the choice of referential expressions (e. g., Grüning and Kibrik, [14]). This too hints towards a closer examination of neural networks for anaphor resolution, since the issues of generation and interpretation can be regarded to be closely related: if, in a certain context, the model of referential choice predicts the usage of a *pronominal* expression for mentioning a particular discourse referent, this might as well be interpreted as evidence for choosing the discourse referent as antecedent for a pronoun occurring in this context.

The work presented below is intended as a first step towards closing this apparent gap in research. While chiefly comparing different machine learning models with respect to the application case of anaphor resolution, Connolly et al. ([6]) neglected a bunch of further important issues, among which are the empirical fine-tuning of the neural network learning parameters, the strategy employed for training data generation, and the sources of evidence to be taken into account, or how to optimally integrate machine-learned classifiers into a fully-fledged anaphor resolution algorithm (see Mitkov, [15]). However, in order to obtain expressive evaluation results that properly compare with the results of other state-of-the-art approaches, these points should be addressed as well.

Two previous studies are taken as the points of departure: [5], describing ROSANA, a classical salience-based and manually knowledge-engineered algorithm for robust pronominal anaphor resolution, and [10], describing the descendant system ROSANA-ML, in which the salience-based preference rankings are substituted by classifiers that are automatically acquired through C4.5 decision tree learning. In the current investigation, a system ROSANA-NN will be designed and evaluated that employs the same general algorithm as ROSANA-ML, but uses neural networks instead of decision trees for antecedent candidate ranking. Thus, as elaborated in [10], the conceptually clean distinction between domain- and genre-independent restrictions and at least partly genre-specific antecedent selection preferences provides an adequate base for the focused application of machine-learned (here: neural network) classifiers as part of a hybrid strategy in which the universal filtering criteria remain manually engineered. While the experiments thus consider the task of robust pronominal anaphor resolution, the general scope of the conducted research is much broader. The fundamental strategy for integrating anaphor interpretation criteria as well as the neural network learning framework developed below apply to the great majority of anaphora types. The paper should thus be conceived as contributing much more than yet another robust pronoun resolver.

The presentation of the work is organized as follows: section 2 provides a general description of the methodology as well as the algorithms and systems

used for training data generation, neural network learning, and classifier application. In section 3, the different experimental stages to be carried out are identified; clearly, they are directly related to the plethora of configuration options of the type of classifiers (here: neural networks trained by a backpropagation algorithm) to be learned. Employing this experimental framework, section 4 then presents the evaluation results and the empirical findings. In section 5, the results of ROSANA-NN will be compared with the results of competing approaches, looking at its decision-tree-based and manually knowledge-engineered ancestors ROSANA-ML and ROSANA as well as at the work of Connolly et al. ([6]). Finally, in section 6, conclusions are drawn and directions of further research are identified.

2 Methodology and Algorithms

According to the employed neural-network-based machine-learning approach, two phases are distinguished. (1) During the *training phase*, based on a training text corpus, a set of feature vectors is generated which consists of feature tuples derived from the *(anaphor, antecedent candidate)* pairs that are still considered during the antecedent *selection* phase of the anaphor resolution algorithm, i. e. pairs that have passed all (strict) antecedent filtering criteria. By employing intellectually gathered key data, these vectors are then classified as either cospecifying or non-cospecifying. In the classifier learning step proper, these training cases are submitted to Mitchell's implementation of the backpropagation algorithm [16],[4], which, by employing a gradient descent learning strategy and a feedforward technique, iteratively adjusts the weights of a multi-layer neural network with the goal to converge towards a classifier properly fitting the training data and suitable for accurately categorizing unseen feature vectors that are of the same signature as the training vectors. (2) In the *application (anaphor resolution) phase*, the learned classifiers are employed for antecedent selection: to discern between more and less plausible candidates, instead of salience factors, neural network classifiers are applied. Thus, as initially motivated, the basic strategy consists in learning the preference criteria only, thus resorting to classical rule-based robust implementations of the antecedent filtering strategies, among which are syntactic disjoint reference and number/gender agreement.

Two algorithms are hence distinguished: (a) the feature vector generation algorithm, which is employed during the training phase, and (b) the anaphor resolution algorithm proper, which specifies the general strategy of the application phase.

2.1 Feature Vector Generation

In figure 1, the specification of the feature vector generation algorithm is given. Step 1, in which different kinds of restrictions for eliminating impossible

[4] See chapter 4 of [16]; the backpropagation implementation has been taken from the webpage `http://www.cs.cmu.edu/~tom/mlbook.html` (December 2004).

1. *Candidate Filtering*: for each anaphoric NP α, determine the set of admissible antecedents γ:
   (a) verify morphosyntactic (number and gender) or lexical agreement with γ;
   (b) if the antecedent candidate γ is intrasentential: apply the robust syntactic disjoint reference filter as specified in [5], figure 4.
2. *Feature vector generation*: for each remaining anaphor-candidate pair (α_i, γ_j):
   (a) generate, according to the feature signature σ under consideration, the feature vector
   $$fv(\alpha_i, \gamma_j) := (n_{\alpha_i}, n_{\gamma_j}, f_1, \ldots, f_{k_\sigma}).$$
   where n_{α_i} and n_{γ_j} are the numbers (unique identifiers, referred to in the key) of the occurrences α_i and γ_j, and $f_1, \ldots, f_{k_\sigma}$ are (individual and relational) features derived from α_i and γ_j with respect to the signature σ;
   (b) write $fv(\alpha_i, \gamma_j)$ to an external training data file.

Fig. 1. ROSANA-NN: feature vector generation

antecedents (in particular, agreement in person/number/gender and syntactic disjoint reference) are applied, is identical with the antecedent filtering phase of the manually designed ROSANA algorithm. In step 2, however, during feature vector generation, the salience ranking of the antecedent candidates is substituted by the mapping of each remaining anaphor-candidate pair (α_i, γ_j) to a feature vector $fv(\alpha_i, \gamma_j)$, the attributes $f_1, \ldots, f_{k_\sigma}$ of which comprise individual and relational features derived from the descriptions of the occurrences α_i and γ_j. The *signature* of the feature vectors, i. e. the inventory of features to be taken into account, has to be chosen carefully in order to fulfill the conditions of robust processing: instead of requiring complete and unambiguous descriptions, they should be computable from potentially partial representations such as fragmentary syntactic parses. (See section 4.1.)

2.2 Anaphor Resolution

The specification of the ROSANA-NN anaphor resolution algorithm proper is given in figure 2. Again, step 1 is identical with the antecedent filtering phase of the manually designed ROSANA algorithm. Step 2, however, is modified. For a particular instance (α_i, γ_j) of anaphor and antecedent candidate, after the computation of the feature vector $fv(\alpha_i, \gamma_j)$, a learned neural network classifier, which might depend upon the the particular type of anaphor to be resolved, is consulted:[5] basically, its result $\Psi_\sigma^{type(\alpha_i)}(fv(\alpha_i, \gamma_j))$ consists in a prediction $\in \{COSPEC, NON\_COSPEC\}$.[6] In the subsequent step, these predictions

[5] By now, due to technical reasons, the classifier application has not been technically integrated with the ROSANA-NN implementation; rather, the consultation is accomplished by looking up externally precomputed classification results. However, the implementation yields outcomes equivalent to those of a fully integrated system.

[6] To put it formally: a *classifier function* $\Psi_\sigma^{type(\alpha_i)} : A_1 \times A_2 \times \ldots \times A_{k_\sigma} \mapsto \{COSPEC, NON\_COSPEC\}$ is applied that maps instances of the underlying signature σ to cospecification/non-cospecification predictions.

1. *Candidate Filtering*: for each anaphoric NP α, determine the set of admissible antecedents γ:
   (a) verify morphosyntactic (number and gender) or lexical agreement with γ;
   (b) if the antecedent candidate γ is intrasentential: apply the robust syntactic disjoint reference filter as specified in [5], figure 4.
2. *Candidate scoring and sorting*:
   (a) for each remaining anaphor-candidate pair (α_i, γ_j):
      i. *consultation of the neural network classifier*: determine the prediction $\Psi_\sigma^{type(\alpha_i)}(fv(\alpha_i, \gamma_j))$ of the learned neural network classifier with respect to the instance $fv(\alpha_i, \gamma_j)$.
   (b) for each anaphor α: sort candidates γ_j according the following criteria:
      – *primary*: candidates γ_j for which $\Psi_\sigma^{type(\alpha)}(fv(\alpha, \gamma_j)) = COSPEC$ are preferred over candidates $\gamma_{j'}$ for which $\Psi_\sigma^{type(\alpha)}(fv(\alpha, \gamma_{j'})) = NON\_COSPEC$;
      – *secondary*: surface nearness.
   (c) sort the anaphors α according to the above criteria applied to their respective best antecedent candidates.
3. *Antecedent Selection*: consider anaphors α in the order determined in step 2c. Suggest antecedent candidates $\gamma_j(\alpha)$ in the order determined in step 2b. Select $\gamma_j(\alpha)$ as candidate if there is no interdependency, i. e. if
   (a) the morphosyntactic features of α and $\gamma_j(\alpha)$ are still compatible,
   (b) for all occurrences $\delta_{\gamma_j(\alpha)}$ and δ_α the coindexing of which with $\gamma_j(\alpha)$ and (respectively) α has been determined in the *current* invocation of the algorithm: the coindexing of $\delta_{\gamma_j(\alpha)}$ and δ_α, which results transitively when choosing $\gamma_j(\alpha)$ as antecedent for α, does neither violate the binding principles nor the i-within-i condition. (see the full specification in [5], figure 4)

Fig. 2. ROSANA-NN: anaphor resolution through backpropagation networks

are employed for ranking the candidate sets of each anaphor: candidates which are (heuristically) classified to cospecify with the anaphor rank higher than candidates that are (heuristically) predicted as non-cospecifying; surface nearness (i. e. word distance) serves as the secondary criterion.[7] There is a final step 3 in which antecedents are selected. The remaining candidates are considered in the order determined by the ranking step; further means are taken to avoid combinations of antecedent decisions that are mutually incompatible (see [5]).

3 Layout of Experiments

3.1 Experimental Degrees of Freedom

There are various experimental degrees of freedom that should be considered:

1. the *sources of evidence (features, feature vector signatures)* upon which to classify a given pair (α_i, γ_j) of anaphor and antecedent candidate;

[7] Among the possible refinements are: further ranking the candidates according to the real value ε yielded by the neural network classification result lookup (see section 4.1), or eliminating candidates which are (heuristically) classified as not cospecifying.

2. the *techniques employed for encoding the input and output space* of the network;
3. the *number κ of internal notes* making up the (here) single hidden neural network layer; this number should be chosen large enough in order to enable the network to learn all relevant regularities of the data space to be modeled; on the other hand, it should not be chosen too large as this might result in an unwanted overfitting of the particular sample data;[8]
4. the *parameters learning rate (η) and momentum (ζ) of Mitchell's backpropagation algorithm* (see [16], p97ff): setting them to low values will drastically slow down network convergence, while choosing them too large might result in missing the sought-for empirical optimum;
5. the *number τ of training epochs*, which should be chosen suitably in order to achieve convergence towards the training data without overfitting them;
6. the settings that determine the *distribution of the training data (data generation mode)*, i. e. the way how positive and negative sample cases are generated based on referentially annotated corpora; this should be addressed by taking into account
7. *the particular way how the classifiers are employed by the anaphor resolution algorithm*, as this determines the distribution of cases relevant during extrinsic classifier application;
8. whether *one general or several anaphor-type-specific classifiers* should be learned.

Thus, there are considerably more dimensions along which one might vary the experimental settings than in case of decision trees (see [10]).

Moreover, in order to obtain results independent of a particular partition of the annotated data into training, validation, and test cases, cross-validation should be carried out:

- at *intrinsic (learned classifier) level*, determining the classifiers' accuracy regarding their predictions $\in \{COSPEC, NON\_COSPEC\}$;
- at *extrinsic (application) level*, determining the anaphor resolution results obtained with the classifiers.

3.2 Annotated Text Corpus and Disciplines of Formal Evaluation

The training and evaluation of the ROSANA-NN system will be carried out on a corpus of 53 referentially annotated news agency press releases, comprising

[8] In general terms, the descriptional capabilities of the representational model and the number of training data should be kept in relation. Regarding neural networks, besides the size of the hidden layer, the chosen feature vector signature as well as the employed encoding strategy determine the number of nodes and, thus, the potential descriptional power of the network. Clearly, the larger the network, the more training data should be available in order to ensure convergence towards a classifier that appropriately generalizes. Notably, this issue of *data fragmentation* is frequently neglected; Ng and Cardie ([11]) mention it briefly with respect to decision tree classifiers, for which they identify the desideratum that each leaf of the learned decision tree should cover roughly the same minimum number of training data instances.

24,886 tokens, 332 third-person non-possessives, and 212 third-person possessive pronouns. In order to support cross-validation, this corpus d_1^{53} has been randomly partitioned into six document sets $ds_i, 1 \leq i \leq 6$ of approximately equal size. In all experiments, the training data generation and the application of the trained system take place on potentially noisy data, i. e. without a-priori intellectual correction of orthographic or syntactic errors, and without any post-editing of the possibly partial or incorrect parses derived by the robust syntactic preprocessor, which is the FDG parser for English of Järvinen and Tapanainen ([17]).

The anaphor resolution performance will be evaluated with respect to two evaluation disciplines. In the *immediate antecedent* (*ia*) discipline, the classical accuracy measure is employed that determines the precision of correct immediate antecedent choices; by further taking into account cases of unresolved pronouns, the respective recall measure is obtained.[9] In the *non-pronominal anchors* (*na*) discipline, antecedents are required to be common or proper nouns, which is particularly relevant for anaphor resolution applications; again, it is distinguished between precision and recall. Thus, the anaphor resolution performance is measured according to the tradeoffs (P_{ia}, R_{ia}) and (P_{na}, R_{na}). For formal definitions and an in-depth discussion of the two measures, the reader is referred to [5].

4 Experiments and Empirical Results

In order to deal with the issues identified in section 3.1, the experiments will be divided into two stages: stage 1, addressing signature optimization, network i/o encoding, and a first, coarse narrowing-down of the data generation settings; stage 2, addressing the issues of identifying the most promising combinations of data generation mode and number of hidden layer nodes, determining the respective empirically optimal numbers of training epochs, and intrinsically as well as extrinsically evaluating the learned classifiers' performance. In order to limit evaluation efforts, cross-validation will be confined to stage 2; regarding signature and i/o encoding, relative empirical performance is expected to be virtually independent from the particular (training, evaluation) data partitioning.

Two of the experimental degrees of freedom identified in section 3.1 will not be considered in detail. The question whether to use one general or two type-specific classifiers for non-possessive vs. possessive pronouns has been settled in favour of the latter option, taking into account that the ROSANA-ML experiments have brought evidence that type-specific classifiers might yield slightly better results if combined with appropriate training data generation strategies (see [10]). Likewise, first experiments have indicated that the backpropagation parameters of learning rate (η) and momentum (ζ) should be kept best at their original settings ($\eta = 0.3$ and $\zeta = 0.3$, see [16], p97ff).

[9] Under the assumption that *all* pronouns are resolved, the precision measure is equivalent to the accuracy measure employed for evaluating classical approaches of, e.g., Lappin and Leass ([1]) and Kennedy and Boguraev ([2]). By allowing for unresolved pronouns, a (P, R) tradeoff is obtained, which corresponds to the evaluation measure employed by Aone and Bennett ([7]).

4.1 Stage 1: Data Generation, Signatures, and I/O Encodings

At the first stage of experiments, the number κ of internal (hidden layer) nodes is set to the fixed value of 20. Each training run is confined to $\tau = 160$ training epochs. The goal consists in determining a promising subset of signatures and data generation modes to be evaluated in full detail at the second experimental stage, where the parameters κ and τ will then be reconsidered.

Data Generation Modes. Six data generation modes are considered, four of which have already been investigated in the ROSANA-ML decision tree experiments. The data generation modes differ with respect to the subset of (α, γ) occurrence pairs used for generating positive (classified as COSPEC) and negative (classified as NON_COSPEC) training cases:

- *standard*: the set of antecedent candidates γ to be paired with a particular anaphor α for generating training vectors $fv(\alpha, \gamma)$ is identical with the set of candidates considered by the ROSANA-NN anaphor resolution algorithm in its canonical configuration; recency filters, which depend upon the type of anaphor to be resolved, apply.
- *no cataphors*: in this case, the same recency filters as under the *standard* setting apply; however, instances of backward anaphora (γ surface-topologically *following* α) are not considered as training samples.[10]
- *no recency filter*: recency filters of the standard setting are switched off; thus, since *all* candidates preceding the anaphor and fulfilling the further filters give rise to a training case, the resulting training set is significantly enlarged; while, from a learning-theoretical point of view, it is considered adequate to mirror the application case distribution as close as possible, this strategy might nevertheless prove reasonable in the (common) case of training data sparsity.
- *no cataphors, no recency filter*: combines the *no cataphors* and *no recency filter* settings.
- *SNL*: a training data generation strategy successfully applied by Soon et al. ([9]); for each anaphor α, at most one positive sample is included in the training set, viz., the feature vector constructed over α and (as far as existent) its *surface-topologically nearest* cospecifying antecedent γ^N; negative samples are constructed by taking into account all (non-cospecifying) occurrences surface-topologically situated between γ^N and α.
- *NC*: a strategy successfully applied by Ng and Cardie ([11], [12]); as in mode *SNL*, the lookback is restricted by a particular cospecifying antecedent γ^{NP}, which, however, this time is required to be *non-pronominal*; any occurrence between α and γ^{NP} gives rise to a further (here: negative or positive) sample;

[10] In fact, the version of the ROSANA-NN algorithm put under scrutiny below also employs a *no cataphors* setting, which might thus be considered as the standard setting proper. However, in order to facilitate comparison, the terminology has been kept identical to the terminology employed in the publications describing the ROSANA-ML results.

thus, the data sets derived by applying mode *NC* subsume the respective data sets constructed under mode *SNL*.[11]

Features and Signatures. The most fundamental question regards the set of attributes, i. e. the signature of the feature vectors from which the classifiers will be learned. The choice is confined to sources of evidence available in the considered environment of robust, knowledge-poor processing. Figure 3 displays the respective inventory of attributes taken into account during the following experiments. *type(o)* denotes the type of the respective occurrence o, in particular

Feature	Examples of Instances	Description	σ_{DT}	σ_b	σ_c	σ_d	σ_e	#IN
type(α)	PER3, POS3	type of anaphor α	•	•	•			16
synfun(α)	subje, trans	syntactic function of α	•	•		•	•	16
synlevel(α)	TOP, SUB	syntactic position of α	•	•	•	•	•	3
number(α)	SG, PL, SGPL	number of α	•	•	•		•	2
gender(α)	MA, FE, MAFE	gender of α	•	•	•		•	3
type(γ)	NAME, PER3	type of candidate γ	•	•		•	•	16
synfun(γ)	subje, trans	syntactic function of γ	•	•		•	•	16
synlevel(γ)	TOP, SUB	syntactic position of γ	•	•	•	•	•	3
number(γ)	SG, PL, SGPL	number of γ	•	•	•		•	2
gender(γ)	MA, FE, MAFE	gender of γ	•	•	•		•	3
dist(α,γ)	INTRA, PREV	sentence distance	•	•	•	•	•	3
dir(α,γ)	ANA, KATA	resumption direction	•	•	•			1
synpar(α,γ)	YES, NO	syntactic role identity?	•	•		•	•	1
syndom(α,γ)	$[\alpha \rightarrow \gamma]$, $[\gamma \rightarrow \alpha]$, no	synt. dominance relation	•	•	•	•	•	3
subject(α)	YES, NO	anaphor α is subject?		•	•	•	•	1
subject(γ)	YES, NO	candidate γ is subject?		•	•	•	•	1
pronoun(γ)	YES, NO	candidate γ is pronoun?		•	•	•	•	1
thenp(γ)	YES, NO	candidate γ is definite NP?		•	•	•	•	1
prostr(α,γ)	YES, NO	α, γ string-id. pronouns?		•	•	•	•	1
synpar*(α,γ)	SuSP, ObSP, NoSP	weak syntactic role identity		•	•	•	•	3
Σ (#IN)			88	96	47	69	79	96

Fig. 3. Inventory of features over which the signatures are defined

PER3/POS3 (third person non-possessive/possessive pronouns), VNOM (common noun phrases), and NAME (proper nouns); regarding the anaphor ($o = \alpha$), the choice is restricted to PER3 and POS3 in the current experiments. The feature *synfun(o)* describes the syntactic function of o. *synlevel(o)* captures a coarse notion of (non-relational) syntactic prominence[12], which is measured by counting the number of principal categories[13] occurring on the path between o and the

[11] Originally, this mode was employed for common NP anaphor resolution.

[12] In contrast to *relational* notions of syntactic prominence, in which the relative position to the other occurrence is taken into account (e.g. *c-command*).

[13] To put it more formally: nodes that, in the sense of the Government and Binding (GB) theory, constitute *binding categories* (see [18]).

root of the respective parse fragment. Features *number(o)* and *gender(o)* capture the respective morphological and lexical characteristics of anaphor α and candidate γ. Furthermore, some relational features are considered: *dist(α,γ)* (sentence distance, distinguishing between three cases: same sentence, previous sentence, two or more sentences away), *dir(α,γ)* (whether γ topologically precedes α or vice versa), *synpar(α,γ)* (identity of syntactic function)[14], and *syndom(α,γ)* (relative syntactic position of the clauses of anaphor and candidate in case of intrasentential anaphora)[15]. These 14 features constitute the signature σ_{DT} that was found to perform best in the ROSANA-ML decision tree experiments (see [10]).

Motivated by the approaches of Ng and Cardie ([11]) and Soon et al. ([9]), this original inventory of ROSANA-ML attributes has been supplemented by six additional promising features dealing with *pronominal* anaphora and deemed relevant for the acquisition of *preference* strategies.[16] The choice is restricted to those attributes that are computable based on the knowledge made available by the robust preprocessors currently used in the ROSANA-NN framework. This excludes from consideration a bunch of features that deal with semantic class information provided by WordNet.[17] Moreover, in the context of pronominal anaphor resolution, some other attributes, such as BOTH_PRONOUNS by [9] or STR_MATCH by [11] boil down to more trivial features, now dealing merely with the antecedent or with pronoun string identity. Proceeding along these lines, the following six features have been added, five of which are binary: *subject(o)*, capturing whether anaphor/candidate occurs in the subject role; *pronoun(γ) / thenp(γ)*, describing whether the candidate is a pronoun or, respectively, a definite NP; the relational feature *prostr(α,γ)*, capturing whether candidate as well as anaphor are pronouns with identical surface form; finally, the ternary relational feature *synpar*(α,γ)* has been introduced, which models a weak version of syntactic parallelism, distinguishing three cases: both anaphor and candidate are subjects; neither of them is a subject; exactly one of them is a subject.

I/O Encodings. According to figure 3, all considered features take values from a particular finite set. Attributes taking only two values, such as *dir(α,γ)* or *synpar(α,γ)*, are binarily encoded by a single network input, assigning an input of 0.1 (TARGET_LOW) in one case and 0.9 (TARGET_HIGH) in the other; attributes defined over a set of more than two values are encoded in an unary way, thus, for instance, resulting in 16 inputs modeling the *synfun(o)* features as 16 syntactic roles are distinguished, activating exactly one input for a given

[14] Thus immediately capturing the role inertia information that has been found to be useful in the classical, manually designed approaches [1,5].

[15] e.g. $[\alpha \rightarrow \gamma]$ describes the case in which the clause of γ is syntactically subordinated to the clause of α.

[16] Ng and Cardie ([11]) and Soon et al. ([9]) have a more general scope as they are dealing with common and proper noun anaphora as well, and they are aiming at learning general coreference resolution strategies, including antecedent *filtering* criteria.

[17] Moreover, in the approaches of Ng and Cardie ([11]) and Soon et al. ([9]), this source of evidence primarily addresses cases in which the available semantic information is non-trivial, i. e. cases in which both anaphor and candidate are *non-pronominal.*.

case and deactivating all other inputs; in the special case of possibly ambiguous attributes such as *number(o)* and *gender(o)*, a canonical powerset encoding scheme is employed. In the rightmost column of figure 3, the number of inputs resulting for the different attributes under this encoding scheme are shown.

In the case of anaphor or coreference resolution, the output encoding happens to be a trivial matter, as the prediction space of the backpropagation network consists only of two elements: *cospecifying* and *non-cospecifying*. During training, the value 0.9 is used to encode the former case, while the value 0.1 is used to model the latter case. During network application, output values > 0.5 are thus interpreted as COSPEC predictions, while values ≤ 0.5 are considered to predict NON_COSPEC cases.

Evaluation Results. Distinguishing between classifiers for third person non-possessive and possessive pronouns, intrinsic classifier accuracies have been determined for each combination of the above-defined six data generation modes and five feature signatures; As no cross-validation shall be carried out at experimental stage 1, considerations are restricted to the particular (training, test) data partition $[d_1^{53} \setminus ds_6, ds_6]$.

It would be beyond the scope of the paper to discuss the results for the $2 \cdot 5 \cdot 6 = 60$ combinations in full detail. However, there are some important findings that shall be briefly summarized as they give rise to focus on one particularly promising signature and three auspicious data generation modes at experimental stage 2. Considerations shall be restricted to signature σ_e and modes *SNL*, *NC*, and *-ca (no cataphors)*, as it turned out that: (1) with only one exception, modes *SNL* and *NC* yield the relatively highest C accuracies; (2) (only) for signatures σ_{DT}, σ_b, and σ_e, both *SNL* and *NC* achieve an outstanding C accuracy well above the 50% level; thus, if one suspects that high C accuracy might be relevant for improving extrinsic (anaphor resolution) performance, these three signatures and two modes are on the short list; (3) regarding the important case of PER3 pronouns, the combination of σ_e and *SNL* achieves a particularly high C accuracy of 0.68; thus, it has been decided to focus on signature σ_e in the subsequent experiments;[18] finally, in order to cover classifiers biased towards C $\cup$ N accuracy, mode *-ca* shall be considered as well as it exhibits high C $\cup$ N accuracy while still yielding a reasonable C accuracy.

4.2 Stage 2: Internal Nodes, Training Epochs, and Cross-Validation

With the goal of systematically narrowing down the remaining set of experimental options, stage 2 deals with: (1) optimizing the number τ of training epochs, which was provisionally limited to 160 at stage 1; (2) selecting an appropriate number κ of internal (hidden layer) nodes, which was provisionally set to the fixed value of 20 at stage 1. In this context, the requirement *systematically* amounts to empirically optimizing the parameters τ and κ based on intrinsic cross-validation runs. Eventually, a small number of particularly promising

[18] Using different signatures for PER3 and POS3 classifiers might be a further option.

configurations shall be identified and subjected to the ultimate discipline of (3) extrinsic (ROSANA-NN anaphor resolution) cross-validation.

Intrinsic Cross-Validation: Training Epochs (τ), Internal Nodes (κ) The issue of avoiding overfitting the training data shall be addressed by an intrinsic cross-validation approach according to which τ is set to the average $\tau^* = \frac{1}{6}\sum_{i=1}^{6}\tau_i$ over the six data partitions. Thus, there are actually two substages of intrinsic cross-validation: the first one employed to determine average values τ^*; the second one carried out to determine the intrinsic classifier performance.

Separate experiments shall be carried out for each combination of data generation mode and considered number κ of internal network nodes. Hidden layers of three sizes will be considered: $\kappa \in \{20, 30, 40\}$. As it is further distinguished between non-possessives vs. possessives and C ∪ N vs. C accuracy, $3 \cdot 3 \cdot 2 \cdot 2 = 36$ intrinsic cv experiments are carried out at both substages.

Figure 4 displays the results of the first substage of intrinsic cross-validation, viz., the average numbers τ^* of epochs to train before a worsening of the respective ((C ∪ N) or C) accuracy on the test data indicates that the learned network begins to overfit the training data. As the above experiments indicated that 160

PER3	$\kappa = 20$		$\kappa = 30$		$\kappa = 40$	
	$\tau^*_{C\cup N}$	τ^*_C	$\tau^*_{C\cup N}$	τ^*_C	$\tau^*_{C\cup N}$	τ^*_C
-ca	260	480	100	340	80	440
SNL	100	560	80	740	100	680
NC	60	700	60	500	60	300

POS3	$\kappa = 20$		$\kappa = 30$		$\kappa = 40$	
	$\tau^*_{C\cup N}$	τ^*_C	$\tau^*_{C\cup N}$	τ^*_C	$\tau^*_{C\cup N}$	τ^*_C
-ca	300	340	200	460	140	500
SNL	40	700	40	500	40	540
NC	160	260	40	240	20	280

Fig. 4. PER3 and POS3 classifiers: average values τ^* (cross-validated)

training cycles might in general be insufficient, the considered interval of epochs is enlarged to $0 \leq \tau \leq 1000$. These results confirm the tendency observed at the above experimental stage 1: without exception, it takes more training cycles to converge towards a classifier with high C accuracy (columns τ^*_C) than towards a classifier with high C ∪ N accuracy (columns $\tau^*_{C\cup N}$), and with only one exception (POS3, *-ca*, $\kappa = 20$), the difference regarding the appropriate number of training cycles is considerable. Moreover, the above preliminary upper bound of $\tau \leq 160$ turns out to be too small to learn empirically optimal classifiers biased towards high C accuracy.

Turning towards substage 2, viz., intrinsic cross-validation proper, the figures obtained for the overall 36 experiments generally confirm the results obtained at stage 1 on the particular (training, test) data partition $[d_1^{53} \setminus ds_6, ds_6]$. For each pronoun type, intrinsic cross-validation results of four particularly promising *(dgm, κ)* combinations are displayed in figure 5;[19] these are the combinations that will be further subjected to extrinsic cross-validation below. Regarding non-possessives, combination *a* is considered because of its outstanding C ∪ N accuracy, while *b* and *c* are chosen because of their high C accuracy; combination *d*

[19] Due to space limitations, results on the particular data partitions are not included.

	Setting	DGM	κ	τ^*	$A_{C\cup N}$	A_C
PER3	*a*	*-ca*	40	80	**0.89**	0.44
	b	*SNL*	30	740	0.85	**0.54**
	c	*NC*	20	700	0.86	**0.62**
	d	*-ca*	40	440	0.87	**0.52**

	Setting	DGM	κ	τ^*	$A_{C\cup N}$	A_C
POS3	*A*	*-ca*	40	140	**0.88**	0.51
	B	*SNL*	30	500	0.81	**0.59**
	C	*NC*	20	260	0.83	**0.58**
	D	*SNL*	30	40	**0.86**	0.45

Fig. 5. PER3 and POS3 classifiers: results of intrinsic cross-validation

is included because it promises a relatively high C accuracy which is expected to be accompanied by relatively high C ∪ N results. Concerning the possessive pronouns, combinations *B* and *C* are considered because of their outstanding C accuracy, while *A* and *D* are selected because they promise high C ∪ N results which are expected to come with a still relatively high C performance.

According to the results, some of the combinations seem to be unattractive as they are quantitatively outperformed by one of their competitors; this holds with respect to the PER3 *b* setting, which is majorized by PER3 *c*, and with respect to POS3 *D*, which is majorized by POS3 *A*. However, beyond the merely quantitative aspects, there are the qualitative issues of data distribution: classifiers should perform well on the particular subset of cases most relevant for anaphor resolution. Hence, these settings shall be further considered anyway, as they might differ substantially regarding the distribution of the correctly classified cases. In fact, one criterion governing the selection of the above combinations has been to consider each data generation mode (*-ca*, *SNL*, and *NC*) and each optimization criterion (A_C and $A_{C\cup N}$) at least once for both pronoun types.

Extrinsic Cross-Validation. In figures 6 and 7, the results of the 6-fold extrinsic (anaphor resolution) cross-validation experiments are displayed. The tables show the results for each particular setting in the discipline of immediate antecedency (*ia*), given as tradeoffs (P_{ia}, R_{ia}).[20] Since, in all but one case, every PER3 and POS3 pronoun is assigned an antecedent, the P_{ia} and R_{ia} figures are virtually identical; thus, in effect, under the current configuration of ROSANA-NN, they boil down to the canonical accuracy measure as described above in section 3.2.[21] As it turned out that the relative performance ranking of the different settings in the *ia* discipline is identical with the performance ranking in the non-pronominal anchors (*na*) discipline, the (P_{na}, R_{na}) tradeoffs are not

[20] As ROSANA-NN intertwines PER3 and POS3 resolution, there might be interdependencies between these two subprocesses, which implies that it is impossible to extrinsically evaluate them separated from each other. In fact, the results given for the non-possessives have been obtained by employing the classifiers pertaining to settings *a*, *b*, *c*, *d* together with the setting *A* classifier as the *standard classifier for possessives*; likewise, for evaluating the possessive settings *A*, *B*, *C*, *D*, the classifier pertaining to setting *a* has been chosen as the *standard classifier for non-possessives*.

[21] Document set ds_4 gives rise to a single exception as it contains an instance of the non-possessive pronoun "*them*" occurring near the beginning of a document for which an antecedent fulfilling the congruence constraint could not be found because the single correct *pl* candidate was erroneously assigned the Number attribute *sg*.

	(P_{ia}, R_{ia})			
	a	b	c	d
(ds1) $[d_1^{53} \setminus ds_1, ds_1]$	$(0.58, 0.58)$	$(0.58, 0.58)$	$(0.53, 0.53)$	$(0.63, 0.63)$
(ds2) $[d_1^{53} \setminus ds_2, ds_2]$	$(0.64, 0.64)$	$(0.59, 0.59)$	$(0.59, 0.59)$	$(0.69, 0.69)$
(ds3) $[d_1^{53} \setminus ds_3, ds_3]$	$(0.67, 0.67)$	$(0.58, 0.58)$	$(0.63, 0.63)$	$(0.60, 0.60)$
(ds4) $[d_1^{53} \setminus ds_4, ds_4]$	$(0.71, 0.70)$	$(0.67, 0.66)$	$(0.70, 0.69)$	$(0.63, 0.63)$
(ds5) $[d_1^{53} \setminus ds_5, ds_5]$	$(0.59, 0.59)$	$(0.59, 0.59)$	$(0.55, 0.55)$	$(0.59, 0.59)$
(ds6) $[d_1^{53} \setminus ds_6, ds_6]$	$(0.63, 0.63)$	$(0.61, 0.61)$	$(0.61, 0.61)$	$(0.57, 0.57)$
(ds1-6): weighted avg.	$(\mathbf{0.64}, \mathbf{0.64})$	$(0.60, 0.60)$	$(0.60, 0.60)$	$(0.62, 0.61)$

Fig. 6. PER3 classifiers, 6-fold extrinsic (anaphor resolution) cv (*ia* discipline)

	(P_{ia}, R_{ia})			
	A	B	C	D
(ds1) $[d_1^{53} \setminus ds_1, ds_1]$	$(0.70, 0.70)$	$(0.58, 0.58)$	$(0.55, 0.55)$	$(0.76, 0.76)$
(ds2) $[d_1^{53} \setminus ds_2, ds_2]$	$(0.67, 0.67)$	$(0.67, 0.67)$	$(0.67, 0.67)$	$(0.75, 0.75)$
(ds3) $[d_1^{53} \setminus ds_3, ds_3]$	$(0.76, 0.76)$	$(0.85, 0.85)$	$(0.73, 0.73)$	$(0.79, 0.79)$
(ds4) $[d_1^{53} \setminus ds_4, ds_4]$	$(0.75, 0.75)$	$(0.64, 0.64)$	$(0.77, 0.77)$	$(0.74, 0.74)$
(ds5) $[d_1^{53} \setminus ds_5, ds_5]$	$(0.70, 0.70)$	$(0.63, 0.63)$	$(0.74, 0.74)$	$(0.74, 0.74)$
(ds6) $[d_1^{53} \setminus ds_6, ds_6]$	$(0.66, 0.66)$	$(0.66, 0.66)$	$(0.63, 0.63)$	$(0.69, 0.69)$
(ds1-6): weighted avg.	$(0.71, 0.71)$	$(0.67, 0.67)$	$(0.69, 0.69)$	$(\mathbf{0.74}, \mathbf{0.74})$

Fig. 7. POS3 classifiers, 6-fold extrinsic (anaphor resolution) cv (*ia* discipline)

depicted at this stage of consideration. In fact, the (P_{ia}, R_{ia}) results should be regarded the proper base of comparison as they immediately capture the extrinsic classifier performance, while the *na* results are sensitive to error chaining effects that should not be ascribed to the classifiers.

Notably and somewhat unexpectedly, it is the classifiers' cumulated $(\mathrm{C} \cup \mathrm{N})$ accuracy that seems to be of higher relevance. According to the determined weighted average results, the extrinsically best performing classifiers correspond to the settings *a*, *d* (PER3), *D*, and *A* (POS3), and are thus the very classifiers that intrinsically score highest on the cumulated $(\mathrm{C} \cup \mathrm{N})$ set of cases (see figure 5). Concerning possessive pronouns, it is interesting to see that setting *D* rather than *A* seems to be the clear extrinsic winner, hence giving evidence that, as suspected above, data distribution is indeed an issue, making the *D* classifier scoring extrinsically higher than the *A* classifier despite the fact that the latter quantitatively outperforms the former at intrinsic level.

If one thus combines the extrinsically highest-scoring classifiers for non-possessives and possessives, viz., the PER3 classifier corresponding to setting *a*, and the POS3 classifier corresponding to setting *D*, the cumulated and averaged extrinsic cross-validation results shown in figure 8 are obtained for ROSANA-NN. The (P_{ia}, R_{ia}) results of $(0.74, 0.74)$ shown for POS3 stem from the very experiment that has given rise to the results shown in column *D* of figure 7. The results of $(0.64, 0.64)$ for PER3 as well remain identical to the results in column *a* of figure 6. Furthermore, figure 8 gives the results in the non-pronominal anchors (*na*)

System	Setting	Corpus	antecedents (P_{ia}, R_{ia})		anchors (P_{na}, R_{na})	
			PER3	POS3	PER3	POS3
ROSANA-NN	*(a,D)*	$cv_6(d_1^{53})$	$(0.64, 0.64)$	$(0.74, 0.74)$	$(0.61, 0.61)$	$(0.64, 0.64)$
ROSANA-ML	$(1_{nc}^{tc}$,h)	$cv_6(d_1^{66})$	$(0.66, 0.66)$	$(0.75, 0.75)$	$(0.62, 0.62)$	$(0.68, 0.68)$
	$(1_{nc}^{tc}$,h)	$[d_1^{31}, d_{32}^{66}]$	$(0.65, 0.64)$	$(0.76, 0.76)$	$(0.62, 0.61)$	$(0.73, 0.73)$
ROSANA	standard	$[d_1^{31}, d_{32}^{66}]$	$(0.71, 0.71)$	$(0.76, 0.76)$	$(0.68, 0.67)$	$(0.66, 0.66)$

Fig. 8. Anaphor resolution results: ROSANA-NN vs. ROSANA-ML and ROSANA

evaluation discipline. Due to error chaining, it is, in general, harder to determine a cospecifying non-pronominal antecedent than an arbitrary antecedent; hence, the (P_{na}, R_{na}) tradeoffs of $(0.61, 0.61)$ for non-possessives and $(0.64, 0.64)$ for possessives are lower than the respective (P_{ia}, R_{ia}) tradeoffs.[22]

5 Comparison

Evaluation results of ROSANA-NN's ancestors are included in figure 8.[23] At first glance, ROSANA-NN seems to perform slightly worse than ROSANA-ML if one takes the *immediate antecedency* tradeoffs as the base of comparison.[24] Importantly, however, it should be taken into account that the ROSANA-NN evaluation data stem from experiments on a redundancy-free corpus, whereas the ROSANA-ML results have been obtained on a corpus exhibiting a certain degree of redundancy, which might be suspected to facilitate the learning task. Thus, given that a more difficult corpus has been employed, the results indicate that ROSANA-NN performs at least similar to its decision-tree-based ancestor.

Compared with its salience-based ancestor ROSANA, ROSANA-NN performs comparably on possessives, whereas it lags significantly behind on non-possessives. This confirms the findings of the C4.5 decision tree experiments (see [10]), according to which non-possessives are harder to deal with by ML means than possessives. The inferior results on non-possessives might be taken as an indicator that still the required sources of evidence are not adequately captured, and that the inventory of features over which the signatures are defined should be appropriately refined.

Finally, comparing the outcomes of the ROSANA-NN evaluation with the results given by Connolly et al. ([6]), it has to be taken into account that they

[22] See [5] for a more elaborate discussion of this issue.

[23] For a proper interpretation of these figures, one should take into account that the employed evaluation corpora differ slightly: $cv_6(d_1^{53})$ refers to the extrinsic cross-validation on the above-considered redundancy-free corpus of 53 news agency press releases; $cv_6(d_1^{66})$ refers to the 6-fold extrinsic cross-validation on the full - to a certain extent redundant - set of 66 news agency press releases as employed for cross-validating ROSANA-ML; $[d_1^{31}, d_{32}^{66}]$ refers to the bipartition of the full set of 66 press releases into a development vs. an evaluation corpus as employed for developing and assessing the original ROSANA system.

[24] This is justified as the (P_{ia}, R_{ia}) tradeoffs immediately capture the extrinsic performance, while the (P_{na}, R_{na}) results are sensitive to error chaining effects - see the above discussion.

consider the harder pronoun resolution task of determining *non-pronominal* antecedents. According to their findings, the two investigated types of backpropagation networks score highest, achieving an extrinsic accuracy of 0.55 (subspace-trained backpropagation networks) and 0.52 (ordinary backpropagation networks) for pronouns; no distinction between non-possessives and possessives is drawn. While a proper comparison should be based on a common evaluation corpus, this might be interpreted as a first indicator that the neural-network-based anaphor resolver considered above performs better, as its cumulated non-possessive $\cup$ possessive accuracy regarding non-pronominal antecedents amounts to 0.62.

6 Conclusion

Taking the previous work on the manually knowledge-engineered anaphor resolution system ROSANA and its successful hybrid, partly decision-tree-based descendant ROSANA-ML as the points of departure, it has been investigated what can be gained by employing backpropagation networks as the machine learning device for automatically determining antecedent preference criteria for pronoun resolution, leaving the filtering criteria to the discretion of the knowledge engineer. This research was motivated by the findings of Connolly et al. ([6]), who gained evidence that, compared to C4.5 decision trees, the standard backpropagation algorithm might be slightly ahead with respect to the task of object (NP) anaphor resolution.

According to the above results, the hybrid neural network approach ROSANA-NN performs similar to its decision-tree-based ancestor ROSANA-ML; given that a more difficult corpus has been employed for evaluation, it might even be the case that ROSANA-NN is slightly ahead. Extrinsic cross-validation on a corpus of 53 press releases has shown that ROSANA-NN achieves an accuracy of 0.64 on third-person non-possessives and of 0.74 on third-person possessives in the evaluation discipline of immediate antecedency. Thus, while these results do not yet allow to conclude that backpropagation networks are the unique best choice, they at least indicate that backpropagation networks are among the most successful machine learning model for anaphor resolution, in as far supporting the findings of Connolly et al. ([6]).

The methodology for systematically dealing with the numerous experimental degrees of freedom regarding the application of backpropagation network classifiers to anaphor resolution constitutes itself a valuable contribution. A two-stage approach has been developed in which signature optimization issues are confined to stage 1, and intrinsic as well as extrinsic cross-validation are confined to stage 2. Overall evaluation efforts thus remain bearable.

Subsequent research should address the question whether the above results generalize to larger text corpora and other text genres. Moreover, it should be instructive to investigate subspace-trained backpropagation networks, as this is the machine learning model that Connolly et al. ([6]) found to perform even better. Furthermore, there are recent advances in the field of rule learning,

e. g. the SLIPPER approach by Singer and Cohen ([19]), which have not been taken into account in the earlier work of Connolly et al. ([6]); it should thus be worthwhile to consider these models as alternatives to neural network and decision tree learning. Regarding the ROSANA-NN algorithm itself, further experiments should address the question whether, by referring to the actually real-valued classification outcome ε, it can be suitably biased towards high precision; respective experiments have proven to be successful under the employment of decision trees (see [10]).[25] Finally, it should be revealing to compare ROSANA-NN, ROSANA-ML, and ROSANA based on a detailed qualitative analysis of their respective so-called competence cases according to the framework proposed in [20].

References

1. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994) 535–561
2. Kennedy, C., Boguraev, B.: Anaphora for everyone: Pronominal anaphora resolution without a parser. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING). (1996) 113–118
3. Baldwin, B.: Cogniac: High precision coreference with limited knowledge and linguistic resources. In Mitkov, R., Boguraev, B., eds.: Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid. (1997) 38–45
4. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of the 17th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal. (1998) 869–875
5. Stuckardt, R.: Design and enhanced evaluation of a robust anaphor resolution algorithm. Computational Linguistics **27**(4) (2001) 479–506
6. Connolly, D., Burger, J.D., Day, D.S.: A machine-learning approach to anaphoric reference. In: Proceedings of the International Conference on New Methods in Language Processing (NEMLAP). (1994)
7. Aone, C., Bennett, S.W.: Evaluating automated and manual acquisition of anaphora resolution strategies. In: Proceedings of the 33rd Annual Meeting of the ACL, Santa Cruz, New Mexico. (1995) 122–129
8. Ge, N., Hale, J., Charniak, E.: A statistical approach to anaphora resolution. In: Proceedings of the Sixth Workshop on Very Large Corpora, Montreal. (1998) 161–170
9. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27**(4) (2001) 521–544
10. Stuckardt, R.: A machine learning approach to preference strategies for anaphor resolution. In Branco, A., McEnery, T., Mitkov, R., eds.: Anaphora Processing: Linguistic, Cognitive, and Computational Modelling, John Benjamins (2005) 47–72
11. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the ACL, Philadelphia. (2002) 104–110

[25] Biasing towards high precision is deemed important for typical applications of anaphor and coreference resolution such as text summarization and question answering.

12. Ng, V., Cardie, C.: Weakly supervised natural language learning without redundant views. In: HLT-NAACL 2003: Proceedings of the Main Conference. (2003) 173–180
13. Olsson, F.: A survey of machine learning for reference resolution in textual discourse. SICS Technical Report T2004:02, Swedish Institute of Computer Science (2004)
14. Grüning, A., Kibrik, A.A.: Modelling referential choice in discourse: A cognitive calculative approach and a neural network approach. In Branco, A., McEnery, T., Mitkov, R., eds.: Anaphora Processing: Linguistic, Cognitive, and Computational Modelling, John Benjamins (2005) 163–198
15. Mitkov, R.: Factors in anaphora resolution: They are not the only things that matter. a case study based on two different approaches. In Mitkov, R., Boguraev, B., eds.: Proceedings of the ACL'97/EACL'97 Workshop on Operational Factors in Practical, Robust Anaphor Resolution for Unrestricted Texts, Madrid. (1997) 14–21
16. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997)
17. Järvinen, T., Tapanainen, P.: A dependency parser for english. Technical Report TR-1, Department of General Linguistics, University of Helsinki (1997)
18. Chomsky, N.: Lectures on Government and Binding. Foris Publications, Dordrecht (1981)
19. Cohen, W.W., Singer, Y.: A simple, fast, and effective rule learner. In: Proceedings of the 16th National Conference on Artificial Intelligence (AAAI), Menlo Park, CA. (1999) 335–342
20. Stuckardt, R.: Three algorithms for competence-oriented anaphor resolution. In: Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 04). (2004) 157–163

Improving Coreference Resolution Using Bridging Reference Resolution and Automatically Acquired Synonyms

Ryohei Sasano[1,⋆], Daisuke Kawahara[2], and Sadao Kurohashi[3]

[1] University of Tokyo, Graduate School of Information Science and Technology
7-3-1 Hongo Bunkyo-ku, Tokyo, 113-8656, Japan
[2] National Institute of Information and Communications Technology, Kyoto
[3] Kyoto University, Graduate School of Infomatics
ryohei@nlp.kuee.kyoto-u.ac.jp

Abstract. We present a knowledge-rich approach to Japanese coreference resolution. In Japanese, proper noun coreference and common noun coreference occupy a central position in coreference relations. To improve coreference resolution for such language, wide-coverage knowledge of synonyms is required. We first acquire knowledge of synonyms from large raw corpus and dictionary definition sentences, and resolve coreference relations based on the knowledge. Furthermore, to boost the performance of coreference resolution, we integrate bridging reference resolution system into coreference resolver.

1 Introduction

In text, expressions that refer to the same entity are repeatedly used. Coreference resolution, which recognizes such expressions, is an important technique for natural language processing. This paper focuses on coreference resolution for Japanese text.

In Japanese, pronouns are not used so much; most anaphors are represented as proper noun phrases or common noun phrases. To resolve coreference for such language, string matching technique is useful, because an anaphor and its antecedent often share strings [1]. Learning-based coreference approaches, which have been intensively studied in recent years [2,3,4], use string matching as features for learning. However, in some cases, coreferential expressions share no string, and string matching technique can not be applied.

Resolving such coreference relations requires knowledge that these two expressions share a same meaning. Then, we first propose a method for extracting synonyms[1] from large raw corpus and dictionary definition sentences, and utilize the synonyms to coreference resolution.

Our target language Japanese also has a characteristic that it has no article. Articles can be a clue for anaphoricity determination, so this characteristic

⋆ Research Fellow of the Japan Society for the Promotion of Science (JSPS).
[1] In this paper, synonyms include acronyms and abbreviations.

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 125–136, 2007.

makes anaphoricity determination difficult. We combine bridging reference resolution with coreference resolution as a clue to determine anaphoricity. Roughly speaking, we consider modified NPs are not anaphoric. But if an NP have a bridging relation, it is considered as anaphoric.

The rest of the paper is organized as follows. In Section 2, we present a method for extracting synonyms from raw corpus and dictionary definition sentences. In Section 3, we present basic strategy for coreference resolution and how to use the extracted synonyms and the result of bridging reference resolution for coreference resolution. We show the experimental results on news paper articles in Section 4 and compare our approaches with some related work in Section 5.

2 Synonym Extraction

It is difficult to recognize coreference relations between absolutely different expressions without knowledge of synonyms. To construct a high-performance coreference resolver, we acquire knowledge of synonyms in advance.

As resources for synonym extraction, we use raw corpus and dictionary definition sentences. The characteristic of synonyms extracted from raw corpus is the ability to respond to new words. However, very familiar synonyms, such as *US* and *America*, is not extracted from parenthesis expressions. Thus, in order to extract very familiar synonyms, we also extract synonyms from dictionaries for humans.

2.1 Synonym Extraction from Parenthesis Expressions

When unfamiliar synonymous expressions are used for the first time in text, the information is often written in text by using parenthesis. In example (1), "*KEDO*", a synonym of "*Chosen Hanto Enerugi Kaihatu Kiko*"(Korean Peninsula Energy Development Organization), is written in the following parenthesis. Therefore, we first extract synonyms from parenthesis expressions that appeared in raw corpus.

(1) *Suzuki Chosen Hanto Energy Kaihatu Kiko* (KEDO)
Suzuki Korean Peninsula Energy Development Organization (KEDO)

taishi-ga yutai-shita.
ambassador retire

(The Korean Peninsula Energy Development Organization (KEDO) Ambassador Suzuki retired.)

Parenthesis is not always used to indicate synonym. For example, parenthesis is sometimes used to indicate attribution of preceding noun phrases such as age or affiliation. Thus, the problem is how to extract parenthesis pairs that indicate synonym.

In order to deal with this problem, we make an assumption that if a pair A and B that is appeared in parenthesis expression "*A(B)*" is a synonym pair, the frequency of the parenthesis expressions is high and the reverse pair "*B(A)*"

Table 1. Thresholds for synonym extraction

	type	two-way	threshold
1	One consists of English letters and the other does not	yes no	1 50
2	One consists of Japanese letters *katakana* and the other does not	yes no	2 300
3	One consists of Chinese characters and the other is the abbreviation [2]	yes no	0 infinite
4	others	yes no	30 infinite

can also appeared in corpus. According to this assumption, we extract synonym pairs from parenthesis expressions as follows:

1. Count the frequency of pairs A and B. B is an expression in a parenthesis and A is the preceding noun phrase, that is "*A (B)*".
2. Set frequency thresholds for several types by observing the frequencies of randomly selected 100 pairs.
3. If the frequency exceeds the thresholds, the pair A and B is judged as a synonym pair.

Table 1 shows the thresholds, which are set not to extract incorrect synonym pairs. When there are also examples of "*A (B)*" besides "*B (A)*", we call this pair as two-way pair, and use the geometric mean of the frequencies against the looser threshold.

We extracted synonym pairs from Japanese newspaper articles in 26 years (12 years of *Mainichi* newspaper and 14 years of *Yomiuri* newspaper). There are about 10 million parenthesis expressions in the newspaper articles.

Table 2 shows the result of extraction. We acquired 2,653 synonym pairs. Almost all of the extracted synonym pairs are correct because we set the threshold not to extract incorrect synonym pairs.

2.2 Synonym Extraction from Dictionary

Secondly, in order to extract very familiar synonyms, we use definition sentences of dictionaries for humans. The following process is carried out for each dictionary entry A.

1. If the definition sentence ends with "*no ryaku*" (abbreviation of) or "*no koto*" (synonym of), we extract the rest of the sentence as a synonym candidate B; otherwise extract whole the sentence as B.
2. If B itself is an entry of dictionaries or enclosed by angle brackets, the pair of A and B is judged as a synonym pair.

We extracted synonyms from *Reikai Shougaku Kokugojiten* [5] and *Iwanami Kokugo Jiten* [6]. As a result, we extracted 402 synonym pairs from dictionary definition sentences. Table 3 shows examples of extracted synonym pairs.

[2] One expression must include all Chinese characters included in the other expression.

Table 2. The result of synonym extraction from parenthesis expressions

type	#	examples
1	1,572	*kokunai sou-seisan* = GDP domestic gross product GDP *Europe rengo* = EU European Union EU
2	732	*jugyo keikaku* = *syllabus* class plan syllabus *shien kigyo* = *sponsor* support company sponsor
3	239	*Gakushu kenkyuu sha* = *Gakken* study pursuit corporation = Gakken *Nihon kogyo ginko* = *Kogin* Japan industrial bank = Kogin
4	110	*ushi kaimenjou noushou* = *kyogyubyo* bovine spongiform encephalopathy mad cow disease *Myanmar* = *Burma* Myanmar Burma
sum	2,653	

Table 3. Examples of extracted synonyms from dictionaries

type of definition sentence	examples	
	entry	extracted synonym
...*-no ryaku*	*fukei* policewoman	*fujin keikan* woman cop
	Niti JP	*Nihon* Japan
...*-no koto*	*Chuugoku* China	*Chuuka Jinmin Kyowakoku* the People's Republic of China
	Bei US	*America* America
others	*Chokou* Yanzi Jiang	*Yousukou* Chang Jiang
	Japan Japan	*Nihon* Nippon

Only 4 synonym pairs extracted from dictionary definition sentences overlapped with the synonym pairs extracted from parenthesis expressions. Therefore, it is reasonable to suppose that we extract very familiar synonyms from definition sentences that were not extracted from parenthesis expressions in raw corpus.

As a whole, we acquired 3,051 synonym pairs from raw corpus and dictionary definition sentences.

3 Strategy for Coreference Resolution

We propose a method to improve coreference resolution using knowledge of synonyms and bridging reference resolution.

3.1 Basic Strategy for Coreference Resolution

The outline of our coreference resolver is as follows:

1. Parse input sentences by using a Japanese parser and recognize named entity.
2. Consider each subsequence of a noun phrase as a possible anaphor if it meets "*Condition 1*".
3. For each anaphor:
   (a) From the position of the anaphor to the beginning of document, consider each noun sequence as antecedent candidate.
   (b) If the anaphor and the antecedent candidate meet "*Condition 2*", judge as coreferential expressions and move to next anaphor.

"*Condition 1*" and "*Condition 2*" are varied between methods. "*Condition 1*" judge the anaphoricity of the subsequence.

We use KNP [7] as a Japanese parser. To recognize named entity, we apply a method proposed by Isozaki and Kazawa [8] that use NE recognizer based on Support Vector Machines.

3.2 Determination of Markables

The first step of coreference resolution is to identify the markables. Markables are noun phrases that related to coreference. We consider how to deal with compound nouns.

Previous work on coreference resolution in Japanese focused on the whole compound noun and cannot deal with this example:

(2) *Lifestyle-no* *chosa-wo jisshi-shita.* *Chosa* *naiyo-wa* ...
lifestyle investigation conduct investigation content

(ϕ conducted an investigation. The content of the investigation was ...)

In this example, the second "*chosa*"(investigation) that is contained in a compound noun "*chosa naiyo*" refers to the preceding "*chosa*". To deal with such a coreference relation, we consider every subsequence of a compound noun as a markable, that is, we consider "*chosa naiyo*", "*chosa*" and "*naiyo*" as a markable for *chosa naiyo*.

But we consider named entities as an exception. Named entities are not divided and handled as a whole.

3.3 Baseline Methods

We consider 3 baseline methods. In all of these methods, "*Condition 2*" is true when the anaphor exactly matches the antecedent candidate. Only "*Condition 1*" (i.e. anaphoricity determination) varies among these 3 baselines.

In a primitive baseline (*baseline 1*), "*Condition 1*" is always true, that is, every noun sequence is considered an anaphor.

For a bit more sophisticated baselines (*baseline 2* and *baseline 3*), we assume that a modified noun phrase is not anaphoric.

Table 4. *Condition 1* for each baseline

	Condition 1
baseline 1	always true
baseline 2	true when the noun sequence is not modified by its preceding nouns in the same phrase
baseline 3	true when the noun sequence has no modifier

(3) a. *Uno shusho-wa Doitsu-ni totyaku-shita. Shusho-wa kuukou-de . . .*
Uno prime minister Germany arrived prime minister airport

(Prime minister Uno arrived in Germany. At the airport the minister . . .)

b. *Uno shusho-wa Doitsu-ni totyaku-shita. Asu Doitsu Shusho-tono. . .*
Uno prime minister Germany arrived Tomorrow German prime minister

(Prime minister Uno arrived in Germany. Tomorrow, with German prime minister . . .)

In example (3a), "*shusho*" (prime minister) in the first and second sentence refer to the same entity, but not in example (3b). This is because the second "*shusho*" in (3b) is modified by "*Doitsu*" (German), and this "*shusho*" is turned out to be a person other than "*Uno shusho*".

We consider that a partial noun sequence of a compound noun is modified by its preceding nouns in the compound noun. For example, for the compound noun "XY", "Y" is considered to be modified by "X", and thus "Y" is regarded as non-anaphoric (in this case, noun sequences "XY" and "X" are regarded as anaphoric).

In both *Baseline 2* and *baseline 3*, modified noun phrases are considered non-anaphoric. These two methods differ in the scope of the considered modifier. In *baseline 2*, "*Condition 1*" is true when the noun sequence is not modified by its preceding nouns in the same noun phrase. On the other hand, in *baseline 3*, "*Condition 1*" is true only when the noun sequence do not have any modifier including clausal modifier and adjective modifier. Table 4 show the "*Condition 1*" for each baseline.

3.4 How to Use Synonym Knowledge

The basic strategy for determining a coreference relation is based on precise string matching between an anaphor and its antecedent candidate. We also make use of synonym knowledge to resolve a coreference relation that cannot be recognized by string matching.

In the synonym knowledge using methods, "*Condition 2*" is true not only when the anaphor exactly matches the antecedent candidate, but also when the anaphor is a synonym of the antecedent candidate.

3.5 How to Use Bridging Reference Resolution

We explain how to use the result of bridging reference resolution to coreference resolution. As mentioned, we do not consider a modified NP anaphoric in *baseline 2* and *baseline 3*. However, in some cases, an modified NP can be anaphoric. To deal with such cases, if two NPs share strings and have a bridging relation to the same entity, we consider the latter NP is anaphoric and has coreference relation to the former.

We use the method for bridging reference resolution proposed by Sasano et al.[9]. This method is based on automatically constructed nominal case frames. Nominal case frames are useful knowledge for resolving bridging reference and represents indispensable entities of the target noun.

(4) *Murayama shusho-wa nento-no kisha kaiken-de shokan-wo*
Murayama prime minister beginning of year press conference impressions

happyo-shita. Nento shokan-no yoshi-wa ika-no tori.
express beginning of year impressions point as follows

(Prime Minister Murayama expressed his impressions at the press conference of the beginning of the year. The point of the impressions is as follows.)

In example (4), the second "*shokan*"(impression) is modified by "*nento*" (beginning of year) and is not considered anaphoric in *baseline 2* or *baseline 3* method. However, "*shokan*"(impression) has a case frame named "AGENT" as shown in Table 5, and its bridging relation to "*shusho*"(prime minister) is recognized (i.e. the system recognize that the impression is the impression of the prime minister). Accordingly, the second "*shokan*" is considered anaphoric and the coreference relation between the first and the second "*shokan*" is recognized.

In the methods using the result of bridging reference resolution, "*Condition 1*" is also true when the anaphor has a bridging relation, and then "*Condition 2*" is true only when the anaphor and it's antecedent candidate have the same referent of bridging.

As another example, although the second "*kekka*"(result) in example (5) is modified by "enquêtet" and is not considered anaphoric in *baseline 2* or *baseline 3* method, bridging reference resolver recognizes the two "*kekka*" refer to same entity "enquête" and the system recognizes the coreference relation between the first and the second "*result*".

(5) 2006 FIFA *world cup-no yushokoku yosou enquête-wo okonatta.*
2006 FIFA world cup winner expectation questionnaire conducted

Kekka-wa Brazil-ga top-datta. Kuwasii enquête kekka-wa HP-de.
result Brazil top detail questionnaire result web page

(The expectation questionnaire about 2006 FIFA world cup win ner was conducted. The top of the questionnaire result was Brazil. The detail of the result appeared in web page.)

Table 5. Examples of nominal case frame

Nominal case frame of "*shokan*"(impression)

case frame	examples	: frequency
AGENT	"*watashi*"(I)	: 24
	"*chiji*"(governor)	: 16
	"*sori*"(prime minister)	: 3
	"*hissha*"(writer)	: 2
	...	: ...

Nominal case frame of "*kekka*"(result)

case frame	examples	: rate
"*koto*	"*chosa*"(investigation)	: 7648
(something)	"*senkyo*"(election)	: 1346
	"*enquête*"(questionnaire)	: 734
	"*jikken*"(experiment)	: 442
	...	: ...

"*koto*" = "*Aru koto-ga moto-ni natte okotta kotogara.*"
(a consequence, issue, or outcome of something)

4 Experiments

We conducted experiments on the Kyoto Corpus Version 4.0 [10]. In the corpus, coreference relations are manually annotated on the articles of *Mainichi* newspaper. We used 322 articles, which comprise 2098 sentences. These sentences have 2872 coreference tags that match our coreference criteria.

We used 3 baseline methods, *baseline 1*, *baseline 2* and *baseline 3*. In addition, for *baseline 2* and *baseline 3*, we also conducted experiments with synonym knowledge and/or bridging reference resolution. Thus, all in all we conducted experiments in 9 different conditions.

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}. \qquad (a)$$

Table 6 shows the results of coreference resolution. F-score is calculated according to (a). *Baseline 1* achieve high recall but lowest precision and f-score. We can say that considering modified NPs as non-anaphoric improves F-score. We can also say that the condition used in *baseline 2*, "*Condition 1*" is true when the noun sequence is not modified by its preceding nouns in the same phrase, achieve best performance.

Furthermore, using knowledge of synonyms and the result of bridging reference resolution improves F-score and the usefulness of them is confirmed, but the effect is limited.

To investigate recall for several coreference types, we randomly selected 200 coreference tags from the Kyoto Corpus and evaluated the result of coreference resolution using *baseline 2* method with synonym knowledge and bridging reference resolution. Table 7 shows the recall for each coreference type.

Table 6. Experimental results of coreference resolution

method	precision	recall	F-score
baseline 1	57.0 (2246/3943)	**78.2** (2246/2872)	65.9
baseline 2	71.7 (2187/3052)	76.1 (2187/2872)	73.8
with bridging	71.5 (2200/3077)	76.6 (2200/2872)	74.0
with synonym	71.7 (2217/3092)	77.2 (2217/2872)	74.3
with syn. & brid.	71.5 (2231/3121)	77.7 (2231/2872)	**74.5**
baseline 3	77.4 (1966/2541)	68.5 (1966/2872)	72.6
with bridging	77.0 (1994/2590)	69.4 (1994/2872)	73.0
with synonym	**77.4** (1997/2581)	69.5 (1997/2872)	73.2
with syn. & brid.	77.0 (2025/2630)	70.5 (2025/2872)	73.6

Table 7. Recall for each coreference type

relations between anaphor & antecedent	recall
1. anaphor's string is contained in antecedent's string	83.5 (142/170)
2. anaphor and its antecedent have a synonymous relation	50.0 (4/8)
3. other coreference types	0.0 (0/22)
sum	73.0 (146/200)

The coreference relations that can be recognized by string matching are well recognized. On the other hand, the relations that need synonym knowledge to recognize are not (the recall is 50.0% (4/8)). However, 7 synonym relations out of 8 are included in automatically acquired knowledge of synonyms, and 3 coreference relations can not recognized only because the anaphors are modified. Therefore we can say that the coverage of the automatically acquired synonyms is not too small for resolving coreference relations between synonymous expressions.

The other types of coreference relations, such as relations between hypernym and hyponym, can not recognize fundamentally by our proposed method. To resolve such relations is our future work.

In order to investigate the cause of erroneous system outputs, we classify erroneous system outputs into 4 categories. Table 8 shows the classified error

Table 8. Error analysis of erroneous system outputs

error type	num
The anaphor and antecedent candidate refer to another entities	52
The possible anaphor is a general noun and not anaphoric	32
The antecedent candidate is a general noun and not anaphoric	7
others	9
sum	100

types of randomly selected 100 erroneous system outputs of *baseline 2* method with synonym knowledge and bridging reference resolution. Major erroneous system outputs were caused by two reasons:

1. *Baseline 2* method does not consider clausal or adjective modifiers.
2. Our system does not consider the generic usage of nouns.

In example (6), though the second "*jishin*"(earthquake) does not have coreference relation to "*Sanriku Harukaoki Jishin*", our system judges the two "*jishin*" refer to same entity because our system does not consider the modifiers "*yoshin-to mirareru*"(thought to be an aftershock).

(6) *Sanriku Harukaoki Jishin-no yoshin-to mirareru jishin-ga hassei-shita.*
Far-off Sanriku Earthquake aftershock thought earthquake occurred

(An earthquake thought to be an aftershock of Far-off Sanriku Earthquake occurred.)

In example (7), although the second "wine" is used in generic usage, our system considers the second "wine" have coreference relation to "French wine" because our system does not consider generic usage of nouns.

(7) *Kare-wa* France-*no* wine-*ga suki-de kare-no ie-niwa* wine cellar-*ga aru.*
he French wine like his house wine cellar have

(He likes French wine and has wine cellar in his house.)

5 Related Work

Murata and Nagao proposed a rule-based coreference resolution method for determining the referents of noun phrases in Japanese sentences by using referential properties, modifiers and possessors [11]. As a result of experiments, they obtained a precision rate of 78.7% and a recall rate of 77.3%.

Their method performed relatively well. This may be because their experiments is constructed on small and supposedly easy corpus. Half of their corpus is occupied by fairy tale that is supposed to be easy to analyze.

Table 9. Comparison with previous work

	precision	recall	F-score
Murata and Nagao	78.7 (89/113)	77.3 (89/115)	78.1
Iida et al.	76.7 (582/759)	65.9 (582/883)	70.9
Proposed	**71.5** (2231/3121)	**77.7** (2231/2872)	**74.5**

Iida et al. proposed a machine learning approach for coreference resolution for Japanese [12]. Their process is similar to the model proposed by Ng and Cardie [13]. As a result of experiments on Japanese newspaper articles, they obtained a precision rate of 76.7% and a recall rate of 65.9%.

Table 9 shows the comparison with previous work and our proposed method. Since they used different data set and coreference criteria for experiments, these scores are not comparable as-is. However, taking into consideration Murata and Nagao uses small and supposedly easy corpus, we can say that our proposed method achieved enough performance.

Though these scores are not comparable as-is, rule-based methods outperformed learning-based methods in Japanese. This may be because recognizing most of coreference relations does not need complicated rules.

Bean and Riloff proposed a noun phrase coreference resolution system that uses information extraction patterns to identify contextual roles and creates four contextual role knowledge sources using unsupervised learning [14]. Experiments showed that the contextual role knowledge improved coreference performance for pronouns but not for noun phrases.

6 Conclusion

We have described a knowledge-rich approach to Japanese coreference resolution. We first proposed a method for acquiring knowledge of synonyms from large raw corpus and definition sentences of dictionaries for humans. Second, we proposed a method for improving coreference resolution by using the automatically acquired synonyms and the result of bridging reference resolution. Using the acquired synonyms and the result of bridging reference resolution boosted the performance of coreference resolution and the effectiveness of our integrated method is confirmed.

References

1. Yang, X., Zhou, G., Su, J., Tan, C.L.: Improving noun phrase coreference resolution by matching strings. In: Proceedings of 1st International Joint Conference of Natural Language Processing (IJCNLP04). (2004) 326–333
2. Soon, W.M., Ng, H.T., Lim, D.C.Y.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27**(4) (2001) 521–544

3. Ng, V.: Machine learning for coreference resolution: From local classification to global ranking. In: Proceedings of the 43rd Annual Meeting of the Asssociation for Computational Linguistics. (2005) 157–164
4. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A mention-synchronous coreference resolution algorithm based on the bell tree. In: Proceedings of the 42nd Meeting of the Association for Computational Linguistics. (2004) 135–142
5. Tajika, J., ed.: Reikai Syogaku Kokugojiten. Sanseido (1997)
6. Nishio, M., Iwabuti, E., Mizutani, S., eds.: Iwanami Kokugo Jiten. Iwanami Shoten (2000)
7. Kurohashi, S., Nagao, M.: A syntactic analysis method of long Japanese sentences based on the detection of conjunctive structures. Computational Linguistics **20**(4) (1994) 507–534
8. Isozaki, H., Kazawa, H.: Efficient support vector classifiers for named entity recognition. In: Proceedings of the 19th International Conference on Computational Linguistics. (2002) 390–396
9. Sasano, R., Kawahara, D., Kurohashi, S.: Automatic construction of nominal case frames and its applicatoin to indirect anaphora resolution. In: Proceedings of the 20th International Conference on Computational Linguistics. (2004) 1201–1207
10. Kawahara, D., Kurohashi, S., Hasida, K.: Construction of a Japanese Relevance-tagged Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation. (2002) 2008–2013
11. Murata, M., Nagao, M.: An estimate of referent of noun phrases in Japanese sentences. In: Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics. (1998) 912–916
12. Iida, R., Inui, K., Matsumoto, Y., Sekine, S.: Noun phrase coreference resolution in japanese most likely candidate antecedents (in japanese). Journal of Information Processing Society of Japan **46**(3) (2005) 831–844
13. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. (2002) 104–111
14. Bean, D., Riloff, E.: Unsupervised learning of contextual role knowledge for coreference resolution. In: Proceedings of the Human Language Technology Conference / North American Chapter of the Assosication for Computational Linguistics Annual Meeting. (2004) 297–304

Evaluating Hybrid Versus Data-Driven Coreference Resolution

Iris Hendrickx[1], Veronique Hoste[2], and Walter Daelemans[1]

[1] University of Antwerp, CNTS - Language Technology Group
Universiteitsplein 1, Antwerp, Belgium
[2] University College Ghent, LT3 - Language and Translation Technology Team
iris.hendrickx@ua.ac.be

Abstract. In this paper, we present a systematic evaluation of a hybrid approach of combined rule-based filtering and machine learning to Dutch coreference resolution. Through the application of a selection of linguistically-motivated negative and positive filters, which we apply in isolation and combined, we study the effect of these filters on precision and recall using two different learning techniques: memory-based learning and maximum entropy modeling. Our results show that by using the hybrid approach, we can reduce up to 92 % of the training material without performance loss. We also show that the filters improve the overall precision of the classifiers leading to higher F-scores on the test set.

1 Introduction

Coreference resolution, resolving different descriptions or names for the same underlying object, is an important text analysis module for further processing and understanding of text, for example in applications like information extraction and question answering.

As an alternative to knowledge-based approaches, corpus-based machine learning techniques have become increasingly popular for the resolution of coreferential relations. In a typical machine learning approach to coreference resolution, information on pairs of noun phrases is represented in a set of feature vectors. Unsupervised learning techniques, e.g. [1], view coreference resolution as a clustering task of combining noun phrases into equivalence classes. Most learning approaches to coreference resolution, however, are supervised learning techniques, for example the C4.5 decision tree learner[2] as in [3], [4] and [5] or the RIPPER rule learner [6] as in [7]. A supervised learning approach requires an annotated corpus from which for each pair of noun phrases a class (coreferential or not coreferential) can be obtained. The pair of NPs is represented by a feature vector containing distance, morphological, lexical, syntactic and semantic information on the candidate anaphor, its candidate antecedent and also on the relation between both. In a postprocessing phase of such an approach, a complete coreference chain has to be built between the pairs of NPs that were classified as being coreferential.

As a consequence of recasting the problem as a classification task, coreference resolution data sets reveal large class imbalances: only a small part of the possible

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 137–150, 2007.

relations between noun phrases (NPs) is coreferential. When trained on such imbalanced data sets, classifiers can exhibit a good performance on the majority class instances but a high error rate on the minority class instances. Always assigning the "non coreferential" class will lead to a highly 'accurate' classifier, which cannot find any coreferential chain in a text.

In order to cope with these class imbalances, different instance selection techniques have been proposed to rebalance the corpus [8,9,7,10,11,12]. In [12], rebalancing is done without any a priori linguistic knowledge about the class to be solved. Most approaches, however, aim to produce better performing classifiers through the application of linguistically motivated filters on the training data before application of the classifier. Through the application of these linguistic filters, part of the problem to be solved, viz. coreference resolution, is solved beforehand and only a small part of the instances is handled by the classifier. Unfortunately, previous literature on these filters is relatively vague on their exact implementation and use and no systematic studies of their impact are provided.

In this paper, we investigate the hybrid approach of combined knowledge-based filtering and machine learning in a more systematic way. We apply a selection of linguistically-motivated negative and positive filters. Negative filters filter out negative instances, positive filters do the same with examples of co-referring NPs. We study the effect of these filters on performance and we investigate how much reduction in training material we can obtain without performance loss. The filters are considered separately and in combination. We use two different machine learning techniques to demonstrate the effects of filtering: a memory-based learning approach and a maximum entropy approach. Most existing learning approaches to coreference resolution can be described as eager decision tree or rule learning approaches; we investigate in this paper how a memory-based learning and a maximum entropy approach tackle the problem of coreference resolution and the problem of the skewness of the data. The experiments are performed on the KNACK-2002 Dutch data set [12].

The remainder of this paper is organized as follows. Section 2 discusses the preparation of the data sets including the selection of positive and negative instances and presents the two machine learning packages we use in the experiments. Section 3 gives an overview of instance selection in the machine learning of coreference resolution literature, and discusses the positive and negative filters. This section also reports on the results obtained for both learners in a hybrid architecture with filters compared to a completely data-driven setting, and to baselines. Section 4 concludes this paper.

2 Experimental Setup

2.1 Data

Our experiments are performed on a Dutch coreferentially annotated corpus, KNACK-2002. KNACK is a Flemish weekly news magazine with articles on national and international current affairs. For the annotation of the corpus, the

MUC-7 [13] manual, the manual from Davies et al. [14] and the work from van Deemter and Kibble [15] were taken as source. The complete corpus consists of 267 documents annotated with coreference information for NPs. 12,546 noun phrases are annotated with coreferential information. For the experiments, 50 documents are randomly selected, of which 25 are used for training and the other half for testing.

For the construction of the initial data sets, we selected all noun phrases, which could be detected after preprocessing the raw text corpora. The following preprocessing steps were taken: tokenization was performed by a rule-based system using regular expressions. Dutch named entity recognition was performed by looking up the entities in lists of location names, person names, organization names and other miscellaneous named entities. We applied a part-of-speech tagger and text chunker for Dutch that use the memory-based tagger MBT [16], trained on the Spoken Dutch Corpus (http://lands.let.ru.nl/cgn). Finally, grammatical relation finding was performed, using a shallow parser to determine the grammatical relation between NP chunks and verbal chunks, e.g. subject, object, etc. The relation finder [17] was trained on the previously mentioned Spoken Dutch Corpus. It offers a fine-grained set of grammatical relations, such as modifiers, verbal complements, heads, direct objects, subjects, predicative complements, indirect objects, reflexive objects, etc. Figure 1 gives an overview of the part-of-speech tags, chunk tags and relation tags for the following KNACK-2002 training sentence.

(1) < COREF ID = "1528" MIN = "conflict" > Het conflict over het grensgebied < /COREF > is zo oud als < COREF ID = "1464" > < COREF ID = "1451" > India < /COREF > en < COREF ID = "1459" > Pakistan < /COREF > < /COREF>.

English: The conflict about the border area is as old as India and Pakistan.

On the basis of the preprocessed texts, instances are created. We create an instance between every NP and its preceding NPs, with a restriction of 20

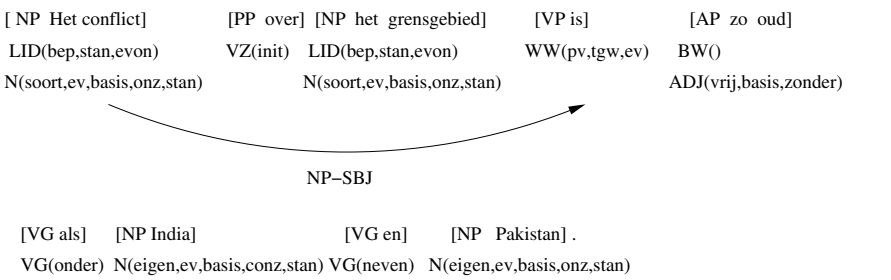


Fig. 1. Part-of-speech tags, chunk tags and relation tags for the example sentence (1)

sentences backwards. A pair of NPs that belongs to the same coreference chain, gets a positive label; all other pairs get a negative label. This is the basic set of instances. In the training set, the positive class accounts for only 8.5% of the total number of 76,920 instances.

Instances describe the relation between a potential anaphor and its antecedent. For each NP pair we create a set of 39 features encoding morphological-lexical, syntactic, semantic, string matching and positional information sources. The overview below gives a short impression of the type of information encoded in the features.

- morphological-lexical
  Is there number agreement between anaphor and antecedent? Is it a definite/indefinite anaphor? Is the anaphor/antecedent a pronoun or proper noun?
- syntactic
  Is the anaphor/antecedent object or subject of the sentence, is the anaphor an apposition?
- positional
  The local context of the anaphor, the distance in NPs and in sentences between the anaphor and antecedent.
- semantic
  Named entity type information. Are the NPs synonyms/hypernyms of each other?
- string matching
  Are the anaphor and antecedent a complete or partial match or alias from each other? Do they share the same head word?

2.2 Learners

Two machine learning techniques are applied to the task of coreference resolution: a memory-based learning algorithm and a maximum entropy learner.

Memory-based learning (a k-nearest neighbor approach) is a lazy learning approach that stores all training data in memory. At classification time, the algorithm classifies new instances by searching for the nearest neighbors to the new instance using a similarity metric, and extrapolating from their class. In our experiments we use the TIMBL [18] software package[1] that implements a version of the k-nn algorithm optimized for working with linguistic datasets and that provides several similarity metrics and variations of the basic algorithm. Lazy learning is claimed to have the right bias for learning language processing problems as it doesn't abstract from exceptions and subregularities as more eager learning approaches do through mechanisms like pruning.

Maximum entropy modeling (a kind of exponential or log linear modeling), on the other hand, is a discriminative statistical machine learning approach [19,20] that derives a conditional probability distribution from labeled training data by

[1] URL:`http://ilk.uvt.nl`

assigning a weight to each feature. We use the entropy modeling software package MAXENT by Zhang Le [21]. Maximum Entropy has been shown to provide good results with language data, and can handle large feature vectors and feature redundancy.

Memory-based learning offers several algorithmic parameters such as number of nearest neighbors, the feature weighting and distance weighting. These parameters can, individually and in combination, affect the functioning of the algorithm. We use a heuristic wrapped-based method to set them automatically for all experiments.

Wrapped progressive sampling (WPS)[22] combines classifier wrapping [23] with progressive sampling of training material [24]. WPS starts with a large pool of experiments, each with one systematically generated recombination of tested algorithmic parameter settings. In the first step of WPS, each attempted setting is applied to a small amount of training material and tested on a small amount of held-out training data. Only the best settings are kept; all others are removed from the pool of competing settings. In subsequent iterations, this step is repeated, retaining only the best-performing settings, with an exponentially growing amount of training and held-out data – until all training data is used or one best setting is left. Selecting the best settings at each step is based on classification score on the held-out data; a simple one-dimensional clustering on the ranked list of scores determines which group of settings is selected for the next iteration. The final selected parameters of the WPS procedure are then used to classify the test set.

We did not optimize the parameters of MAXENT as it was shown in [22] that WPS did not increase the generalization performance of maximum entropy modeling. We train MAXENT with L-BFGS parameter estimation, 100 iterations and a Gaussian prior with mean zero and σ^2 of 1.0.

2.3 Evaluation

Defining the coreference resolution process as a classification problem involves the use of a two-step procedure. In a **first step**, the classifier (in our case TIMBL or MAXENT) decides on the basis of the information learned from the training set whether the combination of a given anaphor and its candidate antecedent in the test set is classified as a coreferential link. Since each NP in the test set is linked with several preceding NPs, this implies that one single anaphor can be linked to more than one antecedent, which for its part can also refer to multiple antecedents, and so on. Therefore, a **second step** is taken, which involves the selection of one coreferential link per anaphor.

In our experiments, the two steps are organized as follows. As a first step, in the classification experiments on the instance level, possibly coreferential NPs are classified as being coreferential or not. For the experiments with both learners, we perform 25-fold cross validation on the training data and we evaluate the results of our experiments by computing micro-averaged precision, recall and F-score at the instance level. For the second step, the experiments on the test set of 25 documents, the performance is also reported in terms of precision, recall and

F-measure, but this time using the MUC scoring program from Vilain et al. [25]. The program looks for the evaluation at equivalence classes, being the transitive closure of a coreference chain. [2]

We can illustrate this testing procedure for the coreferential relation between "he" and "President Bush" in the following test sentence.

(2) **President Bush** met Verhofstadt in Brussels. **He** talked with our prime minister about the situation in the Middle East.

For the NP "he" test instances are built for the NP pairs displayed in Table 1. After application of TIMBL or MAXENT, the result of the **first step** might be that the learner classifies the first instance as non-coreferential and the last two instances as being coreferential. Since we start from the assumption that each NP can only corefer with exactly one other preceding NP, a **second step** is required to make a choice between these two positive instances (he - Verhofstadt) and (he - President Bush).

Table 1. Test instances built for the "he" in example (2)

Antecedent	Anaphor	Classification
Brussels	he	no
Verhofstadt	he	yes
President Bush	he	yes

In a second step, the coreferential chains are built on the basis of the positively classified instances. For this step, different directions can be taken: a "closest-first" approach (eg. [5]) in which the first markable found to be coreferent with the anaphor is the antecedent, or an approach [7] which aims to find the most likely antecedent. This is done by selecting the antecedent with the highest confidence value among the candidate antecedent, or a twin-candidate approach [26,9] in which the antecedent for an anaphor is selected after pairwise comparison of the possible antecedents.

Instead of selecting one single antecedent per anaphor, as in the previously described approaches, we tried to build complete coreference chains for our documents. We will now continue with a description of our selection procedure.

2.4 Antecedent Selection

We used the following counting mechanism to recover the coreference chains in the test documents.

1. Given an instance base with anaphor - antecedent pairs $(\text{ana}_i, \text{ant}_{ij})$, for which $i = 2$ to N and $j = i - 1$ to 0. Select all positive instances for each anaphoric

[2] We did not compute significance scores because the scores given by MUC scoring program are not proper input for significance testing.

Table 2. Example of overlap computation between the grouping of ID 2 and the groupings 5, 8 and 20

Overlap	ID+NP	ID+NP
1	5 Loral Space	2 Loral Space
0.08	8 Globalstar	2 Loral Space
0	20 Lockheed Martin Corp.	2 Loral Space

Table 3. Example output from the antecedent selection script. The table shows the incremental construction of one coreferential chain.

ID + Anaphor <- ID + Antecedent
8 Globalstar <- 43 Globalstar Telecommunications Ltd.
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar <- 103 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar <- 103 Globalstar <- 123 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar <- 103 Globalstar <- 123 Globalstar <- 139 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar <- 103 Globalstar <- 123 Globalstar <- 139 Globalstar <- 144 Globalstar
8 Globalstar <- 43 Globalstar Telecommunications Ltd. <- 64 Globalstar <- 102 Globalstar <- 103 Globalstar <- 123 Globalstar <- 139 Globalstar <- 144 Globalstar <- 211 Globalstar

NP. Then make groupings by adding the positive ant_{ij} to the group of ana_i and by adding ana_i to the group of ant_{ij}.
The following is an example of such a grouping. The numbers represent IDs of anaphors/antecedents. The number before the colon is the ID of the anaphor/antecedent and the other numbers represent the IDs which relate to this anaphor/antecedent.

2: 2 5 6 25 29 36 81 92 99 231 258 259 286
5: 2 5 6 25 29 36 81 92 99 231 258 259 286
6: 2 5 6 25 29 36 81 92 99 231 236 258 259 286
8: 8 43 64 102 103 123 139 144 211 286
20: 20 32 69 79

2. Then compare each ID grouping with the other ID groupings by looking for overlap between two groupings. Select the pairs with an overlap value above a predefined threshold. We selected all pairs with an overlap value above 0.1.

For example, we computed the overlap between the grouping of ID 2 with the groupings of IDs 5, 8 and 20 in the previous example as can be seen in Table 2. For the groupings of 2 and 5, we can observe a complete overlap. Combining ID 8 with ID 2, however, leads to a very weak overlap (only on one ID) and an overlap value of 0.08. And no overlap is found for the combination of ID 20 and ID 2. If we take into account an overlap threshold of 0.1, this implies that the two last NP pairs in the table below will not be selected.

3. For each pair with an overlap value above the threshold, compute the union of these pairs. Table 3 illustrates this procedure.

3 Hybrid Versus Data-Driven Resolution

In order to rebalance the highly skewed data sets, positive and negative filters can be applied to the data. These filters split the basic set of instances in two parts: one parts gets a label automatically assigned by the filter, the other part is classified by a classifier. There are several ways to look at this approach. It can be regarded as a language engineering approach, a preprocessing trick, but it can also be made into a principled approach to creating hybrid knowledge-based and machine learning based systems where both approaches solve the problems they are best at. To be able to do this, a systematic study has to be undertaken of the effect of different possible filters.

3.1 Related Research on Instance Selection

Some of the filters proposed in literature aim exclusively at the reduction of negative instances, reducing the positive class skewness. Strube et al. [8], for example, apply a number of filters, which reduce up to 50% of the negative instances. These filters are all linguistically motivated, e.g. discard an antecedent-anaphor pair (i) if the anaphor is an indefinite NP, (ii) if one entity is embedded into the other, e.g. if the potential anaphor is the head of the potential antecedent NP, (iii) if either pronominal entity has a value other than third person singular or plural in its agreement feature. And Yang et al. [9] use the following filtering algorithm to reduce the number of instances in the training set: (i) add the NPs in the current and previous two sentences and remove the NPs that disagree in number, gender and person in case of pronominal anaphors, (ii) add all the non-pronominal antecedents to the initial candidate set in case of non-pronominal anaphors.

Others such as Ng and Cardie [7] and Harabagiu et al. [10] also try to filter out less important or very easy positive instances to force the learning algorithm to specialize on the more difficult cases. Ng and Cardie [7] propose both negative sample selection (the reduction of the number of negative instances) and positive sample selection (the reduction of the number of positive instances), both under-sampling strategies aiming to create a better coreference resolution system. Given the observation that one antecedent is sufficient to resolve an anaphor, they present a corpus-based method for the selection of easy positive instances, which is inspired by the example selection algorithm introduced in [10]. The assumption is that the

easiest types of coreference relationships to resolve are the ones that occur with high frequencies in the training data. Harabagiu et al. [10] mine by hand three sets of coreference rules for covering positive instances from the training data by finding the coreference knowledge satisfied by the largest number of anaphor-antecedent pairs. The high confidence coreference rules, for example, look for (i) repetitions of the same expression, (ii) appositions or arguments of the same copulative verb, (iii) name alias recognitions, (iv) anaphors and antecedents having the same head. Whenever the conditions for a rule are satisfied, an antecedent for the anaphor is identified and all other pairs involving the same anaphor can be filtered out. Ng and Cardie [7] write an automatic positive sample selection algorithm that coarsely mimics the [10] algorithm by finding a confident antecedent for each anaphor. They show that system performance improves dramatically with positive sample selection. The application of both negative and positive sample selection leads to even better performance. But they mention a drawback in case of negative sample selection: it improves recall but damages precision.

Uryupina [11] distinguishes between four types of markables (pronouns, definites, named entities, and all the other NPs) and proposes different sample selection mechanisms, reflecting the different linguistic behavior of these anaphors. In cross-comparative results with and without instance selection she shows an increase on both speed and performance.

All previous approaches concentrate on instance selection through the application of linguistically motivated filters. In [12], this rebalancing of the data is done without any a priori knowledge about the task to be solved and linked to the specific learning behavior of a lazy learner (TIMBL) and an eager learner (RIPPER). This work shows that both learning approaches behave quite differently in case of skewness of the classes and they also react differently to a change in class distribution.

The described selection approaches provide very few results on the effect of these filters on performance. In case cross-comparative results are provided, this is done in a coarse-grained manner. In the remainder of this paper, we will discuss our selection of filters and investigate in a fine-grained fashion whether these filters contribute to classification performance and how.

3.2 Positive and Negative Filtering

Our hybrid approach works as follows. After instance creation, each instance is matched against the filter rule. The subset of instances that match with the filter rule are labeled by the filter. The other part of the instance set is handled by the classifier. The filter rule is applied to both training and test instances.

In order to assess the effect of filtering on classification results, we investigate the following filters:

- *fdef*: The first filter rule we investigated, filters out all instances containing an indefinite anaphor and assigns a negative label to these instances.
- The filter *fhead* filters out instances in which the anaphor and antecedent are located at a distance of more than three sentences from each other. Instances

beyond the scope of three sentence, which share the same head word, are retained in the data set.
- The filter *fagree* applies to pronouns only and demands agreement between anaphor and antecedent. The filter removes instances in which the antecedent does not have the same number or an incompatible gender type.
- The filter rule *fmatch* is based on [7] and is the only filter that also assigns positive labels. The filter assigns a positive label to an instance that describes an anaphor and antecedent which have a complete string match (discarding determiner information). All other instances containing that same anaphor with another antecedent get a negative label.
- The filter *f3s* restricts the search space for pronouns to three sentences. The filter rule assigns all pronoun-antecedent pairs at a larger sentence distance a negative label. So this filter deliberately can remove part of the positive instances, based on the observation that most pronouns refer to a close-by antecedent.

These filters are also combined: the filter combination 1 combines the four negative filters, whereas the filter combination 2 combines all filters together.

On the one hand, these simple filter rules are aimed at the removal of negative instances to change the balance between positive and negative instances. On the other hand, some of the filter rules also deliberately remove positive instances. As explained in Section 2.3 one anaphor can be preceded by multiple coreferential antecedents, but we only need to resolve one antecedent to build up the coreferential chain. The filters *f3s, fhead* and *fmatch* are based upon this principle. The rule *f3s* is based on the observation that pronouns usually find an antecedent within three sentences. The filter therefore can filter out positive antecedents that are at a larger distance from the anaphor. The filter *fmatch* removes other possible antecedents if one confident antecedent (complete match) has been found. The *fhead* rule assigns a negative label to potentially coreferential NPs at a sentence distance larger than 3, if they do not share the same head word with the anaphor under consideration.

Table 4 gives an overview of the number of training instances that are left to be handled by the classifier after the different filters have been applied. The last column of the table shows the number of positive instances in those training sets. The results show that these filters account for a large part of the data and indeed lead to a less skewed data set, except for the filters *f3s* and *fmatch*. The *fhead* filter has also removed a part of the positive examples but a much larger part of the negatives, leading in the end to a positive effect on the class balance. The combination 2 filter, for example, accounts for 91.7% of the data, only leaving 6,286 instances to be classified by the learner. These instances also have a less skewed distribution.

We go on to investigate whether the skewness of the data is indeed harmful for the classifiers and whether filtering leads to better classification results and a better overall performance.

Table 4. Number of training instances after application of the filters. The second column gives the absolute numbers, whereas the third column shows the percentage of training instances left to be treated by the classifier. The last column shows the skewness of the class distribution in those data sets.

Filter	number	%inst	%pos
normal	76,920	100	8.5
fdef	64,656	84.1	9.2
fagree	66,786	86.8	9.2
f3s	59,183	76.9	7.5
fhead	15,041	19.6	19.5
fmatch	57,479	74.4	8.0
combi1	9,723	12.6	20.3
combi2	6,286	8.3	17.7

3.3 Results

We first consider the results of the 25-fold cross validation experiments on the training set in which we evaluate the performance of the first step of our approach: classifying NP pairs as being coreferential or not. Table 5 shows the micro-averaged F-scores of both classifiers, on the left hand side computed on all training instances (the joint effect of filters and machine learning) and on the right hand side on the subset of instances classified by the learner (the work that is left after the application of the filters). In general, the overall F-scores (in the right column) of the hybrid systems measured are lower than the F-score of the systems without filtering (*default*). Specially the f3s filter has a low score which can be explained by the fact that this rule deliberately labels part of the positive training instances as being negative (the instances where the distance between pronouns and antecedents is larger than three). When we look at the subset of instances classified by a learner, we observe for MAXENT that each filter improves the F-score on the subset. For TIMBL we observe only for some of the filters an improvement. This observation for TIMBL is in line with earlier findings in [27,12].

Another observation relates to skewness. Three of the filters (*fhead, combi1, combi2*) change the class balance between positive and negative instances drastically as shown in Table 4. We can observe for both MAXENT and TIMBL that these three filters lead to the highest classifier F-scores on the cross-validation data. We do not see this clear effect on the test set. We believe this is due to the difference in measurement. On the test set, we measure F-scores at coreference chain level, and we do not need to retrieve all positive instances to build the complete coreference chain.

We now discuss the results on the test set showing MUC-scores computed at the coreference chain level. We computed a baseline score by assigning each NP in the test set its most nearby NP as antecedent. This gives us a baseline score with a high recall of 63.1%, a precision of 22.7% and an F-score of 33.4%.

The results on the test set for MAXENT and TIMBL are shown in Table 6. For TIMBL, we observe that all hybrid systems improve the precision of the system at

Table 5. Summary of micro-averaged F-scores of 25-fold cross validation experiments on the training set for MAXENT and TIMBL with and without the different filters.The left part of the table shows the results of the combined filters and learners, whereas the right part of the table only considers the subset of instances classified by the learners.

	MAXENT	TIMBL	#num.	MAXENT	TIMBL
default	37.6	46.7	76,920	37.6	46.7
fdef	37.6	44.2	64,656	40.0	46.8
fagree	37.9	44.7	66,786	39.5	46.4
f3s	31.6	35.2	59,183	41.5	45.2
fhead	34.8	39.7	15,041	58.3	67.0
fmatch	43.1	43.6	57,479	39.0	39.7
combi1	29.3	31.3	9,723	65.9	70.8
combi2	31.5	30.5	6,286	55.6	54.0

Table 6. MUC scores on the test set of TIMBL and MAXENT with and without the different filters

	TIMBL			MAXENT		
	recall	precision	F-score	recall	precision	F-score
normal	**60.0**	**35.2**	**44.4**	**41.7**	**42.2**	**42.0**
fdef	49.2	46.7	47.9	39.5	46.4	42.7
f3s	58.0	36.8	45.1	51.2	43.8	47.2
fagree	50.2	40.4	44.7	41.3	42.3	41.8
fhead	39.8	60.3	47.9	45.5	42.7	44.1
fmatch	46.7	48.4	47.5	51.2	42.4	46.4
combi1	40.7	46.1	43.2	38.5	51.6	44.1
combi2	**36.7**	**61.0**	**45.8**	**40.0**	**51.8**	**45.1**

the expense of recall. Only in the case of the *combi1* filter this leads to a lower F-score; in all other cases the shift leads to a higher F-score. For MAXENT, all hybrid systems except *fagree* have a higher F-score on the test set than the default system. Each filter produces a higher precision and in the case of *f3s, fhead* and *fmatch* also a higher recall.

4 Concluding Remarks

We have shown that two distinct learning techniques benefit from a combined approach of knowledge-based filtering and machine-learning based classification. We observe that our simple filter rules can provide a large reduction in the number of instances to be classified. The filters improve the overall precision of the system on the test set leading to higher F-scores in almost all experiments.

Most successful is the *combi2 filter* which combines five simple filter rules and leads to a large instance reduction of up to 92%, and produces a better F-score on the test set for both MAXENT and TIMBL.

As future work we plan to investigate the filter rules in contrast to a machine learning approach in which the feature weights correspond to the filters are boosted. It would also be interesting to investigate whether similar filter rules have a similar positive effect for other languages. In contrast to a pure machine learning approach, a hybrid approach has the disadvantage that it may require careful re-engineering of the knowledge-based part for different languages.

References

1. Cardie, C., Wagstaff, K.: Noun phrase coreference as clustering. In: Proceedings of the 1999 joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora. (1999) 82–89
2. Quinlan, J.: C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, CA (1993)
3. Aone, C., Bennett, S.: Evaluating automated and manual acquisition of anaphora resolution strategies. In: Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-1995). (1995) 122–129
4. McCarthy, J.: A Trainable Approach to Coreference Resolution for Information Extraction. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst MA (1996)
5. Soon, W., Ng, H., Lim, D.: A machine learning approach to coreference resolution of noun phrases. Computational Linguistics **27** (2001) 521–544
6. Cohen, W.W.: Fast effective rule induction. In: Proceedings of the 12th International Conference on Machine Learning (ICML-1995). (1995) 115–123
7. Ng, V., Cardie, C.: Combining sample selection and error-driven pruning for machine learning of coreference rules. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). (2002) 55–62
8. Strube, M., Rapp, S., Müller, C.: The influence of minimum edit distance on reference resolution. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP-2002). (2002) 312–319
9. Yang, X., Zhou, G., Su, S., Tan, C.: Coreference resolution using competition learning approach. In: Proceedings of the 41th Annual Meeting of the Association for Compuatiational Linguistics (ACL-03). (2003) 176–183
10. Harabagiu, S., Bunescu, R., Maiorano, S.: Text and knowledge mining for coreference resolution. In: Proceedings of the 2nd Meeting of the North American Chapter of the Association of Computational Linguistics (NAACL-2001). (2001) 55–62
11. Uryupina, O.: Linguistically motivated sample selection for coreference resolution. In: Proceedings of DAARC-2004. (2004)
12. Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, Antwerp University (2005)
13. MUC-7: Muc-7 coreference task definition. version 3.0. In: Proceedings of the Seventh Message Understanding Conference (MUC-7). (1998)
14. Davies, S., Poesio, M., Bruneseaux, F., Romary, L.: Annotating coreference in dialogues: Proposal for a scheme for mate. http://www.hcrc.ed.ac.uk/ poesio/MATE/anno_manual.htm (1998)

15. van Deemter, K., Kibble, R.: On coreferring: Coreference in muc and related annotation schemes. Computational Linguistics **26** (2000) 629–637
16. Daelemans, W., Zavrel, J., Berck, P., Gillis, S.: Mbt: A memory-based part of speech tagger generator. In: Proceedings of the 4th ACL/SIGDAT Workshop on Very Large Corpora. (1996) 14–27
17. Tjong Kim Sang, E., Daelemans, W., Höthker, A.: Reduction of dutch sentences for automatic subtitling. In: Computational Linguistics in the Netherlands 2003. Selected Papers from the Fourteenth CLIN Meeting. (2004) 109–123
18. Daelemans, W., van den Bosch, A.: Memory-based Language Processing. Cambridge University Press (2005)
19. Guiasu, S., Shenitzer, A.: The principle of maximum entropy. The Mathematical Intelligencer **7** (1985)
20. Berger, A., Della Pietra, S., Della Pietra, V.: Maximum Entropy Approach to Natural Language Processing. Computational linguistics **22** (1996)
21. Le, Z.: Maximum Entropy Modeling Toolkit for Python and C++ (version 20041229). Natural Language Processing Lab, Northeastern University, China. (2004)
22. van den Bosch, A.: Wrapped progressive sampling search for optimizing learning algorithm parameters. In: Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence. (2004) 219–226
23. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artificial Intelligence **97** (1997) 273–323
24. F. Provost, D.J., Oates, T.: Efficient progressive sampling. In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining. (1999) 23–32
25. Vilain, M., Burger, J., Aberdeen, J., Connolly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the Sixth Message Understanding Conference (MUC-6). (1995) 45–52
26. Connolly, D., Burger, J., Day, D.: A machine learning approach to anaphoric reference. In: Proceedings of the International Conference on 'New Methods in Language Processing'. (1994)
27. Daelemans, W., van den Bosch, A., Zavrel, J.: Forgetting exceptions is harmful in language learning. Machine Learning **34** (1999) 11–41

Automatic Anaphora Resolution for Norwegian (ARN)

Gordana Ilić Holen

University of Oslo
Department of Literature, Area Studies and European Languages
Postboks 1003, Blindern 0315 Oslo
g.i.holen@ilos.uio.no

Abstract. The ARN system — an Automatic Anaphora Resolution System for Norwegian — is a rule-based anaphora resolution system that was designed on the basis of two existing systems for the English language: Mitkov's Original Approach with its later development MARS, and the RAP system by Lappin and Leass. A substantial group of rules within these systems is based upon a super-rule supported by Centering theory, which gives preference to subjects candidates over objects candidates, and object candidates over candidates within adverbial and prepositional phrases. These rules cannot be applied to Norwegian, due to differences in information structure between Norwegian and English. Although there is a tendency in both languages to avoid conveying new information with the subject, Norwegian goes to much greater lengths to avoid it. This tendency leads to a substantially higher number of sentences with expletive subjects in Norwegian than in English, rendering those subjects unsuitable as antecedent candidates.

Making a complex preference to handle the sex/gender conflict and giving preference to pronominal candidates and candidates in close proximity to the anaphor has proved to be a good strategy for Norwegian.

ARN was designed to resolve the third person pronoun with the exception of pronoun *det* 'it (neut.)', and has achieved an accuracy of 70.5%.

ARN is an anaphora resolution system designed for Norwegian. Most of the current anaphora resolution systems are made for English (e.g. [1,8,7,9,10]). Why does Norwegian, which is so closely related to English, need its own anaphora resolution system? In this paper, I argue that the differences in information structure between Norwegian and English are deep enough that a different approach for Norwegian is needed.

The systems by which ARN was directly inspired — Mitkov's Original Approach (henceforth MOA) [9], MARS [10], and RAP by Lappin and Leass [7] — are all rule-based. They apply sets of rules (*factors*) that award scoring points to antecedent candidates during the resolution of an anaphor, and ultimately propose the candidate with better score as the most likely antecedent. Most of the rules that are common for MARS and RAP are based on the following super-rule:

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 151–166, 2007.

> Give preference to candidates that are subjects over candidates that are direct objects, to direct objects over indirect objects and to indirect objects over other constituents such as adverbial or prepositional phrases.

Four of the seven rules used by RAP are based on this super rule, and RAP is, with 86% correctly resolved anaphora, considered a very successful anaphora resolution system. This super rule does not, however, seem to work for Norwegian. Giving preference to subject candidates has shown to impair ARN's performance, and the same was observed for the penalizing of prepositional phrase candidates. The introduction of other parts of the super rule, such as rewarding object candidates and penalizing candidates that are part of adverbial phrases, revealed to have a minimal impact on the system's performance.

This is a surprising finding considering the fact that Norwegian and English are closely related languages, and the fact that MARS has been successfully applied to such different languages as Polish and Arabic. In the present paper, I propose that this result is due to a different information structure for Norwegian and for English.

1 The Architecture of the System

1.1 Scope

ARN has been designed to resolve Norwegian third person personal pronouns (table 1) with the exception of the pronoun *det* 'it (neut.)'.

Table 1. The third person pronouns in Norwegian

	Singular					Plural
	Human			Non-Human		
	m.	f.	Polite (m. & f.)	m. & f.	*n.*	
Subject form	han	hun	De[1]	den	*det*	de
Object form	han ham	henne	Dem	den	*det*	dem

The problems of resolving the Norwegian *det* 'it (neut.)' are similar to those of resolving English *it*, though at some points *det* poses even greater challenges:

– The anaphor *det* is ambiguous with the definite article *det*:

(1) det røde eplet
the/it red the-apple
the red apple

[1] The pronoun of polite addressing *De* has in the last couple of decades all but disappeared from the language, and can now be seen only in older texts.

– The anaphor *det* is ambiguous with the expletive subject *det*:

(2) Det kom en mann nedover veien.
it came a man down the-road

A man came down the road.

(3) Det kom nedover veien.
It came down the-road

It came down the road.

– And, finally, Norwegian *det* can, just like its English counterpart, have non-nominal antecedents:

(4) John sparka Tom. Det var ikke særlig kult.
john kicked tom it was not particularly cool

John kicked Tom. That wasn't very cool.

As the time-frame of the project did not allow undertaking such an extensive investigation, I have chosen not to attempt the resolution of anaphor *det* 'it (neut.)' from ARN. This pronoun is thus neither included in the total anaphor count that are to be resolved by ARN, nor is it taken into consideration when the percentage of correctly resolved anaphora was calculated.

1.2 Data

The data used in this project consisted of Norwegian newspaper and literary texts taken from The Oslo Corpus [11]. The Corpus uses the following tags:

– Syntactic: @subj *subject*, @obj *object*, @i-obj *indirect object*, @iv *infinite verb* @fv *finite verb*
– Morphological: verb *verb*, pres *present*, subst *noun*, mask *masculine*, ub *non-determined*, fl *plural*
– Semantic: &person *person*, &org *organization*, &sted *location*, &verk *publication*, &hend *event* and &annet *miscellaneous*

The data set also includes a parallel set of files containing the same texts manually tagged for referential chains[2]. These were used for evaluating the ARN system. The entries of this second set contain id-number, token, ending, lemma, morphological class, syntactic function and, for anaphora, the id-number of the antecedent. Here are some examples of entries from the main and the secondary corpus (English translations added in italics):

(5)
```
"<venter>"  WD b1357=M b1358=M b526=S0 b730=S0 b771=S0
            "vente" verb pres @fv
            "vent" subst appell mask ub fl @obj @subj
```

[2] These files were obtained from the BREDT project at the University of Bergen (http://bredt.uib.no/).

(6) `b('292','Utålmodig','dig','utålmodig','a','adv').` *impatiently*
`b('293','venter','ter','vente','v','fv').` *waits*
`b('294','hun','hun','hun','pron','subj','254').` *she*

The files contain 46972 words that were divided into a training corpus consisting of 21800 words (693 anaphora[3]) and a test corpus of 25172 words (939 anaphora) before the work on ARN started. The corpora have an approximately even distribution of newspaper texts and literary texts. In addition, a set of five word lists[4] comprising a total of 1999 words were used to determine whether a given noun denotes a human being. The lists included words denoting agents of temporary activity (133 words), agents of persistent activity (83 words), kinship relations (66 words), professions (463 words) and geographic origin and nationalities (1254 words).

1.3 The Process of Resolution

Upon finding an anaphor, ARN makes a list of possible resolution candidates. This list consists of all nouns and third person pronouns that precede the anaphor in the current sentence and all nouns and third person pronouns in the two preceding sentences, thus defining a three-sentence window within which the anaphora resolution is performed. Each rule or factor applied by ARN rewards or penalizes each candidate with positive or negative points, ranging from -100 to +100. The final score of each candidate in the current sentence is multiplied by 1.0, in the penultimate sentence by 0.75, and that of each candidate in the antepenultimate sentence by 0.50. The candidate with the highest score is selected, and if its score exceeds a threshold — set at zero points in ARN — the candidate is proclaimed antecedent, and if not, it is announced that no appropriate candidate was found.

Resolving anaphora to nominal candidates is complicated by a rather complex gender situation in Norwegian, which has two coexisting systems: a two-gender system (common and neuter) and a three-gender system (masculine, feminine and neuter). The two-gender system is actually a three-gender system in which the female gender has been collapsed into the male gender, giving a common gender that is identical to the male gender. According to the newest official standard of written Norwegian Bokmål language, all the words with feminine gender can take masculine gender and be inflected thereafter. Today, the systems are used interchangeably, and the use of both systems within a single declination paradigm is relatively widespread [3, p. 152]. All three paradigms given in Figure 1 are considered correct.

[3] Here, and in the remainder of the article, the terms *anaphora* and *third person pronouns* will be used interchangeably, as meaning "the third person pronouns excluding the pronoun *det* 'it (neut.)' ".

[4] The first four lists were compiled by the SIMPLE-project (http://cst.dk/simple/index.html), and modified by Andra Björk Jónsdóttir and Lilja Øvrelid. The fifth one is from the The Norwegian Language Council (Språkrådet) (www.sprakradet.no).

en bok	– boken	– bøker – bøkene
common/masc.	*common/masc.*	
ei bok	– boka	
fem.	*fem.*	
en bok	– boka	
masc.	*fem.*	

Fig. 1. A book - the book - books - the books

The gender of indefinite nouns is not marked in the noun itself, but on some of its determiners, while the gender of definite nouns is marked both in the noun itself, and on (most of) its determiners.

2 Factors

The following factors were initially implemented in ARN.

- *Factor 1: Number/Gender/Animacy factor.*
- *Factor 2: The sentence proximity factor.*
- *Factor 3: Boost pronoun.*
- *Factor 4: Subject preference* (discarded from the final version).
- *Factor 5: Direct object preference.*
- *Factor 6: Indirect object preference.*
- *Factor 7: Adverbial phrase penalization.*
- *Factor 8: Prepositional phrase penalization* (discarded from the final version).
- *Factor 9: Syntactic parallelism.*
- *Factor 10: Section heading preference.*
- *Factor 11: Indefiniteness penalization.*

2.1 Factor 1: Number/Gender/Animacy Factor

The Number/Gender/Animacy factor uses a whole range of points (from -100 to +100) to express the probability that a candidate denotes a non-human entity or a person of male or female sex and awards and penalizes the candidates accordingly.

In many cases, but not always, gender corresponds to sex. It is common to apply some sort of morphological filter to deal with candidates that do not match in gender and/or number during the preprocessing. This problem has not received much attention: a good illustration is Kennedy and Boguraev [8] who only mention "a set of morphological filters which eliminate from consideration any discourse referent which disagrees in person, number or gender, with the pronoun". However, their result analysis shows that 35% percent of all the mistakes made by their own anaphora resolution system is due to gender mismatch, and one of their system improvement propositions is including a lexical data-base which includes detailed gender information.

In ARN, this factor has been implemented as a preference rather than a constraint. Instead of filtering out nouns of non-fitting gender, it classifies candidates in animacy/gender groups and rewards or penalizes them accordingly, depending on the pronoun it is resolving.

Nominal candidates have been categorized in four classes according to their gender and animacy.[5] The classification is based on the semantic tags of the Oslo Corpus entries and the five word lists described in section 1.2.

- **Class 1: Nouns that do not denote humans**
  This class contains all nouns that are not tagged as human proper names in the corpus, and that are not found in any of the five lists. As the nouns in this group do not have natural gender, their grammatical gender is taken into account, so that this class has three subclasses, masculine, feminine and neutral (m, f and n in Table 2).
- **Class 2: Nouns that denote humans, with gender that cannot be determined**
  This class contains all the members of the five lists that denote humans of gender that cannot be determined. In addition to these, this class also contains all nouns tagged as human names of unknown gender, such as most of the foreign (non-Norwegian) names.
- **Class 3: Nouns denoting males**
  This class consists of proper names tagged as male and elements of the lists that denote persons that necessarily are male, such as *bror* 'brother', *skjørtejeger* 'womanizer' and *baryton* 'baritone'.
- **Class 4: Nouns denoting females**
  This class includes proper names tagged as female and members of lists that contain nouns that necessarily denote females, such as *fristerinne* 'temptress', *amme* 'wet-nurse' and *talskvinne* 'spokes-woman'.

According to the same classification, the third person pronouns can be classified as:

- A pronoun that refers to non-humans: *den* 'it (m./f.)'.
- A pronoun that may or may not refer to humans: *de* 'they'/*dem* 'them'.
- A pronoun that refers to humans of unknown gender: *De*[6] 'you (court.)'/*Dem* 'you (court.acc.)'.
- A pronoun that refers to males: *han* 'he'/ *ham* 'him'.
- A pronoun that refers to females: *hun* 'she'/ *henne* 'her'.

Both the third person pronouns and the aggregated scores awarded by Factor 1 to candidates belonging to the different classes are given in Table 2.

[5] Animacy is not an unproblematic term. In its main sense, it denotes the attribute of being alive, as opposed to inanimate (non-living) objects. In a narrower sense, it denotes humans as opposed to both inanimate objects and living beings that are not humans. I will use the term in the latter, narrower sense.

[6] The pronoun for polite addressing (*De*) is not normally used in modern Norwegian, but it has been included since the corpus contains some older texts, written at a time when its usage was widespread.

Table 2. Points awarding by Factor 1

Pronouns			*den* 'it(m/f)'	*de* 'they'	*De* 'you'	*han* 'he'	*hun* 'she'
Nouns	Class 1: *Not human*	m	100	0	0	-100	-100
		f	100			-100	-100
		n	0			-100	-100
	Class 2: *Human*		-50	75	75	100	75
	Class 3: *Human male*		-50	75	75	100	0
	Class 4: *Human female*		-50	50	50	0	100

The points in Table 2 were initially set according to grammar-based rules of thumb. For instance, they inlcude a rule according to which, when the resolution of the pronoun *hun* 'she' is at stake, words that denote female persons should be given the maximal sum and words denoting, say, non-living entities of neutral gender should be penalized. These rules were later experimentally adjusted. I will not analyze this factor further here, the object of this short overview is to illustrate that the question of sex and gender is a complex one and deserves more attention.

2.2 Factor 2: The Reference Proximity Factor

The reference proximity factor awards 100 points to all candidates from the sentence which contains the anaphor, 50 points to the candidates from the penultimate sentence, while candidates from the ante-penultimate sentence get no extra points.

2.3 Factor 3: Boost Pronoun

The boost pronoun factor awards the pronominal candidates 75 points. This factor is included because pronominalized entities tend to be more salient [10]. In addition, the antecedent NP can be out of range of the algorithm and pronouns can thus be used as "stepping-stones" between the anaphor and a distant antecedent.

2.4 Centering Factors: Factors 4, 5, 6, 7 and 8

In ARN, the factors based on the centering theory are the following:

- *Factor 4: Subject preference*
- *Factor 5: Direct object preference*
- *Factor 6: Indirect object preference*
- *Factor 7: Adverbial phrase penalization*
- *Factor 8: Prepositional phrase penalization*

Most of these factors have in different forms been applied in MOA, MARS and RAP. In MOA, which does not have access to syntactic information, the first NP, which is assumed to be a subject, gets +1 point from the *First noun phrases/*

Givenness factor. In MARS, *The obliqueness factor* awards the subject NP the maximum +2 points, direct object obtains +1 point, indirect object obtains zero points, and the NPs whose function the parser cannot identify are penalized with -1. Furthermore, NPs that are part of a prepositional phrase are penalized with -1 point. In RAP, these factors are covered by four salience weights: *Subject emphasis* (80 pts), *Accusative emphasis* (50 pts), *Non-adverbial emphasis* (50 pts) and *Indirect object and oblique component emphasis* (40 pts).

As the background for his obliqueness factor Mitkov [10] presents the following hierarchy of grammatical functions:

SUBJECT > DIRECT OBJECT > INDIRECT OBJECT > (OTHER).

It is based on the hierarchy of forward-looking centers (C_f) of the Centering theory by Grosz et al.[4]. An extended C_f hierarchy is given by Brennan et al. [2]:

SUBJECT > OBJECT > OBJECT2 > OTHER SUBCATEGORIZED FUNCTIONS > ADJUNCTS.

However, Brennan et al. do say that they are aware that this hierarchy reflects the surface structure in English and stress the need for more research on languages with different word order, such as German, as well as languages that provide an identifiable topic function, such as Japanese.

Factor 4: Subject Preference Factor (Discarded from the final version) Initially, *The subject preference factor* was meant to award the candidates that are subjects +50, +75 or +100 points. However, experimentation showed that this led to a deterioration of the results in all the files of the training corpus, apart from two files for which the number of correctly resolved anaphora remained the same. The best results were thus achieved by excluding this factor from the system.

I believe that the reason for the failing of this factor is to be found in the differences in information structure between Norwegian and English. As I have previously mentioned, ARN does not resolve the pronoun *det* 'it (neut.)', nor does it try to identify expletive subjects. Not being able to distinguish between a logical and formal subject would perhaps not pose such a significant problem if it were not for the fact that Norwegian uses the expletive pronoun much more than English does, especially in the subject position, leaving many subjects unsuitable as reference candidates. The Norwegian Reference Grammar [3, p. 691] states that the subject is not normally a carrier of new information. In this, Norwegian does not differ from English. However, Norwegian goes to greater lengths to avoid having a subject convey new information. According to Norwegian Reference Grammar [3, p. 1092] example (7) is not the natural answer to the question *Who found the money?*.

(7) Nils fann pengane.
nils found the-money

The natural answer would be (8):

(8) Det var Nils som fann pengane.
it was nils who found the-money

One way of keeping new information away from the subject position is, indeed, using the cleft construction as in (8). Cleft constructions in Norwegian also have the function of highlighting the information in the cleft clause, but this function is secondary to the function of preventing the new information from becoming a subject [3]. This may be a reason for the higher occurrence of cleft forms in Norwegian than in English. Further support for this idea is presented in Gundel [5], who compared a Norwegian text with its English translation and came to the conclusion that clefts are much more commonly used in Norwegian than in English, as only 28% of Norwegian clefts were translated as clefts in English. Although she worked with a single text and a single translator, the results correspond to a more extensive study of clefts, pseudo-clefts and inverted pseudo-clefts for Swedish and English by Johansson [6]. The study, based on 500 tokens of Swedish clefts, found that only 33% of clefts were translated as clefts in English. As there seems to be no differences in distribution restrictions between Norwegian and English clefts, Gundel proposes that the reason for the more frequent use of clefts in Norwegian is that it shows a more consistent mapping between information structure and syntactic structure by making a clear distinction between topic and focus as well as between presupposed and non-presupposed content.

Besides clefts, Norwegian has other constructions that include expletive *det* 'it', such as presentational form (9) and impersonal passive (10).

(9) Det arbeidet en mann i skogen.
it worked a man in the-wood

A man worked in the wood.

(10) Det vart overrekt vinnaren ein pokal.
it became presented the-winner a cup

The winner was presented a cup.

The constructions with formal subjects not only render subjects unsuitable as antecedents, they also influence salience of other parts of the sentence, such as direct and indirect objects and adverbial and prepositional phrases. More indirectly, they also influence the salience of candidates in indefinite form, and thereby also the effect of *Indefiniteness penalization* (Factor 11).

Factor 5: The Direct Object Preference Factor. This factor awards 50 points to the candidates that are direct objects. Including *The direct object preference factor* into the system improved the results on the training corpus by only 0.29 percentage points.

In Norwegian constructions that are used to avoid having the subject convey new information, the role of bringing in new information often falls on the direct object. This is a good reason for favoring direct object NPs, in addition to it being supported by the centering theory. This factor did, indeed, raise the results of the training corpus, but the impact was weaker than expected: since *The subject preference factor* fails, we could expect this factor to at least take

over its impact. The result improvement of 0.29 percentage points reflects four extra correct resolutions and one additional wrong resolution. All four anaphora that became correctly resolved when this factor was introduced to the system replaced antecedents that belonged to the same referential chains. In other words, those resolutions were not actually wrong in the first place.

All we can conclude from the data is that the factor is inconclusive — excluding or including it did not have any significant impact on the training corpus.

Factor 6: The Indirect Object Preference Factor. This factor awards 50 points to candidates that are indirect objects. Including this factor in ARN led to slightly impaired results when ARN was applied to the training corpus: the number of correctly resolved anaphora fell from 512 to 511, or from 73.89% to 73.74%. The single case where ARN rejected a correct resolution was an example where the chosen candidate was a part of the same referential chain as the correct one. This minor deterioration in performance can therefore be dismissed.

Why did this factor have such a negligible effect? There are two possible explanations:

- Awarding positive points to indirect object NPs is a good strategy in half of the cases and counterproductive in the rest of the cases, so the effects cancel out.
- This factor is over-shadowed by other factors that award or penalize the same candidates that would otherwise be preferred by this factor; its influence is therefore not felt. Those factors could, for instance, be *The number/gender/animacy factor* (Factor 1), *The reference proximity factor* (Factor 2) or *Syntactic parallelism* (Factor 9).

Factor 7: Adverbial Phrase Penalization. Candidates that are parts of adverbial phrases are penalized with -50 points. Applied on the training corpus this factor did not change the result. Penalizing factor with any other sum or awarding candidates positive points impaired the result.

Factor 8: Prepositional Phrase Penalization (Discarded from the final version). Initially, this factor was meant to penalize candidates that are part of prepositional phrases (by -50 or -25 points). However, experiments showed that the best results were achieved by giving a weak preference (+25 points) to this type of candidates. When this version of the factor was included in ARN and applied to the training corpus, the results improved by 0.43 percentage points. The factor was however excluded from the system although its performance was better than the performance of some of the other factors that were retained. There are two reasons for this decision: Firstly, resolution analyses showed that in all the cases where the inclusion of Factor 8 led to the correct resolution, the newly proposed candidate belonged to the same referential chain as the old one. The improvement of the result can thus be considered accidental; Secondly, in contrast to the failing of Factor 4, I could not find any explanation for why giving a preference to candidates that are part of a prepositional phrase should be beneficial for an anaphora resolution system.

Why did the Centering Factors Perform So Poorly? It was clear from ARN's performance that including the *The subject preference factor* (Factor 4) led to impairment of the results. However, including of the rest of the factors of this group led to much less conclusive results: their presence or absence led only to minor fluctuations in the results of less then 0.5 percentage points per factor. There can be several reasons for this, including:

- **The problem of *det* 'it (neut.)'**
  ARN does not resolve the pronoun *det* 'it (neut.)', and does not try to disambiguate it from the expletive subject of the same form. Since we do not know how many of the subjects are an expletive *det* 'it', we do not know anything about the distribution of either presentational and topicalized constructions, or impersonal passives. This distribution influences not only the subject's suitability as a potential antecedent, but also the performance of all the Centering theory-based factors, i.e. Factors 5, 6, 7 and 8 and, to some extent, Factor 11.
- **Reference chains**
  If the proposed candidate itself is an anaphor, ARN can check whether the candidate has the same antecedent as the anaphor, and, in a case of match, proclaim the resolution correct. This means that if anaphor a_1 refers to anaphor a_2, and a_2 refers to noun n_1, then if the anaphor a_1 is resolved to the noun n_1, this resolution is rendered correct even though the referential bond a_1-n_1 is not tagged in the secondary corpus. In this way, ARN makes short referential chains of three links ($a_1 \rightarrow a_2 \rightarrow n_1$). The limited length of these chains poses a problem in the evaluation of the factors' impact on the systems. There were several instances where applying a factor apparently led to incorrect resolution, but where the antecedent selected by ARN and the antecedent manually tagged as correct in the secondary corpus, belonged to the same referential chain. The cases of the reverse situations were also numerous, making the evaluation of a single factor's contribution to the system imprecise.
- **The problem of data**
  The size of the data set has been limited by the size of the manually annotated corpus that has been essential for ARN's training and evaluation, and all the results should be seen in the light of this limited data set. More data and more experiments are needed to get a more nuanced picture of the centering factors' contributions.

For these reasons I have decided to exclude only two factors from ARN – *Subject preference* (Factor 4) and *Prepositional phrase penalization* (Factor 8) – while other factors are retained in anticipation of a new version of the system. The final version of ARN contains thus nine rules.

2.5 Factor 9: Syntactic Parallelism

The syntactic parallelism factor awards 50 points to the candidates fulfilling the same syntactic role as the anaphor.

The inclusion of this factor into ARN led to an improvement of the system's performance in most of the files of the training corpus. The overall improvement was 1.15% percentage points which is one of the higher contributions of a single factor to the system. Interestingly, this factor's contribution was much higher in the previous version of ARN in which Factor 8 was also present, when including Factor 9 improved the result by 4.47 percentage points. The cases where this factor was decisive in choosing the correct candidate were divided between the ones in which both the candidate and anaphor were subjects and the cases where both were prepositional phrase supplement, which could explain the strong influence of the later discarded Factor 8 on this factor.

2.6 Factor 10: Section Heading Preference

The section heading preference factor awards 50 points to the candidates that also appear in the section heading. When the factor was introduced to the system, it led to only one additional correctly resolved anaphor. The explanation for the low impact of this factor could be that it was applicable only to one file of the training corpus, the one that contains newspaper articles. Although it is by far the largest file in the corpus, covering approximately half of it, it is also the file with the lowest percentage of third person pronouns (1.16% words compared to 4.87% in the rest of the files).

2.7 Factor 11: Indefiniteness Penalization

The indefiniteness penalization factor penalizes the candidates that are in the indefinite form with -25 points. Including this factor in ARN led to a minimal improvement of 0.3 percentage points.

Why this small impact? Factor 11 was initially introduced to promote definite NPs because they tend to be the theme of discourse. However, in the constructions that include expletive *det* 'it', this strategy is counter-productive. In these sentences, the direct object has to be in the indefinite form and penalizing direct objects in this setting is a wrong strategy. The positive effect of factor 11 on the rest of the sentences seems to be canceled out by its negative effect on sentences with expletive *det* 'it'.

3 Results and Factor Evaluation

On the training corpus, ARN correctly resolved 73.73% anaphora, which is only 3.23% higher than the 70.50% later achieved on an unknown (test) corpus.

3.1 Comparison with Baseline Models

ARN has been compared with the following baseline models:

- B1: The closest candidate (i.e. the closest noun) is proposed as antecedent.
- B2: The closest candidate that matches in gender and number is proposed as antecedent.

Table 3. Results of applying ARN to a test corpus

FILE NO.	Number of words	Number of anaphora	Correctly resolved	Incorrectly resolved	No approp. candidates	% correctly resolved
1	9290	91	67	23	1	73.62637
10	1988	92	42	39	11	45.65213
11	1993	70	51	18	1	72.85714
12	1976	114	83	29	2	72.80701
13	1998	60	38	18	4	63.33333
14	1976	171	141	29	1	82.45614
15	1978	134	85	44	5	63.43283
16	1966	122	94	27	1	77.04918
17	1988	85	61	24	0	71.76471
TOTAL	25153	939	662	251	26	70.50053

When the baseline model B1 was applied to the test corpus, it resolved correctly 22.36% of the anaphora, while model B2 achieved a result of 45.56%. Both of these results are thus considerably lower than ARN's 70.50%.

3.2 Relative Importance and Decision Power

The *Relative importance* measure [10] indicates a single factor's contribution to the system. Relative importance (RI_K) of a factor K is defined through Equation 1.

Equation 1

$$RI_K = \frac{SR - SR_K}{SR}$$

where SR is the success rate[7] of the system and SR_K is the success rate of the system when the factor K is excluded. In other words, RI_K is a measurement of how much the system loses in performance when the factor K is removed. The RI measures of the factors in ARN are given in Table 4.

When evaluated by RI_K, the results of the three first factors (*Number/Gender/Animacy factor*, *The reference proximity factor*, *Boost pronoun factor*), stand out so much from the rest of factors that they can be said to constitute the core of ARN. In other words, ARN could achieve reasonably good results using only these three factors.

Another measure proposed by Mitkov [10] is the *Decision Power* (DP_K). For a factor K that awards positive points only, the decision power is defined as the ratio between the number of chosen antecedents awarded points by the factor K (S_K), and the total number of candidates awarded points by factor K (A_K).

[7] Success rate for an anaphora resolution algorithm is defined as a ratio between number of successfully resolved anaphora and number of all anaphora [10].

Table 4. The relative importance (RI) of factors applied in ARN

K	RI_K
Factor 1: Number/Gender/Animacy factor	0.139
Factor 2: The sentence proximity factor	0.083
Factor 3: Boost pronoun	0.066
Factor 5: Direct object preference	0.003
Factor 6: Indirect object preference	0.003
Factor 7: Adverbial phrase penalization	0.003
Factor 9: Syntactic parallelism	0.012
Factor 10: Section heading preference	0.000
Factor 11: Indefiniteness penalization	0.018

Equation 2

$$DP_K = \frac{S_K}{A_K}$$

Decision power measures a single factor's impact without paying attention to its influence on the system. For instance, a factor that only once awards points to a candidate gets the maximal DP of 1 if that candidate gets chosen. On the other side, if a factor is based on a quality that all candidates have, such as proximity to the anaphor, it necessarily gets a high A_K value and consequently low DP.

For factors that give negative points, DP is defined in Equation 3:

Equation 3

$$DP_K = \frac{NonS_K}{A_K}$$

where $NonS_K$ is the number of cases where a candidate penalized by factor K has *not* been selected as the antecedent; A_K is again the total number of candidates penalized by the factor.

The definition Mitkov gives for DP of impeding factors does, however, present some problems, as it is biased in favor of impeding factors. In the test corpus of ARN, there are over 12000 candidates, and only 939 anaphora, which means that all candidates have a much higher chance of not being chosen as an antecedent than of being. The DP of the impeding factors is therefore high compared to DPs of promoting factors and should only be compared to DPs of the other impeding factors. The DP of the factors applied in ARN is given in Table 5.[8]

The DP measurement confirms the importance of *Number/Gender/Animacy factor* (Factor 1) and *Boost pronoun factor* (Factor 3), but also emphasizes the value of *Indirect object preference* (Factor 6) and *Section heading preference* (Factor 10). This is especially interesting in connection with Factor 10, which

[8] As Factor 1 awards both positive and negative points to the candidates, the only way to compute its DP was to split it into two factors, marked with $_+$ and $_-$ indexes.

Table 5. The factors' decision power (DP)

K	A_K	S_K	$NonS_K$	DP_K
Factor 1: Number/Gender/Animacy factor$_+$	3816	733	–	0.192
Factor 1: Number/Gender/Animacy factor$_-$	6552	–	6522	0.995
Factor 2: The reference proximity factor	8763	808	–	0.092
Factor 3: Boost pronoun	1449	580	–	0.400
Factor 5: Direct object preference	1447	82	–	0.057
Factor 6: Indirect object preference	57	7	–	0.123
Factor 7: Adverbial phrase penalization	461	–	455	0.987
Factor 9: Syntactic parallelism	4963	589	–	0.119
Factor 10: Section heading preference	94	14	–	0.149
Factor 11: Indefiniteness penalization	3398	–	3338	0.982

according to the *Relative Importance* measurement (RI_{10}= 0.000) does not contribute to the system at all. The *Decision Power* measurement confirms that it is a solid factor worth keeping.

4 Conclusion

This paper presents ARN — a rule-based anaphora resolution system for Norwegian, designed to resolve the third person pronouns with the exception of the pronoun *det* 'it(neut.)'.

The application of a set of rules based on the Centering theory's forward-looking centers (C_f) hierarchy, which is central to the RAP and MARS anaphora resolution systems, did not have a corresponding positive effect on the resolution of Norwegian pronouns. Most notably, giving preference to subject candidates brought about a clear impairment of the results. I propose that this is due to differences in information structure between Norwegian and English. The particularly strong tendency in Norwegian for not conveying new information with the subject leads to a much higher occurrence of sentences with expletive subject *det* 'it', which are quite unsuitable as antecedent candidates.

In order to be able to give a more conclusive analysis, it would be necessary to conduct experiments on a larger data set. It would also be very illuminating to include the resolution of the anaphor *det* 'it' into the system, or at least determine the ratio of expletive subjects to the anaphora *det* in the data set.

Moreover, instead of introducing a morphological filter for gender, I suggest that a more complex preference that would account for the conflict between grammatical gender of the candidates that refer to human beings and the sex of the persons they denote should be used. The Decision Power (DP) and Relative Importance (RI) measurements have also confirmed the central role of rules that promote candidates close to an anaphor as well as those that promote pronominal candidates.

Acknowledgments. The design and implementation of ARN was done as a partial fulfillment of the Candidata philologiæ thesis (≈ Master thesis) at the

Department of Linguistic and Nordic studies, University of Oslo. I wish to thank my supervisor, professor Janne Bondi Johannessen of the The Text Laboratory, ILN, UiO, for her valuable feedback. I also wish to thank three anonymous referees for their useful comments.

References

1. Baldwin, B.: CogNIAC: A Discourse Processing Engine. Ph. D. thesis, University of Pennsylvania Department of Computer and Information Sciences (1995)
2. Brennan, S., Friedman, M. V. and Pollard, C. J.: A centering approach to pronouns. In Proceedings of the 25th Annual Meeting of the Association of Computational Linguistics, Standford (1987)
3. Faarlund, J. T., Lie, S., and Vannebo, K. I.: Norsk referansegrammatikk. Oslo: Universitetsforlaget (1997)
4. Grosz, B. J., Joshi, A. K., and Weinstein, S.: Centering: a framework for modeling the local coherence of discourse. Computational Linguistics 21(2) (1995) 203-226
5. Gundel, J. K.: Information structure and the use of cleft sentences in English and Norwegian. In Hasselgård, H., Johansson, S. K. A.,Behrens, B. and Fabricius-Hansen, C. (Eds.), Information Structure in a Cross-linguistic Perspective, Volume 39 of Language and Computers, Amsterdam: Rodopi (2002) 113-129
6. Johansson, M.: Clefts in contrast. A contrastive study of English and Swedish texts and translations. Linguistics 39, (2001) 547-582.
7. Lappin, S. and Leass, H.J.: An Algorithm for Pronominal Anaphora Resolution, Computational Linguistics 20(4) (1994) 535-561
8. Kennedy, C. and Boguraev, B.: Anaphora for everyone: Pronominal anaphora resolution without a parser. In The Proceedings of the 16th Inter- national Conference on Computational Linguistics (COLING'96), Copenhagen, Denmark (1996) 113-118
9. Mitkov, R.: Robust pronoun resolution with limited knowledge. In COLING-ACL (1998) 869-875
10. Mitkov, R.: Anaphora Resolution. Pearson Education (2002)
11. The Oslo Corpus of Tagged, Norwegian Texts. The Text Laboratory, ILN, University of Oslo. http://www.tekstlab.uio.no/norsk/bokmaal/english.html

"Who Are We Talking About?" Tracking the Referent in a Question Answering Series

Matteo Negri and Milen Kouylekov

ITC-irst, Centro per la Ricerca Scientifica e Tecnologica,
Via Sommarive 18, Povo-Trento, Italy
{negri,kouylekov}@itc.it

Abstract. The capability of handling anaphora is becoming a key feature for Question Answering systems, as it can play a crucial role at different stages of the QA loop. At the question processing stage, on which this paper is focused, it enhances the treatment of follow-up questions, allowing for a more natural interaction with the user. As much as the QA task evolves towards a realistic dialogue-based scenario, one of the concrete problems raised by follow-up questions is tracking their actual referent. Each question may in fact refer to the topic of the session, to an answer given to an earlier question, or to a new entity it introduces in the dialogue. Focusing on the referent traceability problem, we present and experiment with a possible data-driven solution which exploits simple features of the input question and its surrounding context (the target of the session, and the previous questions) to inform the next phases of the QA process.

1 Introduction

Question Answering (QA) is the task of automatically returning a textual expression, extracted from a large document collection, in response to a question asked in natural language. Along the years, the QA challenge has presented systems with increasingly complex subtasks, which require knowledge-intensive NLP techniques and advanced reasoning capabilities. These include: *i)* handling a broad variety of question typologies; *ii)* extracting answers from multilingual document collections; *iii)* presenting the user with a concise and justified output; *iv)* handling spatial and temporal restrictions to address complex information needs that exist within a richer user context.

In the framework of the main evaluation campaigns (*i.e.* TREC[1], CLEF[2], and NTCIR[3]), the complexity of the input questions has followed this trend. Adhering to the roadmap for QA research [1], the test set of the TREC QA

[1] http://trec.nist.gov
[2] http://www.clef-campaign.org/
[3] http://research.nii.ac.jp/ntcir/

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 167–178, 2007.

target id="136" text="Shiite"

- q id="136.1" *Who was the first Imam of the Shiite sect of Islam?*
- q id="136.2" *Where is his tomb?*
- q id="136.3" *What was this person's relationship to the Prophet Mohammad?*
- q id="136.4" *Who was the third Imam of Shiite Muslims?*
- q id="136.5" *When did he die?*
- q id="136.6" *What portion of Muslims are Shiite?*
- q id="136.7" *What Shiite leaders were killed in Pakistan?*
- q id="136.8" *Other*

Fig. 1. A series from the TREC-2005 main QA task

task has moved from a list of simple isolated *factoid* questions (TREC 2000 and 2001), to the heterogeneous question typologies used in the last editions.

Since 2004 [19], systems participating in TREC are presented with "factoid[4]", "list[5]", and "other[6]" questions grouped into series. Each series has the target of a definition associated with it (including people, organizations, events, and other entities), which provides the initial context of the session. Each question in a series asks for additional information either about the target, or about the context originated by answers given to the preceding questions in the session. The main objective of this evaluation setting is to provide an abstraction of an information dialog in which the user is trying to define the target. In a real-world scenario, in fact, questions are not asked in isolation, but rather in a cohesive manner that often involves a sequence of related questions to meet the user's information needs.

One of the complexity factors in the proposed evaluation scenario is related to the presence of anaphoric expressions within the input questions. These may refer either to the target of the series, or to the answer given to a previous question. The example reported in Figure 1 provides a clear picture of the situation. While some questions in the series (*i.e.* questions 136.1, 136.4, 136.6, and 136.7) are non anaphoric and easier to handle, others (*i.e.* questions 136.2, 136.3, 136.5) show a shift of the referent, realized by means of anaphoric expressions (*i.e.* the pronouns "*his*" and "*he*", and the nominal construct "*this person*"), which makes them more difficult to manage. In this framework, where question interpretation has to be situated in a particular context as the QA session proceeds, anaphora resolution and referent tracking capabilities emerge as key features of a QA system. Focusing on the referent traceability problem, the remainder of this paper presents a data-driven solution which exploits simple features of the input

[4] "Factoid" questions are fact-based, short-answer (typically a noun phrase or a simple verb phrase) questions such as *"Where is Bollywood located?"*.

[5] "List" questions ask for different instances of a particular kind of information to be returned, such as *"List the names of chewing gums"*.

[6] "Other" questions require a non redundant list of facts that have not already been discovered by previous answers, about a given topic.

question and its surrounding context (*i.e.* the target of the session, and the previous questions) to inform the following phases of the QA process.

The paper is structured as follows. Section 2 briefly overviews significant related works dealing with coreference resolution, and the integration of anaphora processing components in a QA architecture. It's worth noting that, usually addressing a different problem (*i.e.* how to exploit anaphora resolution techniques to enhance answer extraction), the approaches reported in literature still fall short from providing support to tackle the referent traceability problem. Section 3 describes the overall strategy we adopt to address the task, situating it in the framework of a QA architecture. Section 4 presents the main features of the learning approach we experimented with, and reports the results of a preliminary evaluation carried out over the TREC 2005 QA test set. Section 5 concludes the paper proposing directions for future work.

2 Related Works

Relevant related works belong to two main research areas, namely coreference resolution and QA.

Coreference Resolution. From a methodological point of view, many insights come from the successful application of machine learning techniques to coreference resolution (*i.e.* the task of determining if two expressions in a document both refer to the same entity). Machine learning approaches proved to be effective, operating primarily by recasting the problem as a binary classification task. Specifically, a pair of NPs is classified as co-referring or not based on constraints that are learned from an annotated corpus. [16], for instance, describes a system which exploits 12 surface-level features to induce a decision tree for NP coreference resolution on the MUC6 [11] and MUC7 [12] data. During training, positive examples are generated in the form of feature vectors for appropriate pairs of markables. Training examples are then given to a learning algorithm to build the classifier. At test time, all markables in a new document are determined, and potential pairs of co-referring markables are presented to the classifier to determine the appropriate antecedent of each mention. [13] extends such approach by *i*) adding some precision-oriented extra-linguistic modifications to the machine learning framework, and *ii*) providing the learning algorithm with many additional linguistic knowledge sources. The classifier learned from the resulting extended feature set, which contains 53 lexical, semantic, grammatical, and knowledge-based features, is proved to achieve a performance that is comparable to hand-crafted systems. Similar approaches, with slight variations, are reported in [17] (which adapts the learning features to deal with pronouns with non-NP-antecedents in a spoken dialogue scenario), [22] (which proposes a twin-candidate learning model to present the preference relationship between the antecedent candidates more reliably), and [7] (which proposes a coreference resolution technique based on finding the best path from the root to a leaf in a Bell tree).

Question Answering. From the QA perspective, several earlier studies demonstrate the effectiveness of exploiting anaphora resolution techniques to boost the performance of a QA system ([10], [9], [18], [21]). In most of these works the positive impact of anaphora processing is demonstrated only with respect to the document retrieval and answer extraction phases, where the largest number of relevant passages (document retrieval) and answer candidates (answer extraction) have to be processed by the system. In this direction, [10] and [9] describe a mechanism for resolving pronouns based on a maximum entropy model, and evaluate the contribution of coreference resolution to different NLP applications including QA. As pointed out by the authors, there are two main advantages of applying coreference resolution in the QA framework. On the one hand, the number of relevant answer candidates retrieved by the system is slightly increased, leading to an overall performance improvement of +1.5% in terms of Mean Reciprocal Rank (MRR[7]). On the other hand, leaving aside the TREC scenario, where the required output has to be extracted from the document collection without any modification, the application of coreference resolution can play an important role in terms of answer presentation. Filling in pronouns' referents in the extracted text can in fact make it much more coherent (and responsive) to the reader. Similar conclusions are drawn in [18], which reports significant performance improvements (up to 25% in terms of returned answers) when pronominal anaphora resolution is carried out over the documents returned by the search engine.

Up to date, however, few works addressed the issues raised by dialogue-like series at a question processing stage, and the potential usefulness of anaphora processing at this level has been substantially disregarded. Most of the systems in the TREC exercise [5], [19] simply replace question's pronouns with the series target, checking at most for morpho-syntactic compatibility (*e.g.* number and gender agreement). [2] and [4] take a step further towards a more theoretically grounded approach to the problem. [2] proposes and motivates through examples a semantic-rich discourse representation that captures both discourse roles of a question and discourse relations between questions in a QA series. Three types of discourse transitions (*i.e. Topic Extension*, *Topic Exploration*, and *Topic Shift*) are analysed to determine how the context (preceding questions and answers) can be used to create a "mental map" of user information needs, and support both question interpretation and answer extraction. The proposal of a fine-grained analysis of discourse structure, however, is not corroborated by any experimental evidence. Finally, [4] presents possible extensions of a German QA system to handle several types of anaphora and ellipses occurring in a dialogue-style user interaction. The proposed solution is based on deep linguistic processing of both the questions and the original text collection. Standard coreference resolution techniques are then applied to the users' inputs and the answers that the system

[7] MRR is one of the evaluation metrics used at TREC, where a ranked list of up to 5 responses per question has to be returned by the system. The score assigned to an individual question is the reciprocal of the rank at which the first correct answer was returned, or 0 of no correct response was returned. The score for the entire run is then the mean over the set of questions in the test.

produces. Also in this case, only a general algorithm is reported, without any evaluation of system's performance.

Building on the good results provided by machine learning approaches to coreference resolution, and in light of the limited attention given to their potential usefulness at a question processing level, the following sections present a possible data-driven solution to the referent traceability problem in a dialogue-like QA scenario.

3 Dealing with QA Series

As mentioned above, in the past TREC editions the typical approach to anaphoric questions was a simple rewriting procedure, based on the substitution of question's pronouns with the target of the series. Following this approach, once the input has been transformed into a question whose interpretation does not depend on the surrounding context, the analysis is carried out in the same way as it is done with non-anaphoric questions. Unfortunately, this solution is no longer feasible under the conditions of a "natural" dialogue-like series. In this situation (as it is reflected by the TREC 2005 and 2006 test sets), we must take into account that:

1. The target may be a complex or vague concept, which is not easy to describe with a single noun or named entity (*e.g. "Russian submarine Kursk sinks"*, *"France wins World Cup in soccer"*, *"Plane clips cable wires in Italian resort"*, *"First 2000 Bush-Gore presidential debate"*);
2. Follow-up questions do not necessarily refer to the series target, but also to answers given to previous questions.

On the one side, the complexity of the target makes the question reformulation procedure a hard and error-prone task. Consider, for instance, the target/question pair:

T= *"Plane clips cable wires in Italian resort"*
Q= *"When did <u>the accident</u> occur?"*

Rewriting automatically the question Q, in order to make it independent from the surrounding context, is a complex generation exercise, which requires the substitution of the definite NP *"the accident"* with a reasonable reformulation of the target.

On the other side, a mechanism to track the actual referent of each question as the series proceeds becomes necessary to avoid errors. For instance, given the target/question pair (taken from the series reported in Figure 1):

T= *"Shiite"*
...
Q= *"Where is <u>his</u> tomb?"*

the substitution of the pronoun *"his"* with the target would produce the question *"Where is Shiite tomb?"*, which is inevitably doomed to failure.

In light of these considerations, our solution adopts a bag-of-words approach which gets round the question reformulation problem, still remaining capable to track the appropriate referents as a series proceeds. The underlying assumption is that the correct analysis of anaphoric input questions does not necessarily depend on their reformulation. Nevertheless, since the possibility of successfully answering a question depends on retrieving at least one relevant document containing the right answer, we should pay the utmost attention to the query formulation phase. Only at this stage, in fact, the correct resolution of anaphoric expressions becomes necessary in order to: *i*) *maximize* the possibility of retrieving relevant passages (using the appropriate search terms), and *ii*) *minimize* the possibility of introducing noise (*i.e.* wrong terms) in the search query.

According to the bag-of-words approach we adopt, once the referent of an anaphoric question has been selected (between the target and the answer to an earlier question), its terms are combined with the question keywords into a boolean query to the document collection. In case of reference shifts, a new referent remains valid until a non-anaphoric question referring to a new entity is found. Our hypothesis is that combining the right keywords to search the document collection is safer than rewriting the input question in a "complete" (self-contained) form, and easier than producing the semantic-rich discourse representation proposed by [2].

An example of QA architecture extended with a module for question referent selection is reported in Figure 2. The example describes a system based on a rather standard three-components backbone (similar to the one thoroughly described in [8], and [6]). Such backbone includes: a *question processing component*, which is in charge of the linguistic analysis of the input questions, a *search*

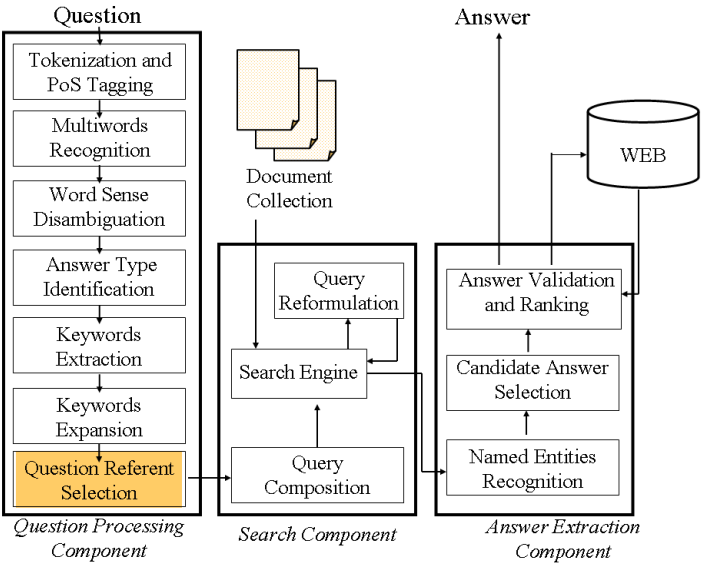


Fig. 2. Architecture of a QA system

component, which performs query composition and document retrieval, and an *answer extraction component*, which extracts the final answer from the retrieved text passages.

Following the proposed approach, the **question processing component** carries out the analysis of each question as it is: no question reformulation is performed, and anaphoric expressions are left unresolved. Specifically, the linguistic analysis of the input question is performed sequentially by a cascade of modules which finally return:

- the "Answer Type" of the question (the semantic category of the expected answer, *e.g.* "PERSON", "LOCATION", "MOVIE_TITLE");
- a list of relevant question keywords expanded with morphological derivations and synonyms;
- the referent of the question.

Then, in the document retrieval phase, the **search component** is in charge of managing the query keywords, combining them into a boolean query to extract the best matching text portions from the target document collection. In case of anaphoric questions, the search query will contain both the terms coming from the question, and those coming from the selected referent. In case none, or too few documents are retrieved, query relaxation loops are performed by discarding the keywords with lowest priority. Priorities are assigned considering different types of features of the query terms, including capitalization, part of speech, and WordNet sense. The source of a term (*i.e.* the question, the series target, or the answer given to a previous question) is also considered: keywords extracted from the input question have a lower priority than those extracted from the selected referent.

At the end of the process, the **answer extraction component** analyses the retrieved text portions to return a final answer. The process is carried out in three steps. Firstly, a named entity recognition module identifies all the entities that match the answer type category. Then, a filtering process is carried out to reduce the list of possible answer candidates (for instance on the basis of their frequency and their distance from the question keywords within the text portions retrieved by the search engine). Finally, an answer validation module is in charge of ranking the selected candidates (a Web-based statistical approach assigns a score to each candidate), and returning the final answer.

Adopting the proposed approach the problem of dealing with anaphoric questions in a TREC-like QA series is "reduced" to the problem of finding the appropriate referent of each question. The next section describes how this task is accomplished by means of a classifier opportunely trained.

4 A Data-Driven Approach to Question Referent Selection

4.1 Training Corpus

To develop our referent tracking module, a training corpus has been manually annotated starting from the test set of the TREC 2005 main QA task (75 series, for a total of 530 questions).

Each question q_i, in a series S having T as a target, has been classified with respect to the three typologies we want to deal with, namely:

- *Type0*: non anaphoric questions;
- *Type1*: anaphoric questions referring to T;
- *Type2*: anaphoric questions referring to the answer given to the preceding question in S (referred to as q_{i-1}).

Such classification is still rough, as it ignores other more complex phenomena involved in the evolution of a dialogue-like series. For instance, it does not take into account those cases in which, instead of a single concept, the referent of a question is a set of entities represented by the answers given to two or more preceding questions, or by a combination of the target of the session with one or more answers previously returned (see Section 5 for some examples). However, our coarse-grained classification is motivated by the fact that the variability of referent shift phenomena was still poorly represented in the TREC 2005 QA test set. In our training corpus, for instance, only two questions out of 530 refer to a set of entities derived from previous turns in the session. In addition, even with respect to this coarse-grained distinction in three typologies, the distribution of the examples in our training corpus is rather unbalanced, with 129 questions assigned to *Type0*, 385 assigned to *Type1*, and only 16 assigned to *Type2*. On the one side, this unbalanced distribution is likely to reflect a real-world situation; on the other side, the possible impact on the learning process should not be neglected. In fact, unless very discriminative features will be selected, the large amount of *Type1* questions will probably influence the classifier's performance, determining a bias towards this class.

4.2 Learning Features

The classifier has been trained considering the following features of the target T, the current question q_i, and the previous question q_{i-1} in a series S:

1. T features:
   - **T-has-verbs** (0, 1): set to 1 if T contains at least one verb, 0 otherwise.
   - **T-capitalized** (0, 1): set to 1 if all the words in T are capitalized.
   - **T-contains-LifeForms** (0, 1): set to 1 if T contains at least one hyponym of "life_form" in WordNet [3].
   - **T-first-LifeForm-position** (integer): position of the first hyponym of "life_form" in T.

2. q_i Features:
   - q_i**-number-in-Session** (integer): q_i position in S.
   - q_i**-Type** (1,..., 3): factoid=1, list=2, other=3.
   - q_i**-StartsWith** (0,..., 7): "who"=0, "where"=1, "when"=2, "which"=3, "what"=4 "how+adj"=5, "how+verb"=6 other=7.
   - q_i**-contains-Target-tokens** (0, 1): set to 1 if q_i contains at least one token present in T.

- q_i**-contains-Target-lemmas** (0, 1): set to 1 if q_i contains at least one lemma of a term in T.
- q_i**-contains-prons** (0, 1): set to 1 if q_i contains at least one pronoun.
- q_i**-contains-pers-prons** (0, 1): set to 1 if q_i contains at least one personal pronoun.
- q_i**-contains-ThisPlusTargetWord** (0, 1): set to 1 if q_i contains the word "this" followed by a term present in T.
- q_i**-contains-ThisPlusWord** (0, 1): set to 1 if q_i contains the word "this" followed by any term non present in T.
- q_i**-contains-ThisPlusQ1Noun** (0, 1): set to 1 if q_i contains at least one noun present in q_{i-1}.
- q_i**-contains-LifeForms** (0, 1): set to 1 if q_i contains at least one hyponym of "life_form" in WordNet.
- q_i**-first-LifeForm-position** (integer): position of the first hyponym of "life_form" in q_i.

3. q_{i-1} Features:
   - q_{i-1}**-Type** (1,..., 3): factoid=1, list=2, other=3.
   - q_{i-1}**-StartsWith** (0,..., 7): "who"=0, "where"=1, "when"=2, "which"=3, "what"=4 "how+adj"=5, "how+verb"=6 other=7.
   - q_{i-1}**-contains-pers-pron** (0, 1): set to 1 if q_{i-1} contains at least one personal pronoun.
   - q_{i-1}**-contains-LifeForms** (0, 1): set to 1 if q_{i-1} contains at least one WordNet hyponym of "life_form".
   - q_{i-1}**-first-LifeForm-position** (integer): position of the first hyponym of "life_form" in q_{i-1}.

4.3 Evaluation

A preliminary experiment has been carried out training a classifier using a decision tree algorithm (namely the Weka [20] J48 algorithm[15]). At first glance, the results of this evaluation are very encouraging. The 10-fold cross-validation over the training set resulted in fact in 505 out of 530 instances correctly classified (corresponding to an accuracy of 95.28%). However, as can be observed from the confusion matrix reported in Table 1, the performance over the three classes is not uniform, reflecting the unbalanced distribution of the examples in the training. In particular, unsurprisingly, only 4 instances out of 16 (25%) of the class less represented in the training set (*Type2*) are correctly recognized.

We believe that further experiments over a larger training corpus will allow us to improve these results.

5 Conclusions and Future Work

One of the specific issues raised by follow-up questions in a dialogue-like QA series is tracking the referent shifts that often occur as the series proceeds. Focusing

Table 1. Confusion matrix

	Type0	Type1	Type2
Type0	82	12	0
Type1	0	344	1
Type2	0	12	4

target id="120" text="Rose Crumb"

- q id="120.1" *What was her occupation?*
- q id="120.2" *Where was she from?*
- q id="120.3" *What organization did she found?*
- q id="120.4" *When did she found it?*
- q id="120.5" *What awards has she received?*
- q id="120.6" *How old was she when she won the awards?*
- q id="120.7" *Other*

target id="137" text="Kinmen Island"

- q id="137.1" *What is the former name of Kinmen?*
- q id="137.2" *What country governs Kinmen?*
- q id="137.3" *What other island groups are controlled by this government?*
- q id="137.4" *In the 1950's, who regularly bombarded Kinmen*
- q id="137.5" *How far is Kinmen from this country?*
- q id="137.6" *Of the two governments involved over Kinmen, which has air superiority?*
- q id="137.7" *Other*

Fig. 3. Other examples from the TREC-2005 main QA task

on this problem, the objective of our work was to define a working methodology to deal with the most frequent (and more tractable) type of referent shift, which occurs when the target of a question is represented by the answer given to a preceding question in the series. In this direction, we presented a preliminary experiment with a classifier trained over the TREC 2005 question set. Results confirm the viability of the proposed approach, even though they reflect the fact that the training corpus adopted is unbalanced with respect to the phenomena we want to model. Building on these results, future research will concentrate both on improving the simple learning approach here proposed, and on a deeper theoretical analysis of the more complex linguistic phenomena occurring in a typical dialogue-like QA series.

From the learning perspective, since there is still room for improvement, future developments will follow two complementary directions, namely: *i)* a more accurate feature selection, trying to find a set of more discriminative features; and *ii)* learning the model from a larger set of training examples (*e.g.* including the TREC 2006 series), trying to balance the number of examples of the three typologies we are dealing with.

From a theoretical perspective, further improvements may be achieved by a more comprehensive analysis of the evolution of a series, and the phenomena involved in the process. At a first stage, such analysis will consider a more complex type of referent shift, which occurs when the target of a question is a set of entities instead of a single concept. Such complex referents can be represented either by the answers given to two or more preceding questions in the series, or by a combination of the series target with one or more previous answers. Two examples of this situation are represented by the series reported in Figure 3. In the first one, the correct referent of question 120.4 is obtained by combining the series target (*i.e. "Rose Crumb"*) with the answer given to question 120.3 (*i.e. "Hospice of Clallam County"*). In the second one, the correct referent of question 137.6 is obtained by combining the answers given to questions 137.2 (*i.e. "Taiwan"*) and 137.4 (*i.e. "China"*).

References

1. Burger, J., Cardie, C., Chaudhri, V., Gaizauskas, R., Harabagiu, S., Israel, D., Jacquemin, C., Lin, C-Y., Maiorano, S., Miller, G., Moldovan, D., Ogden, B., Prager, J., Riloff, E., Singhal, A., Shrihari, R., Strzalkowski, T., Voorhees, E., Weishedel, R.: Issues, Tasks and Program Structures to Roadmap Research in Question & Answering (Q&A). Document Understanding Conferences Roadmapping Documents (2001)
2. Chai, J.Y., Jin, R.: Discourse Structure for Context Question Answering. Proceedings of the HLT-NAACL 2004 Workshop on Pragmatics of Question Answering. Boston, MA, May 6-7 (2004)
3. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
4. Fliedner, G.: Towards Natural Interactive Question Answering. Proceedings of the 5th International Conference on Language Resources and Evaluation - LREC-2006. Genoa, Italy, May 22-28 (2006)
5. Harabagiu, S., Hickl, A., Lehmann, J., Moldovan, D.: Experiments with Interactive Question-Answering. Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics. University of Michigan, Ann Arbor, USA, June 25-30 (2005).
6. Kouylekov, M., Magnini, B., Negri, M., Tanev, H.: ITC-irst at TREC-2003: the DIOGENE QA system
Proceedings of TREC 2003 (2004)
7. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A Mention-Synchronous Coreference Resolution Algorithm Based on the Bell Tree. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. Barcelona, Spain, July 21 - 26 (2004)
8. Magnini, B., Negri, M., Prevete, R., Tanev, H.: Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC-2002. Proceedings of TREC 2002 (2003)
9. Morton, T.: Coreference for NLP Applications. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong-Kong, October 3-6 (2000)
10. Morton, T.: Using coreference for question answering Proceedings of the ACL 1999 Workshop on Coreference and Its Applications. University of Maryland, College Park, MD. USA, June 22 (1999)

11. MUC-6: Proceedings of the Sixth Message Understanding Conference (MUC-6), San Francisco, CA. Morgan Kaufmann (1995).
12. MUC-7: Proceedings of the Seventh Message Understanding Conference (MUC-7), San Francisco, CA. Morgan Kaufmann (1998).
13. Ng, V., Cardie, C.: Improving Machine Learning Approaches to Coreference Resolution. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, July 6-12 (2002)
14. Novischi, A., Moldovan, D.: Question Answering with Lexical Chains Propagating Verb Arguments. Proceedings of COLING-ACL 2006. Sydney, Australia, July 17-21 (2006)
15. Quinlan, R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
16. Soon, W.M., Ng, H.T., Lim, D.C.Y: A Machine Learning Approach to Coreference Resolution of Noun Phrases. Computational Linguistics, 27(4):521-544. (2001)
17. Strube, M., Muller, C.: A Machine Learning Approach to Pronoun Resolution in Spoken Dialogue. Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan, July 7-12 (2003)
18. Vicedo, J.L., Ferrández, A.: Importance of Pronominal Anaphora Resolution in Question Answering Systems. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Hong-Kong, October 3-6, (2000)
19. Voorhees, E.: Overview of the TREC 2004 Question Answering Track. TREC 2004 Proceedings (2005)
20. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann (2005)
21. Watson, R., Preiss, J., Briscoe, T.: The Contribution of Domain-Independent Robust Pronominal Anaphora Resolution to Open-Domain Question Answering. Proceedings of the Symposium on Reference Resolution and its Applications to Question Answering and Summarization. Venice, Italy June 23-25 (2003)
22. Yang, X., Zhou, G.D., Su, J., Tan, C. L.: Coreference Resolution Using Competition Learning Approach Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Sapporo, Japan, July 7-12 (2003)

Anaphora Resolution: To What Extent Does It Help NLP Applications?

Ruslan Mitkov, Richard Evans, Constantin Orăsan, Le An Ha, and Viktor Pekar

University of Wolverhampton
Research Group in Computational Linguistics
Research Institute in Information and Language Processing
Stafford Street, Wolverhampton, WV1 1SB, United Kingdom
r.mitkov@wlv.ac.uk

Abstract. Papers discussing anaphora resolution algorithms or systems usually focus on the intrinsic evaluation of the algorithm/system and not on the issue of *extrinsic evaluation*. In the context of anaphora resolution, extrinsic evaluation concerns the impact of an anaphora resolution module on a larger NLP system of which it is part. In this paper we explore the extent to which the well-known anaphora resolution system MARS [1] can improve the performance of three NLP applications: text summarisation, term extraction and text categorisation. On the basis of the results so far we conclude that the deployment of anaphora resolution has a positive albeit limited impact.

1 Introduction

Papers discussing anaphora resolution algorithms or systems[1] usually describe the work of the algorithm or the system. In the majority of cases, they also report evaluation results related to its performance. This type of evaluation is known as *intrinsic* evaluation and accounts for the performance of the algorithm/system which is measured in terms of metrics such as recall, precision or success rate [3,1]. On the other hand, *extrinsic evaluation* in the context of anaphora resolution concerns the impact of an anaphora resolution module on a larger NLP system of which it is part. In this paper we address the issue of extrinsic evaluation in anaphora resolution and explore for the first time the extent to which our anaphora resolution system MARS [1] can improve the performance of three NLP applications: text summarisation, term extraction and text categorisation.

Section 2 of this paper will introduce Mitkov's original knowledge-poor algorithm, whereas section 3 will discuss its fully automatic implementations: the early version (hereafter referred to as MARS02) and the recent version (MARS06). Section 4 will outline the evaluation data used in our experiments and Section 5 will report the evaluation results when deploying MARS in three

[1] For definition of the distinction between anaphora resolution algorithms and anaphora resolution systems, see [2].

A. Branco (Ed.): DAARC 2007, LNAI 4410, pp. 179–190, 2007.

different NLP applications: text summarisation, term extraction and text categorisation. Section 6 will provide a discussion of the evaluation results and finally section 7 will offer concluding remarks.

2 Brief Overview of Mitkov's Original Knowledge-Poor Pronoun Resolution Algorithm

MARS is based on Mitkov's [4,5] robust, knowledge-poor approach to pronoun resolution which was motivated by the pressing need in the 1990s for anaphora resolution algorithms operating robustly in real-world, knowledge-poorer environments in order to meet the demands of practical NLP systems. The first version of the algorithm was reported in [4] as an inexpensive, fast and yet reliable alternative to the labour-intensive and time-consuming construction of a knowledge-based system.[2] This project was also an example of how anaphors can be resolved quite successfully (at least in a specific genre, namely computer/technical manuals) without any sophisticated linguistic knowledge, even without parsing. In addition, the evaluation showed that the basic set of factors employed (referred to as 'indicators', see below) can work well not only for English, but also for other languages.

Mitkov's approach relies on a list of preferences known as *antecedent indicators*. The approach operates as follows: it works from the output of a text processed by a part-of-speech tagger and an NP extractor, identifies noun phrases which precede the anaphor within a distance of 2 sentences[3], checks them for gender and number agreement with the anaphor and then applies the indicators to the remaining candidates by assigning them a positive or negative score (2, 1, 0 or -1). The noun phrase[4] with the highest composite score is proposed as antecedent.

The antecedent indicators are applied to all NPs which have passed the gender and number filter.[5] These indicators can act in either a *boosting* or an *impeding* capacity. The boosting indicators apply a positive score to an NP, reflecting a positive likelihood that it is the antecedent of the current pronoun. In contrast, the impeding ones apply a negative score to an NP, reflecting a lack of confidence that it is the antecedent of the current pronoun. Most of the indicators are genre-independent and related to coherence phenomena (such as salience and distance)

[2] The approach has become better known through a later updated publication [5].

[3] Subsequent versions of the approach have used search scopes of different lengths (e.g. 2, 3 or 4 sentences), but the original algorithm only considered two sentences prior to the sentence containing the anaphor. The NP patterns are limited to the identification of base NPs and do not include complex or embedded phrases.

[4] The approach only handles pronominal anaphors whose antecedents are noun phrases.

[5] The approach takes into consideration the fact that in English there are certain collective nouns which do not agree in number with their antecedents (e.g. *government, team, parliament* etc. These entities and can be referred to by plural pronouns; equally some plural nouns such as *data* can be referred to by *it*) and are thus exempted from the agreement test.

or to structural matches, whereas others are genre-specific.[6] The boosting and impeding indicators are described in detail in [5]. The work presented in [1] provides some additional detail on the indicators used by the algorithm.

The aforementioned antecedent indicators are preferences and not absolute factors. There might be cases where one or more of the antecedent indicators do not 'point' to the correct antecedent. For instance, in the sentence 'Insert the cassette into the VCR making sure it is turned on', the indicator *prepositional noun phrases* would penalise the correct antecedent. When all preferences (antecedent indicators) are taken into account, however, the right antecedent is still likely to be tracked down - in the above example, the *prepositional noun phrases* heuristic stands a good chance of being overturned by the *collocation match* heuristics since the collocation *The VCR is turned on* is likely to appear previously in the text, as it is a typical construction in video technical manuals.

The antecedent indicators have proved to be reasonably efficient in identifying the correct antecedent and the results show that for the genre of technical manuals they may be no less accurate than syntax- and centering-based methods (see [5]). The approach described is not dependent on any theories or assumptions; in particular, it does not operate on the assumption that the subject of the previous utterance is the highest-ranking candidate for the backward-looking center - an approach which can sometimes lead to incorrect results.[7]

3 Outline of MARS

Mitkov's algorithm was enhanced and developed into the fully-automatic pronoun resolution system referred to as MARS.[8] The initial, as well as the current implementations of MARS, which both employ the FDG shallow parser [6] as their main pre-processing tool, are based on a revised version of the original algorithm.

3.1 Early Version of MARS

The initial implementation of MARS [1] followed Mitkov's original approach closely, the main differences being (i) the addition of three new indicators and (ii) a change in the way some of the indicators were implemented or computed due to the available pre-processing tools. Later, MARS also incorporated a program for automatically recognising instances of anaphoric or pleonastic pronouns [7] and intra-sentential syntax filters. This early version of MARS is referred to as MARS02 in the evaluation below.

[6] Typical of the genre of user guides.

[7] For instance, subject-favouring methods or methods relying heavily on syntactic parallelism would incorrectly propose *the utility* as the antecedent of *it* in the sentence 'The utility shows you *the LIST file* on your terminal for a format similar to that in which it will be printed' as it would prefer the subject as the most salient candidate. The *indicating verbs* preference of Mitkov's approach, however, would prefer the correct antecedent *the LIST file*.

[8] MARS stands for Mitkov's Anaphora Resolution System.

The system operates in five phases. In *phase 1*, the text to be processed is parsed syntactically, using Conexor's FDG Parser [6] which returns the parts of speech, morphological lemmas, syntactic functions, grammatical number, and dependency relations between tokens in the text, facilitating complex noun phrase (NP) extraction.

In *phase 2*, anaphoric pronouns are identified and non-anaphoric and non-nominal instances of it are filtered using the machine learning method described in [7].

In *phase 3*, for each pronoun identified as anaphoric, candidate NPs are extracted from the heading of the section in which the pronoun appears, and from the current and preceding two sentences within the paragraph under consideration. Once identified, these candidates are subjected to further morphological and syntactic tests. Extracted candidates are expected to obey a number of constraints if they are to enter the *set of competing candidates*, i.e. the candidates that are to be considered further. Competing candidates are required to agree with the pronoun in terms of number and gender, as was the case in the original algorithm. They must also obey syntactic constraints [1].

In *phase 4*, 14 preferential and impeding factors are applied to the sets of competing candidates. On application, each factor applies a numerical score to each candidate, reflecting the extent of the system's confidence about whether the candidate is the antecedent of the current pronoun. In the implemented system, certain practical issues led to the weights assigned by indicators being computed in a different way from that described in the original algorithm. The full details of these differences are beyond the scope of the current paper, but they are described in detail in [1]. In addition, three new indicators were added, one of which (*syntactic parallelism*) exploits new, previously unavailable features of the pre-processing software.

Finally, in *phase 5*, the candidate with the highest composite score is selected as the antecedent of the pronoun. Ties are resolved by selecting the most recent highest scoring candidate.

3.2 Recent Version of MARS

The more recent version of MARS incorporates several advancements over the system described in [1]. These improvements introduce the inclusion of more precise and strict number and gender agreement, and the addition of one indicator which employs the modelling of selectional restrictions. This recent version is referred to as MARS06 in the evaluation below.

MARS was improved to cater for several frequent causes of apparent number disagreement. These consist of (i) collective nouns, (ii) NPs whose gender is under-specified, (iii) quantified nouns/indefinite pronouns, and (iv) organisation names. These cases were handled by a combination of gazetteers, the integration of an animacy recognition module [8], and named entity recognition [9]. Patterns were used to identify the occurrence of quantified NPs in the parsed text. MARS's recognition of the gender of NP candidates has also been improved. In

addition to gazetteers, a NER system is used to recognise person names and the system for NP animacy recognition is deployed.

Following work such as that described in [10], a new salience indicator was implemented. The selectional preference indicator processes a pronoun, exploits functional dependency information provided by the FDG Parser, and extracts a pair consisting of the verb on which the pronoun depends and the functional role of the pronoun. The selectional preference for the verb is then modeled by means of a distributional approach and used to obtain the likelihood for each candidate NP that it is a potential argument of the verb given that it has the same role as the pronoun. The selectional preference model is used to determine the most likely candidate in the set, and this candidate is awarded a boosting score of +1.

4 Evaluation Data

In this paper, a corpus of newspaper articles published in *New Scientist* was used. There were several reasons for selecting texts from this magazine. First, texts were required which contain a relatively high number of pronouns and are ideally not too different from texts from the technical domain, for which MARS was initially designed. We decided against using technical documents for two reasons. Firstly, we wanted to see how well MARS performs on texts from a different domain and secondly, technical documents are rather long and unsuitable for some types of manual annotation (e.g. coreferential links and automatic summarisation annotation, as explained later in this section). By contrast, the texts from New Scientist were preferred because they were short enough to be manually annotated and were suitable for all the extrinsic evaluation tasks performed.

Fifty-five texts from New Scientist distributed in the BNC were included in our corpus. These texts contained almost 1,200 third person pronouns and over 48,000 words. Before selecting the texts, a filter was applied to ensure that very short (under 2 kilobytes including the SGML annotation) and very long (over 15 kilobytes including the SGML annotation) texts were not included in the corpus. The reason for filtering out texts that are too short is that such texts could not be used in automatic summarisation (see extrinsic evaluation, section 5.1), whereas texts which were too long cannot be reliably annotated. The texts from New Scientist also proved appropriate for the other two extrinsic evaluation tasks investigated in this paper. They are scientific texts that contain a relatively large number of terms and are therefore appropriate for the application of automatic term extraction methods. Further, as they address different clearly identifiable topics, it is feasible for them to be categorised by automatic means.

All the texts in the corpus were annotated with several layers of annotation using PALinkA, a multipurpose annotation tool [11]. First, the texts were annotated for coreferential links using the methodology for NP coreference described in [12]. Once all the markables were identified by the annotators, the head of each one was manually marked in order to facilitate evaluation. Six files, accounting for about 10% of the corpus, were annotated by two annotators in order to assess

the inter-annotator agreement. Using the method described in [13], sets of coreference chains were derived from each annotated file and each pair of elements in a chain was used to produce an exhaustive set of the coreferential links annotated in a document. The set of coreference links derived from the annotations of one annotator is then considered to be the key while the set derived from the annotations made by the other annotator is considered to be the response. On the annotated data used in the current study, evaluation revealed an average F-score of 0.6601 between the two annotators.

In order to be able to run the extrinsic evaluation with regard to text summarisation, the corpus was also annotated with information about important sentences. The methodology applied to annotate the important sentences in a text is the one described in [14] and entailed identification of 15% of the text as essential and of a further 15% of the text as important. In this way, it was possible to evaluate automatic summaries at two different compression rates: 15% and 30%.

In order to evaluate the effect of MARS on automatic term extraction, a reader with good general knowledge was asked to read the same texts annotated with coreference information, and identify terms appearing in each text. These sets of terms served as the gold standard in the evaluation experiment on the basis of which precision, recall and F-measure are computed.

To evaluate automatic classification, each text was annotated with a relevant label derived from the New Scientist web site: "health", "earth", "fundamentals", "being human", "living world", "opinion" and "sex and cloning". The texts from the BNC were not labeled for their original category and so they had to be assigned manually by our annotators. Texts for which none of these labels seemed wholly suitable were assigned the category "other". The 55 texts annotated with coreference information and important sentences were not sufficient to train and test a classifier, and so a further 120 texts from New Scientist were selected and annotated with information about their class. Given that annotation for coreference and summarisation is difficult and time consuming, these 120 were not also annotated with this information.

5 Evaluation

Papers discussing anaphora resolution usually describe the work of the algorithm or the system. In the majority of cases, they also report evaluation results related to the performance of the algorithm/system which is known as intrinsic evaluation and which accounts for its performance. In this paper we shall not discuss the performance of MARS in terms of intrinsic evaluation and for the first time in the literature, we shall seek to focus on the extrinsic evaluation with a view to establishing the extent to which the deployment of our anaphora resolution system, MARS [1], can improve the performance of various NLP applications. For details on the intrinsic evaluation of MARS, the reader is referred to [1] and [15] where the success rate of MARS is reported to range from 45% to 65% depending on the evaluation data. As pointed out in both papers, the

performance of fully automatic anaphora resolution systems is markedly inferior to the performance of algorithms which benefit from post-edited input, i.e., from "perfect" pre-processing. Despite the comparatively low figures reported, MARS still fares as one of the best performing systems operating from the input of a shallow parser. Over the test data described in the present study, MARS02 performs with an average success rate of 0.4663, whereas MARS06 has an average success rate of 0.4947.

5.1 Summarisation

We evaluate the potential usefulness of anaphora resolution (as performed by MARS) in term-based summarisation which operates on the premise that it is possible to determine the importance of a sentence on the basis of the words it contains. The most common way of achieving this is to weight all the words in a text and calculate the score of a sentence by adding the weights of the words from it. In this way, a summary can be produced by extracting the sentences

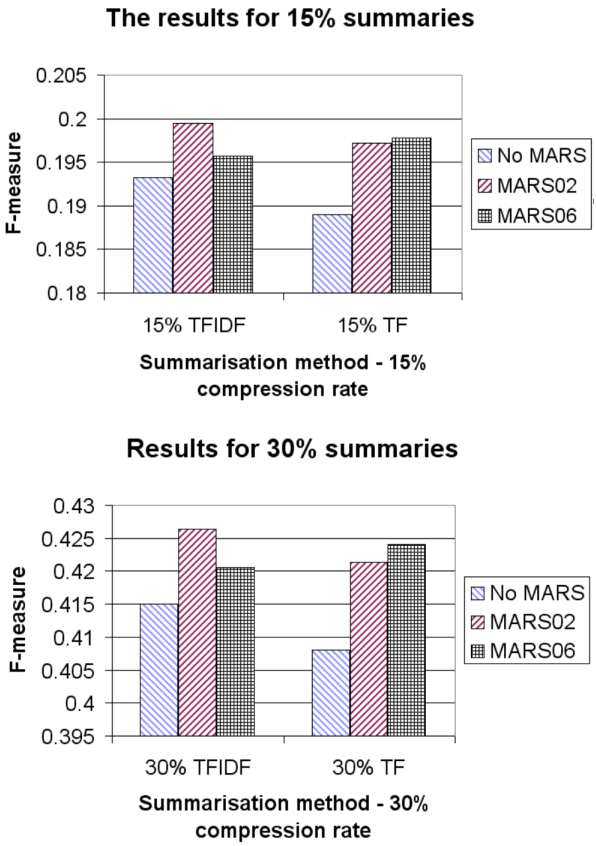


Fig. 1. The results of the automatic summarisation evaluation

with the highest scores until the desired length is reached. In order to calculate the importance of sentences several statistical measures can be applied [16], but the majority of them require that at least the frequency of the word in the document is known. For this reason, words which are referred to by pronouns do not have their weight correctly calculated. For the purpose of this extrinsic evaluation, we integrated MARS into a term-based summariser in an attempt to produce more accurate word weightings and as a result to improve the quality of the summaries produced.

In this paper two term weighting methods are investigated: term frequency and TF*IDF. The corpus used in this evaluation is the one described in Section 4 and the evaluation measures are precision, recall and f-measure. As explained earlier the corpus is annotated for 15% and 30% extracts, so the evaluation was performed for both compression rates. Figure 1 presents the results of the evaluation.

For both compression rates the value of F-measure increases when an anaphora resolution method is used by the summarisation method, but this increase is not statistically significant according to the paired t-test at the .05 level. For term frequency the results are better when the recent version of MARS is used, whereas for TF*IDF the best results are obtained by the early version of MARS.

5.2 Term Extraction

To examine the extent to which the employment of MARS could improve the performance of an automatic term extraction method, we compare the performances of a term extraction engine when run on various versions of a text: the original one and the ones in which pronouns are replaced by the antecedents proposed by MARS02 and MARS06. The term extraction method used is based on a hybrid approach which combines statistical and lexical-syntactic filters similar to [17] and [18]. First n-grams satisfying the POS pattern [AN]*NP?[AN]*N are collected, and then their TF*IDF scores are calculated. Candidates with a frequency count greater than one and TF*IDF score greater than 0.4[9] are selected. The set of New Scientist texts from the BNC is used as the document collection in the calculation of TF*IDF.

In our experiment, the term extraction engine extracts terms from three versions of a text: the original text, the text processed by MARS02, and the one processed by MARS06.[10] For each version of a text, precision, recall, and F-measure are calculated using the gold standard described in Section 4. The average F-measures are shown in Figure 2.

For both versions of MARS, there are improvements in the performance of the term extraction engine (measured using F-measure) although the improvements are not statistically significant, according to the paired t-test. MARS06 does not seem to boost the performance of term extraction over MARS02. MARS02 improves both precision and recall, whereas the main improvement engendered by MARS06 is in terms of recall. In 41% of the texts processed by MARS02, there

[9] These thresholds were determined empirically.

[10] By "processed" we mean that pronouns in the text have been replaced by the antecedents proposed by MARS.

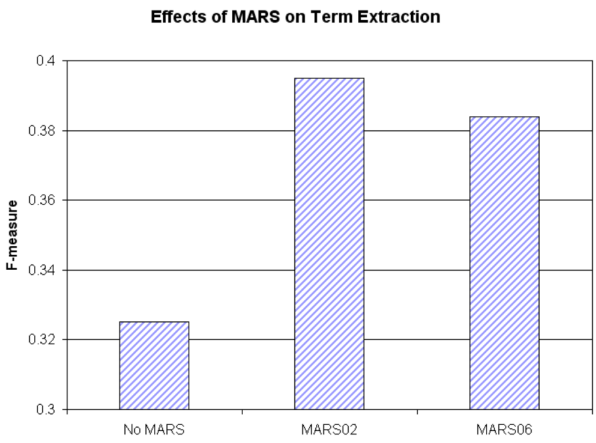


Fig. 2. The effects of two versions of MARS on automatic term extraction

is improvement in F-measures. Declining F-measures are observed in 33% of the texts and there is no change of the F-measure in the rest of the texts (26%).

5.3 Text Classification

In this experiment we examined the influence of anaphora resolution on the quality of automatic text classification. We experimented with four different text classification methods: k nearest neighbours (kNN), Naïve Bayes (NB), Rocchio, Maximum Entropy (MaxEnt), and Support Vector Machines(SVM).[11] We assess the quality of classification models that are learned from documents, in which pronouns are substituted for the noun phrases recognised as their antecedents by the MARS system. The model is then tested on documents where pronouns have been similarly replaced for antecedent noun phrases. Three different models are included in the experiment. The first two are learned from a document collection on which prior anaphora resolution has been applied: one using the early version of MARS (MARS02) and another that uses its new version incorporating proposed improvements (MARS06). The third model is induced from the same document collection without first performing anaphora resolution on it.

During the evaluation, the document collection described in Section 4 was randomly split into ten parts, one part to be used for testing and the rest for training of the model. Each model was evaluated in terms of F-measure, averaged over ten such runs. Figure 3 describes the results of the experiments.

The results show that the use of either version of MARS consistently yields improved classification effectiveness in comparison with the baseline model. MARS02 achieved the best results on two classification methods (kNN and Rocchio), while MARS06 was the best on the other three (NB, MaxEnt, and SVM).

[11] We used the implementations of these methods distributed with the Rainbow text classification toolkit [19].

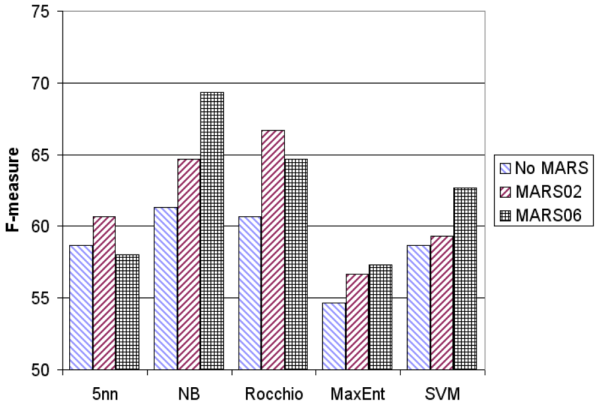


Fig. 3. The effect of anaphora resolution on the accuracy of text categorisation

Although consistent, the improvement on the baseline is not considerable (max. 8% with MARS06, using NB); in none of the compared pairs could statistical significance according to an independent sample t-test be established to the .05 level.

The experiment did not show any considerable differences in the performance of MARS02 and MARS06. MARS06 showed the greatest improvement of 4.66% (on NB), but was worse than MARS02 by 2.66% (on kNN). None of these differences is statistically significant according to the independent sample t-test.

6 Discussion

We aimed to establish the extent to which a task such as anaphora resolution could be useful in other NLP applications. From the results reported above, it is obvious that deployment of MARS has different impacts on the performance of each of the applications experimented with: text summarisation, term extraction and text classification. By and large the deployment of MARS has a positive but at the same time limited impact. In summarisation the deployment of MARS on the evaluation data results in improvement of the F-measure, although this is not significant. Term extraction also benefits from incorporation of MARS as a preprocessing module, but again, the improvement is not statistically significant. In both applications there are different cases where each version of MARS has a more favourable impact. In text categorisation, MARS provides a statistically significant improvement to one of the classification methods (kNN), and statistically insignificant improvement to two other classification methods. However, with respect to the MaxEnt method, performance deteriorates when MARS is employed.

One observation worth making on the basis of the experiments conducted is that the slight improvement in performance of MARS06 as opposed to MARS02 does not necessarily result in improvement of the performance of the application

in which it is employed. It would be interesting to see whether a dramatic improvement in performance of the resolution of anaphors would lead to a marked improvement of the NLP application that exploits it.

To this end, our next step is to establish whether there is any threshold to be achieved in order for anaphora resolution to be considered beneficial in that it almost always enhances the performance of the above applications, possibly bringing a statistically significant improvement. The results of preliminary studies in the area of text summarisation applied to a collection of scientific texts have already been established [20].

7 Conclusion

This paper covers for the first time, to the best of our knowledge, the issue of extrinsic evaluation in the context of anaphora resolution for more than one NLP application. In particular, we explore the extent to which our well-known anaphora resolution system, MARS, can improve the performance of three NLP applications (text summarisation, term extraction and text categorisation). On the basis of the results so far we conclude that the deployment of anaphora resolution has a positive albeit limited impact.

References

1. Mitkov, R., Evans, R., Orasan, C.: A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In: Proceedings of CICLing-2002, Mexico City, Mexico (February 2002) 168 – 186
2. Mitkov, R.: Anaphora resolution. Longman (2002)
3. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics **20**(4) (1994) 535 – 562
4. Mitkov, R.: Pronoun resolution: the practical alternative. In: Proceedings of the Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Lancaster, UK (1996)
5. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of the 18th International Conference on Computational Linguistics (COLING'98/ACL'98), Montreal, Quebec, Canada (August 10 - 14 1998) 867 – 875
6. Tapanainen, P., Järvinen, T.: A non-projective dependency parser. In: Proceedings of the 5th Conference of Applied Natural Language Processing, Washington D.C., USA (March 31 - April 3 1997) 64 – 71
7. Evans, R.: Applying machine learning toward an automatic classification of *It*. Literary and Linguistic Computing **16**(1) (2001) 45 – 57
8. Orăsan, C., Evans, R.: Learning to identify animate references. In Daelemans, W., Zajac, R., eds.: Proceedings of CoNLL-2001, Toulouse, France (July, 6 – 7 2001) 129–136
9. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: Proceedings of ACL02. (2002)
10. Muñoz, R., Saiz-Noeda, M., Montoyo, A.: Semantic information in anaphora resolution. In: Proceedings of PorTAL 2002. (2002) 63 – 70

11. Orăsan, C.: PALinkA: a highly customizable tool for discourse annotation. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, Sapporo, Japan (July, 5 -6 2003) 39 – 43
12. Hasler, L., Orăsan, C., Naumann, K.: NPs for Events: Experiments in Coreference Annotation. In: Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation (LREC2006), Genoa, Italy (24 – 26 May 2006) 1167 – 1172
13. Vilain, M., Burger, J., Aberdeen, J., Connoly, D., Hirschman, L.: A model-theoretic coreference scoring scheme. In: Proceedings of the 6th Message Understanding Conference (MUC-6), San Francisco, California, USA (1995) 45–52
14. Hasler, L., Orăsan, C., Mitkov, R.: Building better corpora for summarisation. In: Proceedings of Corpus Linguistics 2003, Lancaster, UK (March, 28 – 31 2003) 309 – 319
15. Mitkov, R., Hallett, C.: Comparing pronoun resolution algorithms. Journal of Computational Intelligence (forthcoming)
16. Orăsan, C., Pekar, V., Hasler, L.: A comparison of summarisation methods based on term specificity estimation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC2004), Lisbon, Portugal (May 26 – 28 2004) 1037 – 1041
17. Justeson, J.S., Katz, S.L.: Technical terminology: some linguistic properties and an algorithm for identification in text. Journal of Natural Language Engineering **3**(2) (1996) 259–289
18. Hulth, A.: Reducing false positives by expert combination in automatic keyword indexing. In: Proceedings of RANLP 2003, Borovetz, Bulgaria (September 2003) 197–203
19. McCallum, A.K.: Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering. http://www.cs.cmu.edu/ mccallum/bow (1996)
20. Orăsan, C.: Comparative evaluation of modular automatic summarisation systems using CAST. PhD thesis, University of Wolverhampton (2006)

Author Index