Sobha Lalitha Devi
António Branco
Ruslan Mitkov (Eds.)

LNAI 5847

Anaphora Processing and Applications

7th Discourse Anaphora and Anaphor Resolution Colloquium, DAARC 2009
Goa, India, November 2009
Proceedings

Springer

Lecture Notes in Artificial Intelligence 5847

Edited by R. Goebel, J. Siekmann, and W. Wahlster

Subseries of Lecture Notes in Computer Science

Sobha Lalitha Devi
António Branco
Ruslan Mitkov (Eds.)

Anaphora Processing and Applications

7th Discourse Anaphora
and Anaphor Resolution Colloquium, DAARC 2009
Goa, India, November 5-6, 2009
Proceedings

Springer

Volume Editors

Sobha Lalitha Devi
Anna University – K.B. Chandrasekhar Research Centre
MIT Campus of Anna University, Chromepet, Chennai 600044, India
E-mail: sobha@au-kbc.org

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

Ruslan Mitkov
University of Wolverhampton
School of Humanities, Languages and Social Studies
Research Group in Computational Linguistics
Wolverhampton WV1 1SB, UK
E-mail: r.mitkov@wlv.ac.uk

Library of Congress Control Number: 2009936009

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 3-642-04974-5 Springer Berlin Heidelberg New York
ISBN-13 978-3-642-04974-3 Springer Berlin Heidelberg New York

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12775965 06/3180 5 4 3 2 1 0

Preface

Distribution of anaphora in natural language and the complexity of its resolution have resulted in a wide range of disciplines focusing their research on this grammatical phenomenon. It has emerged as one of the most productive topics of multi- and inter-disciplinary research such as cognitive science, artificial intelligence and human language technology, theoretical, cognitive, corpus and computational linguistics, philosophy of language, psycholinguistics and cognitive psychology. Anaphora plays a major role in understanding a language and also accounts for the cohesion of a text. Correct interpretation of anaphora is necessary in all high-level natural language processing applications.

Given the growing importance of the study of anaphora in the last few decades, it has emerged as the frontier area of research. This is evident from the high-quality submissions received for the 7th DAARC from where the 10 excellent reports on research findings are selected for this volume. These are the regular papers that were presented at DAARC.

Initiated in 1996 at Lancaster University and taken over in 2002 by the University of Lisbon, and moved out of Europe for the first time in 2009 to Goa, India, the DAARC series established itself as a specialised and competitive forum for the presentation of the latest results on anaphora processing, ranging from theoretical linguistic approaches through psycholinguistic and cognitive work to corpus studies and computational modelling. The series is unique in that it covers this research subject from a wide variety of multidisciplinary perspectives while keeping a strong focus on automatic anaphor resolution and its applications.

The programme of the 7th DAARC was selected from 37 initial submissions. It included 19 oral presentations and 8 posters from over 50 authors coming from 14 countries: Belgium, Czech Republic, Denmark, Germany, India, Norway, Portugal, Russia, Romania, Spain, The Netherlands, Taiwan, UK and the USA. The submissions were anonymised and submitted to a selection process by which each received three evaluation reports by experts from the programme committee listed below.

On behalf of the Organising Committee, we would like to thank all the authors who contributed with their papers for the present volume and all the colleagues in the Programme 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.

Chennai, August 2009

Sobha Lalitha Devi
António Branco
Ruslan Mitkov

Organisation

The 7th DAARC colloquium was organized by AU-KBC Research Centre, Anna University Chennai, Computational Linguistics Research Group.

Organising Committee

António Branco	University of Lisbon
Sobha Lalitha Devi	Anna University Chennai
Ruslan Mitkov	University of Wolverhampton

Programme Committee

Alfons Maes	Tilburg University
Andrej Kibrik	Russian Academy of Sciences
Andrew Kehler	University of California San Diego
Anke Holler	University of Goettingen
António Branco	University of Lisbon
Christer Johansson	Bergen University
Claire Cardie	Cornell University
Constantin Orasan	University of Wolverhampton
Costanza Navarretta	University of Copenhagen
Dan Cristea	University of Iasi, Romania
Elsi Kaiser	University of Southern California
Eric Reuland	Utrecht Institute of Linguistics
Francis Cornish	University of Toulouse-Le Mirail
Georgiana Puscasu	University of Wolverhampton
Graeme Hirst	University of Toronto
Iris Hendrickx	Antwerp University
Jeanette Gundel	University of Minnesota
Jeffrey Runner	University of Rochester
Joel Tetreault	Education Testing Service, USA
Jos Van Berkum	Max Planck Institute for Psycholinguistics
José Augusto Leitão	University of Coimbra
Kavi Narayana Murthy	Hyderabad Central University
Klaus Von Heusinger	Konstanz University
Lars Hellan	Norwegian University of Science and Technology
Maria Mercedes Piñango	Yale University
Marta Recasens	University of Barcelona
Martin Everaert	Utrecht Institute of Linguistics
Massimo Poesio	University of Essex
Patricio Martinez Barco	University of Alicante

Peter Bosch	University of Osnabrueck
Petra Schumacher	University of Mainz
Renata Vieira	PUC Rio Grande do Sul
Richard Evans	University of Wolverhampton
Robert Dale	Macquarie University
Roland Stuckardt	University of Frankfurt am Main
Ruslan Mitkov	University of Wolverhampton
Sergey Avrutin	Utrecht Institute of Linguistics
Shalom Lappin	King's College, London
Sivaji Bandopadhyaya	Jadavpur University
Sobha Lalitha Devi	Anna University Chennai
Tony Sanford	Glasgow University
Veronique Hoste	Gent University
Vincent Ng	University of Texas at Dallas
Yan Huang	University of Auckland

Table of Contents

Why Would a Robot Make Use of Pronouns? An Evolutionary Investigation of the Emergence of Pronominal Anaphora

Dan Cristea[1,2], Emanuel Dima[1], and Corina Dima[1]

[1] Alexandru Ioan University of Iasi, Faculty of Computer Science,
16, Berthelot St., 700486 Iasi, Romania
[2] Romanian Academy, Institute of Computer Science,
The Iasi branch Blvd. Carol I 22A, 700505, Iasi, Romania
{dcristea,dimae,cdima}@info.uaic.ro

Abstract. We investigate whether and in what conditions pronominal anaphora could be acquired by intelligent agents as a means to express recently mentioned entities. The use of pronouns is conditioned by the existence of a memory recording the object previously in focus. The approach follows an evolutionary paradigm of language acquisition. Experiments show that pronouns can be easily included in a vocabulary of a community of 10 agents dialoguing on a static scene and that, generally, they enhance the communication success.

Keywords: Anaphora, Language emergence, Artificial agents, Simulation, Language games.

1 Introduction

In this paper we study the acquisition of pronominal anaphora by intelligent agents in locally situated communication. The research represents a step in the attempt to decipher the acquisition of language in communities of humans.

Recently, the historical interest to decipher the origins of language seems to be reopened. Progress in this direction comes from approaches over language evolution, especially the experiments towards lexicon acquisition and grammar acquisition. Models of language acquisition [7,2] hypothesise that language users gradually build their language skills in order to optimise their communicative success and expressiveness, as triggered by the need to raise the communication success and to reduce the cognitive effort needed for semantic interpretation.

The Talking Heads experiments [6,8] have already shown that a shared lexicon can be developed inside a community of agents which are motivated to communicate. The participants in the experiments are intelligent humanoid robots, able to move, see and interpret the reality around them (very simple scenes of objects), as well as to point to specific objects. They are programmed to play language guessing games in which a speaker and a hearer, members of a community of agents, should acquire a common understanding of a situation which is visually

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 1–14, 2009.

shared by both participants in the dialogue. After tens of thousands of such games, played in pairs by members of the community, a vocabulary that gives names to concepts which are needed to differentiate the properties of objects spontaneously arises. The vocabulary is shared by the majority of agents and is relatively stable at perturbing influences caused by population growth, decline, or mixing with other smaller groups. It was shown [11] that using multi-words instead of single words could reduce the size of the lexicon, therefore yielding a more efficient communication system.

The next step deals with the acquisition of grammar. Van Trijp [12] recently showed how rudiments of grammar can be developed as a result of interactions. The studied aspects touched the capacity of agents of inventing grammatical markers for indicating event structures, the formation of semantic roles, and the combination of markers into larger argument structure constructions through pattern formation. A formalism used to model grammaticality in the evolutionary approach is Fluid Construction Grammar [10].

In the present paper we are interested to see if the acquisition of pronominal anaphora inside a community of intelligent agents can be empirically modelled by following an evolutionary approach and, if so, to point out which are the minimal cognitive requirements that allow the use of pronominal anaphors, how many interactions would be necessary for a pronoun to appear in the vocabulary of a community, and what is the communication gain if pronouns are used.

The world is extremely complex and a trial to bring it into the laboratory in order to model language acquisition in a natural setting is unrealistic. Languages have evolved to manage the complexity of the world around us and it is clearly an error to consider that first the humans gained a sophisticated cognitive apparatus and only after that moment they started to invent the language. It is known that the evolution of the human brain is closely correlated to the acquisition of language [9]. We are, therefore, forced to simplify the model we employ.

A few simple rules of a linguistic game are explained in section 2. In section 3 we ground the use of pronouns on a minimal cognitive infrastructure. In section 4, some scene settings displaying increasing difficulty of comprehension are introduced. The dialogue experiments are conducted in these settings and results are presented in section 5. Finally, section 6 summarises some conclusions.

2 The Experiments Framework

We organized a number of game-based experiments during which the agents were expected to achieve rudiments of discourse-level performance. For that, a Java framework, called Uruk [3], in which games can be easily defined and which allows for any number of experiments, has been developed. By properly manipulating a number of parameters, Uruk can be made to support the description of the closed worlds, of the agents' cognitive capacities and their lexical memories, as well as of their dialogues.

We consider a world as being composed of objects with properties (shape, colour, position, etc.), and a scene is a world with a specified configuration

of objects. The framework offers two ways to generate a scene: by manually describing all the objects populating it as well as all their properties, or by generating it randomly. The number of objects generated in a random scene can be set within specified lower and upper bounds.

We do not speak about software agents *stricto sensu* [5]; in our framework they are not mobile, are not autonomous, and do not react to the change of their environment. However, based on a learning process, they will arrive to possess rudiments of language. The language is acquired through interactions in a community of similarly equipped agents, although not necessarily identical. The only environment the agents can interact with is a scene which can be perceived through a number of perception channels. Each channel targets a specific property of an object. The acuity of agents on these channels can vary at will, agents being able to discretise a 0-to-1 range of continuous values, on each channel, in a set of categories, with different granularities. We model this way the natural diversity within a community. In much the same way, an average human being can perceive in the continuous spectre of colours only 7 important ones, while a painter distinguishes 100 nuances, for which she has names.

A game is a specified protocol of interaction between two agents. An example of such a protocol is the "guessing game" [8], where one agent chooses an object in the scene, generates an utterance that describes it, and a second agent must guess the object described by the utterance (without knowing which was the chosen object). If the object is correctly indicated, the trust of both agents in the proper usage of the words describing the conceptualisation of the object increases. If the object is not guessed, a repairing strategy is applied: the speaker points to the object he has chosen and the trust it has in the words used to name it decreases, while the hearer either learns the new words or associates a greater level of trust for this expression to describe the conceptualisation of the object. After a large number of games of this kind, played in pairs by agents, the community shares a common vocabulary and associates words to categories (concepts) with a high level of trust.

The game develops as follows. The first turn belongs to the speaker. He silently chooses an object, which will be known as the *focus*, from the objects that are present in the scene and runs a conceptualisation algorithm to find all sets of categories that can unambiguously distinguish the focus among all the other objects (e.g. *red square* in a scene in which there is only one red square among other objects). From the found list of sets of categories a lexicalisation algorithm then selects just one set of categories for which the most appropriate lexical expression can be formed.

The lexicon of the agent is a set of 3-tuples of the form (category, word, confidence). Such a 3-tuple is an association between a category and a word, weighted by a relative confidence factor. Let's note that the correspondence between the conceptual space and the lexicon of an agent can be a many-to-many mapping between categories and words (one word can be ambiguous because it can designate more categories, and a category can be named by a set of synonymous words). When producing an utterance (as a speaker), an agent

needs to find out the word that describes best a certain category among its set of synonyms, and vice-versa, when deciphering an utterance (as a hearer) she has to associate the most plausible concept to an ambiguous word. The confidence factor is used in precisely this situation: the agent scans its lexicon and selects the association with the maximum confidence factor between all the associations with the designated category (in speaking) and word (in hearing).

In order to find the optimal set of words describing the distinctive categories, the lexicalisation algorithm computes scores for each found set of discriminating categories as the average of the best associations (words with the highest confidence factors that the agent knows for the chosen categories). Then the set of words with the highest score is selected.

The winning expression is the shortest one (in number of items) when only one such set expression exists. When more than one winning sets have the same size, the one with the maximum confidence score is chosen. This strategy implements the *principle of economy of expression*. Only when there is more than one set with the maximum confidence and the same minimum size, the winning expression will be decided randomly. The chosen expression is then "spoken", i.e. transferred to the hearer.

The second turn of a multi-game belongs to the hearer who tries to decompose the transferred string and to find an object in the environment that would correspond to the description. The "heard" string is first tokenized by the hearer into elementary words. Then the agent matches each one of these words with the most likely category, by scanning her own lexical memory and selecting the category that is associated with a word that is present in the description with the highest confidence.

In the best scenario, the agent would exactly match the words and reproduce the original set of characteristics. Based on this set, the hearer is able to judge the identity of an object in the scene, which, luckily, will be the same as the intended focus in the speaker's mind. In an average simulation, though, some words could be unknown to the hearer and some could be associated to a different category than the one used by the speaker. In any case, the result of this interpretation step is a (possibly empty) set of characteristics. The hearer will now find the objects in the world that possess all these characteristics. When more than one object could be a possible target, the agent will simply choose one at random. If the hearer has a plausible candidate, she will point to it, or otherwise, when the decoding gives an empty set of objects, an "I don't know" kind of message is issued.

Next, if there is one object pointed to by the hearer, it is compared against the object chosen by the speaker (the focus). Based on this comparison, the game can succeed (the indicated object by the hearer is indeed the focus chosen by the speaker) or fail (no object identified, or the one indicated is a wrong one). Both conceptual spaces of the speaker and the hearer are influenced by the result of the game.

In most of the success cases both agents have the same word-category associations, and their confidence in these associations will be increased by a relatively

large quota. It could happen, however, to finish a game with success, although the interpretation of some of the words used by the agents is different. In this case the agents managed to communicate, but only by chance, and the increase of confidence will be misleading. It is expected that this erroneous conclusion will be penalized later in other interactions.

In the fail case, the object answered by the hearer being different than the focus chosen by the speaker, both agents conclude that the words they have used to name categories were inappropriate and so they decrease the confidence of these word-category pairs. The negative quota of penalty is somewhat lower than the positive quota used in a successful game, such that a successful game outweighs a failed one.

The next step is the learning phase that takes place only in case of fail. A game can fail because the hearer didn't know the words used by the speaker, so the hearer is supposed to improve her word-category associations. The speaker shows to the hearer the identity of the intended focus object. With the knowledge of the true object and of the words that describe it, the hearer retraces the operations made by the speaker and computes the list of discriminating sets of categories. As she knows that the number of words equals the number of categories, she only retains the sets that have the same size as the list of words. However, as she doesn't know the order in which these categories were used, she stores the words in her internal memory, as every possible association between the words used and the distinctive categories. The confidence factor for each association is set as the maximum possible confidence divided by the number of associations.

The final step of the game is purging of the lexicons. During this phase all associations whose confidence decreased below a minimum threshold are removed from the agents' memories.

The overall theoretical complexity of the algorithms used for finding the name of an object and for learning is high. Finding the name of an object is exponential on the number of channels. This happens because all subsets of the categories that an object falls into are computed, in order to lexically describe it. The learning algorithm is factorial on the number of lexemes used by the speaker. However, this unacceptable theoretical complexity does not harm too much the running time, because the number of perception channels of an agent is fixed and usually small (the current experiments set it at 5). Also, the distinctive categories used to identify an object are bounded by the number of perception channels of the agent.

3 The Inception of Pronouns

Anaphora represents the relationship between a term (called "anaphor") and another one (called "antecedent"), when the interpretation of the anaphor is in a certain way determined by the interpretation of the antecedent [4]. When the anaphor refers the same entity as the antecedent, we say that the anaphor and the antecedent are coreferential. When the surface realisation of the anaphor is that of a pronoun, the coreference relation also fulfils other functions:

– it brings conciseness in the communication by avoiding direct repetitions of a previous expression, thus contributing to the economy of expression – a central principle in the communication between intelligent agents;

– it maintains the attention focused on a central entity by referring it with extremely economical evoking means. Indeed only entities which already have a central position in the attention could be referred by pronouns and, once referred, their central position is further emphasised.

Anaphora, as a discourse phenomenon, presupposes non-trivial cognitive capacities. The one we are concerned with in this paper is the capacity of memorising the element in focus. This capacity is so central and elementary that we decided to consider it as being provided by a dedicated "perception" channel – actually a memory channel. Indeed, both cognitive aspects of distinguishing between right and left (to give a common example of perception) and of remembering that a certain object was in focus recently involve primitive cognitive functions. The lack of memory would make a dialogue impossible, the same way as the lack of spatial perception abilities would make the recognition of spatial relations impossible.

The focussing memory is modelled through a channel called **previous-focus,** with two values [**true**, **false**]. Excepting for the first utterance of the dialogue when there is no previously focused entity, on each subsequent utterance there is one entity (object) which is remembered as being the focus of the previous game. As such, each object in the scene has a value of **false** on the **previous-focus** channel, except for the object which has been in focus previously and whose corresponding value on this channel is **true**.

It is clear that modelling the memory of objects previously in focus as a one-place channel (only one object can have the value "true") is a severe simplification, which we accept in this initial shape of the experiments. In reality, human agents are known to be able to record many more discourse entities already mentioned, on which pronouns can afterwards be anchored. The differentiating properties (of pronouns) can include features like animacy, gender and number. Using these properties, as well as the syntactic structure and the discourse structure, a sentence could include more than just one pronoun, each referring unambiguously to a different antecedent.

We are not concerned here to model the cognitive processes that make possible the recognition of objects. We simply assume that the agents have either the capacity to distinguish the focussed object among the other objects, based on its intrinsic properties, or that the agent eye-tracks the objects from a first position to a second one between games. In the first case the object would have to be identified again based on the memorized specific features (the memory channel resembling more a memory cell), while in the second case the identity of the focussed object would have to be maintained rather than regained from memory.

The type of games we are interested in when modelling anaphoric phenomena are multi-turn, such that one entity which has been already in focus previously, could be referred again later. In this study, we are targeting only pronominal

anaphors. If we want an agent to develop the ability of using pronouns, the dialogue should include a sequence of utterances in which an entity is mentioned more than once.

4 The Settings

The problem we are concerned with is *when* and *why* intelligent agents would develop linguistic abilities for using anaphoric means in communication and how anaphora could complete a conceptualisation.

It is clear that an agent could have at least two reasons for choosing to name an object by a pronoun:

– because it uses less words (for instance, *it* instead of *the left red circle*);
– because this way the OLD (therefore, the entity previously in focus) is explicitly signalled, maintaining it there.

On the other hand, an agent has also at least one reason why not using a pronoun:

– because it could introduce an ambiguity.

The use of pronouns should emerge naturally during the experiments, solely by modelling these contrary tendencies. It should, therefore, not be enforced (given programmatically).

To model the acquisition of pronominal anaphora, four different settings have been used, which we believe present an ascending degree of complexity. All are anchored on a two-turn game. What make the difference between these settings are the changes in the scene of the second turn as compared to the first, as well as the chosen focus.

In the first setting (Fig. 1) both games are played in the same scene and the co-speaker will focus in the second game the same object that the speaker focussed in the first.

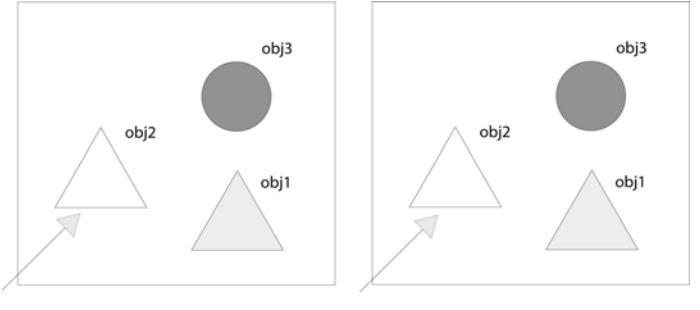

Fig. 1. Setting 1
Turn 1: A names obj_2 by *low left*
Turn 2: B names obj_2 by *that*

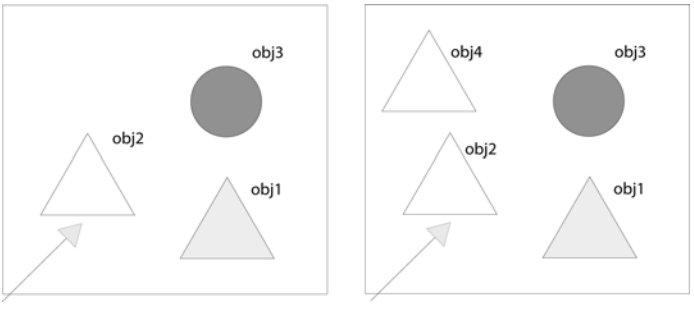

Fig. 2. Setting 2
Turn 1: A names obj_2 by *low left*
Turn 2: B names obj_2 by *that*

In the second setting (Fig. 2) new objects are introduced in the scene of the second game, while the focus remains unchanged.

In the third setting (Fig. 3), the objects in the second game's scene are shuffled (their spatial properties like horizontal and vertical position are randomly changed). The co-speaker will keep the focus on the same object, although it might have changed its position.

In the fourth setting (see Fig. 4), the scene of the second game is again a shuffled version of the scene in the first game and the focus can no longer be identified by any of the attributes used in the first game. In this particular scene, the agents do not distinguish colours or shapes so the objects can be identified only through position and anaphoric means.

All experiments have been run with the following parameters: number of agents: 10; number of multi-games: 5000; number of objects in the scenes: between 5 and 10; channels: "hpos" (horizontal position), "vpos" (vertical position), "color", "shape", "previous-focus"; channels granularity (number of distinguishable values

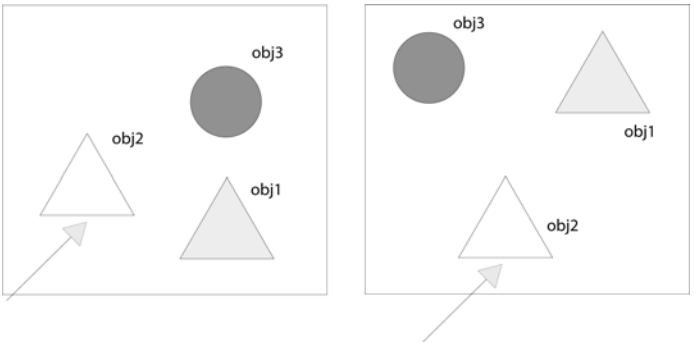

Fig. 3. Setting 3
Turn 1: A names obj_2 by *low left*
Turn 2: B names obj_2 by *that*

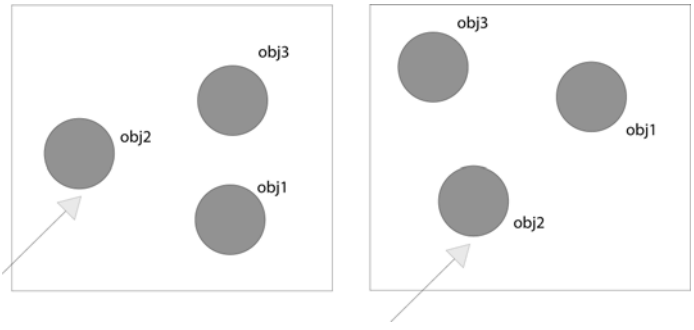

Fig. 4. Setting 4

Turn 1: A names obj_2 by *left*
Turn 2: B names obj_2 by *that*

on channels): from 2 to 4. A number of 4 values on the "hpos" channel, for instance, could mean: "extreme-left", "left", "right" and "extreme-right". As mentioned, the "previous-focus" channel has 2 values: "true" and "false".

As we see, in every multi-game the focus is maintained on the same object in both turns. Let us notice that it makes no difference who the speaker is in the second game. Only for the sake of displaying a dialogue we considered the second utterance as produced by the co-speaker.

5 The Results

Fig. 5–9. display the average success rates along 10 series of 2500 multi-games, for different configurations of objects and settings. The success rate is considered to be, at a certain moment in time, the percent of game successes in the previous 100 games.

Fig. 5 and 6 show the success rate in setting 1 with scenes counting 5 and 8 objects, while Fig. 7–9 display the success rate in settings 2-4 when there are 8 objects in the scene. In all experiments, only multi-games which reported success after the first turn have been retained, as we were interested here only in the acquisition of pronouns (mentioned only in the second turn in each multi-game) and not in a stabilisation of a lexicon in general. So, if at the end of the first turn, agent B does not recognise the object indicated by agent A, the game is stopped and disregarded. In all figures, the (black) line above reports the percent of general success rate (after the second turn), while the (gray) line below reports the success rate that is due to the use of pronouns.

The abruptly growing shapes of the lines above, in all four settings, show that, very quickly, the agents acquire a common understanding of the objects in the scene (to be more precise – over the object in the focus). In general, after 300-400 games, the success rate stabilises to 100%. However, as the (gray) lower lines show, in fewer cases this common understanding is due to the use of pronouns.

Fig. 5. Setting 1 – with 5 objects

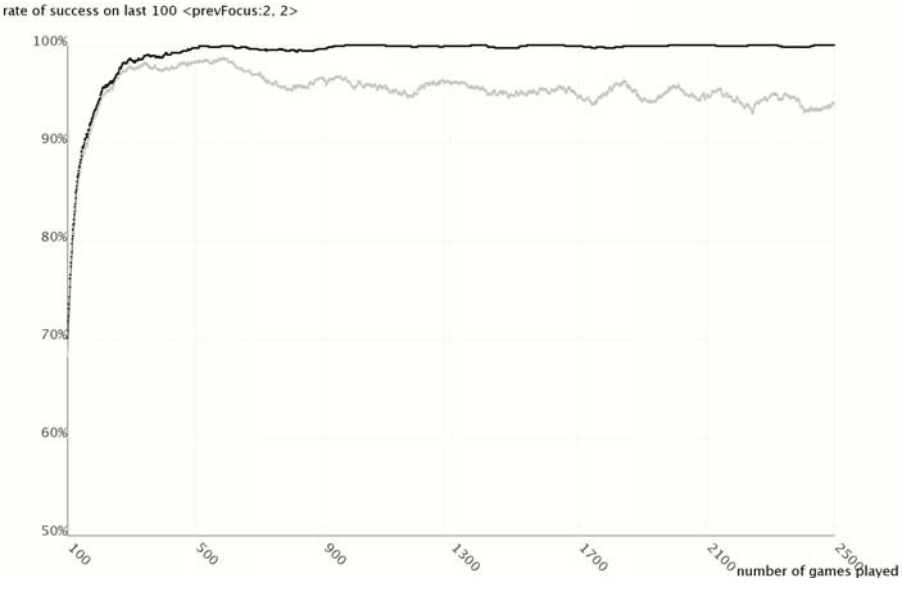

Fig. 6. Setting 1 – with 8 objects

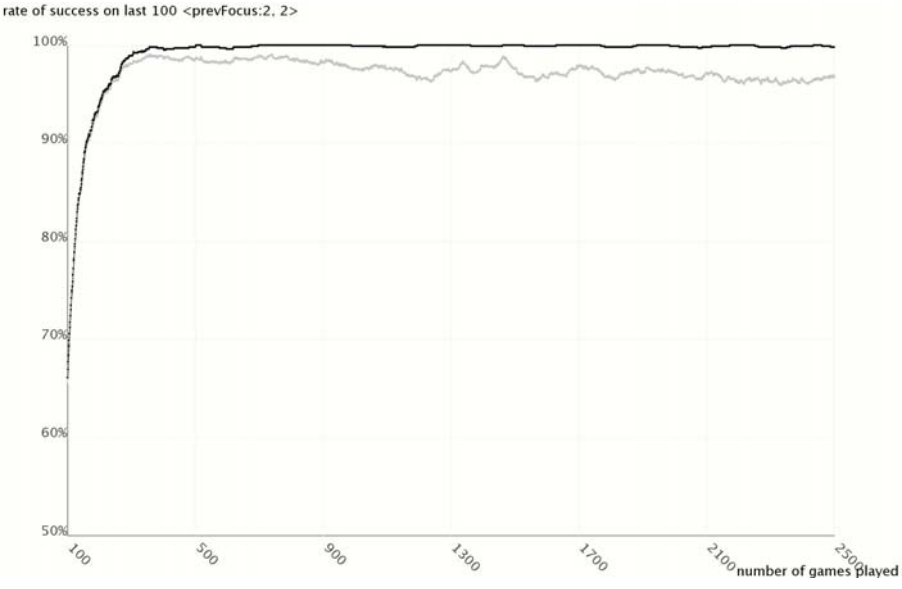

Fig. 7. Setting 2 – with 8 objects

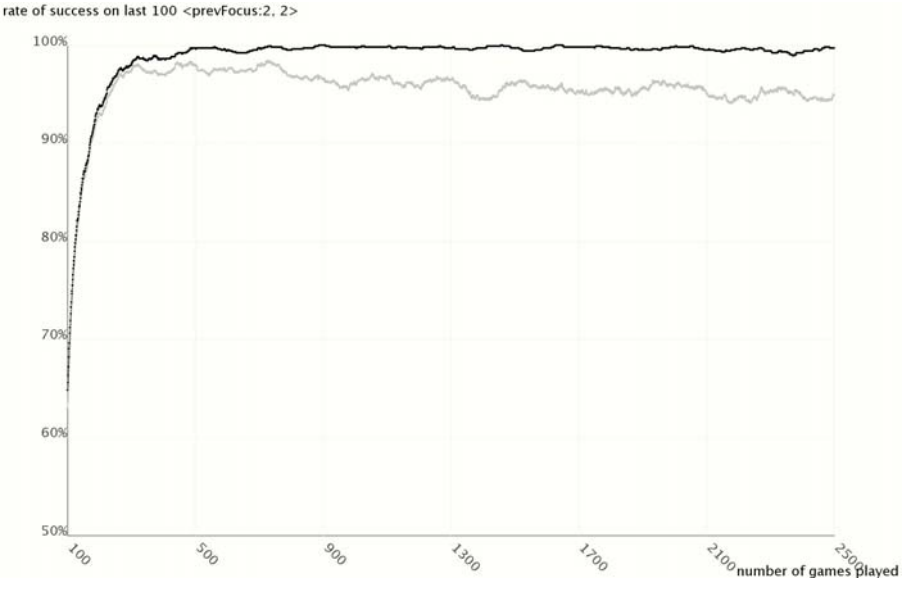

Fig. 8. Setting 3 – with 8 objects

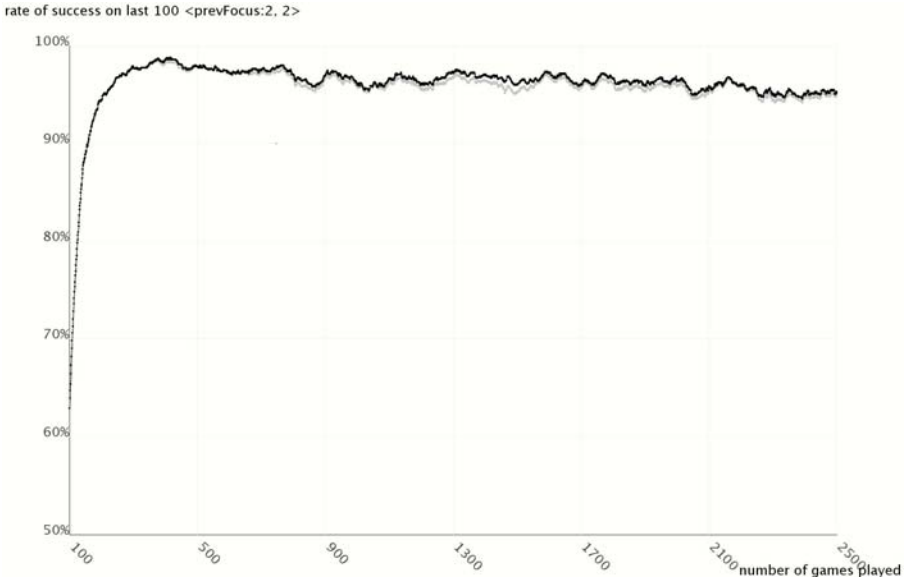

Fig. 9. Setting 4 – with 8 objects

This should not be interpreted as an indication of the fact that the use of pronouns reduces the success rate, but that in some cases other referential expressions than pronouns are also used to identify an object which has been previously in focus (for example, in setting 3, Fig. 3, B can use *down* instead of *that*).

However, if we compare Fig. 5 and Fig. 6, we see that when the number of objects is larger, the need to use pronouns also goes up. This is clearly due to the fact that a greater agglomeration of objects in the scene makes their identification based on other features than being recently in focus more ambiguous. Indeed, the agents chose randomly among the shortest best known categorisations which one to use for identifying the object in focus from those able to individualise it unambiguously. If all possible utterances have the same length and the same confidence (the confidence of a linguistic expression is calculated as the mean value of the confidence of the words used to utter the corresponding categorisation) one of them is chosen randomly. Choosing the shortest utterance is the only bonus that favours the economy of expression, therefore the use of pronouns. The graphs show that when there are more objects in the scene, being recently in focus remains the conceptual feature with the highest confidence.

An interesting thing is revealed by the graph in Fig. 9: the two lines representing global success rate and pronoun-based success rate are practically identical. This means that when the situation is very complex, in almost all cases the agents prefer to use a pronoun to identify an already mentioned object.

Finally, we were interested to see what happens when we impose the use of pronouns. Fig. 10 shows two lines, both drawn for setting 1: the lower (black) line represents the normal use of pronouns in the case of success in the second

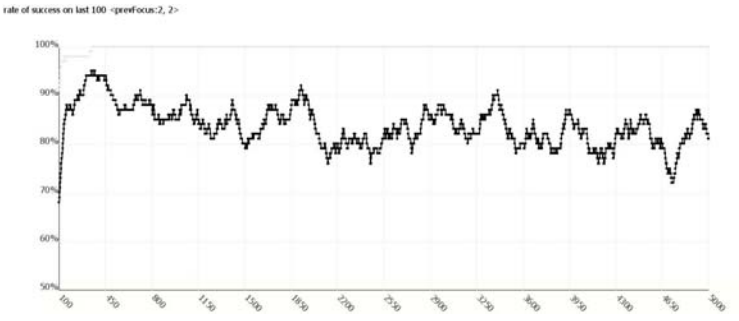

Fig. 10. Imposing the use of pronouns – setting 1

turn, while the upper (gray) line represents the success rate in the second turn when we enforced the use of pronouns. The experiment shows that the particular conditions of this setting make superfluous the need for more than one channel (in this case "previous-focus") to identify the focus. Indeed, practically, each time a pronoun is used the success is guaranteed.

6 Conclusions

In this paper we have advocated that the acquisition of pronouns in language can follow an evolutionist pattern therefore pronouns can appear in language as a natural, spontaneous process, driven by the necessity of the agents to acquire common understanding over a situation.

The study does not show, however, that this is the only way in which pronouns could have appeared in natural languages. It simply shows a possibility. Although the experiment was successful, it is also not realistic to claim that our model of representing anaphoric phenomena and their emergence in a community of artificial agents copies the natural way in which anaphora appeared in languages. We can only try to guess the cause of anaphora's natural inception. The limits of the experiment are obvious.

We have used a paradigm in which a community of agents communicate. A common agreement over a focussed object in a scene is rewarded by an enhancement of the trust in both the conceptualisation used and the linguistic means to express it. After a number of experiments, a certain lexicon is acquired by the community.

The model we used has considered the existence of a memory channel remembering the object recently in focus. When such a channel is open, the identification of an object already mentioned, and which should be mentioned again, can be made quicker and with less ambiguity because it implies less categorisation. The linguistic expression of this economic categorisation is the pronoun. The experiments show a clear tendency of the agents to enhance their linguistic ability to use pronouns in more and more complex contexts.

When the number of objects in the scene increases, the chance that the "previous-focus" channel is the only channel that uniquely identifies an object is very high and therefore the use of pronoun becomes dominant.

In the future it would be interesting to study what semantic features attract the specialisation of pronouns. Can the categories of male/female, animate/inanimate and singular/plural, as they are used to differentiate pronominal forms in most languages, be generalised? Could a class of experiments intended to put in evidence the different semantic features of anaphoric expressions be imagined within the limited worlds of the 'talking heads'? Another thing that we don't know yet is which are the levers that can be triggered to restrain the proliferation of lexical forms of pronouns in the community of agents, as in most natural languages there are very few synonyms to express one category of pronouns.

Amongst the details of improving the framework, apart from the simulations in which parallel machines shall be used as hardware support to speed up the execution, there are also some other theoretical questions to investigate. Open problems are the connection between the short-term memory and the anaphora, the possibility of simulating distinct pronominal categories like the natural-language forms that are connected with gender (which can not only be masculine or feminine but also "classes of hunting weapons, canines, things that are shiny" [1]), number and others, as well as a resolution algorithm for these cases of ambiguity.

References

1. Boroditsky, L.: How does our language shape the way we think? In: What's Next: Dispatches on the Future of Science. Vintage, London (2009)
2. Briscoe, T.: Linguistic evolution through language acquisition: formal and computational models. Cambridge Univ. Press, Cambridge (1999)
3. Dima, E.: Anaphoric Phenomen. In: Evolving Lexical Languages. Master thesis, Department of Computer Science, Alexandru Ioan Cuza University, Iasi (2009)
4. Lust, B., Reidel, D.: Introduction to Studies in the Acquisition of Anaphora (1986)
5. Nwana, H.S.: Software Agents: An overview. Cambridge University Press, Cambridge (1996)
6. Steels, L.: Self-organizing vocabularies. In: Langton, C.G. (ed.) Proceeding of Alife V. (1997)
7. Steels, L.: The synthetic modeling of language origins. Evolution of Communication 1(1), 1–35 (1997)
8. Steels, L.: The Talking Heads Experiment. Words and Meanings. Vol. 1, Special pre-edition for LABORATORIUM, Antwerpen (1999)
9. Steels, L.: Intelligence with representation. Philosophical Transactions of the Royal Society A 361, 2381–2395. 51 (2003)
10. Steels, L.: Fluid Construction Grammar. Ellezelles Course Notes (draft version). Sony Computer Science Laboratory Paris, Artificial Intelligence Laboratory, VUB, Brussel (2008)
11. Van Looveren, J.: Design and Performance of Pre-Grammatical Language Games. Ph.D. thesis, Vrije Universiteit Brussel, Brussels. 53,155 (2005)
12. Van Trijp, R.: Analogy and Multi-Level Selection in the Formation of a Case Grammar. A Case Study in Fluid Construction Grammar, Ph.D. thesis. University of Antwerpen (2008)

Automatic Recognition of the Function of Singular Neuter Pronouns in Texts and Spoken Data

Costanza Navarretta*

Centre for Language Technology, University of Copenhagen,
Njalsgade 140-142, build. 25, 2300 Copenhagen S, Denmark
costanza@hum.ku.dk

Abstract. We describe the results of unsupervised (clustering) and supervised (classification) learning experiments with the purpose of recognising the function of singular neuter pronouns in Danish corpora of written and spoken language. Danish singular neuter pronouns comprise personal and demonstrative pronouns. They are very frequent and have many functions such as non-referential, cataphoric, deictic and anaphoric. The antecedents of discourse anaphoric singular neuter pronouns can be nominal phrases of different gender and number, verbal phrases, adjectival phrases, clauses or discourse segments of different size and they can refer to individual and abstract entities. Danish neuter pronouns occur in more constructions and have different distributions than the corresponding English pronouns *it*, *this* and *that*. The results of the classification experiments show a significant improvement of the performance with respect to the baseline in all types of data. The best results were obtained on text data, while the worst results were achieved on free-conversational, multi-party dialogues.

Keywords: Singular neuter pronouns, Pronominal functions, Machine learning, Individual and Abstract anaphora, Text and Spoken corpora, Annotation.

1 Introduction

In this paper we describe the results of unsupervised (clustering) and supervised (classification) learning experiments with the purpose of recognising the function of singular neuter personal and demonstrative pronouns (sn-pronouns henceforth) in Danish corpora of written and spoken language. Therefore, we will relate our work to relevant work done on English and Dutch data. Danish sn-pronouns are very frequent and have many functions such as non-referential (expletive henceforth), cataphoric, deictic and anaphoric. The antecedents of discourse anaphoric sn-pronouns can be nominal phrases of different gender and number, verbal phrases, adjectival phrases, clauses or discourse segments of different size and they can refer to individual and abstract entities (individual and

* Thanks to Sussi Olsen, Hanne Fersøe and Patrizia Paggio, University of Copenhagen.

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 15–28, 2009.

abstract anaphors, respectively). Danish sn-pronouns occur in more constructions and have different distributions than the corresponding English pronouns *it*, *this* and *that*.

The first step towards the resolution of the anaphoric occurrences of sn-pronouns is their identification and classification with respect to their type of antecedent, see also [7], and this is the subject of the paper. The main goals of our work have been the following: i) to test how well unsupervised and supervised learning algorithms identify the function of Danish sn-pronouns in texts and spoken data; ii) to individuate the information which is most useful to this task; iii) to evaluate the function classification provided in the annotated corpora which we used.

We start by discussing related work in section 2; then we present the data which we have used in section 3; we describe our machine learning experiments and discuss the obtained results in section 4; finally we conclude and present work still to be done in section 5.

2 Related Work

To our knowledge there is no previous work to automatically recognise the function of Danish sn-pronouns. Some algorithms to resolve English pronominal anaphora presuppose pre-editing of the data to allow for the exclusion of non-referential and cataphoric occurrences of pronouns, other algorithms include the identification of some of the pronominal functions[1].

When full-parsing of data is not possible or desirable, filtering mechanisms and selectional preferences are applied to the data to identify the main functions of pronominal occurrences and exclude some of them from the resolution algorithms, see among others [8,18,23,19].

The resolution of the English pronouns *it*, *this* and *that* in English dialogues has been addressed in [7,3,25,19]. Eckert and Strube's algorithm [7] relies on complex knowledge about language and discourse structure and identifies individual and abstract occurrences of third-person singular pronouns in English on the basis of the context in which the pronouns occur and of their type (personal or demonstrative). The algorithm has only been tested manually and non-anaphoric occurrences of the pronouns were excluded from the test. The same method has been partly adapted and incorporated in an algorithm for resolving Danish discourse pronominal anaphora [20,21]. Also the Danish algorithm has only been tested manually, relies on many knowledge sources and accounts only for pronominal anaphoric occurrences. Byron's PHORA-algorithm [3] resolves the occurrences of *it*, *this* and *that* in domain-specific dialogues. It is implemented and relies on semantic knowledge and a speech act model. An other implemented algorithm for resolving the same English pronouns is described in [25]. This algorithm

[1] A comparison of the most known resolution algorithms including information on how much pre-editing and pre-processing they require can be found in [17].

relies on various types of linguistic information extracted from the Penn Treebank. Finally a machine learning approach for identifying and resolving third-person singular pronouns in English is proposed in [19]. The algorithm has been trained and tested on five dialogues, which were annotated for this task, and relies exclusively on the corpus annotation. The algorithm is exposed to all occurrences of *it*, but the non-anaphoric occurrences were pre-annotated in the data in order to trigger all types of negative preferences which allowed the system to sort them out. The results of this algorithm are much lower than those obtained by the algorithms relying on complex linguistic and discourse structure knowledge.

A machine learning approach for recognising non-referential occurrences of the English *it* in a text corpus is presented in [1]. In this approach some of the rules implemented in rule-based systems are generalised via word patterns which are added to the system as features. The system also uses external knowledge sources in the form of two word lists containing weather verbs and idioms. The system achieved the best results using 25 features (precision was 82% and recall 72% on the given corpus).

The classification of referential and non-referential uses of the Dutch pronoun *het* (it) in two text corpora is described in [12]. The classification comprises the following uses of this pronoun: individual and abstract anaphoric, non-referential, anticipatory subject and anticipatory object. The reported results of the classification give an improvement of approx. 30% for all distinctions with respect to the baseline (the most frequent class). In [12] the authors also measure the effects of the classification on a machine learning based coreference resolution system.

Our research is inspired by most of these approaches, especially the work described in [7,20,1,12]. The novelty of our approach, apart from the language which we investigate, consists in the following:

- we use both texts and spoken data of various types;
- we deal with personal and demonstrative pronouns as well as weak and strong pronouns in spoken data (prosodic information about stress is included);
- we rely on a very fine-grained classification of the functions of Danish sn-pronouns which covers all occurrences of these pronouns in both texts and spoken data.

In these experiments we only use n-grams of words and, on texts, very basic linguistic information. We start from the raw data (no annotation at all) and investigate to which extent machine learning algorithms (first unsupervised then supervised) can be useful to identify the function of sn-pronominal occurrences. In the supervised experiments we first consider n-grams of words and the classification of sn-pronouns in the data, then we test the learning algorithms adding to the words in the texts lemma and POS information. In this we follow the strategy proposed by [6] which consists in testing various machine learning algorithms and types of linguistic information to find the most appropriate datasets and algorithms to resolve NLP tasks.

3 The Data

In written Danish sn-pronouns comprise the pronoun *det* (it/this/that), which is ambiguous with respect to its pronominal type, and the demonstrative pronoun *dette* (this). In spoken language they comprise the unstressed personal pronoun *det* (it), the stressed demonstrative pronouns *d'et* (this/that), *d'et her* (this) and *d'et der* (that). The stressed demonstrative pronoun *d'ette* occurs very seldom in spoken language (there were only two occurrences of it in our data and they both referred to an individual entity).

3.1 The Corpora

The corpora we use have been collected and annotated by many research groups for different purposes. Thus they are very heterogeneous.

The written corpora comprise general language texts [14], legal texts and literary texts [16]. They consist of 86,832 running words. The spoken language corpora comprise transcriptions of monologues and two-party dialogues from the DanPASS corpus [10], which is a Danish version of the MAPTASK corpus, multi-party verbose dialogues from the LANCHART corpus [9] and interviews from Danish television (LANCHART+TV henceforth). The monologues consist of 23,957 running words; the DanPASS dialogues contain 33,971 words and the LANCHART+TV consists of 26,304 words.

3.2 The Annotation

All texts contain automatically acquired POS-tag and lemma information. Most of the spoken corpora are also POS-tagged, but with different tagsets. The texts contain structural information such as chapters, sections and paragraphs, while the transcriptions of spoken language contain information about speakers' turns and timestamps with respect to the audio files[2]. All sn-pronouns in the spoken data are marked with stress information. The DanPASS data also contain rich prosodic information.

In all corpora sn-pronouns and their functions are marked. (Co)reference chains of the anaphoric sn-pronouns are also annotated together with other linguistically relevant information, such as the syntactic type of the antecedent, the semantic type of the referent and the referential relation type, see [22].

The corpora are available in the XML-format produced by the PALinkA annotation tool [24]. The classification of the function of sn-pronouns provided in the data is very fine-grained. It comprises the following classes:

- expletive (all non-referential uses);
- cataphoric (the pronoun precedes the linguistic expression necessary to its interpretation);
- deictic (the pronoun refers to something in the physical word);

[2] All the transcriptions were provided in the PRAAT TextGrid format (http://www.praat.org).

- individual anaphoric;
- individual vague anaphoric (the individual antecedents are implicit in discourse);
- abstract anaphoric;
- abstract vague anaphoric (the abstract antecedents are implicit in discourse);
- textual deictic (the anaphors refer to, but are not coreferential with, preceding linguistic expressions [15]);
- abandoned (the pronouns occur in unfinished and abandoned utterances[3]).

80% of the corpora were annotated independently by two expert annotators and then the two annotations were compared. The remaining 20% of the data were only coded by one annotator and revised by the other. In case of disagreement the two annotators decided together which annotation to adopt. In difficult cases a third linguist was consulted to choose an annotation. The annotators could listen to the audio files when coding the spoken data.

Inter-coder agreement was measured in terms of *kappa* scores [5,4] on the first subset of the annotated data (most of the text corpora and the DanPASS dialogues).

Table 1 shows the *kappa*-scores for the most frequent pronominal functions as they are reported in [22].

Table 1. Inter-coder agreement as *kappa* scores

Function	Text corpora	DanPASS dialogues
expletive	0.83	0.77
cataphor	0.73	0.72
individual	0.90	0.88
individual vague	0.92	0.92
abstract	0.89	0.84
abstract vague	0.8	0.84
textual deictic	0.91	0.89

4 The Experiments

The learning experiments have been run in the WEKA system [26] which permits testing and comparing a variety of algorithms. It also provides an interface with which to explore the data and the learning results. We ran the experiments on four datasets automatically extracted from the annotated corpora and translated into the *arff*-format required by WEKA. The four datasets we distinguish in our experiments are the following:

1. the texts
2. the DanPASS monologues
3. the DanPASS dialogues
4. the LANCHART+TV dialogues.

[3] These are also called disfluencies in the literature.

Table 2. Sn-pronouns and their functions in the data

Pronoun	Expl	IndAna	AbsAna	VagIA	VagAA	Catap	Deict	TDeic	Aband	Total
Texts										
det	345	152	130	8	10	58	1	4	0	708
dette	0	23	71	0	4	0	0	0	0	98
all	345	175	201	8	14	58	1	4	0	816
DanPASS Monologues										
unstressed	22	107	27	14	1	14	0	0	25	210
stressed	1	74	10	8	13	11	1	0	12	130
all	23	181	37	22	14	25	1	0	37	340
DanPASS Dialogues										
unstressed	34	177	100	25	5	17	0	4	72	434
stressed	10	121	111	22	7	22	7	3	31	334
all	44	298	211	47	12	39	7	7	103	768
LANCHART+TV										
unstressed	124	301	199	56	16	128	8	5	138	975
stressed	0	69	93	10	7	32	1	2	46	260
all	124	370	292	66	23	160	9	7	184	1235

The sn-pronouns and their functions in the four datasets are given in Table 2. The following abbreviations are used in the table: *Expl* for expletive, *IndAna* for individual anaphor, *AbsAna* for abstract anaphor , *VagIA* for vague individual anaphor, *VagAA* for vague abstract anaphor, *Catap* for cataphor, *Deict* for deictic, *TDeic* for textual-deictic, *Aband* for abandoned.

4.1 Clustering Experiments

Clustering was run on the raw data, but the pronominal function information in the annotated data was used to evaluate the obtained clusters. The best results in terms of the highest number of recognised clusters and "correctness"[4] were achieved by the WEKA EM (Expectation Maximisation). Clustering was tested on n-grams of varying size. The best results on the text data were achieved with a window of one word preceding and two words following the sn-pronouns. Five clusters were returned and they were bound to individual anaphor, expletive, cataphor, abstract anaphor and no-class. Correctness was 37.5 %. The best results on the DanPASS monologues were obtained using a window of 2 words preceding and 3 words following the sn-pronouns. Five clusters were recognised which were bound to the functions individual anaphor, abandoned, vague abstract anaphor, expletive and abstract anaphor. Correctness was 41.5%. On the

[4] Correctness is calculated by WEKA in the test phase by assigning to each cluster the pronominal function which in the evaluation data is attributed to the largest number of items in that cluster. The function assignment is optimised with respect to the recognised clusters. A *no-class* tag is assigned to clusters whose items have functions which have already been assigned to other clusters. Finally, correctness is calculated for the clusters which have been assigned a function.

DanPASS dialogues the best results were obtained with a window of 2 words preceding and following the sn-pronouns. The pronouns from the DanPASS dialogue data were grouped into 4 clusters (abandoned, individual anaphor, vague abstract and cataphor) and correctness was 43.5%. On the LANCHART+TV data the best results were achieved with a window of two words preceding and four words following the sn-pronouns. The algorithm returned 3 clusters connected to the functions individual anaphor, abstract anaphor and expletive. Correctness was 29.5 %.

The fact that clustering gives the best results on the text data confirms that it is harder to process transcriptions of spoken data than written data because other information available in spoken language is not included in the transcriptions.

From the experiments we can conclude that unsupervised learning on datasets of the size we are working with does not provide satisfactory results for the task of recognising such fine-grained functions of sn-pronouns (too few clusters were identified and correctness was too low).

4.2 Classification on Words

In the classification experiments we trained several classifiers on data extracted from the corpora. The pronominal function annotated in the corpora was used both for training and testing the classifiers. We started running various classifiers on n-grams as in the clustering experiments, then we run them on the data enriched with various types of information. The latter experiments have only been run on text data. In all cases the results were tested using 10-fold cross-validation. As baseline in our evaluation we used the results provided by the WEKA ZeroR class that predicts the most frequent attribute value for a nominal class (accuracy is the frequency of the most used category). The Weka algorithms which we have tested are: Naive Bayes, SMO, IBK, LBR, KStar, NBTree, LADTree and Rotation Forest. The algorithms were tested on windows of various sizes (going from the largest one: 3 words before and 5 words after the sn-pronouns to the smallest one: 1 word before and 2 words after the sn-pronouns).

For texts the best results were achieved by the WEKA NBTree class (it generates a decision tree with Naive Bayes classifiers at the leaves) and the dataset comprised three words before and five words after the sn-pronouns. For monologues the best results were obtained by the SMO class (Sequential Minimal Optimization) run on a window of one word before and three words after the sn-pronouns. For all dialogues the best results were achieved using a window of 2 words preceding and 3 words following the sn-pronouns. On the DanPASS dialogue data the algorithm that gave the best results was the WEKA SMO class, while for the LANCHART+TV data the best results were obtained by the KStar[5] class. The results of the classification algorithms in terms of Precision, Recall and F-measure are in Table 3. The table shows the baseline and the three best results obtained for each dataset by various algorithms.

[5] KStar is an instance-based classifier which uses an entropy-based distance function.

Table 3. Classification results: words and pronominal function

Algorithm	Precision	Recall	F-measure
Texts			
Baseline	18.3	42.8	25.7
NBTree	62.3	65.4	62.4
NaiveBayes	61.1	64.4	61.4
RotationForest	60.7	63.5	60.4
Monologues			
Baseline	28.3	53.2	37
SMO	64.3	66.8	64.7
KStar	63.2	66.5	61.3
IBK	59.6	63.5	60.9
DDialogues			
Baseline	15.1	38.8	21.7
SMO	54.5	57.2	55.4
NaiveBayes	52.9	56.6	53.2
RotationForest	49.9	53.4	50
LDialogues			
Baseline	9	30	13.8
KStar	33.4	35.4	32.9
NBTree	32.9	36.6	32.8
SMO	32.3	33.6	32.7

Figures 1, 2, 3 and 4 show the confusion matrices produced by the algorithms that performed best on each of the four datasets.

From the confusion matrices it is evident that the performance of classification is bound to the frequency of the various types of item in the data: occurrences of pronouns with frequently used functions are better classified than occurrences of pronouns with seldomly occurring functions such as textual deictic, deictic and, in some datasets, vague anaphor. Thus the confusion matrices reflect the differences in the distribution of the pronominal functions in the various datasets.

From the confusion matrices it can also be seen that cataphors and individual and abstract anaphors are often confused with expletives. Distinguishing between cataphors and expletives was also problematic for the annotators especially in texts, but they did not have any problem in distinguishing expletives from anaphoric uses of the personal pronouns. Classification also confused a number of individual and abstract anaphora in the texts. This was in few cases also a problematic issue for humans because of the ambiguity of the data. Vague anaphors were often not recognised as such, but this is understandable because they often occur in the same contexts as non-vague anaphors. Finally most classes were mixed up in the LANCHART+TV data.

In Table 4 the results obtained for each category by the best performing algorithms on the four datasets are given.

The results of all the experiments indicate that the classification algorithms give significantly better results than the baseline, although the results obtained on multi-party dialogues were much worse than those obtained on the other

```
  a   b   c   d   e   f   g   h   <-- classified as
316   4   9  16   0   0   0   0 |   a = explet
 35  11   8   4   0   0   0   0 |   b = cataphor
 48   1  78  46   0   2   0   0 |   c = indiv
 49   5  28 119   0   0   0   0 |   d = abstr-ana
  7   1   2   4   0   0   0   0 |   e = abstr-vague
  2   0   0   3   0   3   0   0 |   f = indiv-vague
  0   0   1   0   0   0   0   0 |   g = deictic
  0   0   1   3   0   0   0   0 |   h = textual-deictic
```

Fig. 1. Confusion matrix for texts

```
  a   b   c   d   e   f   g   h   <-- classified as
 13   0   0   0   0   0   0  10 |   a = explet
  0 173   4   2   0   2   0   0 |   b = indiv
  0  17   6   0   0   1   1   0 |   c = cataphor
  0  15   1   6   0   0   0   0 |   d = indiv-vague
  0   0   0   0   0   0   1   0 |   e = deictic
  0   7   2   2   0  26   0   0 |   f = abstr-ana
  0   4   2   1   0   0   7   0 |   g = abstr-vague
  6   0   0   0   0   0   0  31 |   h = abandoned
```

Fig. 2. Confusion matrix for monologues

```
  a   b   c   d   e   f   g   h   <-- classified as
  6  12   1   1   0   1   0   2 |   a = explet
  7 156   4   2   0   2   0  10 |   b = indiv
  1  15   6   0   0   1   1   1 |   c = cataphor
  1  13   1   5   0   0   0   2 |   d = indiv-vague
  0   0   0   0   0   0   1   0 |   e = deictic
  0   7   2   2   0  26   0   0 |   f = abstr-ana
  0   4   2   1   0   0   7   0 |   g = abstr-vague
  2  10   1   1   0   2   0  21 |   h = abandoned
```

Fig. 3. Confusion matrix for DanPASS dialogues

```
 a   b   c   d   e   f   g   h   i   <-- classified as
21   0  13  20   2  63   0   0   5 |   a = explet
 2   1   8   0   0  10   0   0   2 |   b = abstr-vague
 6   3 124  17   6 109   1   3  23 |   c = abstr-ana
13   0  32  23   0  76   1   0  15 |   d = cataphor
 3   0   6   1   4  47   0   0   5 |   e = indiv-vague
18   2  68  26   8 218   2   0  28 |   f = indiv
 0   0   0   2   0   3   2   0   2 |   g = deictic
 0   0   3   0   0   4   0   0   0 |   h = textual-deictic
 6   0  44   9   3  77   1   0  44 |   i = abandoned
```

Fig. 4. Confusion matrix for LANCHART+TV dialogues

Table 4. Classification results per category

Function	Precision	Recall	F-measure
NBTree on Texts			
expletive	69.1	91.6	78.8
cataphor	50	19	27.5
individual anaphor	61.4	44.6	51.7
abstract anaphor	61	59.2	60.1
vague abstract anaphor	0	0	0
vague individual anaphor	60	37.5	46.2
deictic	0	0	0
textual deictic	0	0	0
SMO on Monologues			
expletive	35.3	26.1	30
cataphor	35.3	24	28.6
individual anaphor	71.9	86.2	78.4
abstract anaphor	81.3	70.3	75.4
vague abstract anaphor	77.8	50	60.9
vague individual anaphor	41.7	22.7	29.4
deictic	0	0	0
abandoned	58.3	56.8	57.5
SMO on DanPASS dialogues			
expletive	42.1	36.4	39
cataphor	27.8	12.8	17.5
individual anaphor	58.1	73.2	64.8
abstract anaphor	68.6	68.2	68.4
vague abstract anaphor	0	0	0
vague individual anaphor	23.3	14.9	18.2
deictic	33.3	14.3	20
textual deictic	0	0	0
abandoned	56	49.5	52.6
KStar on lanchart dialogues			
expletive	30.4	16.9	21.8
cataphor	23.5	14.4	17.8
individual anaphor	35.9	58.9	44.6
abstract anaphor	41.6	42.5	42
vague abstract anaphor	16.7	4.3	6.9
vague individual anaphor	17.4	6.1	9
deictic	28.6	22.2	25
textual deictic	0	0	0
abandoned	35.5	23.9	28.6

data. The results with respect to the baseline for the texts, the monologues and the DanPASS dialogues show an improvement of 36.4%, 30.7% and 33.7%, respectively, with respect to the baseline, while the improvement for the lanchart+tv dialogues is only 19.1%.

Although these results cannot be directly compared with the results reported for the classification of the functions of the Dutch *het* in [12], the magnitude of

the improvement with respect to the baseline in the two experiments is similar, except for the results obtained on the LANCHART+TV dialogues which are not as good as the other results. Considering the fact that we look at more categories and more types of data than it was the case in the Dutch experiments, the results we have obtained are positive.

The reasons for the bad results obtained on the LANCHART+TV dialogues compared with the results obtained for the DANPASS data are many. The most important are, in our opinion, the following. Firstly these dialogues are free-conversational and include four discourse participants, while the DANPASS dialogues are two-party MAPTASK dialogues which are much more homogeneous. Secondly the quality of the transcription of the DANPASS dialogues is much higher than that of the transcription of the LANCHART dialogues. In the latter transcriptions there were a number of errors which we did not correct, and the timestamps in the speakers' tracks were not always precisely marked. Because we used these timestamps to automatically determine the order in which simultaneous speech had to be represented in the format required by PALINKA, there are probably a number of errors in the data. Finally, the distribution of the pronominal function types in the LANCHART dialogues is different from that in the other datasets, and the automatic treatment of multi-party dialogues should include information of various type such as the physical objects in the space where the conversation take place, including the discourse participants and adjacency pairs. This type of information was not available for the LANCHART corpus.

The F-measure for the recognition of expletives on the basis of the annotation of the pronominal function is 78.8% in the texts, 30% in the DANPASS monologues, 39% in the DANPASS dialogues and, finally, 32.9% in the LANCHART+TV. Only the measures obtained for the texts are satisfactorily and near to those obtained in [1] where a lot of features and two word lists were used for identifying non-referential from referential uses of *it*.

In the light of the obtained classification results, we are now revising some of the annotations of the function of pronouns. This is especially the case for the cataphoric function.

4.3 Classification of Pronouns in Texts Enriched with POS and Lemma Information

In these experiments we run classification on the texts adding to the words lemma and POS information. A window of one word preceding and three words following the sn-pronouns was used in order to reduce the size of the data.

The best results obtained by various classifiers on n-grams of words, of words+lemma, of words+POS and of words+lemma+POS are in Table 5.

These results indicate that adding lemma and POS information increases the performance of classification, but these improvements are not significant[6].

[6] In the experiments significance was calculated as corrected resampled t-test via the WEKA experimenter [26].

Table 5. Classification results: words and linguistic features

Data	Algorithm	Precision	Recall	F-measure
All	Baseline	18.3	42.8	25.7
word	Rotation Forest	60.7	63.3	60.5
word+lemma	NBTree	61.4	63.9	62
word+POS	RotationF	62.4	64	61.5
word+lemma+POS	SMO	61.3	64.3	62.1

The precision of the POS tagger (the Brill tagger [2] trained on the Danish Parole corpus [11]) used to tag the textual data is approx. 97%. The precision of the CST lemmatiser [13] which was used on the texts is also approx. 97%.

Using manually corrected annotation may improve the classification results.

5 Conclusions and Future Work

In the paper we have described unsupervised and supervised machine learning experiments with the purpose of recognising the function of Danish sn-pronouns in texts and spoken data of various type.

The results of our clustering experiments indicate that unsupervised learning on datasets of the size we are working with does not provide satisfactory results for the task of recognising so fine-grained functions of sn-pronouns as those provided in the annotation because too few clusters are identified and correctness is too low.

The results of classification using simple n-grams and the annotation of the function of sn-pronouns gave an improvement with respect to the baseline of 36.4% on text data, 37.9% on the DanPASS monologues and 43.1% on the DanPASS dialogues and 19.1% on the lanchart+tv dialogues. Our results for the first three datasets are better than those reported for a Dutch sn-pronoun by [12]. These results indicate that classifiers can be useful to tag the function of pronouns in texts, monologues and some types of dialogues, although the data cannot be used without manual correction.

We also run the classification experiments on the text data adding lemma and POS information to the n-grams . The added linguistic information improved the performance of the classifiers on the data, but the improvement is not significant.

An analysis of the human classification of the function of pronouns in the light of the results of classification indicates that the definition of the cataphoric function is problematic, and that vague anaphors are in many cases difficult to identify automatically. We are now revising some of the annotations in the light of the classification results.

In future we will include in the data the syntactic information extracted from a large computational lexicon which contains some of the information which is useful to identify expletive, abstract and individual anaphoric uses of pronouns and test whether classification improves on our datasets enriched with this type of information.

References

1. Boyd, A., Gegg-Harrison, W., Byron, D.: Identifying non-referential it: a machine learning approach incorporating linguistically motivated patterns. In: Proceedings of the ACL Workshop on Feature Engineering for Machine Learning in NLP, Ann Arbor Michigan, June 2005, pp. 40–47 (2005)
2. Brill, E.: Transformation-Based Error-Driven Learning and Natural Language Processing. A Case Study in Part of Speech Tagging. Computational Linguistics 21(4), 543–565 (1995)
3. Byron, D.K.: Resolving pronominal reference to abstract entities. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 80–87 (2002)
4. Carletta, J.: Assessing agreement on classification tasks: the kappa statistics. Computational Linguistics 22(2), 249–254 (1996)
5. Cohen, J.: Weighted kappa; nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bullettin 70, 213–220 (1968)
6. Daelemans, W., Hoste, V., De Meulder, F., Naudts, B.: Combined optimization of feature selection and algorithm parameters in machine learning of language. In: Lavrač, N., Gamberger, D., Todorovski, L., Blockeel, H. (eds.) ECML 2003. LNCS (LNAI), vol. 2837, pp. 84–95. Springer, Heidelberg (2003)
7. Eckert, M., Strube, M.: Dialogue acts, synchronising units and anaphora resolution. Journal of Semantics 17(1), 51–89 (2001)
8. Evans, R.: A comparison of Rule-Based and Machine Learning Methods for Identifying Non-nominal It. In: Christodoulakis, D.N. (ed.) NLP 2000. LNCS (LNAI), vol. 1835, pp. 233–240. Springer, Heidelberg (2000)
9. Gregersen, F.: The LANCHART Corpus of Spoken Danish. Report from a corpus in progress. In: Current Trends in Research on Spoken Language in the Nordic Countries, pp. 130–143. Oulu University Press (2007)
10. Grønnum, N.: DanPASS - A Danish Phonetically Annotated Spontaneous Speech Corpus. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy (May 2006)
11. Hansen, D.H.: Træning og brug af Brill-taggeren på danske tekster. Ontoquery technical report, Center for Sprogteknologi, Copenhagen (2000)
12. Hoste, V., Hendrickx, I., Daelemans, W.: Disambiguation of the Neuter Pronoun and Its Effect on Pronominal Coreference Resolution. In: Matoušek, V., Mautner, P. (eds.) TSD 2007. LNCS (LNAI), vol. 4629, pp. 48–55. Springer, Heidelberg (2007)
13. Jongejan, B., Hansen, D.H.: The CST Lemmatiser Technical report, Centre for Language Technology (2001)
14. Keson, B., Norling-Christensen, O.: PAROLE-DK. Technical report, Det Danske Sprog- og Litteraturselskab (1998), http://korpus.dsl.dk/e-resurser/parole-korpus.php
15. Lyons, J.: Semantics, vol. I-II. Cambridge University Press, Cambridge (1977)
16. Maegaard, B., Offersgaard, L., Henriksen, L., Jansen, H., Lepetit, X., Navarretta, C., Povlsen, C.: The MULINCO corpus and corpus platform. In: Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, J., Tapias, D. (eds.) Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genova, Italy, May 2006, pp. 2148–2153 (2006)

17. Mitkov, R., Hallett, C.: Comparing Pronoun Resolution Algorithms. Computational Intelligence 23(2), 262–297 (2007)
18. Mitkov, R., Evans, R., Orasan, C.: A New, Fully Automatic Version of Mitkov's Knowledge-Poor Pronoun Resolution Method. In: Gelbukh, A. (ed.) CICLing 2002. LNCS, vol. 2276, pp. 168–186. Springer, Heidelberg (2002)
19. Müller, C.: Resolving it, this and that in unrestricted multi-party dialog. In: Proceedings of ACL 2007, pp. 816–823. Prague (2007)
20. Navarretta, C.: The use and resolution of Intersentential Pronominal Anaphora in Danish Discourse. Ph.D. thesis. University of Copenhagen (February 2002)
21. Navarretta, C.: Resolving individual and abstract anaphora in texts and dialogues. In: Proceedings of the 20th International Conference of Computational Linguistics, COLING 2004, Geneva, Switzerland, pp. 233–239 (2004)
22. Navarretta, C., Olsen, S.: Annotating abstract pronominal anaphora in the DAD project. In: Proceedings of LREC 2008, Marrakesh, Morocco (May 2008)
23. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, August 2002, pp. 730–736 (2002)
24. Orasan, C.: PALinkA: a highly customizable tool for discourse annotation. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialog, Sapporo, pp. 39–43 (2003)
25. Strube, M., Müller, C.: A machine learning approach to pronoun resolution in spoken dialogue. In: Proceedings of the ACL 2003, pp. 168–175 (2003)
26. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco (2005)

A Deeper Look into Features for Coreference Resolution

Marta Recasens[1] and Eduard Hovy[2]

[1] CLiC - University of Barcelona,
Gran Via 585, Barcelona, Spain
[2] Information Sciences Institute,
4676 Admiralty Way, Marina del Rey CA, USA
mrecasens@ub.edu, hovy@isi.edu

Abstract. All automated coreference resolution systems consider a number of features, such as head noun, NP type, gender, or number. Although the particular features used is one of the key factors for determining performance, they have not received much attention, especially for languages other than English. This paper delves into a considerable number of pairwise comparison features for coreference, including old and novel features, with a special focus on the Spanish language. We consider the contribution of each of the features as well as the interaction between them. In addition, given the problem of class imbalance in coreference resolution, we analyze the effect of sample selection. From the experiments with TiMBL (Tilburg Memory-Based Learner) on the AnCora corpus, interesting conclusions are drawn from both linguistic and computational perspectives.

Keywords: Coreference resolution, Machine learning, Features.

1 Introduction

Coreference resolution, the task of identifying which mentions in a text point to the same discourse entity, has been shown to be beneficial in many NLP applications such as Information Extraction [6], Text Summarization [13], Question Answering [8], and Machine Translation. These systems need to identify the different pieces of information concerning the same referent, produce coherent and fluent summaries, disambiguate the references to an entity, and solve anaphoric pronouns.

Given that many different types of information – ranging from morphology to pragmatics – play a role in coreference resolution, machine learning approaches [12,9] seem to be a promising way to combine and weigh the relevant factors, overcoming the limitations of constraint-based approaches [4,7], which might fail to capture global patterns of coreference relations as they occur in real data. Learning-based approaches decompose the task of coreference resolution into two steps: (i) classification, in which a classifier is trained on a corpus to learn the probability that a pair of NPs are coreferent or not; and (ii) clustering, in which the pairwise links identified at the first stage are merged to form distinct coreference chains.

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 29–42, 2009.

This paper focuses on the classification stage and, in particular, on (i) the features that are used to build the feature vector that represents a pair of mentions,[1] and (ii) the selection of positive and negative training instances. The choice of the information encoded in the feature vectors is of utmost importance as they are the basis on which the machine learning algorithm learns the pairwise coreference model. Likewise, given the highly skewed distribution of coreferent vs. non-coreferent classes, we will consider whether sample selection is helpful. The more accurate the classification is, the more accurate the clustering will be.

The goal of this paper is to provide an in-depth study of the pairwise comparison stage in order to decrease as much as possible the number of errors that are passed on to the second stage of coreference resolution. Although there have been some studies in this respect [14,1,3], they are few, oriented to the English or Dutch language, and dependent on poorly annotated corpora. To our knowledge, no previous studies compared systematically a large number of features relying on gold standard corpora, and experiments with sample selection have been only based on small corpora. For the first time, we consider the degree of variance of the learnt model on new data sets by reporting confidence intervals for precision, recall, and F-score measures.

The paper is organized as follows. In the next section, we review previous work. In Section 3, we list our set of 47 features and argue the linguistic motivations behind them. These features are tested by carrying out different machine learning experiments with TiMBL in Section 4, where the effect of sample selection is also assessed. Finally, main conclusions are drawn in Section 5.

2 Previous Work

Be it in the form of hand-crafted heuristics or feature vectors, what kind of knowledge is represented is a key factor for the success of coreference resolution. Although theoretical studies point out numerous linguistic factors relevant for the task, computational systems usually rely on a small number of shallow features, especially after the burst of statistical approaches. In learning-based approaches, the relative importance of the factors is not manually coded but inferred automatically from an annotated corpus. Training instances for machine learning systems are feature vectors representing two mentions (m_1 and m_2) and a label ('coreferent' or 'non-coreferent') allowing the classifier to learn to predict, given a new pair of NPs, whether they do or do not corefer.

The feature set representing m_1 and m_2 that was employed in the decision tree learning algorithm of [12] has been taken as a starting point by most subsequent systems. It consists of only 12 surface-level features (all boolean except for the first): (i) sentence distance, (ii) m_1 is a pronoun, (iii) m_2 is a pronoun, (iv) string match (after discarding determiners), (v) m_2 is a definite NP, (vi) m_2 is a demonstrative NP, (vii) number agreement, (viii) WordNet semantic class

[1] This paper restricts to computing features over a pair of mentions – without considering a more global approach – hence *pairwise comparison features*.

agreement,[2] (ix) gender agreement, (x) both m_1 and m_2 are proper nouns (capitalized), (xi) m_1 is an alias of m_2 or vice versa, and (xii) m_1 is an apposition to m_2. The strongest indicators of coreference turned out to be string match, alias and appositive.

Ng and Cardie [9] expanded the feature set of [12] from 12 to a deeper set of 53, including a broader range of lexical, grammatical, and semantic features such as substring match, comparison of the prenominal modifiers of both mentions, animacy match, WordNet distance, whether one or both mentions are pronouns, definite, embedded, part of a quoted string, subject function, and so on. The incorporation of additional knowledge succeeds at improving performance but only after manual feature selection, which points out the importance of removing irrelevant features that might be misleading. Surprisingly, however, some of the features in the hand-selected feature set do not seem very relevant from a linguistic point of view, like string match for pronominal mentions.

More recent attempts have explored some additional features to further enrich the set of [9]: backward features describing the antecedent of the candidate antecedent [16], semantic information from Wikipedia, WordNet and semantic roles [10], and most notably, Uryupina's [14] thesis, which investigates the possibility of incorporating sophisticated linguistic knowledge into a data-driven coreference resolution system trained on the MUC-7 corpus. Her extension of the feature set up to a total of 351 nominal features (1096 boolean/continuous) leads to a consistent improvement in the system's performance, thus supporting the hypothesis that complex linguistic factors of NPs are a valuable source of information. At the same time, however, [14] recognizes that by focusing on the addition of sophisticated features she overlooked the resolution strategy and some phenomena might be over-represented in her feature set.

Bengtson and Roth [1] show that with a high-quality set of features, a simple pairwise model can outperform systems built with complex models on the ACE dataset. This clearly supports our stress on paying close attention to designing a strong, linguistically motivated set of features, which requires a detailed analysis of each feature individually as well as of the interaction between them. Some of the features we include, like modifiers match, are also tested by [1] and, interestingly, our ablation study comes to the same conclusion: almost all the features help, although some more than others.

Hoste's [3] work is concerned with optimization issues such as feature and sample selection, and she stresses their effect on classifier performance. The study we present is in line with [14,1,3] but introduces a number of novelties. First, the object language is Spanish, which presents some differences as far as coreference is concerned. Second, we use a different corpus, AnCora, which is twenty times as large as MUC and, unlike ACE, it includes a non-restricted set of entity types. Third, the coreference annotation of the AnCora corpus sticks to a linguistic definition of the identity relationship more accurate than that behind the MUC or ACE guidelines. Fourth, we do not rely on the (far from perfect) output of

[2] Possible semantic classes for an NP are *female, male, person, organization, location, date, time, money, percent*, and *object*.

preprocessing modules but take advantage of the gold standard annotations in the AnCora corpus in order to focus on their real effect on coreference resolution.

3 Pairwise Comparison Features

The success of machine learning systems depends largely on the feature set employed. Learning algorithms need to be provided with an adequate representation of the data, that is to say, a representation that includes the "relevant" information, to infer the best model from an annotated corpus. Identifying the constraints on when two NPs can corefer is a complex linguistic problem that remains still open. Hence, there is a necessity for an in-depth study of features for coreference resolution from both a computational and a linguistic perspective. This section makes a contribution in this respect by considering a total of 47 features, making explicit the rationale behind them.

– **Classical features** (Table 1)**.** The features that have been shown to obtain better results in previous works [12,9,5] capture the most basic information on which coreference depends, but form a reduced feature set that does not account for all kinds of coreference relations.
  - PRON_m_1 and PRON_m_2 specify whether the mentions are pronouns since these show different patterns of coreference, e.g., gender agreement is of utmost importance for pronouns but might be violated by non-pronouns [3].
  - HEAD_MATCH is the top classical feature for coreference, since lexical repetition is a common coreference device.
  - WORDNET_MATCH uses the Spanish EuroWordNet[3] and is true if any of the synset's synonyms of one mention matches any of the synset's synonyms of the other mention.
  - NP type plays an important role because not all NP types have the same capability to introduce an entity into the text for the first time, and not all NP types have the same capability to refer to a previous mention in the text.
  - The fact that in newspaper texts there is usually at least one person and a location about which something is said accounts for the relevance of the NE type feature, since NE types like *person* and *organization* are more likely to corefer and be coreferred than others.
  - SUPERTYPE_MATCH compares the first hypernym of each mention found in EuroWordNet.
  - As a consequence of the key role played by gender and number in anaphora resolution, GENDER_AGR and NUMBER_AGR have been inherited by coreference systems. See below, however, for finer distinctions.

[3] Nominal synsets are part of the semantic annotation of AnCora. EuroWordNet covers 55% of the nouns in the corpus.

- The rationale behind QUOTES is that a mention in quotes identifies a mention that is part of direct speech, e.g., if it is a first- or second-person pronoun, its antecedent will be found in the immediate discourse.

Table 1. Classical features

Feature	Definition	Value
PRON_m_1	m_1 is a pronoun	true, false
PRON_m_2	m_2 is a pronoun	true, false
HEAD_MATCH	Head match	true, false, ?[a]
WORDNET_MATCH	EuroWordNet match	true, false, ?[a]
NP_m_1	m_1 NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NP_m_2	m_2 NP type	common, proper, article, indefinite, possessive, relative, demonstrative, numeral, interrogative, personal, exclamative
NE_m_1	m_1 NE type	person, organization, location, date, number, other, null
NE_m_2	m_2 NE type	person, organization, location, date, number, other, null
NE_MATCH	NE match	true, false, ?[b]
SUPERTYPE_MATCH	Supertype match	true, false, ?[a]
GENDER_AGR	Gender agreement	true, false
NUMBER_AGR	Number agreement	true, false
ACRONYM	m_2 is an acronym of m_1	true, false, ?[c]
QUOTES	m_2 is in quotes	true, false
FUNCTION_m_1	m_1 function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
FUNCTION_m_2	m_2 function	subject, d-obj, i-obj, adjunct, prep-obj, attribute, pred-comp, agent, sent-adjunct, no function
COUNT_m_1	m_1 count	#times m_1 appears in the text
COUNT_m_2	m_2 count	#times m_2 appears in the text
SENT_DIST	Sentence distance	#sentences between m_1 and m_2
MENTION_DIST	Mention distance	#NPs between m_1 and m_2
WORD_DIST	Mention distance	#words between m_1 and m_2

[a] Not applicable. This feature is only applicable if neither m_1 nor m_2 are pronominal or conjoined.

[b] Not applicable. This feature is only applicable if both mentions are NEs.

[c] Not applicable. This feature is only applicable if m_2 is an acronym.

– **Language-specific features** (Table 2)**.** There are some language-specific issues that have a direct effect on the way coreference relations occur in a language. In the case of Spanish, we need to take into account elliptical subjects, grammatical gender, and nouns used attributively.

  - There is a need to identify elliptical pronouns in Spanish because, unlike overt pronouns, they get their number from the verb, have no gender, and always appear in subject position, as shown in (1), where the elliptical subject pronoun is marked with ∅ and with the corresponding pronoun in brackets in the English translation.

    (1) Klebánov manifestó que ∅ no puede garantizar el éxito al cien por cien.
    'Klebánov stated that *(he)* cannot guarantee 100% success.'

Table 2. Language-specific features

Feature	Definition	Value
ELLIP_m_1	m_1 is an elliptical pronoun	true, false
ELLIP_m_2	m_2 is an elliptical pronoun	true, false
GENDER_PRON	Gender agreement restricted to pronouns	true, false, ?
GENDER_MASCFEM	Gender agreement restricted to masc./fem.	true, false, ?
GENDER_PERSON	Gender agreement restricted to persons	true, false, ?
ATTRIBa_m_1	m_1 is attributive type A	true, false
ATTRIBa_m_2	m_2 is attributive type A	true, false
ATTRIBb_m_1	m_1 is attributive type B	true, false
ATTRIBb_m_2	m_2 is attributive type B	true, false

Table 3. Corpus-specific features

Feature	Definition	Value
NOMPRED_m_1	m_1 is a nominal predicate	true, false
NOMPRED_m_2	m_2 is a nominal predicate	true, false
APPOS_m_1	m_1 is an apposition	true, false
APPOS_m_2	m_2 is an apposition	true, false
PRONTYPE_m_1	m_1 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
PRONTYPE_m_2	m_2 pronoun type	elliptical, 3-person, non-3-person, demonstrative, possessive, indefinite, numeric, other, ?
EMBEDDED	m_2 is embedded in m_1	true, false
MODIF_m_1	m_1 has modifiers	true, false
MODIF_m_2	m_2 has modifiers	true, false

Table 4. Novel features

Feature	Definition	Value
FUNCTION_TRANS	Function transition	100 different values (e.g., subject_subject, subject_d-obj)
COUNTER_MATCH	Counter match	true, false, ?
MODIF_MATCH	Modifiers match	true, false, ?
VERB_MATCH	Verb match	true, false, ?
NUMBER_PRON	Number agreement restricted to pronouns	true, false, ?
TREE-DEPTH_m_1	m_1 parse tree depth	#nodes in the parse tree from m_1 up to the top
TREE-DEPTH_m_2	m_2 parse tree depth	#nodes in the parse tree from m_2 up to the top
DOC_LENGTH	Document length	#tokens in the document

- Since Spanish has grammatical gender, two non-pronominal nouns with different gender might still corefer, e.g., *el incremento* 'the increase' (masc.) and *la subida* 'the rise' (fem.). Gender agreement is an appropriate constraint only for pronouns.
- GENDER_MASCFEM does not consider those NPs that are not marked for gender (e.g. elliptical pronouns, companies).
- GENDER_PERSON separates natural from grammatical gender by only comparing the gender if one of the mentions is an NE-person.[4]
- Attributive NPs[5] are non-referential, hence non-markables. ATTRIBa and ATTRIBb identify two Spanish constructions where these NPs usually occur:

  Type A. Common, singular NPs following the preposition *de* 'of', e.g., *educación* 'education' in *sistema de educación* 'education system.'

  Type B. Proper nouns immediately following a generic name, e.g., *Mayor* 'Main' in *calle Mayor* 'Main Street'.

– **Corpus-specific features** (Table 3)**.** The definition of coreference in the AnCora corpus differs from that of the MUC and ACE corpora in that it separates identity from other kinds of relation such as apposition, predication, or bound anaphora. This is in line with van Deemter and Kibble's [15] criticism of MUC. Predicative and attributive NPs do not have a referential function but an attributive one, qualifying an already introduced entity. They should not be allowed to corefer with other NPs. Consequently, the use we make of nominal-predicate and appositive features is the opposite to that made by systems trained on the MUC or ACE corpora [12,5]. Besides, the fact that AnCora contains gold standard annotation from the morphological

[4] Animals are not included since they are not explicitly identitifed as NEs.
[5] *Attributively* used NPs qualify another noun.

to the semantic levels makes it possible to include additional features that rely on such rich information.

- We employ NOMPRED to filter out predicative mentions.
- We employ APPOS to filter out attributively used mentions.
- Gold standard syntactic annotation makes it possible to assess the efficacy of the EMBEDDED and MODIF features in isolation from any other source of error. First, a nested NP cannot corefer with the embedding one. Second, depending on the position a mention occupies in the coreference chain, it is more or less likely that it is modified.

– **Novel features** (Table 4)**.** We suggest some novel features that we believe relevant and that the rich annotation of AnCora enables.

- FUNCTION_TRANS is included because although FUNCTION_m_1 and FUNCTION_m_2 already encode the function of each mention separately, there may be information in their joint behaviour.[6] E.g., *subject_subject* can be relevant since two consecutive subjects are likely to corefer:

(2) [...] explicó *Alonso, quien anunció la voluntad de Telefónica Media de unirse a grandes productoras iberoamericanas*. Por otra parte, *Alonso* justificó el aplazamiento.
'[...] explained *Alonso, who announced the will of Telefónica Media to join large Latin American production companies*. On the other hand, *Alonso* justified the postponement.'

- COUNTER_MATCH prevents two mentions that contain a different numeral to corefer (e.g., *134 millones de euros* '134 million euros' and *194 millones de euros* '194 million euros'), as they point to a different number of referents.
- Modifiers introduce extra information that might imply a change in the referential scope of a mention (e.g., *las elecciones generales* 'the general elections' and *las elecciones autonómicas* 'the regional elections'). Thus, when both mentions are modified, the synonyms and immediate hypernym of the head of each modifying phrase are extracted from EuroWordNet for each mention. MODIF_MATCH is true if one of them matches between the two mentions.
- The verb, as the head of the sentence, imposes restrictions on its arguments. In (3), the verb *participate* selects for a volitional agent, and the fact that the two subjects complement the same verb hints at their coreference link. VERB_MATCH is true if either the two verbal lemmas or any synonym or immediate hypernym from EuroWordNet match.

(3) *Un centenar de artistas* participará en el acto [...] el acto se abrirá con un brindis en el que participarán *todos los protagonistas de la velada*.
'*One hundred artists* will participate in the ceremony [...] the ceremony will open with a toast in which *all the protagonists of the evening gathering* will participate.'

[6] The idea of including conjoined features is also exploited by [1,5].

Table 5. Characteristics of the AnCora-Es datasets

	Training set	Test set
# Words	298 974	23 022
# Entities	64 421	4 893
# Mentions	88 875	6 759
# NEs	25 758	2 023
# Nominals	53 158	4 006
# Pronominals	9 959	730

- NUMBER_PRON is included since non-pronominal mentions that disagree in number might still corefer.
- DOC_LENGTH can be helpful since the longer the document, the more coreferent mentions, and a wider range of patterns might be allowed.

4 Experimental Evaluation

This section describes our experiments with the features presented in Section 3 as well as with different compositions of the training and test data sets. We finally assess the reliability of the most appropriate pairwise comparison model.

Data: The experiments are based on the AnCora-Es corpus [11], a corpus of newspaper and newswire articles. It is the largest Spanish corpus annotated, among other levels of linguistic information, with PoS tags, syntactic constituents and functions, named entities, nominal WordNet synsets, and coreference links.[7] We split randomly the freely available labelled data into a training set of 300k words and a test set of 23k words. See Table 5 for a description.

Learning algorithm: We use TiMBL, the Tilburg memory-based learning classifier [2], which is a descendant of the k-nearest neighbor approach. It is based on analogical reasoning: the behavior of new instances is predicted by extrapolating from the similarity between (old) stored representations and the new instances. This makes TiMBL particularly appropriate for training a coreference resolution model, as the feature space tends to be very sparse and it is very hard to find universal rules that work all the time. In addition, TiMBL outputs the information gain of each feature – very useful for studies on feature selection – and allows the user easily to experiment with different feature sets by obscuring specified features. Given that the training stage is done without abstraction but by simply storing training instances in memory, it is considerably faster than other machine learning algorithms.

[7] AnCora is freely available from http://clic.ub.edu/ancora

Table 6. Distribution of representative and balanced data sets

	Training set		Test set	
	Representative	Balanced	Representative	Balanced
Positive instances	105 920		8 234	
Negative instances	425 942	123 335	32 369	9 399

We select parameters to optimize TiMBL on a held-out development set. The distance metric parameter is set to overlap, and the number of nearest neighbors (k parameter) is set to 5 in Section 4.1, and to 1 in Section 4.2.[8]

4.1 Sample Selection

When creating the training instances, we run into the problem of class imbalance: there are many more negative examples than positive ones. Positive training instances are created by pairing each coreferent NP with all preceding mentions in the same coreference chain. If we generate negative examples for all the preceding non-coreferent mentions, which would conform to the real distribution, then the number of positive instances is only about 7% [3]. In order to reduce the vast number of negative instances, previous approaches usually take only those mentions between two coreferent mentions, or they limit the number of previous sentences from which negative mentions are taken. Negative instances have so far been created only for those mentions that are coreferent. In a real task, however, the system must decide on the coreferentiality of all mentions.

In order to investigate the impact of keeping the highly skewed class distribution in the training set, we create two versions for each data set: a representative one, which approximates the natural class distribution, and a balanced one, which results from down-sampling negative examples. The total number of negatives is limited by taking only 5 non-coreferent mentions randomly selected among the previous mentions (back to the beginning of the document). The difference is that in the balanced sample, non-coreferent mentions are selected for each coreferent mention, whereas in the representative sample they are selected for all mentions in the document. See Table 6 for statistics of the training and test sets.

Combining each training data set with each test set gives four possible combinations (Table 7) and we compute the performance of each of the models. The output of the experiments is evaluated in terms of precision (P), recall (R) and F-score (F). Although the best performance is obtained when testing the model on the balanced sample (models B and D), making a balanced test set involves

[8] When training the model on the full feature vectors, the best results are obtained when TiMBL uses 5 nearest neighbors for extrapolation. However, because of the strong skew in the class space, in some of the hill-climbing experiments we can only use 1 nearest neighbor. Otherwise, with 5 neighbors the majority of neighbors are of the negative class for all the test cases, and the positive class is never predicted (recall=0).

Table 7. Effect of sample selection on performance

	Training set	Test set	P	R	F
Model A	Representative	Representative	84.73	73.44	78.68
Model B	Representative	Balanced	88.43	73.44	80.24
Model C	Balanced	Representative	66.28	80.24	72.60
Model D	Balanced	Balanced	83.46	87.32	85.34

knowledge about the different classes in the test set, which is not available in non-experimental situations. Therefore, being realistic, we must carry out the evaluation on a data set that follows the natural class distribution. We focus our attention on models A and C.

Down-sampling on the training set increases R but at the cost of a too dramatic decrease in P. Because of the smaller number of negative instances in the training, it is more likely for an instance to be classified as positive, which harms P and F. As observed by [3], we can conclude that down-sampling does not lead to an increase in TiMBL, and so we opt for using model A.

4.2 Feature Selection

This section considers the informativeness of the features presented in Section 3. We carry out two different feature selection experiments: (i) an ablation study, and (ii) a hill-climbing forward selection.

In the first experiment, we test each feature by running TiMBL on different subsets of the 47 features, each time removing a different one. The majority of features have low informativeness, as no single feature brings about a statistically significant loss in performance when omitted.[9] Even the removal of HEAD_MATCH, which is reported in the literature as one of the key features in coreference resolution, causes a statistically non-significant decrease of .15 in F. We conclude that some other features together learn what HEAD_MATCH learns on its own. Features that individually make no contribution are ones that filter referentiality, of the kind *ATTRIBb*_m_2, and ones characterising m_1, such as PRON_m_1. Finally, some features, in particular the distance and numeric measures, seem even to harm performance. However, there is a complex interaction between the different features. If we train a model that omits all features that seem irrelevant and harmful at the individual level, then performance on the test set decreases. This is in line with the ablation study performed by [1], who concludes that all features help, although some more than others.

Forward selection is a greedy approach that consists of incrementally adding new features – one at a time – and eliminating a feature whenever it causes a drop in performance. Features are chosen for inclusion according to their information gain values, as produced by TiMBL, most informative earliest. Table 8 shows the

[9] Statistical significance is tested with a one-way ANOVA followed by a Tukey's post-hoc test.

Table 8. Results of the forward selection procedure

Feature vector	P	R	F	Feature vector	P	R	F
HEAD_MATCH	92.94	17.43	29.35	COUNTER_ MATCH	81.76	63.64	71.57
PRON_ m_2	57.58†	61.14*	59.30	MODIF_ m_1	81.08	64.67	71.95
ELLIP_ m_2	65.22*	53.04†	58.50	PRONTYPE_ m_1	81.70	64.84	72.30
-ELLIP_m_1	89.74*	34.09†	49.41	GENDER_ AGR	81.60	65.12	72.44
WORDNET_ MATCH	65.22	53.04	58.50	NOMPRED_ m_1	81.89	65.04	72.50
NE_ MATCH	65.22	53.04	58.50	GENDER_ PERSON	87.95*	64.78	74.61
-PRON_m_1	86.73*	38.74†	53.56	FUNCTION_ m_2	87.06	65.96	75.06
NUMBER_ PRON	69.04*	58.20*	63.16	FUNCTION_ m_1	85.88†	69.82*	77.02
-GENDER_PRON	86.64*	37.39†	52.24	QUOTES	85.83	70.11	77.18
VERB_ MATCH	80.31*	55.53†	65.66	COUNT_ m_2	85.62	70.73	77.47
SUPERTYPE_ MATCH	80.22	55.56	65.65	COUNT_ m_1	84.57	71.35	77.40
MODIF_ m_2	78.18	61.68*	68.96	NE_ m_1	83.82	72.48	77.74
NUMBER_ AGR	79.94	61.81	69.71	ACRONYM	83.99	72.46	77.80
ATTRIBb_ m_2	80.08	61.85	69.80	NE_ m_2	83.48	73.14	77.97
ATTRIBa_ m_2	80.14	61.84	69.81	NP_ m_2	82.81	73.55	77.91
ATTRIBa_ m_1	80.22	61.83	69.84	NP_ m_1	82.27	74.05	77.94
ATTRIBb_ m_1	80.23	61.82	69.83	FUNCTION_ TRANS	82.29	73.94	77.89
EMBEDDED	80.33	61.78	69.84	TREE-DEPTH_ m_2	80.54	72.98	76.57
GENDER_ MASCFEM	81.33	62.96	70.98	-TREE-DEPTH_ m_1	78.25†	72.52	75.27
APPOS_ m_1	81.46	62.96	71.02	-SENT_ DIST	78.17†	72.16	75.05
APPOS_ m_2	81.44	62.95	71.01	-DOC_ LENGTH	79.36*	70.36†	74.79
MODIF_ MATCH	81.35	63.10	71.08	MENTION_ DIST	79.52	72.10	75.63
NOMPRED_ m_2	81.38	63.37	71.26	WORD_ DIST	79.14	71.73	75.25
PRONTYPE_ m_2	81.70	63.59	71.52				

results of the selection process. In the first row, the model is trained on a single (the most informative) feature. From there on, one additional feature is added in each row; initial "-" marks the harmful features that are discarded (provide a statistically significant decrease in either P or R, and F). P and R scores that represent statistically significant gains and drops with respect to the previous feature vector are marked with an asterisk (*) and a dagger (†), respectively. Although F-score keeps rising steadily in general terms, informative features with a statistically significant improvement in P are usually accompanied by a significant decrease in R, and vice versa.

The results show several interesting tendencies. Although HEAD_MATCH is the most relevant feature, it obtains a very low R, as it cannot handle coreference relationships involving pronouns or relations between full NPs that do not share the same head. Therefore, when PRON_ m_2 is added, R is highly boosted. With only these two features, P, R and F reach scores near the 60s. The rest of the features make a small – yet important in sum – contribution. Most of the features have a beneficial effect on performance, which provides evidence for the value of building a feature vector that includes linguistically motivated features. This includes some of the novel features we argue for, such as NUMBER_PRON and VERB_MATCH. Surprisingly, distance features seem to be harmful. However, if we train again the full model with the k parameter set to 5 and we leave out the numeric features, F does not increase but goes down. Again, the complex interaction between the features is manifested.

4.3 Model Reliability

In closing this section, we would like to stress an issue to which attention is hardly ever paid: the need for computing the reliability of a model's performance. Because of the intrinsic variability in any data set, the performance of a model trained on one training set but tested on another will never be maximal. In addition to the two experiments varying feature and sample selection reported above, we actually carried out numerous other analyses of different combinations. Every change in the sample selection resulted in a change of the feature ranking produced by TiMBL. For example, starting the hill-climbing experiment with a different feature would also lead to a different result, with a different set of features deemed harmful. Similarly, changing the test set will result in different performance of even the same model. For this reason, we believe that merely reporting system performances is not enough. It should become common practice to inspect evaluations taken over different test sets and to report the model's *averaged* performance, i.e., its F, R, and P scores, each bounded by confidence intervals.

To this end, we split randomly the test set into six subsets and evaluated each output. Then we computed the mean, variance, standard deviation, and confidence intervals of the six results of each P, R, and F-score. The exact performance of our pairwise comparison model for coreference (model A in Table 7) is 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F.

5 Conclusion

This paper focused on the classification stage of an automated coreference resolution system for Spanish. In the pairwise classification stage, the probability that a pair of NPs are or are not coreferent was learnt from a corpus. The more accurate this stage is, the more accurate the subsequent clustering stage will be. Our detailed study of the informativeness of a considerable number of pairwise comparison features and the effect of sample selection added to the few literature [14,1,3] on these two issues.

We provided a list of 47 features for coreference pairwise comparison and discussed the linguistic motivations behind each one: well-studied features included in most coreference resolution systems, language-specific ones, corpus-specific ones, as well as extra features that we considered interesting to test. Different machine learning experiments were carried out using the TiMBL memory-based learner. The features were shown to be weakly informative on their own, but to support complex and unpredictable interactions. In contrast with previous work, many of the features relied on gold standard annotations, pointing out the need for automatic tools for ellipticals detection and deep parsing.

Concerning the selection of the training instances, down-sampling was discarded as it did not improve performance in TiMBL. Instead, better results were obtained when the training data followed the same distribution as the real-world data, achieving 81.91±4.25 P, 69.57±8.13 R, and 75.12±6.47 F-score. Finally, we pointed out the importance of reporting confidence intervals in order to show the degree of variance that the learnt model carries.

Acknowledgements. We are indebted to M. Antònia Martí for her helpful comments. This research was supported by the FPU Grant (AP2006-00994) from the Spanish Ministry of Education and Science, and the Lang2World (TIN2006-15265-C06-06) and Ancora-Nom (FFI2008-02691-E/FILO) projects.

References

1. Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: Proceedings of EMNLP, pp. 294–303 (2008)
2. Daelemans, W., Bosch, A.V.: Memory-Based Language Processing. Cambridge University Press, Cambridge (2005)
3. Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. Ph.D thesis, University of Antwerp (2005)
4. Lappin, S., Leass, H.J.: An algorithm for pronominal anaphora resolution. Computational Linguistics 20(4), 535–561 (1994)
5. Luo, X., Ittycheriah, A., Jing, H., Kambhatla, N., Roukos, S.: A mention-synchronous coreference resolution algorithm based on the Bell tree. In: Proceedings of ACL, pp. 21–26 (2004)
6. McCarthy, J.F., Lehnert, W.G.: Using decision trees for coreference resolution. In: Proceedings of IJCAI, pp. 1050–1055 (1995)
7. Mitkov, R.: Robust pronoun resolution with limited knowledge. In: Proceedings of ACL-COLING, pp. 869–875 (1998)
8. Morton, T.S.: Using coreference in question answering. In: Proceedings of the 8th Text REtrieval Conference, pp. 85–89 (1999)
9. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of ACL, pp. 104–111 (2002)
10. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In: Proceedings of HLT-NAACL, pp. 192–199 (2006)
11. Recasens, M., Martí, M.A.: AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation (to appear)
12. 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)
13. Steinberger, J., Poesio, M., Kabadjov, M.A., Jeek, K.: Two uses of anaphora resolution in summarization. Information Processing and Management: an International Journal 43(6), 1663–1680 (2007)
14. Uryupina, O.: Knowledge Acquisition for Coreference Resolution. Ph.D thesis, Saarland University (2007)
15. Van Deemter, K., Kibble, R.: On coreferring: Coreference in MUC and related annotation schemes. Computational Linguistics 26(4), 629–637 (2000)
16. Yang, X., Su, J., Zhou, G., Tan, C.L.: Improving pronoun resolution by incorporating coreferential information of candidates. In: Proceedings of ACL, pp. 127–134 (2004)

Coreference Resolution on Blogs and Commented News

Iris Hendrickx[1] and Veronique Hoste[1,2]

[1] LT3 - Language and Translation Technology Team, University College Ghent, Groot-Brittanniëlaan 45, Ghent, Belgium
[2] Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281(S9), Ghent, Belgium
iris.hendrickx@hogent.be, ve.hoste@ugent.be

Abstract. We focus on automatic coreference resolution for blogs and news articles with user comments as part of a project on opinion mining. We aim to study the effect of the genre shift from edited, structured newspaper text to unedited, unstructured blog data. We compare our coreference resolution system on three data sets: newspaper articles, mixed newspaper articles and reader comments, and blog data. As can be expected the performance of the automatic coreference resolution system drops drastically when tested on unedited text. We describe the characteristics of the different data sets and we examine the typical errors made by the resolution system.

Keywords: Coreference resolution, Blogs, Machine learning.

1 Introduction

One of the major challenges in an ever more globalizing world, in which the rise of the internet has led to a tremendous information and opinion overload, is the development of techniques which can assist humans in managing and exploiting this information wealth. Whereas, until recently, the international natural language processing research community mainly focused on the "factual" aspects of content analysis , we can observe a growing interest in the analysis of attitude and affect in textual sources. As messages (consumer reviews, blogs, e-mails, short messages, etc.) are becoming more prevalent on the Internet than edited (newswire) texts, it becomes crucial to develop robust technologies to extract not only the factual information, but also opinions, evaluations, beliefs and speculations from text.

In blogs, opinion sites, message boards, chats and forums, people can describe their personal experiences and opinions on about anything. People write about their personal life and express their opinions through writing blogs; they actively participate in discussions around the news by participating in forums or by posting comments on texts written by others. Newspapers have engaged in these trends: they no longer just publish their news articles online, but they offer their readers the opportunity to participate and publish their own comments

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 43–53, 2009.

and opinions about an article. News is also much more interactive as it is not published once a day as is the case with printed newspapers, but news stories are updated every time an event evolves. In case of major events, some newspapers even start live blogs offering people direct communication with the journalists present at the scene of the event.

As people are so productive in expressing their opinions on the web nowadays, their generated content is not only useful for anyone who has to make everyday decisions (like which brand to choose, which movie to go to, which hotel to choose), companies as well are anxious to understand how their services and products are perceived. Given the enormous amount of potentially interesting information, which is impossible to handle manually by media analysts, an automatic procedure is required which offers a digest of opinions on a certain product, service or company. This media reviewing procedure creates a variety of opportunities for individuals and organizations: to support companies in product and service benchmarking, to support market and competitor intelligence, in customer complaint management, in customer relation management, in advertising (associate advertisements with user-generated content), as decision support for political organizations, etc.

In order to support media analysts in their analysis of trends and opinions, automatic extraction tools are needed which are able to reliably detect the three basic components of an opinion [14]: (i) an *opinion holder*, viz. the person, institution, government, etc. that holds a specific opinion on a particular object, (ii) the *target*, i.e. a product, person, event, organization, topic, or even an opinion on which an opinion is expressed [20] and an *opinion* i.e. a view, attitude, or appraisal on an object from an opinion holder. The opinion classification could be tertiary (sentiment polarity classification) [28] or scalable (sentiment strength detection). Both the identification of the opinion holder and the target involve coreference resolution [21].

Coreferential resolution between the mentioned entities in the text and across different texts plays an important role in automatic opinion mining. We focus on automatic coreference resolution for blogs and news articles with user comments as part of a project on opinion mining for Dutch. We aim to study the effect of the genre shift from edited structured newspaper text to unedited, unstructured blog data. We compare our coreference system on three data sets: newspaper articles, mixed newspaper articles and reader comments, and blog data. Blogs can be seen as online diaries expressing the personal opinions of the blog author. They are often written in a style that resembles spoken language. Published news articles on the other hand are highly structured, factual and edited. On the point of referring expressions, blogs contain much more personal pronouns than newspaper text [16].

In the next Section we first describe related work on coreference resolution and opinion mining. Section 3 gives a detailed overview of the three data sets we use and describes the characteristics of the different text genres. In Section 4, we explain our experimental setup. Section 5 presents our results which are further discussed in the section. Section 6 presents some concluding remarks.

2 Related Work

Nicolov et al. [18] investigated the effect of coreference resolution for the task of product opinion mining in blog data. As text from a blog often contains topic drifts, they propose to use snippets of texts around a product name instead of full blog posts as a starting point for opinion extraction. In their study, they showed that information on coreference relations can improve their opinion mining system with approximately 10%.

The work of Stoyanov and Cardie [21] studies coreference resolution for opinion summarization. The authors focus on identifying opinion holders and resolving coreference relations between them. They work with partially annotated data in which only the opinion holder's coreferential information is annotated. They propose a new algorithm that can handle partially supervised clustering of this type of data. Choi et al. [4] and Bethard et al. [2] present closely related work, yet they aim at another type of relations. They study the recognition of entities and the relations between opinion holders and entities which by themselves represent opinions or beliefs. According to Kobayashi et al. [13], opinion mining and anaphora resolution can be considered as similar type of tasks: one can view linking an opinion to a source as linking an anaphor to an antecedent.

From a methodological point of view, coreference resolution on blog data could also benefit from prior work on coreference in dialogue. Strube and Müller [23] describe a machine learning approach to the resolution of third person pronouns in spoken dialogue which uses a set of additional features which are specifically designed to handle spoken dialogue data (e.g. type of antecedent, verb's preference for arguments of a particular type). Their results show that these additional features are mainly beneficial for recall. Jain et al. [11] describe a rule-based system for handling anaphora in multi-person dialogues. The system integrates different constraints and heuristics, some of which are tailored to dialogues, but they do not evaluate the added value of these specific constraints and heuristics. Luo et al. [15] focus on coreference resolution in conversational documents (2007 ACE data) which incorporate speaker and turn information. They propose to use this metadata information to compute a group of binary features and show that this metadata information improves the ACE-value for broadcast conversation and telephone conversation documents. Given the (highly) unstructured nature of both dialogues and blogs, the insights from coreference resolution on dialogue data can be useful for coreference resolution on blogs. Our present study, however, is mainly focused on investigating the effect of genre shift; in the near future, we plan to investigate feature construction typically tailored to blog texts.

3 Data

In the present study, we aim to investigate the effect of the genre shift from edited structured newspaper text to unedited, unstructured blog data. In order to do so, we compared our coreference system on three data sets, namely newspaper articles, mixed newspaper articles and reader comments, and blog data.

As data set of *published news text* we used the KNACK 2002 data set which contains 267 Dutch news articles manually annotated with part-of-speech, named entities and coreferential information between noun phrases [9]. In the experiments presented here, we only use the manually annotated coreference links. For part-of-speech tags and named entities we use automatically predicted labels produced by automatic taggers as detailed in Section 4.

In WordNet 3.0 [7] a *blog* is defined as "a shared online journal where people can post diary entries about their personal experiences and hobbies; postings on a blog are usually in chronological order". A corpus of blogs has typical characteristics in terms of its content, structure and temporal aspects [16]. The author of a blog writes about his or her personal life often addressing many diverse topics and expresses individual comments, ideas and thoughts. The internal structure of a blog is a series of pieces of texts (posts). Timelines are an important feature of blogs as each post in the blog has a time stamp and the most recent posts are listed first. Blogs should not be seen as personal, isolated generated content, but rather as part of a network: blog posts contain links to other pages and many blogs offer readers the possibility to post reactions, making a blog interactive. As blogs are not edited they contain more spelling errors, ungrammatical sentences, and they deviate from newspaper text in terms of the use of capitalization, abbreviations and punctuation marks denoting emphasis or emoticons (like :D) or duration effects (like ...).

Example 1 (Excerpt from news comments. Each comment has an author and time stamp.).

```
wtf is twitter
Drinkyoghurt | 31-03-09 | 00:35
---
Duh, its just texting  to a site so your friends can read
em there.  What'dya mean detour?
And sorry if I explain it wrong, that's cuz I don't give
a shit.
Ozdorp | 31-03-09 | 00:52
---
Those extra hours of training at school makes  teenagers
smarter apparently....
Paramada | 31-03-09 | 01:23
---
Twitter  doesn't stand a chance. They offer the same functio-
nality as  SMS (with respect to character limitation) plus
some functions that the rest of the internet (Google, Digg,
RSS) deals with in a much better way.  If you want to be
popular, do it with a suitable media method like Hyves.
(Not a fan either but I do have an account to get rid of
all that ridiculous 'JOIN HYVES' spam.)
Canterwood | 31-03-09 | 16:52
---
```

Our third source of data consists of *newspaper articles and reader comments* and is a mixture of text produced by professional writers and user-generated unedited text. The reader comments have the form of posts with a time stamp and are mostly displayed in chronological order. Both types of text address the same topic, but differ highly in style and are opposites in many aspects such as formal versus informal, factual versus personal, edited versus unedited. Contrary to most blog posts which usually address all kinds of topics and thoughts, the reader comments of a news article have a focused topic. The posted reactions to news articles on news source websites have the same informal writing style and structural characteristics as the blog data.

As an evaluation set, we collected 5 news articles with reader comments from an online newspaper and 15 blog posts. These were also manually annotated with coreferential information. The blog posts were collected from two blogs on Belgian cities and are written by multiple authors. The content of the blog posts varies from personal stories about a certain event to more informative blog posts describing upcoming events in the city.

We selected five news articles and accompanying comments. The selected news articles themselves are rather short, no longer than 20 sentences. The number of reader comments per article ranges from 88 to 123 different comments. In general these comments are short, majority containing atmost one or two sentences. The language use strongly resembles chat or spoken language. As an example of this type of data, we translated an excerpt of the comments on a Dutch news article stating that adolescents are not enthusiastic about Twitter as shown in Example 1. We consider each news article and the accompanying reader comments as one single document. This is a practical choice, many of the comments refer to the entities mentioned in the news article. However we do notice that our single document view is somewhat simplistic and not all characteristics of the data are well captured in our representation.

Table 1 gives an overview of the size of the different test sets. It mainly reveals that there are no differences in sentence length between the fairly structured blog data and the published news texts. The data set with the newspaper articles and reader comments, however, contains shorter sentences. Table 2 presents information about the type and quantity of anaphors in the different test sets. Our observations confirm the findings published in [16]; the blogs and commented news both contain relatively more pronouns. Here we focus on a quantitative overview of the number of pronouns which are not part of a coreference chain, presents a similar tendency: 61% of the pronouns in the data set containing the newspaper

Table 1. Data statistics: number of tokens, sentences and average sentence length per data set

Test set	#Documents	#Tokens	#Sent.	Av. sent. length
Published news texts	25	111,117	576	19.3
News and comments	5	14,276	937	15.2
Blogs	15	5,689	289	19.7

Table 2. Proportion of pronominal, common noun and proper noun coreferential NPs. Number of pronouns which are not part of a coreference chain.

Test set	No coreference	Coreference			
	Pronouns	Pronouns	Proper N.	Common N.	All
Published News texts	178	282	426	492	1200
News and comments	610	390	200	537	996
Blogs	101	214	100	269	583

articles with reader comments does not refer to a preceding antecedent, whereas this percentage is much lower for the other two data sets (Published: 32.1% and Blogs: 38.7%).

4 Experimental Setup

The coreference resolution system takes a machine learning approach following the example of a.o. Soon et al. [19], Ng and Cardie [17] and is based on previous work of Hoste [10] for Dutch. Coreference resolution is seen as a classification task in which each pair of noun phrases in a text is classified as having a coreferential relation or not. For each pair of noun phrases, a feature vector is created denoting the characteristics of the pair of noun phrases and their relation.

To create the feature vectors, we first process the text. First, tokenisation is performed by a rule-based system using regular expressions. Part-of-speech tagging and text chunking is performed by the memory-based tagger MBT [5]. For the grammatical relation finding which determines which chunk has which grammatical relation to which verbal chunk (e.g. subject, object, etc.) a memory-based relation finder is used [24]. We also use a automatic Named Entity Recognition system, MBT trained on Dutch data set of the CoNNL 2002 shared task [25]. Besides these predicted labels (persons, organizations, locations, miscellaneous names), the system performs a look up names in gazetteer lists to supplement the automatic system, and to refine the predicted label *person* to *female* or *male*.

Several information sources contribute to a correct resolution of coreferential relations: morphological, lexical, syntactic, semantic and positional information and also world knowledge. In order to come to a correct resolution of coreferential relations, existing systems, e.g. [8,3,19,22], use a combination of these information sources. For our coreference resolution system, we extract the following types of features: string overlap, distance between the noun phrases, overlap in grammatical role and named entity type, synonym/hypernym relation lookup in WordNet, morphological suffix information and local context of each of the noun phrases. For a more detailed description of the feature construction, we refer to [10].

We train different systems for different types of referring expressions. This allows us to optimize the system for each type of expression separately. Furthermore, splitting the treatment of the expressions can also help to focus on the errors separately for each referring expression made by the resolution system.

We create three separate systems for pronouns, named entities and common nouns and optimize the machine learning classifier for each type separately. As machine learning algorithm we used memory-based learning as implemented in the software package Timbl [6]. We optimized the algorithmic parameters and feature weighting for each system with a heuristic search method that iteratively tries to find an optimal parameter setting for the data set at hand [26].

The experiments on the three different data sets are set up in the following way. We split the KNACK data set into a training set of 242 articles and a held out set of 25 articles for testing. The blog data set and news comments data set were only used for testing and not for training. We train our coreference resolution system on the KNACK training data and test it on each of the three different test sets. We measure the performance of our system using the MUC [27] and the B-Cubed [1] scoring software.

5 Results

We present the results of our coreference resolution on the three data sets in Table 3. We computed precision, recall and F-score using the MUC scoring and recall computed with the B-cubed method. As can be expected, the performance of the coreference resolution systems drops significantly for the blog and news with comments test sets. The results on the blog material are the lowest. The MUC scores and B-cubed scores show the same tendencies.

Table 3. Results of the coreference resolution system on the three different data sets: Edited newspaper text, blog data and news with reader comments. Scores computed with the MUC and B-cubed scoring methods.

	MUC scoring			B-cubed
Test set	Recall	Precision	F-score	Recall
Published News texts	44.7	66.8	53.6	52.3
Blogs	18.9	40.0	25.7	43.5
News and comments	26.7	42.7	32.8	48.7

5.1 Error Analysis

On the basis of a shallow manual error analysis on three texts of each corpus, we were able to detect typical errors that are made on the different data sets. The most problematic classes are the following:

– **Pronouns erroneously being classified as coreferential:** For the published newspaper texts, we could observe a large number of pleonastic pronouns which were linked with a preceding noun phrase. The pleonastic pronoun was always the neutral third person singular pronoun “het”. The news and reader comments data set reveals the same tendency, but in this data set it is not restricted to the neutral third person singular pronoun. Also personal pronouns

like “je” (you) or “zij” (they) are often used when referring to people in general and not to a specific entity mentioned in the text. e.g.

(1) Du: Als **het** met dat coördinatiecentrum slecht afloopt (...)
(En: If it doesn’t end well with that coordination centre...)

– **Incomplete detection of noun phrases:** All data sets share the problem of the incomplete detection of noun phrases which leads to partial detection of coreferential relations. e.g. in sentence 2 below, only part of the NP is recognized, viz. “Mevrouw”.

(2) Du: **Mevrouw** Spiritus Dasesse zet heel geëmancipeerd haar meisjesnaam voorop (...)
(En: Mrs Spiritus Dasesse puts her maiden name first)

– **Problems with the current feature vector:** For all data sets, the feature vector sometimes does not provide enough disambiguating information to distinguish between a positive and negative classification. e.g. in example 3, “elkaar” is erroneously linked to “de 190 miljoen euro”. The feature vector given below was used as the basis for the positive classification.

(3) Du: Hij herhaalt dus alweer dat hij tegen half januari **de 190 miljoen euro** bij **elkaar** heeft (...)
(En: He repeats again that he’ll have the 190 milioen euro by mid-January)
(7 519 1088) (elkaar) (7 518 1083) (de 190 miljoen euro) 0 1 miljoen euro bij TW(hoofd,prenom,stan) N(soort,ev,basis,zijd, stan) VZ(init) heeft om in WW(pv,tgw,met-t) VZ(init) VZ(init) dist_lt_two appo_no jpron_yes 0 0 0 0 num_na 0 0 0 0 0 0 0 0 I-OBJ I-OBJ I-OBJ person 0 0 0 0 0 zijdig 3p refl def_yes 0 0 0 0 NEG POS

– **Errors that need ’world knowledge’ or sophisticated information resources:** For some of the coreferential links a specialized resource such as an ontology or a database with gathered facts is needed to resolve the ambiguity. The abbreviation “MP” (minister president) in example 4 refers to earlier mentions in the text like “Balkenende” and “JPB”. To resolve these coreferential links one needs to know the name of the current minister president of the Netherlands. Our training material is not helpful because it is older than the comments news articles and blogs, so the names referring to ’minister president’ in the training material are different than the ones in this test material.

(4) Du: Het is in Nederland een grote rotzooi en **onze MP** maar praten over normen en waarden.
(En: The Netherlands is a big mess and our MP just talks about values.)

6 Conclusion

The work presented here can be seen as a first step towards a automatic coreference resolution system that will be integrated in an online automatic extraction tool for media analysis. Here we focused on examining the differences in language use between texts from (printed) newspapers and mixed newspaper articles and reader comments and blog data. We studied the characteristics of the three different data sets in Section 3. We experimented with an automatic coreference resolution system trained on edited newspaper text and compared it's performance on the three different text types. As expected, our results show that the performance of our automatic coreference resolution system drops significantly when confronted with unedited text. Next we examined in more detail the type of errors made by the system and the possible causes of these errors.

An obvious method to improve the coreference resolution system is to train not only on newspaper articles, but also on a data set consisting of spoken language or annotated blogs and commented news data. However, we believe that adding training material will not be sufficient to resolve all problems. In an adapted version of our coreference system we also plan to add additional features. We would like to add factual information gathered from the web or from available corpora. Finding facts is a method that is regularly applied in question-answering systems e.g. [12]. This type of information can be seen as a resource of 'world knowledge' and help to resolve ambiguities like the one illustrated in example 4.

The discussion on related work on dialogues already suggested that information on turn-taking can be valuable. We expect this to be true for blogs and reader comments as well. Especially for pronouns in the commented news data set, explicit information about turn-taking can help our system to resolve pronouns that refer to the author or to authors of previous comments. Because our system already has a separate trained module for pronominal anaphors, it will be relatively easy to adjust the system on this point.

Acknowledgements. The work presented here was conducted within the framework of the DuOMAn (Dutch Online Media Analysis) project which is funded by the Dutch-Flemish STEVIN research program. We would like to thank the anonymous reviewers for their helpful comments.

References

1. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: Proceedings of the First International Conference on Language Resources and Evaluation Workshop on Linguistic Coreference, pp. 563–566 (1998)
2. Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., Jurafsky, D.: Automatic extraction of opinion propositions and their holders. In: AAAI Spring Symposium on Exploring Attitude and Affect in Text, pp. 22–24 (2004)
3. 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, pp. 82–89 (1999)

4. Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics (2006)
5. Daelemans, W., Zavrel, J., Van Den Bosch, A., Van Der Sloot, K.: Memory based tagger, version 2.0, reference guide. Technical Report ILK Technical Report - ILK 03-13, Tilburg University (2003)
6. Daelemans, W., Zavrel, J., Van der Sloot, K., Van den Bosch, A.: TiMBL: Tilburg Memory Based Learner, version 6.1, reference manual. Technical Report 07-07, ILK, Tilburg University (2007)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
8. Fisher, D., Soderland, S., McCarthy, J., Feng, F., Lehnert, W.: Description of the umass system as used for MUC-6. In: Proceedings of the Sixth Message Understanding Conference (MUC-6), pp. 127–140 (1995)
9. Hoste, V., De Pauw, G.: Knack-2002: a richly annotated corpus of dutch written text. In: The fifth international conference on Language Resources and Evaluation, LREC (2006)
10. Hoste, V.: Optimization Issues in Machine Learning of Coreference Resolution. PhD thesis, Antwerp University (2005)
11. Jain, P., Mital, M.R., Kumar, S., Mukerjee, A., Raina, A.M.: Anaphora resolution in multi-person dialogues. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (2004)
12. Jijkoun, V., De Rijke, M., Mur, J.: Information extraction for question answering: Improving recall through syntactic patterns. In: Coling 2004, pp. 1284–1290 (2004)
13. Kobayashi, N., Iida, R., Inui, K., Matsumoto, Y.: Opinion extraction using a learning-based anaphora resolution technique. In: Second International Joint Conference on Natural Language Processing: Companion Volume including Posters/Demos and tutorial abstracts, pp. 175–180 (2005)
14. Liu, B.: Web Data Mining. Exploring Hyperlinks, Contents and Usage Data. Springer, Heidelberg (2006)
15. Luo, X., Florian, R., Ward, T.: Improving coreference resolution by using conversational metadata. In: Proceedings of NAACL HLT 2009, pp. 201–204 (2009)
16. Misnhe, G.: Applied Text Analytics for Blogs. PhD thesis. University of Amsterdam, Amsterdam, The Netherlands (2007)
17. Ng, V., Cardie, C.: Improving machine learning approaches to coreference resolution. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 104–111 (2002)
18. Nicolov, N., Salvetti, F., Ivanova, S.: Sentiment analysis: Does coreference matter? In: Proceedings of the Symposium on Affective Language in Human and Machine, Aberdeen, UK (2008)
19. 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)
20. Stoyanov, V., Cardie, C.: Topic identification for fine-grained opinion analysis. In: Proceedings of the Conference on Computational Linguistics, COLING 2008 (2008)
21. Stoyanov, V., Cardie, C.: Partially supervised coreference resolution for opinion summarization through structured rule learning. In: Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP), pp. 336–344. Association for Computational Linguistics (2006)
22. 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), pp. 312–319 (2002)

23. Strube, M., Müller, C.: A machine learning approach to pronoun resolution in spoken dialogue. In: Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 168–175 (2003)
24. Tjong Kim Sang, E.F., 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, pp. 109–123 (2004)
25. Tjong Kim Sang, E.F.: Introduction to the conll-2002 shared task: Language-independent named entity recognition. In: Roth, D., van den Antal, B. (eds.) Proceedings of CoNLL 2002, pp. 155–158 (2002)
26. Van Den Bosch, A.: Wrapped progressive sampling search for optimizing learning algorithm parameters. In: Proceedings of the 16th Belgian-Dutch Conference on Artificial Intelligence, pp. 219–226 (2004)
27. 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), pp. 45–52 (1995)
28. Wilson, T., Wiebe, J., Hoffman, P.: Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis. Computational Linguistics (2008)

Identification of Similar Documents Using Coherent Chunks

Sobha Lalitha Devi, Sankar Kuppan, Kavitha Venkataswamy, and Pattabhi R.K. Rao

AU-KBC Research Centre, MIT Campus of Anna University, Chennai, India
{sobha,sankar,kavitha,pattabhi}@au-kbc.org

Abstract. We focus on automatically finding similar documents using coherent chunks. The similarity between the documents is determined by identifying the coherent chunks present in them. We apply linguistic rules in identifying the coherent chunks and uses Vector Space Model (VSM) in determining the similarity among documents. We have taken patent documents from USPTO[1] for this work. This method of using coherent chunks for identifying similar documents has shown encouraging results.

Keywords: Coherence, Document similarity, Coreference.

1 Introduction

The work presented in this paper has two parts a) identifying the coherent chunks and b) finding the cross document coherent chunk similarity. The present work analyses the patent documents and identifies whether the documents are similar. Two or more sentences are coherent if they are semantically connected. A set of such coherent sentences are called a coherent chunk. Two objects are said to be similar, when they have common properties between them. For example, two geometrical figures are said to be similar if they have the same shape. Hence similarity is a measure of degree of resemblance between two objects. Two documents are said to be similar if any one of the following holds

i) the events described are same
ii) the topics described are same
iii) the concepts defined are same
iv) the events and the topics are related to each other

For finding coherent chunks in a document, we propose a set of linguistic rules that could identify the connection between adjacent sentences in a document. Based on these rules the connected sentences are identified and that determines a coherent chunk. The coherent chunks give information that is required for the identification of similarity.

The paper is further organized as follows. The following section describes the coherence analyser. In section 3, the similarity analyser is described. In Section 4, the experiments and results are discussed. Section 5 gives the conclusion.

[1] United States Patent and Trademark Office. http://www.uspto.gov/

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 54–68, 2009.

2 Coherence Analyser

In order to interpret a sequence of sentences fully, one must know how the sentences cohere; that is, one must be able to infer implicit relationships as well as non-relationships between the sentences.

Consider the following example:

(1) The car broke down in the middle of the forest and we reached late. Its engine had a problem.

Such sentences appear coherent because it is easy to infer how the second sentence is related to the first.

Contrast this with the following sentence:

(2) The car broke down in the middle of the forest and we reached late. The city has heavy traffic in the evening.

The above sequence of sentences is not coherent since there is no obvious connection between the two sentences. One would say that there is no relationship between the sentences. Furthermore, because the second utterance violates an expectation of discourse coherence [19,9,6], the sentences seem inappropriate since there are no linguistic clues (for example, prefacing the second sentence with *always*) marking it as a topic change. Coherence makes a text semantically meaningful and it is achieved through syntactical features such as anaphors *he, she, it, that, this,* connectives *but, whereas*, repetitions of noun phrases and as well as presuppositions and implications connected to general world knowledge. The identification and specification of sets of linguistic relationships between sentences forms the basis for many computational models of discourse [20,14,9,3].

In this section we present a linguistic method for recognizing coherent relationships between sentences. The coherence clues present in the sentence are directly visible when we go through the flow of the document. Here we consider four features for identifying the coherent chunk such as connectives, anaphors, noun reappearance and thesaurus relationships. The analysis of these clues and the rules developed for identifying the chunks are explained in detail below. For the analysis we have taken data from patents in the domain of electronics which are freely available in the USPTO site.

2.1 Connectives

Little words like *whereas, but, moreover* etc. form the list of connectives. They are a finite set and what they refer will vary in syntactic form. When parsing through a document, the relationship among adjacent sentence is determined by the type of connective that is used. Here we look into connectives which are explicitly marked. Consider the following sentences.

(3) a. This invention relates to an optical fiber connector arrangement.
b. In particular, the invention relates to optical fiber connectors having hermaphroditic bayonet couplings.

c. These couplings have two diametrically opposed connection arms, the arms of one connector fitting in the spacing between the arms of another identical connector to enable the two connectors to be coupled together.
d. Relative rotation between one connector and other identical connector then enables a bayonet connection to be established.
e. There are two possible orientations in which it can be attempted to couple the connectors, and only one orientation is correct (unless there is only one fiber located centrally in the connector).
f. *However*, connectors of this type may be located in areas with adverse weather or light conditions, so that it may be difficult to obtain the correct alignment.

In the above sentences the connective *however* in (3f) is connecting the sentence with (3e). But semantically it is connecting all the six sentences. Here *however* is occurring in the sentence initial position. Hence (3f) and (3e) are coherent chunks. There are connectives such as *accordingly, again, also, besides* which cannot come in the initial position of the sentence, but they indicate the connectedness of a sentence with the previous sentence. Furthermore, the appearance of the connective words such as *consequently, finally, furthermore,* at the beginning or middle of a sentence was found to be highly cohesive with the previous sentence. There are connectives which can appear sentence medially. Hence we classify connectives according to its position of occurrence in a sentence. The position of the connective denotes whether it is connecting to a sentence or a clause or a phrase.

2.2 Anaphora

The most common type of anaphor such as *he, she, it, these, that etc.* are taken into account for analysis and resolution. The major classifications in anaphors are the first, second and third person pronouns. First and second person singular and plural are commonly used as deictic, though they are used in anaphoric form in discourse. Anaphora resolution refers to the problem of determining the noun phrase (NP) that refers to an anaphor in a document. The noun phrase to which the anaphor refers is called its antecedent. There are many approaches to solve this problem such as rule based, statistical and machine learning based approaches. Consider the following example.

(4) a. In particular, the invention relates to optical fiber connectors having hermaphroditic bayonet couplings.
b. These couplings have two diametrically opposed connection arms, the arms of one connector fitting in the spacing between the arms of another identical connector to enable the two connectors to be coupled together.

Here *these* has a referent in the previous sentence *hermaphroditic bayonet couplings* and they make the two sentence cohere.

2.3 Noun Reappearance

The reappearance of NEs in adjacent sentences is an indication of connectedness. When such adjacent sentences are found, they form coherent chunks. Two adjacent sentences are said to be coherent when both the sentences contain one or more reappearing nouns.

(5) a. A *pin and a receptacle* together form a connector for coupling optical elements carried thereby.
b. Advantageously, *the pins and receptacles* of this invention are of simple configuration.

Here the noun phrase *pin and a receptacle* is repeated in both the sentences. This shows that both the sentences are coherent.

2.4 Thesaurus Relationship

The relationship between words across sentences can be used to find semantically related words. The appearance of related words is an indication of its coherence. The relationship between words could be identified using an ontology. Here we have developed an ontology for electronic devices and it has devices as the main node. The subnodes contain the components and the applications of the device. Other than this ontology we use the WordNet ontology for identifying the related words. WordNet covers most of the sense relationships of any noun and verb. To find the semantic neighborhood between adjacent sentences, most of the lexical relationships such as synonyms, hyponyms, hypernyms, meronyms, holonyms and gradation can be used [4]. Hence, semantically related terms are captured through this process.

(6) a. It is accordingly an object of the invention to provide a *plug connector* for a fiber optic cable, which overcomes the above-mentioned disadvantages of the prior art devices and methods of this general type and which results in a connector with a protected connector configuration that can be prefabricated.
b. With the foregoing and other objects in view there is provided, in accordance with the invention, *a connector* for a fiber optic cable, comprising: a plug pin assembly defining a plug-in axis and enclosing an end of an optical fiber, the plug pin assembly having a forward end relative to a plug-in direction and a rear end.

In this example the *connector* in sentence (6b) relates to *plug connector* in sentence (6a).

2.5 Coherence Finding Algorithm

The algorithm has four modules of resolution for coherent chunk identification. In the algorithm the connective identification is done first since this gives more

information about coherent chunks, the next the anaphors are resolved. In the third module the noun reappearance is identified and finally we identify the thesaurus relationship. If anyone of the above is present in a sequence of sentence we consider that sequence as coherent.

Connectives Resolution: Identify the connectives in the set of sentences and find its position in the sentence. Depending on the position of the connective it could be identified whether it is connecting to the previous sentence, clause or phrase. High preference is given to the connective which connects the previous sentence.

Anaphora Resolution: The anaphora resolution system we use work on the salience factors arrived at by linguistic analysis of the corpus, preference rules and semantic disambiguator. The input to the system is a fully parsed output from FDG parser and the output from named entity resolution(NER). The linguistic information taken from FDG parser is Subject, Direct object and Indirect object. We use salience factors and its weight for the resolution of anaphors. The scores are discussed in detail below.

(a) The current sentence gets a score of 100 and it reduces by 10 for each preceding sentence till it reaches the fifth sentence. The system considers five sentences for identifying the antecedent. Current sentence is the sentence containing the anaphor.
(b) The analysis showed that the subject could be the most probable antecedent for the anaphor. The subject noun phrase is given a score of 80.
(c) The direct object of a sentence gets a score of 50.
(d) The indirect object of a sentence gets a score of 40.

The NE tags of the anaphor and the NPs are considered for feature agreements. For example if the tag of anaphor is Individual then the NP with Individual tags alone are considered. We have other feature agreements such as anaphor with Group can have antecedent candidates with Organization. We have also used the pronoun information from the parser for identification of pronouns and pronouns number such as singular or plural. Another rule that is used: Incase if two NP become the probable candidates with same salience score and the agreement is also same, then the NP which is nearer to the anaphor is considered as the antecedent.

Noun Reappearance: Named Entities are used for noun reappearance identification. Named Entity Recognition (NER) is the task of identifying and classifying the rigid designators such as person, place, organization, products, devices etc, in a given document. NER can be visualized as a sequence labeling task and thus can be done with machine learning algorithms supporting sequence labeling task. Our method uses Conditional Random Fields (CRFs) for learning from the corpus and tagging new sentences. CRFs is a machine learning algorithm suitable for sequence labeling task. CRFs extracts features from the training data using the Templates supplied, and learns from the training data the suitable scaling factors for each of those features. We have trained the system with 1,00,000 words and tested the system.

Thesaurus Relationship: We have developed an ontology for electronic goods and that has relationship marked according to the relation of a part with the whole. For example, if we take diode then it will be connected to all equipments which require a diode. We also use WordNet which covers most of the sense relationships. Hence, semantically related terms are captured using the ontology and WordNet.

2.6 System Architecture

The system works as follows: The pre-processor processes the input documents for sentence splitting, morphological analysis, POS tagging, NP chunking and Parsing. The patent documents are preprocessed to enrich the sentences with required syntactic information. The documents are preprocessed using Brill's Tagger [2] for POS tagging and fn-TBL [15] for text chunking. On this preprocessed text, Named Entities are identified using a Named Entity Recognizer as explained in section 2.5. The input to the coherence analyzer is the pre processed text. In the coherent analyzer we use the Connective module first to the pre-processed input. Here we take five sentences above the sentence containing the connective. The sentences which are having the referent for the connective are taken as coherent chunks. Then the anaphora resolution module is called and the anaphors present are identified and resolved. The NPs are taken from five sentences above the sentence in which the anaphor occurs. The agreement is checked using the NER and the FDG output. The sentences which have the antecedent of an anaphor are considered as a coherent chunk. If there is repetition of the same NE more than once in consecutive sentences then they are considered as coherent. Here we look upto the fifth sentence. The ontology is used to identify the whole and part relationship. The sentences which have this relationship are taken as coherent sentences. The sentences which satisfy one of the above rules are considered as a coherent chunk. If all the four rules are applicable then it is considered as a highly coherent chunk.

3 Identifying Similar Chunks across Documents

3.1 What Is Similarity?

Two documents are said to be similar if they describe about a same event or subject or entity. Similar documents are not identical documents. For example a document N1 describes about a bomb blast incident in a city and document N2 also describes about the same bomb blast incident, its cause and investigation details, then N1 and N2 are said to be similar. But if document N3 talks of terrorism in general and explains bomb blast as one of the actions in terrorism and not a particular incident which N1 describes, then documents N1 and N3 are dissimilar.

The task of finding document similarity differs from the task of document clustering. Clustering is a task of categorization of documents based on domain

or field. In the above example, documents N1, N2 and N3 can be said to be a cluster of the crime domain. When documents are similar they share common noun, verb phrases and named entities. While in document clustering, sharing of named entities and noun phrases is not essential. Similarly, similar documents are not identical documents. Identical documents have exactly same content. The task of recognising text entilement is different from identifying similar documents. In text entilement, it has to be identified whether text T2 is inferred from text T1 (RTE-5, TAC 2009). Identification of similar documents would help in recognising text entilement across documents.

3.2 Related Work

Dekang Lin [11] defines similarity from the information theoretic perspective and it is applicable if the domain has probabilistic model. Many similarity measures were developed, such as information content [21], mutual information [8], Dice coefficient [5], cosine coefficient [5], distance-based measurements [10,16], and feature contrast model [23]. McGill surveyed and compared 67 similarity measures used in information retrieval [13].

In the last decade, there has been significant amount of work done on finding similarity of documents and organizing the documents according to their content. Similarity of documents are identified using different methods such as Self-Organizing Maps (SOMs) [12,18], based on Ontologies and taxonomy [7,21], Vector Space Model (VSM) with similarity measures like Dice similarity, Jaccard's similarity, cosine similarity [22]. Bagga et al. [1] have used VSM in their work for finding coreferences across the documents in English.

There are many statistical techniques such as Support Vector Machines (SVM), Vector Space Model (VSM), Latent Semantic Analysis (LSA) used in document processing. SVM is popularly used for document clustering. LSA is used in problems where dimension reduction could be done. It is generally believed that LSA could be used for similarity identification because there is a misconception that similarity identification is also the same as dimensionality reduction. The drawback of LSA is that the reduced dimension matrix is difficult to interpret semantically and is not suitable for identificaiton of similar documents.

3.3 Identification of Similar Documents

In the present work, we have used Vector Space Model (VSM). In VSM, each document is represented by a vector of terms. A vector is a set of elements or objects having magnitude and direction. When the documents are represented as vectors, the words (or terms) in the documents constitute the vector. This is called as vector of terms, also called as document vector. In VSM, the vector of terms specifies the number of times each term occurs in the document (the term frequencies).These term frequency counts are weighted to reflect the global importance of each term with-in the whole set of documents. The weighting function used is the inverse document frequency (idf). If a term t occurs in n documents in the collection then the idf is the inverse of $log\ n$. This vector of

weighted counts is called a "bag of words" representation. Words such as "stop words" (or function words) are not included in the representation. For example the document D1, D2 has the following texts.

D1: "A fiber optic connector having a plug portion with first and second ends, the first end receiving a first fiber optic cable(s) providing optical energy and a second fiber optic cable(s) for bidirectional optical data communications."

D2: "An optical fibre connector is shown comprising a jack mounted to a mother board and a plug mounted to a daughter board. The plug has a slidable insert retained by thrust lances against a shoulder of the plug housing such that the insert can be inserted into a cavity of the jack."

The terms which constitute the document vector for D1 and D2 are *against, insert, thrust, housing, providing, cable, slidable, mounted, bidirectional, jack, data, communications, retained, energy, shown, comprising, daughter, cavity, shoulder, connector, optic, fibre, receiving, board, mother, optical, fiber, plug, portion.* The weights for these terms are calculated, as product of term frequency in the document and inverse document frequency of the terms. Terms are taken in the x-axis and documents on the y-axis.

Similarity between documents is a function of commonality and differences in the documents. The most useful measure for finding this function of commonality and differences is the cosine similarity measure. Cosine similarity measure between two documents is the scalar product of the two document vectors.

Let S_1 and S_2 be the term vectors representing the documents D1 and D2, then their similarity is given by equation 1 as shown below.

$$Sim(S_1, S_2) = \sum_{t_j} (W_{1j} \times W_{2j}) \tag{1}$$

where,

t_j is a term present in both vectors S_1 and S_2.
W_{1j} is the weight of term t_j in S_1 and
W_{2j} is the weight of term t_j in S_2.

The weight of term t_j in the vector S_1 is calculated by the formula given by equation 2, below.

$$W_{ij} = \frac{(tf \times log(\frac{N}{df}))}{\sqrt{(S_{i1}^2 + S_{i2}^2 + \ldots + S_{in}^2)}} \tag{2}$$

where,

tf = term frequency of term t_j
N=total number of documents in the collection
df = number of documents in the collection that the term t_j occurs in.

The denominator $\sqrt{(S_{i1}^2 + S_{i2}^2 + \ldots + S_{in}^2)}$ is the cosine normalization factor. This cosine normalization factor is the Euclidean length of the vector S_i, where

i is the document number in the collection and S_{in}^2 is the square of the product of $(tf \times log(\frac{N}{df}))$ for term t_n in the vector S_i.

For the purpose of identifying similar chunks, the coherent chunks obtained from the coherence chunker are taken and represented in the form of vectors, in a vector space model (VSM). For the task of cross-document similarity, the words within each coherent chunk are considered as terms for building the language model [17].

The main feature in this work of finding similarity across the documents is that, instead of taking the terms from the whole document, only the terms in the coherent chunks for each corresponding section of the document are taken to build the document vector. For finding the similarity between sections within a single document, each coherent chunk identified by the coherence analyser are considered as invidual documents and document vectors are built.

The other feature of this work is that when terms inside the coherent chunks are taken, the terms are not the mere words (separated by white spaces), but it is a named entity (NE) or a noun phrase (NP) or a verb phrase (VP). The whole NE/NP/VP is a single term. For example in the above said documents D1 and D2, the document vector would constitute the terms *fiber optic connector, plug portion, fiber optic cable(s), optical data communications, optical energy, optical fibre connector, mother board, daughter board, jack, mounted*. This way of taking the terms for building the document vector, has helped in identifying similarity between the sections within the document and also across the document. This has helped in reducing false positives. Instead of taking the whole NE as a single term, if each word is taken as single term, the documents which are not actually similar would be shown as similar. This inherently helps in capturing the contextual information. For example, for document D2, there are chances of documents describing mother and daughter relationship also to be shown as similar.

While using VSM in conjunction with cosine similarity measure, the important factor that affects the results of the similarity identification, is setting a good threshold point. In the cosine similarity measure, we obtain the cosine score in the range of 0 (zero) to 1 (one). The score of zero means the two documents being compared are totally different and have no common features. The score of 1 (one) means the documents are completely identical documents, not just similar. A score nearer to 1 means the documents being compared are similar. This score for similarity varies for each document collection. There is no fixed value for this score, to decide the similarity. The score value most suitable for a particular document collection set, has to be identified by doing experiments, by varying the score value and seeing the results. This can be done by doing empirical studies on the data. Here we have performed experimental studies and based on the findings of those experimental studies, we arrived at threshold score of 0.70. For any two documents being compared, if the similarity score crosses the threshold of 0.70, then those two documents are said to be similar.

4 Experiments, Results and Discussion

The experiments were performed on US patent documents, obtained from USPTO (US Patent and Trademark Office) website. Here we automatically identify the coherent chunks within the single patent document and related chunks across the patent documents. A patent document has mainly five sections, i) Abstract ii) Claim iii) Prior Art iv) Summary and v) Detailed Description. Here we have considered 68 patent documents from electronics domain for these experiments.

4.1 Identification of Coherent Chunks

Now the documents are analysed for identifying the coherent chunks using the rules explained in section 2. The coherence analyser chunks the sentences in the document into several coherent chunks. Table 1 shows the performance of each rules and different rules when used together in the coherent analyser. Table 2 shows the number of occurrences, where the rules overlap.

In Table 1, the recall for Rule 1, Rule 2 and Rule 3 are 100%, as in these rules we look for the connectives, anaphors and named entities respectively to chunk the sentences. We use Ontology and WordNet to chunk sentences based on the relationship of the noun phrases in Rule 4. The recall is low since the ontology is not that robust. As the patent documents have more named entities, the number of times the Rule 3 has got triggered is very high. Similarly, as the Rule 4 checks for relation between the sentences, the occurrence of this rule is also very high.

Table 1. Rule-wise Performance of the Coherence Analyser

Rules	No. of chunks the rule should be applied	No. of chunks the rules are applied by the system	No. of chunks where rules are correctly applied	Recall (%)	Precision (%)
Rule 1	615	615	586	100.00	95.28
Rule 2	273	273	258	100.00	94.51
Rule 3	6943	6943	6246	100.00	89.96
Rule 4	6166	5461	4817	88.57	88.21
Rule 1 & 2	6	6	5	100.00	83.33
Rule 1 & 3	411	411	401	100.00	97.57
Rule 1 & 4	383	350	324	91.38	92.57
Rule 2 & 3	132	119	102	90.15	85.71
Rule 2 & 4	131	123	114	93.89	92.68
Rule 3 & 4	4421	4113	3865	93.03	93.97
Rule 1, 2 & 3	1	1	1	100.00	100.00
Rule 1, 2 & 4	4	3	2	75.00	66.67
Rule 1, 3 & 4	259	245	232	94.59	94.69
Rule 2, 3 & 4	102	95	88	93.14	92.63
Rule 1, 2, 3 & 4	0	0	0	0.00	0.00

In Rule 1, the connectives, which have other grammatical functions, mislead this rule. For example, connectives such as *that*, occurs as a complimentizer as well as determiner. There are also chunks, where the topic changes with the sentence starting with connective as shown in example (7).

(7) a. This positions the latch projection 92 of the receptacle assembly 4 behind the transverse rib 38 on the housing body 10.
b. *Whenever* an optical fibre assembly is inserted, or such that the biassing force is selectively applicable by a user.

Here though *whenever* is a connective, these sentences can not be chunked.

In Rule 2, pronouns such as *it*, which also occur as a pleonastic, mislead the rule.

(8) a. One fiber optic connector is a so-called fiber optic "SMA" connector that conforms to certain optical characteristics such as insertion loss characteristics, and standard mechanical characteristics such as thread sizes and diameters of connector mating regions.
b. *It* can be difficult to determine the corresponding fiber optic cables and connectors.

As shown in example (8), *it* occurs as a pleonastic element and not as an anaphor.

The performance of Rule 3 depends on the performance of NER. For sentences which are chunked by Rule 3, even if the sentences have the same named entities occurring in consecutive sentence, they may refer to different real word entities. This reduces the precision. The Rule 4 chunks the related sentences. There are chunks, which do not hold relation as in example (9) and this affects the precision.

(9) a. *Light* weight cabeling is always prefered for a good circuit system.
b. A rotatable connector comprising: a first optical coupler adapted to mount to a first rotatable component, said first optical coupler including a first light emitter for *illumination*.

In Table 2, the overlap of Rule 3 and Rule 4 is very high, as Rule 3 and Rule 4 have triggered in many instances and Rule 4 looks for the chunks with relation and in these chunks reappearing of named entities are common.

The overlap of the rules is pictorially represented in Fig. 1. The overall performance of the coherent chunk analyzer is shown in the Table 3.

Table 2. Overlap of Rules

Overlap	Rule 1	Rule 2	Rule 3	Rule 4
Rule 1	-	6	411	350
Rule 2	6	-	119	123
Rule 3	411	119	-	4113
Rule 4	350	123	4113	-

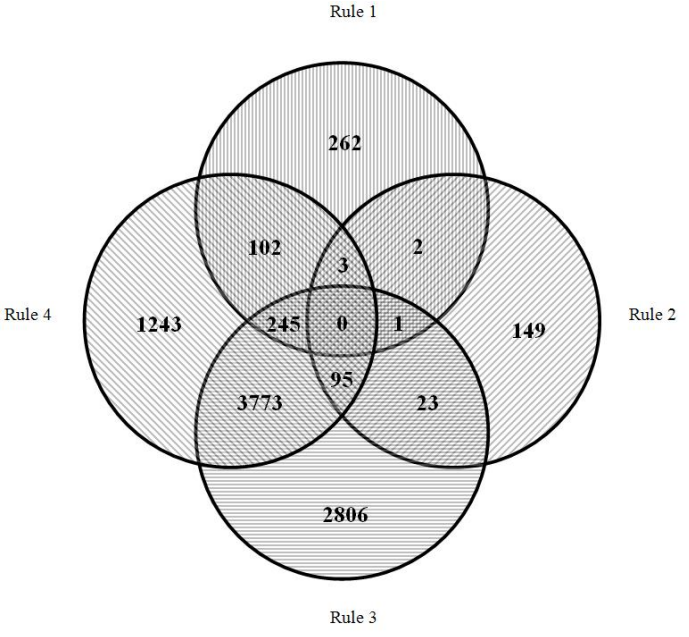

Fig. 1. Venn diagram – showing the rule overlaps

Table 3. Results for Coherent Chunks

Total Chunks in the Document	No. of Chunks identified by the system	No. of Chunks correctly identified by the system	Recall (%)	Precision (%)
1400	1360	1251	97.14	91.99

4.2 Experiments Using Similarity Analyser

The coherent chunks identified by the coherent analyser are taken as the input to this task. Similarity analysis is done as two experiments. In the first experiment we compare these chunks across documents against the same section chunks of the documents and in the second experiment we compare the chunks across documents with different section chunks apart from the same section chunks.

Table 4 show the results of the first experiment and Table 5 show the results of second experiment.

The analysis of the results from Table 4 shows that when Prior Art chunk similarity is very high across documents, it was observed that those patents are from the same inventor. If the Claim is having high similarity score then the Patents are observed to be similar.

Table 4. Similarity analysis when compared across documents against same section chunks

Type of Chunks from section used for similarity measure	Total No. of Chunks	No. of similar chunks identified across documents	Actual No. of similar Chunks	No. of similar chunks correctly identified across documents	Recall (%)	Precision (%)
Abstract	67	5	15	5	33.33	100.00
Claim	68	33	55	29	60.00	87.88
Prior Art	58*	12	30	9	40.00	75.00
Summary	48*	30	45	22	66.67	73.33
Description	68	19	35	12	54.28	63.16
Average					50.856	79.87

* Some patents do not have Prior Art and Summary as separate sections.

In Table 4, we observe that the recall for the *abstract* section is 33.33%. The *abstract* section of the documents has maximum word limit constraint and hence authors would use optimum number of words to describe their invention. Therefore the similarity score obtained would be lower due to less number of common words. The similarity score does not exceed the threshold fixed by us. The threshold for the similarity is deduced by considering all the sections of the documents and not for a specific section. Hence the actual chunks which are similar also get filtered out, this reduces the recall. If the threshold can be

Table 5. Similarity analysis when compared across documents against all section chunks

Type of Chunks from section used for similarity measure	Total No. of Chunks	No. of similar chunks identified across documents	Actual No. of similar Chunks	No. of similar chunks correctly identified across documents	Recall (%)	Precision (%)
Abstract	67	17	40	13	42.50	76.47
Claim	68	42	55	36	76.36	85.71
Prior Art	58*	25	50	17	50.00	68.00
Summary	48*	37	48	30	77.08	81.08
Description	68	45	65	32	69.23	71.11
Average					63.03	76.47

* Some patents do not have Prior Art and Summary as separate sections.

calculated for each section, this reduction in recall could be taken care. In Table 5 also, we find a similar phenomenon happening for the *abstract* section.

In Table 4, we observe that the precision is 63.16 for *description* section. This can be attributed to the sense disambiguation problem. This section describes several electronic or electrical parts used in the devices invented. Even the names used for the parts are same for example "bus", they do not refer to the same real world entity. Hence they are different. Here the sense is domain specific and we would require domain specific thesaurus to disambiguate. From Table 4, we can obtain how many documents are similar, with respect to each section of the document. For example, the similarity analyser gives 5 documents to be similar with respect to *abstract*. From Table 5, we observe that the documents are highly inter-related across the sections. We can infer that the documents describe about devices, which have very similar usage by the end user, but their built and manufacturing is different. For example the documents describe about a connector device used to connect two fiber optic cables. Here the connectors are manufactured differently.

5 Conclusion

We have used coherent chunks in a document to identify the similarity between the different sections in a patent and across patents. Five major sections from patents are taken for similarity analysis and compared with 68 patents. We observe that identifying coherent chunks in the documents and then using the coherent chunks for identifying the similarity gives more accurate results. We obtain precision of 91.99% and a recall of 97.14% for the identification of coherent chunks. Similarity analyzer gives an average precision of 79% and an average recall of 63%.

Acknowledgements. We would like to thank Vijay Sundar Ram for all the inputs related to the analysis of the results.

References

1. Bagga, A., Baldwin, B.: Entity-Based Cross-Document Coreferencing Using the Vector Space Model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (COLING-ACL 1998), pp. 79–85 (1998)
2. Brill, E.: Some Advances in transformation Based Part of Speech Tagging. In: Proceedings of the Twelfth International Conference on Artificial Intelligence (AAAI 1994), Seattle, WA (1994)
3. Cohen, R.: A Computational Model for the Analysis of Arguments. Ph.D. Thesis and Tech. Rep. 151, University of Toronto (1983)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database, pp. 1–12. MIT Press, Cambridge (1998)
5. Frakes, W.B., Baeza-Yates, R. (eds.): Information Retrieval, Data Structure and Algorithms. Prentice Hall, Englewood Cliffs (1992)

6. Grosz, B.J., Joshi, A.K., Weinstein, S.: Providing a Unified Account of Definite Noun Phrases in Discourse. ACL, June 1983, pp. 44–50. MIT Press, Cambridge (1983)
7. Gruber, T.R.: A translation approach to portable ontologies. Knowledge Acquisition 5(2), 199–220 (1993)
8. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of ACL 1990, pp. 268–275 (1990)
9. Hobbs, J.R.: On the Coherence and Structure of Discourse. In: Polanyi, L. (ed.) The Structure of Discourse, Ablex Publishing Corporation, Greenwich (1985); Forthcoming. Also: CSLI (Stanford) Report No. CSLI-85-37, October (1985)
10. Lee, J.H., Kim, M.H., Lee, Y.J.: Information retrieval based on conceptual distance in is-a hierarchies. Journal of Documentation 49(2), 188–207 (1989)
11. Lin, D.: An Information-Theoretic Definition of Similarity. In: Proceedings of International Conference on Machine Learning, Madison, Wisconsin (July 1998)
12. Kohonen, T., Kaski, S., Lagus, K., Salojarvi, J., Honkela, J., Paatero, V., Saarela, A.: Self organisation of a massive document collection. IEEE Transactions on Neural Networks 11(3), 574–585 (2000)
13. McGill, M.: An evaluation of factors affecting document ranking by information retrieval systems. Project report, Syracuse University School of Information Studies (1979)
14. McKeown, K.R.: Generating Natural Language Text in Response to Questions about Database Structure. PhD Thesis, University of Pennsylvania, Philadelphia (1982)
15. Ngai, G., Florian, R.: Transformation-Based Learning in the Fast Lane. In: Proceedings of the NAACL 2001, Pittsburgh, PA, pp. 40–47 (2001)
16. Rada, R., Mili, H., Bicknell, E., Blettner, M.: Development and application of a metric on semantic nets. IEEE Transaction on Systems, Man, and Cybernetics 19(1), 17–30 (1989)
17. Rao Pattabhi, R.K., Sobha, L., Bagga, A.: Multilingual cross-document coreferencing. In: Proceedings of 6th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC), Portugal, pp. 115–119 (2007)
18. Rauber, A., Merkl, D.: The SOMLib digital library system. In: Abiteboul, S., Vercoustre, A.-M. (eds.) ECDL 1999. LNCS, vol. 1696, pp. 323–341. Springer, Heidelberg (1999)
19. Reichman, R.: Conversational Coherency. Cognitive Science 2(4), 283–328 (1978)
20. Reichman-Adar, R.: Extended Person-Machine Interfaces. Artificial Intelligence 22(2), 157–218 (1984)
21. Resnik, P.: Using information content to evaluate semantic similarity in taxonomy. In: Proceedings of IJCAI, pp. 448–453 (1995)
22. Salton, G.: Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, Reading (1989)
23. Tversky, A.: Features of similarity. Psychological Review 84, 327–352 (1977)

Binding without Identity: Towards a Unified Semantics for Bound and Exempt Anaphors

Eric Reuland and Yoad Winter*

Utrecht Institute of Linguistics OTS,
Janskerkhof 13, 3512 BL Utrecht, The Netherlands
{y.winter,e.reuland}@uu.nl

Abstract. Expressions such as English *himself* are interpreted as locally bound anaphors in certain syntactic environments and are exempt from the binding conditions in others. This article provides a unified semantics for *himself* in both of these uses. Their difference is reduced to the interaction with the syntactic environment. The semantics is based on an extension of the treatment of pronominals in variable-free semantics. The adoption of variable-free semantics is inspired by the existence of proxy-readings, which motivate an analysis based on Skolem functions. It is explained why certain anaphor types allow proxy-readings whereas others do not.

Keywords: Anaphors, Exemption, Proxy-readings, Skolem functions, Variable-free semantics.

1 Introduction

One of the intriguing properties of the English anaphoric system is that members of one and the same class of elements –*himself*, and the other members of its paradigm – must be locally bound (they are subject to condition A of the binding theory) in one set of environments, and is exempt from this local binding requirement in other environments. In such environments the antecedent need not be local. In certain cases, a linguistic antecedent may even be absent and a 'logophoric' interpretation obtains. This contrast has been discussed, among others,in [9], [12,13] (henceforth R&R) and [15,16,18]. A typical set of environments where the contrast shows up is given in (1) and (2):

(1) **Alice* expected [the king to invite *herself* for a drink]

(2) a. *Alice* expected [the king to invite [the Rabbit and *herself*] for a drink]
b. *Alice* expected [the king to invite [no one but *herself*] for a drink]

* The names of the authors appear in alphabetical order. The authors would like to thank Jakub Dotlačil and Anna Volkova for their help in providing the Czech and Russian facts.

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 69–79, 2009.

(1) illustrates the canonical case of a condition A violation. *Herself* is an anaphor and must be bound in its local domain, roughly the minimal clause containing it (see [2] for a precise statement of the canonical binding theory). *Alice* is the only potential antecedent of *herself*, but is outside the latter's binding domain. Hence, (1) is ill-formed. However, in (2) *Alice* is even farther away from the anaphor than in (1), yet here *Alice* can serve as an antecedent for *herself*, and these sentences are well-formed. This is problematic for the canonical binding theory, not only technically, but also conceptually. As R&R show, these and other facts - for instance, the differences in distribution between simplex anaphors (henceforth SE-anaphors) and complex anaphors (SELF-anaphors) like Dutch *zich* and *zichzelf* and their cognates in other languages – follow if conditions A and B are essentially seen as conditions on predicates and stated as follows:

(3) *Conditions:*
A: A reflexive-marked syntactic predicate is reflexive
B: A reflexive semantic predicate is reflexive-marked

These conditions are based on the following definitions (from R&R):

(4) *Definitions:*
a. The *syntactic predicate* of (a head) P is P, all its syntactic arguments and an external argument of P (subject)
The *syntactic arguments* of P are the projections assigned Θ-role or Case by P[1]
b. The *semantic predicate* of P is P and all its arguments at the relevant semantic level
c. A predicate is *reflexive* iff two of its arguments are bound by the same λ-operator[2]
d. A predicate (of P) is *reflexive-marked* iff either P is lexically reflexive or one of P's arguments is a SELF-anaphor

For the moment we will focus on condition A, and the definitions in (4a, c, and d). It is easily seen that in (1) *herself* is a syntactic argument of the predicate formed of *invite*. Therefore, it reflexive-marks it. Condition A, then, requires the predicate to be reflexive. This requirement cannot be met due to a feature mismatch between *the king* and *herself*, hence the sentence is ruled out. In (2a,b) *herself* is not a syntactic argument of *invite*. Rather, it is properly contained in one (*the Rabbit and herself* and *no one but herself*, respectively). Consequently, condition A does not apply, and the predicate is not required to be reflexive. Here the anaphor is *exempt*, to use Pollard and Sag's term. Hence, no violation of condition A ensues. Where syntactic principles do not enforce

[1] The reference to P includes P's extended projection in the sense of [4].
[2] The original definition is stated in terms of coindexing. *Being bound by the same operator* follows the definition of binding in [11]. Note that this is not strictly speaking compatible with the variable-free approach to be adopted in the body of this article, but for the present purpose this can be ignored.

an interpretation as a reflexivizer, its eventual interpretation will be determined by general semantic and discourse principles. As is shown extensively by Pollard and Sag, *himself/herself* may end up being bound by a higher c-commanding antecedent, and, if none is available by a suitably prominent discourse entity, receive a logophoric interpretation.

Summarizing, condition A expresses that the SELF- anaphor enforces reflexivity of the predicate in (1), but not in (2). The questions we need to address concern the way in which condition A is syntactically implemented, and how this affects the semantic interpretation.

2 The Syntax of Reflexive-Marking

In a parsimonious theory of grammar, locality of binding should follow from the same general principles that give rise to locality in other domains. Ideally, the grammar should contain no statements specific to binding, except for a definition of binding itself. Similarly, the interpretation of anaphors in various syntactic contexts should be determined by the same semantic principles applying in a uniform way. That is, the semantics of anaphoric expressions should be as general as possible.

R&R [13] do not discuss a specific syntactic mechanism for reflexive-marking. R&R [12], however, suggested that the mechanism involves covert syntactic movement, with SELF moving onto the predicate head by head-movement. In line with earlier approaches such as [6], we assume that *himself* is syntactically complex, with SELF being an N projecting an NP, and *him* occupying a position in the left periphery as in (5a), where F is a functional projection, which we can assume to be Person.[3] When inserted in the proper configuration, SELF can be attracted by the head of the predicate, and adjoin to the latter, as in (5b), which is transparently reflected in nominalizations such as *self-hatred* or *self-admiration*.

(5) a. $[_{FP}$ him $[_{NP}$ SELF]]
b. DP $[_{VP}$ $[_{V}$ SELF V] $[_{DP}$ him $[_{NP}$ (SELF)]]]

This idea is elaborated in [15,16,18]. It is easy to see that the syntactic conditions on exemption follow without further stipulation if the relation between SELF and the predicate is indeed covert syntactic movement. If so, it is sensitive to standard island conditions on movement, specifically head-movement. In (2a) moving SELF onto *invite* would violate the coordinate structure constraint, in (2) the adjunct island constraint (independently of whether these constraints can be reduced to more fundamental principles of grammar). Hence SELF cannot

[3] R&R assume that *him* is in the D-position. For reasons discussed in [19] its position is in a functional projection below D, but for present purposes this issue can be put aside.

move onto the predicate in (2a,b) and require it to be reflexive, which explains the exemption.[4]

A crucial claim of this approach is that English has only one expression *himself*. And indeed, this is the most parsimonious way to derive the complementarity of exempt and bound uses of *himself*. However, this claim naturally leads to the question of its semantics. Can we find a unified semantics of *himself* covering both its bound and its exempt uses? The matter is discussed in [12] and subsequently, in [16,18]. The idea pursued there is that SELF is a predicate expressing identity. When adjoined to the predicate head SELF's extension is intersected with the extension of the predicate head, which intuitively conveys the intended meaning. It is less clear what happens in exempt positions. Since SELF is an identity predicate, one of its arguments will be the pronominal. The other argument will have to pick up its value from the context. However, it is not trivial to express the semantics in a compositional way. In the next section, we will present a compositional semantics for *himself*. As we will show, this semantics is extendable to different types of reflexive markers. Many languages, for instance, use body-part reflexives. As we will see, our approach naturally applies to such elements.

The basis for our semantics, however, is provided by the solution to another puzzle, which we will introduce first.

3 Binding and Proxies

Our treatment of the semantics of *himself* is inspired by one of the well-known properties of reflexive pronouns: their ability to have "proxy readings". This is illustrated in (6) [7]:

(6) (Upon a visit in a wax museum:) All of a sudden Ringo started undressing himself.

Himself in (6) can refer to the "real" Ringo, but also to a statue of the Ringo denoted by the subject. As Jackendoff argues, the availability of proxy interpretations of reflexives (6) must be related to a general property of language: the ability to refer to various "proxies" of an individual concept. In that respect, the reflexive in (6) is not different from non-anaphoric NPs, which can also refer to "non-canonical" proxies (cf. *Ringo/the man is made of stone, whereas Yoko/the woman is made of wax*) (see also [20] for pertinent discussion).

Jackendoff has argued that there is an asymmetry between NPs and anaphors in their ability to carry a proxy reading, and claims that in (7) we cannot have an interpretation where *Ringo* is the proxy and *himself* the person.

(7) *Ringo* fell on top of *him*self

[4] It would lead us beyond the scope of the present article to go over all the cases of exemption. See [13] for detailed discussion. The reduction of reflexive-marking to SELF-movement has a range of non-trivial empirical consequences, which we cannot possibly go into here. They are discussed in detail in [19].

However, it is not at all difficult to create contexts where such an interpretation is easily accessible. Consider a play where some actor plays a younger Ringo, and Ringo plays an older fan. It is no problem to interpret the sentence *Ringo stumbled and fell on top of himself* as true when the actor stumbled and fell on top of the real Ringo. Thus, the availability of proxy-readings represents a general property of expressions for individual concepts. Hence the following generalization is expected to hold:

(8) Generalization: *The range of available proxies for a bound pronoun is the same range of proxies as for its antecedent.*

Thus, while strict identity between the referents of a pronoun and its antecedent is not mandatory even under binding, identity of the candidate proxy referents for the two expressions is mandatory. This generalization reflects the following observation: non-reflexive bound pronouns allow a proxy interpretation. For instance:

(9) All of a sudden, every pop icon started taking off the shirt he was wearing.

In the wax museum context of (6), sentence (9) has a bound reading where the pop icons took the shirts off their respective statues.

This leads to the question of the proper semantics of pronouns. In what one may call the standard analysis of pronouns and anaphors, as summarized in for instance [5], pronouns and anaphors are essentially variables. Their interpretation is given on the basis of assignment functions.

The fact that pronouns have proxy readings does not come naturally in the standard analysis of pronouns as variables. However, we will show that it is naturally accommodated in an extension of [8] variable-free semantics. Her approach to pronouns dispenses with assignment functions, and also with indices. Pronominals are interpreted as the identity function on entities. It is this conception that provides the basis for generalizing over the bound and exempt uses of *himself*.

To capture the availability of proxy-readings as in sentence (9), we propose the following natural modification in Jacobson's use of functions. Non-reflexive pronouns like *he*, instead of simply denoting the identity function on entities, as in [8], denote a *Skolem function*: a function from entities to entities that takes a relation as a parameter. The formal definition is given in (10).

(10) *A function f of type* (ee) *with a relational parameter R is a* ***Skolem function*** *if for every entity x:* $R(x, f_R(x))$ *holds.*

We propose that the context provides a *proxy relation* (PR),describing the possible proxies $\lambda y.PR(x,y)$ of any entity x referred to. This parameter determines the range for each possible entity argument of the Skolem function. We stipulate that any proxy relation must be *reflexive*. This guarantees availability of the standard interpretation, also in cases like (6) and (9), where the referents

for the pronoun and its antecedent entity are strictly identical. This happens because when the relation R is reflexive, one of the Skolem functions f_R is the identity function. Thus, our account generalizes Jacobson's use of functions from entities to entities. Sentence (9) is now analyzed as (9'):

(9') $\forall x[pop\_icon(x) \rightarrow take\ off\ (x, the\ shirt\ \_f_{PR}(x)\_\ was\ wearing)]$

Thus, for every pop icon x, the Skolem function f_{PR} picks up one of x's proxies in the set $\lambda y.PR(x,y)$, possibly x itself. Deriving this analysis is straightforward within Jacobson's framework.

4 Binding of SELF-anaphors

What do these considerations imply for reflexive pronouns, in particular SELF-anaphors? As indicated in (5), we decompose the anaphor *himself* into a pronoun *him*, and *self*. Since pronominals need not be bound, the relative binding requirement of *himself* must reside in the *self*-part.We treat English *self* (as its cognates in other languages) as a relational noun, denoting a relation between entities and their proxies (with the identity relation as the limiting case). This requirement amounts to assuming that *self* denotes a reflexive relation: an entity x can have more than one "self" in addition to x. In the decompositional semantics of *herself*, *self* replaces the contextual proxy-relation of the bare pronoun *her*.

A noun phrase like *Ringo's better self* is not substantially different from any other NP with a relational noun (e.g. *Ringo's better parent*), where the former NP may refer to one of Ringo's "better" proxies in the context of utterance. As noted in section 2, *self* can incorporate [3] with nouns and nominalized transitive verbs. In this, it is similar to other relational nouns. For instance,

(11) a. *self-hater* denotes the predicate $\lambda x.hate(x, \uparrow self\ (x))$
(x is a self-hater if x hates the property (indicated by the $\uparrow$- operator) coupled with x's proxies)
b. *parent-hater* denotes the predicate $\lambda x.hate(x,\uparrow parent(x))$
(x is a parent-hater if x hates the property coupled with x's parents)

The only substantial difference we assume between *self* and other relational nouns is a syntactic one. The noun *self* is able to combine with Skolem functions denoted by non-reflexive pronouns independently of genitive case (viz. *his self/himself* vs. *his parent/*him parent*). There are two ways in which this can happen:

i. The unmarked option – the noun *self* composes with the Skolem function denoted by the pronoun through the binding mechanism. The noun *self* covertly incorporates into the transitive predicate (as happens overtly in *self-hater*) and contributes a proxy relation to the non-reflexive pronoun through Jacobson's Z function in its "proxied" version:

(12) $Z^{PR} = \lambda$R.λf.λx.R $(x, f_{PR}(x))$

In this version of the Z function, it provides the Skolem function f with its parameter. The denotation of a VP like *undress himself* in (6) is obtained using the structure *self-undress him*, analyzed as in (13):

(13) Z^{self} (*undress*)(*him*)= Z^{self} (*undress*)(f)
= λx. undress(x, $f_{self}(x)$) = $\lambda x.x$ undressed one of x's self proxies (by definition of f as a Skolem function)

ii. A marked option – the noun *self* composes with the Skolem function directly. We assume that this marked option is only available in exempt positions, when the incorporation with the predicate is syntactically blocked, as discussed in section 2. e.g. *Max boasted that the queen invited [Lucie and himself] for a drink.* When formation of self-V is syntactically disallowed (as for instance by the Coordinate Structure Constraint), direct composition with the Skolem function leads to the analysis in (14):

(14) himself = f_{self} = a function mapping every entity x to one of its proxies in *self*(x)

Unlike what happens in the unmarked option, now there is no binding that is made necessary by *self*'s composition. As a result, the exempt reading of *himself* allows it to be interpreted as either bound or free, similarly to the non-reflexive pronoun *him*.

Hence, *self* has the same semantics in both cases. The difference resides in how the instruction associated with its semantics is applied. This, in turn, is determined solely by the syntactic context. The crucial advantage of Jacobson's approach is that it makes available an argument for *self* that has the proper type in both the bound and the exempt case.

Remains the question of why the option with *self*-movement is the unmarked case. In [15,16], it is argued that the simplest reason resides in a general economy principle, to the effect that encoding binding dependencies in the syntax is cheapest, hence preferred. For current purposes this suffices, see [19] for more extensive discussion.

5 Simplex Anaphors and Proxy-Interpretation

Dutch (like the Scandinavian languages) has two anaphors, a SE-anaphor *zich*, and a SELF-anaphor *zichzelf*. *Zich* is an anaphor in the sense that it must be bound (although as its locality condition is less strict, see [17] for discussion). R&R [13] analyze it as a pronominal that is under-specified. It is not specified for *gender* and *number*, but only for the feature 3^{rd} *person*. Given this, one would expect that it allows proxy-readings like any pronominal. However, there is a clear contrast between the following sentences:

(15) a. Ringo begon zich te wassen.
"Ringo started to wash" *no proxy reading*

b. Ringo begon zichzelf te wassen.
"Ringo started to wash himself" *proxy reading possible*

That is, again in the situation of the wax museum, suppose Ringo finds his statue dirty and wants to do something about it, we can say (15b), but not (15a). What is the source of the difference between the pronouns *zich* and *zichzelf*? As in the case of exemption discussed above, we would like to find the answer in the syntactic environment, keeping the semantics uniform.

We will relate this to the lexical status of reflexives, coming back to the definition in (1d). A predicate can be reflexive-marked either extrinsically (by a SELF-anaphor), or lexically. Reinhart [10,14] present a theory of operations on argument structure (for details we refer to the works cited). One of the options these works allow is a lexical operation, reducing the internal argument and bundling its thematic role with the role of the external argument, as in (16):

(16) Internal Reduction/Bundling R_s:
a. V_{acc} (θ_1, θ_2) $\rightarrow$ R_s(V) ($\theta_{1,2}$) (where $\theta_{1,2}$ stands for the bundling of θ_1 and θ_2)
b. R_s (V)($\theta_{1,2}$) $\longleftrightarrow$ $\theta_{1,2}$ (λx (V (x,x)))

The operation is available for a subclass of transitive/accusative assigning verbs, including verbs such as *wash*, *shave*, etc. Thus, English *wash* has two related entries, one intransitive and inherently reflexive in *John washed*, the other transitive in *John washed Mary*/*John washed himself*. The same contrast applies to Dutch. One of the cross-linguistic variables is how the reduction operation affects accusative Case. In languages of the English type the reduction also affects accusative Case. The reduced entry is not an accusative Case assigner, hence no further operation is necessary. In Dutch, with a somewhat richer Case system than English, structural accusative is preserved under reduction. Hence an element has to be inserted checking this Case. Crucially, the element to be inserted should not be interpreted as an independent syntactic argument. For reasons discussed in [15], *zich*'s feature deficiency allows it to form one syntactic object with the subject, technically a *chain*.

Thus, our analysis of the chain in the intransitive usage of *waste* ("washed") in (15a) yields the interpretation in (17):

(17) || [Jan,zich] || = f_{PR}(jan) = one of Jan's proxies.

That is, the proxy-function interpreting *zich* "skips" the predicate *waste*, and applies directly to the subject. The resulting interpretation is indistinguishable from the "simple" denotation "jan" of the name *Jan*, given the generalization (8) that any referential NP can be interpreted as any member of the relative set of proxies. Thus, interpreting *zich* by a proxy-function meshes well with the syntactic structure, without further assumptions being necessary.

By contrast, in (15b), similarly to (6), the reflexive pronoun fills in a separate (object) argument position of a transitive verb (here, the transitive reading of

waste). As a result, the analysis of (15b) is similar to the binding with the English sentence *Jan washed himself*.

So, *zich* reflects what one may call *strict binding*. This tallies with the fact that in intentional contexts *zich* only allows a *de se* interpretation. In terms of Cherchia's [1] discussion, *Pavarotti zag zich in de film, maar realiseerde zich niet dat hij het zelf was* "Pavarotti saw SE in the movie, but didn't realize it was he himself" has the air of contradiction, but the result of replacing *zich* by *zichzelf* is fine. For completeness sake, note that due to the defective nature of *zich*, the complex form *zichzelf* is always bound. For details, we refer to [19].

Lexical reflexivization as in Dutch or English is limited to a subclass of transitive verbs. In other languages, bundling applies productively. As argued by Reinhart [10,14], in such languages bundling applies in the syntax. That is, verbs to be reflexivized project two syntactic arguments. The class of languages of this type includes French, Italian, Czech and others. Interestingly, bundling in the syntax is compatible with the availability of proxy readings. Russian, which has only restricted lexical reflexivization marked by the *s'a*-affix and Czech form a nice minimal pair as illustrated in (18) and (19).

(18) a. nedavno, posetivšij muzej, Ringo pomyls'a
(=Ringo, *statue)
recently, having visited the museum, Ringo washed-aff
b. nedavno, posetivšij muzej, Ringo pomyl seb'a
(=statue, ?Ringo)
recently, having visited the museum, Ringo washed himself

(19) a. Ringo se začal prohlížet (=statue, Ringo)
Ringo started to look at himself
b. Ringo mluvil o svém vzhledu (=statue, Ringo)
Ringo talked about his appearance

This contrast follows if syntactically projected argument positions have the same semantic status as pronouns.

6 Extending the Approach

Many languages have yet different strategies of reflexivization. In the language sample studied in [21], the most frequent reflexivization strategy used so-called body-part (BP) reflexives, as for instance in Basque (20) which uses the expression *his head* as an anaphor.

(20) a. aitak bere burua hil du
father+ERG 3SGPOSS head+NOMDEF kill have+3SG+3SG
The father killed himself
b. bere buruan txapela ipini du
3SGPOSS head+LOCDEF cap+NOM put have+3SG+3SG
He put the cap on his head

As illustrated in (20b) this expression can still be used in its literal meaning as well. The question is how the reflexivizing use of *his head* can be understood.

What body-part expressions have in common with *self* is that they are inherently relational. Just like any *self* is some individual's self, a body-part belongs to some individual's body. Pursuing the analysis established in section 4, we will claim that the head of the BP is able to combine with Skolem functions denoted by the non-reflexive pronoun in its POSS position (null or overt). If so, the denotation of a VP like *V PronBP* is obtained using the structure *BP-V him*, which is analyzed just like *undress himself* in (6). The relevant interpretation is, therefore, given in (21):

(21) Z^{BP} (V)(Pron BP) = Z^{BP} (V)(f) = λ x. V(x, f_{BP}(x))
= λ x.x V-ed one of x's body's proxies (by definition of f as a Skolem function)

It is an intriguing question to what extent and under what conditions body-part anaphors are subject to similar exemption effects as English SELF-anaphors. This is a matter for further investigation.

7 Conclusion

Our extension of variable-free semantics allows us to naturally accommodate proxy-readings. It generalizes over proxy-readings for pronominals and anaphors. It allows us to unify the semantics of bound and exempt anaphors, and it provides a natural extension from SELF-anaphors to body-part reflexives. Finally, it allows us to unify the semantics of *zich* where it tails a chain to check a residual case with the general semantics or pronouns. It provides us with a principled means to further investigate the cross-linguistic parameters determining the availability of proxy-readings.

References

1. Chierchia, G.: Anaphora and attitudes de se. In: Bartsch, R., van Benthem, J., van Emde Boas, P. (eds.) Semantics and contextual expression. Foris, Dordrecht (1989)
2. Chomsky, N.: Lectures on Government and Binding. Foris, Dordrecht (1981)
3. Geenhoven, V., Veerle, M.L.: On the property analysis of opaque complements. Lingua 115 (2005)
4. Grimshaw, J.: Extended projections. Ms. Brandeis University (1991)
5. Heim, I., Kratzer, A.: Semantics in Generative Grammar. Blackwell Publishers Ltd., Oxford (1998)
6. Helke, M.: On reflexives in English. Linguistics 106, 5–23 (1973)
7. Jackendoff, R.: Mme. Tussaud meets the Binding Theory. Natural Language and Linguistic Theory 10, 1–33 (1992)
8. Jacobson, P.: Towards a variable-free semantics. Linguistic and Philosophy 22, 117–184 (1999)

9. Pollard, C., Sag, I.: Anaphors in English and the scope of the Binding theory. Linguistic Inquiry 23, 261–305 (1992)
10. Reinhart, T.: The Theta System - an Overview. Theoretical Linguistics 28(3), 229–290 (2002)
11. Reinhart, T.: Interface Strategies: Optimal and Costly Computations. MIT Press, Cambridge (2006)
12. Reinhart, T., Reuland, E.: Anaphors and Logophors: An Argument Structure Perspective. In: Koster, J., Reuland, E. (eds.) Long Distance Anaphora, pp. 283–321. Cambridge University Press, Cambridge (1991)
13. Reinhart, T., Reuland, E.: Reflexivity. Linguistic Inquiry 24(4), 657–720 (1993)
14. Reinhart, T., Siloni, T.: The Lexicon-Syntax Parameter: Reflexivization and Other Arity Operations. Linguistic Inquiry, 389–436 (2005)
15. Reuland, E.: Primitives of Binding. Linguistic Inquiry 32(2), 439–492 (2001)
16. Reuland, E.: Binding Conditions: How are they Derived? In: Müller, S. (ed.) Proceedings of the 12th International Conference on Head-Driven Phrase Structure Grammar, Stanford CSLI (2005a)
17. Reuland, E.: Long Distance Anaphors in Germanic Languages. In: Everaert, M., van Riemsdijk, H. (eds.) The Syntactic Compendium. Blackwell, Oxford (2005b)
18. Reuland, E.: Anaphoric dependencies: How are they encoded? Towards a derivation-based typology. In: König, E., Gast, V. (eds.) Reciprocals and Reflexives – Cross-linguistic and theoretical explorations, pp. 502–559. Mouton de Gruyter, Berlin (2008)
19. Reuland, E.: Anaphora and Language Design Cambridge. MIT Press, MA (Under contract)
20. Safir, K.: The Syntax of Anaphora. Oxford University Press, Oxford (2004)
21. Schladt, M.: The typology and grammaticalization of reflexives. In: Frajzyngier, Z., Curl, T. (eds.) Reflexives: Forms and Functions. Benjamins, Amsterdam (2000)

The Doubly Marked Reflexive in Chinese

Alexis Dimitriadis and Min Que

Utrecht institute of Linguistics OTS, Utrecht University
Janskerkhof 13, 3512 BL Utrecht, The Netherlands
{a.dimitriadis,m.que}@uu.nl

Abstract. We discuss an unusual reflexive construction in which the Chinese reflexive *ziji* appears twice, once before the verb and once after. We demonstrate that this is a distinct construct with its own rules of construal and interpretation; it is not, for example, a combination of a simple *ziji* reflexive and an adverbial intensifier. Notably, their locality properties are also different: Double *ziji* does not tolerate non-local readings. We argue that while *ziji* is (or can be) a logophor [1], double *ziji* is an ordinary Principle A anaphor with all the properties and restrictions that this implies.

Keywords: Reflexives, Chinese, Binding theory.

1 Introduction

The well-known Chinese reflexives *ziji* and *ta-ziji* are anaphors functioning as the internal argument of the reflexive, typically the object (1a).[1] But Chinese also allows an unusual variant, which to our knowledge has not been discussed in the theoretical linguistic literature to date, in which *(ta-)ziji* appears twice, preverbally as well as postverbally (1b,c).

(1) a. Lisi hen ziji / ta-ziji
Lisi hate self / 3sg-self
'Lisi hates himself'
b. Lisi ziji hen ziji.
Lisi self hates self
'Lisi hates himself.'
c. Q: What's the matter with John?
A: Ta(-ziji) hen ziji.
3sg-self hates self
'He hates himself.'

As the above examples show, the double *ziji* construction can co-occur with an overt subject. *Ta-ziji* can be used instead of *ziji,* in one or both positions in various combinations.

While *ziji* can be used prenominally as an intensifier (emphatic), we will show below that the construction in (1) is more than the simple co-occurrence of an intensifier and the ordinary reflexive *ziji.* The construction is unusual in that the reflexive is marked in two places (with an anaphor in object position and with

[1] For discussion of additional variants of *ziji* and their uses, see [2], [3].

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 80–90, 2009.

an adverbial modifier), a pattern of reflexive marking which is relatively rare but not unattested; for example, Kannada reflexives are marked by means of both verbal morphology and an anaphor in object position [4]. But the binding options for the double *ziji* construction are also different: Unlike simple reflexive *ziji,* double *ziji* is obligatorily locally bound. We will show that while simple *ziji* is a typical logophoric anaphor, double *ziji* is an ordinary anaphor and behaves as predicted by Principle A of binding theory.

2 Syntax of the Double-*ziji* Construction

As we have already seen, the double reflexive construction can be used in sentences either with or without an overt nominal subject. The following examples show that (a) a sentence can have a nominal topic doubled by a subject pronoun; (b-d) the *ta-ziji* form can be used pre- or post-verbally in various combinations, together with a nominal subject.

(2) Q: What's the matter with John?
   a. Yuehan ta hen (ta-)ziji.
      John he hate 3sg-self
      'John hates himself.'
   b. Yuehan hen (ta-)ziji
      John hate 3sg-self
   c. Yuehan ziji hen (ta-)ziji.
      John self hate 3sg-self
   d. Yuehan ta-ziji hen (ta-)ziji
      John 3sg-self hate 3sg-self

The double reflexive can also be used with a null subject, or impersonally.

(3) a. Q: What's the matter with John?
      Ta(-ziji) hen ziji.
      3sg-self hate self
      'He hates himself.'
   b. Q: What is John doing?
      Ziji da ziji.
      self hit self
      'He's hitting himself.'

(4) a. Buyao ziji hen ziji.
      Don't self hate self
      'Don't hate yourself' (Imperative)
   b. Ziji hen ziji shi buhao de.[2]
      self hate self be not.good DE
      'It's not good to hate one's self' (Impersonal)

[2] We "gloss" certain particles of Chinese as themselves, e.g., we gloss *de* as *DE, dou* as *DOU,* etc., since their analysis is both contested and irrelevant to our topic. Other non-obvious glosses used in this paper: CL = classifier; DEM = demonstrative; PRT = particle; Perf = perfective; Prog = progressive; PL = plural.

The first occurrence of *ziji* is not a subject, but a VP-adjoined ("adverbial") element appearing inside the verb phrase. This is evident since *ziji* appears to the right of auxiliaries and of the distributor *dou,* which marks the edge of the VP:

(5) Xuesheng-men dou ziji biaoyang-le ziji.
student-PL DOU self praise-Perf self
'The students each praised themselves'

Since *ziji* can in fact be used as an intensifier, we need to address the question of whether this construction might be combination of an ordinary reflexive and an ordinary intensifier, comparable in status to the following English example:

(6) Even John himself criticized himself.

Here the first instance of *himself* does not express any identity of participants, i.e., is not a reflexive, but is an "adnominal" intensifier. We follow the terminology of Gast [5] and classify intensifiers as *adnominal* and *adverbial,* depending on their syntactic attachment. Chinese *ziji* can have both functions:

(7) (Source: Daniel Hole, TDIR)[3]
a. "Adnominal" intensifier:
Buzhang ziji hui lai huanying women.
minister self will come welcome 1pl
'The minister himself will welcome us'
b. "Adverbial exclusive" intensifier:
Nei-wei mingxing bing mei you ziji xie tade zizhuan.
DEM-CL star PRT not have self write his/her autobiography
'The movie star did not write his autobiography himself'

Adverbial intensifiers, like the preverbal part of double *ziji,* appear after the distributive element *dou.* This means that we cannot easily distinguish the two on the basis of syntactic position alone. (Cf. example (5)).

(8) a. Xuesheng-men dou ziji zuofan.
student-PL DOU self cook
'Students cook by themselves (nobody else cooks for them)'
b. Xuesheng-men dou ziji dasao fangjian.
student-PL DOU self clean room
'Students clean their rooms by themselves (not by others)'
c. * Xuesheng-men ziji dou biaoyang-le Lisi.
student-PL self DOU praise-Perf Lisi
'The students praised Lisi by themselves'

We can show, however, that the double-*ziji* construction does not involve an intensifier. First, the meanings associated with intensifier uses of *ziji* are absent in a double-*ziji* example like (9), which does not mean "Zhangsan (by) himself praised himself." (Lisa Cheng, personal communication).

[3] TDIR is the Typological Database of Intensifiers and Reflexives [3].

(9) Zhangsan ziji biaoyang-le ziji
Zhangsan self praise-Perf self
'Zhangsan praised himself'
[Does not mean "Zhangsan (by) himself praised himself."]

Additionally, the double-ziji construction can be used in discourse contexts where an adverbial intensifier is ruled out, as in (A2) below; note that because the question is about Mulan, the intensifier in (A1) is ungrammatical.

(10) Q: Mulan zai gan shenme?
Mulan Prog do what
'What is Mulan doing?'
A1: Mulan zai (*ziji) mai tudou.
Mulan Prog self buy potato
'John (*himself) is buying potatoes.'
A2: Mulan zai (ziji) da-ban ziji.
Mulan Prog self dress.up self
'Mulan is getting dressed up.'

2.1 Transitivity

In classifying reflexive constructions, an important distinction is between those that involve an anaphor with reflexive meaning (as in English) and those that involve a verbal morpheme or adverbial that creates an intransitive predicate [6]. We will term the former *argument reflexives* and the latter *verbal reflexives*.[4] For our purposes the important distinction is not whether the exponent of reflexivity is morphologically bound to the verb, but whether the reflexive predicate involves a transitive verb (whose object is occupied by the reflexive anaphor) or an intransitive one. In some cases, morphologically free reflexives are in fact detransitivizing operators, and should be classified as verbal predicates. The French reflexive clitic *se,* for example, appears to be a cliticized pronoun, hence an argument reflexive; but as [7] already showed, on closer inspection it turns out to be a verbal detransitivizer.

(11) Jean se lave.
John self washes
'John washes'

Conversely, [8] shows that the reflexive morpheme *dzi* in Chichewa, although morphologically incorporated in the verb (where it appears between the verb root and the tense marker), is in fact an incorporated pronoun rather than a detransitivizer. The reciprocal suffix *-ana,* on the other hand, is a detransitivizer.

Since the double *ziji* construction involves an adverbial modifier, then, we consider whether the construction (as a whole) may act as a detransitivizer. We will show that infact it does not: Double-*ziji* reflexives are still syntactically transitive.

[4] [6] uses the name "NP reflexives" for the first category.

While there are numerous language-specific tests of transitivity, we use the object-comparative test of Zec [9], which has wide cross-linguistic applicability.[5] We first illustrate the test in English. Consider example (12), which is ambiguous between a subject comparison reading (irrelevant to our purposes) and the object comparison reading in (b).[6]

(12) John hates Bill more than George.
  a. Subject comparison (irrelevant to transitivity)
     John hates Bill more than George hates Bill
  b. Object comparison
     John hates Bill more than John hates George

If we construct a similar comparative with the reflexive *washes himself,* as in (13), the object comparison reading continues to be available. (Again we ignore the irrelevant subject comparison readings). But if we use the "covert reflexive" sentence *John washes,* as in (14), the object comparative reading disappears:

(13) John washes himself more than George.
  a. Subject comparison, strict or sloppy
     John washes himself more than George washes John/himself
  b. Object comparison: Shows that *washes himself* is transitive
     John washes himself more than he washes George

(14) John washes more than George.
  a. Subject comparison:
     John washes himself more than George washes himself.
  b. Object comparison: Impossible, showing that *washes* is intransitive.
     * John washes himself more he (John) washes George.

The reason is that object comparison requires a transitive antecedent (so that the properties of its object can be compared with the properties of *George*). The covert reflexive in (14) is evidently intransitive, and fails to give the object comparative reading. Equivalent results are found for the detransitivizing reflexives discussed above.

If we now apply this test to Chinese, we find that simple *ziji* reflexives, as well as double *ziji,* do not involve detransitivization. The object comparison reading is available with both of them.[7]

(15) **Transitives**
  Zhangsan hen Lisi bi Wangwu duo
  Zhangsan hate Lisi BI Wangwu more
  'Zhangsan hates Lisi more than Wangwu'
  a. ... more than Wangwu hates Lisi (subject comparison; irrelevant)
  b. ... more than Zhangsan hates Wangwu (object comparison)

[5] Zec's test was adapted to Chichewa by Mchombo [8].
[6] When applying this test to languages with morphological case, Accusative case on *George* may result in unambiguous object comparison.
[7] We thank Meiyi Bao for providing judgements.

(16) **Regular reflexives**
Zhangsan hen ziji bi Wangwu duo
Zhangsan hate self BI Wangwu more
'Zhangsan hates himself more than Wangwu'
Subject comparison (irrelevant to transitivity):
a. * ... more than Wangwu hates Wangwu (sloppy)
b. ... more than Wangwu hates Zhangsan (strict)
Object comparison: Shows that *hen ziji* is transitive
c. ... more than Zhangsan hates Wangwu

(17) **Double reflexives**
Zhangsan ziji hen ziji bi Wangwu duo
Zhangsan self hate self BI Wangwu more
'Zhangsan hates himself more than Wangwu'
Subject comparison:
a. * ... more than Wangwu hates Wangwu (sloppy)
b. ... more than Wangwu hates Zhangsan (strict)
Object comparison: *ziji hen ziji* is transitive
c. ... more than Zhangsan hates Wangwu

3 Locality Conditions

The best-studied aspect of the reflexive *ziji* are arguably the structural conditions on its acceptable antecedents. Simple *ziji* allows a range of long-distance and logophoric construals, as discussed in the following section. The double-ziji construction contrasts markedly with ordinary ziji reflexives.

3.1 Background: Locality and Long-Distance Anaphora with *ziji*

In this short paper we focus on understanding of the double *ziji* construction; for the other Chinese anaphors we will take as our starting point the analysis of Huang and Liu [1], who give a nice summary of the literature concerning their patterns of locality and construal.

Chinese is generally acknowledged to have two reflexive anaphors based on *ziji*: The invariant anaphor *ziji* 'self', and *ta-ziji* 'himself/herself', which shows person and number agreement. *Taziji* is, broadly speaking, a normal Principle-A anaphor; it must be locally bound. *Ziji* allows long-distance and "logophoric" construals. This is shown in example (18a).[8] The antecedent of *ziji* need not be the subject of the main clause, nor does it need to be in the clause immediately dominating the clause where *ziji* appears (example (18b)).

(18) a. *Long-distance readings:*
Zhangsan$_Z$ renwei [Lisi$_L$ hen ziji$_{Z/L}$ / ta-ziji$_{*Z/L}$]
Zhangsan think Lisi hate self / 3sg-self
'Zhangsan$_Z$ thinks Lisi$_L$ hates himself$_L$ / him$_Z$'

[8] The examples in this section are from [1], unless otherwise noted.

b. Zhangsan$_Z$ renwei Lisi$_L$ zhidao [Wangwu$_W$ hen ziji$_{Z/L/W}$]
Zhangsan think Lisi know Wangwu hate
'Zhangsan thinks Lisi knows that Wangwu hates Zhangsan/Lisi/Wangu'

Two other well-studied properties of long-distance *ziji* are subject orientation (19a) and its susceptibility to so-called *blocking effects*.[9] As example (19b) shows, the presence of a potential antecedent with contrasting person features will block coreference with a compatible, but more distant antecedent.

(19) a. *Subject orientation:*
Zhangsan$_Z$ song (gei) Lisi$_L$ yi-zhang ziji$_{Z/*L}$-de xiangpian.
Zhangsan give to Lisi one-CL self-DE picture
'Zhangsan$_Z$ gives Lisi$_L$ a picture of himself$_{Z/*L}$.'
b. *Blocking effects:*
Zhangsan$_Z$ renwei [ni$_Y$ hen ziji$_{*Z/Y}$]
Zhangsan think 2sg hate self
'Zhangsan thinks that you hate yourself.'

For completeness, we mention here that the antecedent of *ziji* need not be overtly present. *Ziji* can also refer to the speaker, or to other sufficiently prominent discourse participants:

(20) *Reference to the speaker:*
Zhe-ge xiangfa, chule ziji, zhiyou sang-ge ren zancheng.
this-CL idea besides self only three-CL people agree
'As for this idea, besides myself, only three other people agree.'
([11]/[12], cited in [1])

3.2 Double *ziji* Is Not a Long-Distance Anaphor

When we consider the allowable construals of the double-*ziji* construction, we find a very different pattern: The subject and object of the reflexive predicate (*da* 'hit' in the following) are obligatorily coreferential. The readings of sentence (21) are fairly straightforward: the antecedent of the reflexive can only be Lisi. In sentence (22), however, we have more construal options: this example might describe situations in which the hitter was Zangsan, Lisi, or even a third person; but in all cases the hitter must be hitting himself (or herself).

(21) Zhangsan$_Z$ renwei Lisi$_L$ ziji da-le ziji$_{*Z/L}$
Zhangsan think Lisi self hit-Perf self
'Zhangsan thinks Lisi$_L$ hit himself$_L$'
(22) Zhangsan$_Z$ renwei Lisi$_L$ zhidao ta-ziji da-le ziji
Zhangsan think Lisi knows 3sg-self hit-Perf self
'Zhangsan$_Z$ thinks Lisi$_L$ knows that [Zhangsan/Lisi/X hit himself].'
Ok: Z hit Z / L hit L / X hit X;
Bad: *Z hit L / *L hit Z / *X hit Z / etc.

[9] See [10] for detailed discussion.

The reason is not some sort of unusual long-distance anaphora: In example (22), *ta* is apparently a pronoun rather than part of the reflexive; it can take any referent suitable for a pronoun in this position, but in each case the predicate *hit* must be reflexively construed.[10] This is also supported by the fact that it is possible to insert a pause after the pronoun *ta.*

For comparison, we provide the readings of the equivalent simplex reflexive. The pronoun can be bound or unbound, and the reflexive takes the usual (well-documented) local or long-distance readings.[11]

(23) Zhangsan$_Z$ renwei Lisi$_L$ zhidao ta da-le ziji
Zhangsan think Lisi knows he hit-Perf self
'Zhangsan$_Z$ thinks Lisi$_L$ knows [Zhangsan/Lisi/X hit Z/L/himself]'
(All combinations ok, except apparently for *'Zhangsan hit Lisi')

Our interpretation is also supported by the fact that such examples behave as if immmune to blocking effects: Each of the following sentences can be about any available referent compatible with the phi-features of the pronoun, as long as the most embedded predicate is reflexive.

(24) a. Zhangsan$_Z$ renwei wo$_i$ zhidao ta ziji da-le ziji.
Zhangsan think 1sg know 3sg self hit-Perf self
'Zhangsan thinks I know (Z hit Z) / (X hit X)'
b. Zhangsan$_Z$ renwei wo$_i$ zhidao wo-ziji da-le ziji$_{*Z/i}$.
Zhangsan think 1sg know 1sg-self hit-Perf self
'Zhangsan thinks I know I hit myself/*him'

The explanation should be clear: The subject of the most embedded predicate is a pronoun, which serves as the local antecedent of the reflexive; hence there is no long-distance anaphora and no opportunity for intervention.

4 Explaining the Binding of Double *ziji*

The binding behaviour of simple *ziji* is quite subtle and complicated, and much of it has been explained by appeal to blocking effects. Might not the behaviour of double *ziji* also be due to blocking effects? To answer this question, we begin with another construction involving two instances of *ziji.*

4.1 Two Possessive *ziji*'s

It is known ([13], cited in [1]) that sentences involving two independent possessor reflexives show interaction effects: In (25), the two instances of *ziji* may have different antecedents as long as at least one of them is locally bound.

[10] Alternately, we might consider this example to involve a null subject; but again the embedded predicate must be reflexively interpreted.

[11] A third-person pronoun blocks long-distance anaphora when it is used deictically; here, we assume a context that allows us to interpret the pronoun non-deictically.

(25) Zhangsan renwei Lisi zhidao [Wangwu ba ziji$_1$ de shu song-gei le
ZS think LS know WW BA self DE book gave-to Perf
ziji$_2$ de pengyou]
self DE friend
'ZS thinks that LS knows that WW gave ziji$_1$'s book to ziji$_2$'s friend'

Allowed readings:

a. Both reflexives may co-refer to Zhangsan, Lisi, or Wangwu.
b. If one *ziji* is local (= Wangwu), the other can have a long-distance reading (either Zhangsan or Lisi).
c. It is ungrammatical for one *ziji* to refer to Zhangsan and the other to Lisi (in either order).

Note that these examples do not involve the double-*ziji* construction: We have to do here with a sentence containing two NP positions, both of them possessors, which are independently expressed in terms of a possessive. In other words, this example contains two instances of reflexivization, rather than one instance involving two overt markers.

Pan analyzes this as a case of blocking: A third-person NP (Lisi) blocks binding only when it is itself a long-distance binder of *ziji*. This must be contrasted with the usual cases of blocking, which involve an intervener with contrasting phi-features, or with deictic reference. For comparison, we repeat an example of ordinary blocking:

(24b) Zhangsan$_Z$ renwei wo$_i$ zhidao wo-ziji da-le ziji$_{*Z/i}$.
Zhangsan think 1sg know 1sg-self hit-Perf self
'Zhangsan thinks I know I hit myself/*him'

What kind of intervention account would account for the double reflexive? In Pan's account, the *antecedent* of one reflexive becomes an intervener, blocking an even higher NP from becoming an antecedent of the other reflexive. Local anaphora is never blocked. This mechanism cannot account for the construal of double *ziji*: With a double reflexive, the subject and object of the verb are necessarily coreferential; we can never have one local and another non-local one. If we were to assume that the first *ziji* has an antecedent (which is questionable, given that it is not the subject of the clause but an adverbial), we should still be able to obtain readings where the subject is local and the object takes a long-distance interpretation. But such readings are uniformly unavailable.

Since no potential intervener exists in the relevant examples, our only option would be to treat the first *ziji* itself as an intervener for the second, as suggested to us by Ken Safir (personal communication). Such a mechanism might descriptively make the right predictions, but it would be a completely new kind of intervention: There is no feature clash, and blocking would be triggered even though the first *ziji* is not long-distance bound, and is not even in an argument position. We conclude that an analysis in terms of interveners, if not entirely untenable, is not particularly plausible.

4.2 Logophoricity

The double-ziji construction always expresses reflexive action of the local subject, even if this is a pronoun or even a null subject (PRO) with arbitrary reference. To better characterize its behaviour, consider the following construals available for the single and double reflexive when used with the grooming verb *daban* 'dress up'.

(26) Mulan bu xihuan chuipeng ziji.
Mulan not like brag.about self
a. Mulan$_i$ doesn't like [PRO$_i$ to brag about herself$_i$].
b. Mulan$_i$ doesn't like [(others=PRO$_j$) to brag about her$_i$].
c. * Mulan$_i$ doesn't like [(others=PRO$_j$) to brag about themselves$_j$].

(27) Mulan bu xihuan ziji chuipeng ziji.
Mulan not like self brag.about self
a. Mulan$_i$ doesn't like [PRO$_i$ to brag about herself$_i$].
b. * Mulan$_i$ doesn't like [(others=PRO$_j$) to brag about her$_i$].
c. Mulan$_i$ doesn't like [(others=PRO$_j$) to brag about themselves$_j$].

The readings in (a) and (b) should come as no surprise: When the (null) subject of *chuipeng* 'brag' is coreferent with *Mulan,* either type of reflexive can be used; and when the subject is construed to mean other, arbitrary persons, simple *ziji* can still refer to Mulan, giving rise to a long-distance construal that is impossible with double *ziji.*

The readings in (c), however, show something new: Simple *ziji* cannot be used as a local reflexive in this context, but the double reflexive can. We propose that the reason for this is the arbitrary referent of the embedded subject in readings (b) and (c), combined with the fact that (simple) *ziji* in (26) is a logophor: The arbitrary referent is not sufficiently prominent to be a logophoric antecedent, and this reading is ruled out. We propose that double *ziji* is not a logophor at all, but an ordinary anaphor similar to the English reflexive. Ordinary anaphors do not impose prominence requirements on their antecedent, and the reading in (27c) is licit since PRO$_j$ is a suitable antecedent for an ordinary anaphor.

5 Conclusion

We have seen that the double-ziji construction is an independent reflexive with its own distinctive properties, which to our knowledge have not previously been discussed in the theoretical literature. In addition to the double locus of marking, the construction differs from simple *ziji* reflexives in behaving like a plain anaphor (i.e., being subject to Binding Principle A), while *ziji* is a logophor. This conclusion presupposes that anaphors and logophors are inherently different; it is not immediately compatible, for example, with the unified account of Reinhart and Reuland [14], who propose that a single class of referentially defective elements behave as anaphors when they appear in argument position, but as logophors ("exempt anaphors") when they appear as adjuncts.

Acknowledgements. We thank Lisa Cheng, Umberto Ansaldo, Eric Reuland, Ken Safir and Meiyi Bao for discussion, judgements and bibliographic references.

References

1. Huang, C.T.J., Liu, C.S.L.: Logophoricity, attitudes and ziji at the interface. In: Cole, P., Harmon, G., Huang, C.T.J. (eds.) Long-Distance Reflexives. Syntax and Semantics, vol. 33, pp. 141–195. Academic Press, San Diego (2001)
2. Liu, C.S.L.: Pure reflexivity, pure identity, focus and Chinese ziji-benshen. Journal of East Asian Linguistics 12, 19–58 (2003)
3. Gast, V., Hole, D., König, E., Siemund, P., Töpper, S.: Typological Database of Intensifiers and Reflexives, version 2.0. Freie Universität Berlin (2007), http://noam2.anglistik.fu-berlin.de/~gast/tdir/ (retrieved July 2009)
4. Lidz, J.: Morphological reflexive marking: Evidence from Kannada. Linguistic Inquiry 26(4), 705–710 (1993)
5. Gast, V.: The Grammar of Identity: Intensifiers and Reflexives as Expressions of the Identity Function. PhD thesis, Freie Universität Berlin (2002)
6. Faltz, L.M.: Reflexivization: A Study in Universal Syntax. PhD thesis, University of California at Berkeley (1977); Published by Garland (1985)
7. Kayne, R.: French Syntax: The Transformational Cycle. MITPress, Cambridge (1975)
8. Mchombo, S.A.: On the binding of the reflexive and the reciprocal in Chichewa. In: Mchombo, S.A. (ed.) Theoretical Aspects of Bantu Grammar, pp. 181–207. CSLI Publications, Stanford (1993)
9. Zec, D.: Objects in Serbo-Croatian. In: Niepokuj, M., Clay, M.V., Nikiforidou, V., Feder, D. (eds.) Proceedings of the 11th Annual Meeting of the Berkeley Linguistics Society, pp. 358–371 (1985)
10. Cole, P., Harmon, G., Huang, C.T.J.: Long-distance reflexives: The state of the art (-). In: Cole, P., Harmon, G., Huang, C.T.J. (eds.) Long-Distance Reflexives. Syntax and Semantics, vol. 33, pp. xiii–xlv. Academic Press, San Diego (2001)
11. Yu, X.F.W.: Challenging Chinese reflexive data. The Linguistics Review 9, 285–294 (1992)
12. Yu, X.F.W.: A Study of Chinese Reflexives. PhD thesis, University of London (1996)
13. Pan, H.H.: Constraints on Reflexivization in Mandarin Chinese. Garland, New York (1997)
14. Reinhart, T., Reuland, E.: Anaphors and logophors: An argument structure perspective. In: Long-Distance Anaphora. Cambridge University Press, Cambridge (1991)

Definiteness Marking Shows Late Effects during Discourse Processing: Evidence from ERPs

Petra B. Schumacher

University of Mainz, Department of English and Linguistics,
Jakob-Welder-Weg 18, 55099 Mainz, Germany
petra.schumacher@uni-mainz.de

Abstract. This paper investigates the processing of indefinite and definite noun phrases in discourse. It presents data from an Event-Related brain Potential (ERP) study that contrasted definite and indefinite noun phrases following three distinct context sentences. The data suggest that coherence considerations influence early processing stages, while morphological definiteness features only affect later stages during reference resolution. In addition, the processing of a definite determiner (prior to encountering the subsequent noun) exerts processing demands, supporting the functional contribution of definiteness marking. Supplementary data from a plausibility questionnaire and two completion studies are also presented. The findings are discussed with respect to a neurocognitive model of reference resolution.

Keywords: Definiteness, Referential processing, Event-related brain potentials, Inferences.

1 Introduction

Anaphora represents a central notion in discourse processing and contributes to both discourse coherence – by linking up with prior knowledge ("dependency formation") – and discourse progression – by introducing new information and condensing or reorganizing discourse representation structure ("discourse updating"). Systematic psycholinguistic investigation of referential processing using event-related brain potentials (ERPs) has recently identified two distinct electrophysiological correlates for these two discourse mechanisms: a negative voltage deflection peaking around 400 ms after the onset of an anaphoric expression (*N400*) whose amplitude varies as a function of increasing processing difficulties during the establishment of a dependency; and a later positive-going ERP signature with an onset latency around approximately 550 ms (*late positivity*) that reflects processing costs arising from the introduction of new discourse units or the modification of previously established discourse representation structures (cf. e.g. [4,5,6,26]).

This paper presents a neurocognitive model of discourse processing that is guided by recent ERP findings from the comprehension of anaphors. It investigates the processing of noun phrases as a function of their definiteness marking

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 91–106, 2009.

and contextual salience using the ERP methodology. Two views on the role of definiteness marking are tested utilizing the neurocognitive model's predictive power. The ERP results are discussed within the proposed model and alternative interpretations of the ERP effects are considered on the basis of a series of offline measures (a plausibility rating and two completion studies).

1.1 Definiteness

Languages often distinguish between definite and indefinite expressions. Definite entities are generally considered to refer to a particular entity or set in the discourse model and definiteness marking serves the functional purpose of pointing to a particular discourse referent in the discourse model. Definiteness can be encoded via a definite article *(the melon)*, a demonstrative *(that melon)* and others. Rigid designators – like indexicals or concepts common to all situations such as *the time, the weather,* etc. – are considered to be inherently definite. Indefiniteness in contrast marks the introduction of an entity that does not have a particular referent in the discourse model yet.

In the research literature, definiteness has been discussed with respect to the notions of familiarity, specificity, salience, uniqueness, and identifiability. Russell provided a logical semantics account of the concept of uniqueness to distinguish between definite and indefinite noun phrases [25]. Accordingly, an indefinite introduces some entity x that has the property associated with the content of the noun, while a definite introduces a unique entity x with that property. The function of definiteness is therefore described as identifying one specific entity x for which the particular properties are true. It has further been suggested that definiteness introduces a presupposition of existence [10,27]. Hawkins introduced the notion of inclusiveness as the core characterization of definiteness [12]. According to this approach, definiteness functions as a marker for all entities to which the properties associated with the content of the noun phrase apply. However, not every definite noun phrase refers to one and only one particular referent nor does it necessarily refer to an entity previously introduced into the discourse model. For instance, the referent of *Smith's murderer* in *Smith's murderer is insane* might not have been identified yet. However, the concept of an individual who is characterized as being Smith's murderer can be established [8]. Nevertheless, the overarching notion is that a definite marker signals the uniqueness or familiarity of the corresponding entity. Thus regardless of how definiteness is conceived of, the majority of accounts ascribe a functional characterization to definiteness, roughly requiring the establishment of a linkage with a particular discourse referent (that is further specified as given or prominent or unique).

In contrast, an indefinite marker has traditionally been viewed as an existential quantifier. Heim, for instance, proposed a framework in which indefinite expressions introduce new entities into discourse space (i.e. novelty constraint), while definite expressions serve the function of referring to already established discourse referents (i.e. familiarity constraint) [13]. An obvious problem of this latter characterization, however, emerges with inferentially linked definite expressions

(yielding accommodation in the absence of an explicitly available discourse referent as described below).

This paper investigates the consequences of definite and indefinite marking for language comprehension. If the primary function of indefiniteness is to introduce an independent discourse referent, while definiteness signals the identifiability and familiarity of a particular entity, this should be reflected in the functional and temporal architecture of the language processor.

2 Discourse Comprehension: A Neurocognitive Approach

To this end, we turn to electrophysiological measures during language comprehension. ERPs are voltage fluctuations that reflect the brain's spontaneous electrical activity, which occurs in response to a sensory, motor, or cognitive stimulus event and is measured non-invasively by electrodes applied to the scalp. In general, ERPs provide a high temporal resolution, which allows us to thoroughly sketch the time course of language processing, and they reveal functionally distinct processes. In this paper, I want to focus primarily on two ERP signatures that have been identified as markers of discourse integration: the N400 and the late positivity. In addition, a left anterior negativity (LAN) is briefly discussed with respect to working memory demands.

2.1 N400 Effect and Dependency Formation / Linking

The N400, a negative-going deflection peaking around 400 ms after the onset of a critical stimulus event, is a well-known correlate of lexical-semantic processing. Its amplitude is inversely related to the degree of plausibility and goodness-of-fit (for a comprehensive overview see [19]). With respect to referential processing, the amplitude of the N400 has been correlated with the degree of difficulty of referent identification (see e.g. [28]) and is contingent on the associative strength of the antecedent/anchor expression [4] or the accessibility and prominence of the antecedent/anchor in discourse representation [6]. Generally, the more difficult the establishment of a link with discourse representation is, the more enhanced is the N400-amplitude.

2.2 Late Positivity and Discourse Updating

Processing cost arising from the updating of discourse representation structure has been associated with a late positivity that has an onset latency around 550 ms after stimulus-onset. This signature emerges for instance when an independent discourse representation must be created; this is the case for so-called indirect anaphors that depend on an inference-based relation with information in discourse representation, yet require the introduction of an independent discourse representation (to be available for future referential processes) – see [4], and for independent evidence from bare quantifiers [16]. Enriched composition as in reference transfer (e.g. *The ham sandwich from table four wants to pay.*),

metaphor comprehension, or thematic-semantic enrichment when a prior event representation must be updated towards a more specific event also evoke a late positivity (e.g. a *finding dead* event becomes a *shooting* event in the course of integrating the instrument *the pistol* into discourse as in the following example: *After a student was found dead in the park, a little girl discovered the pistol behind a tree*)(cf. [5,7,26]). These data suggest that creating new discourse referents and updating or modifying prior discourse representation structures exert similar processing demands reflected in a late positivity.

2.3 LAN and Working Memory

A left anterior negativity (LAN) has been identified as a reflex of working memory demands (cf. e.g. [18]). With respect to integration at the level of discourse, a LAN has been reported for discourse-dependent pronominals, suggesting that discourse integration increases working memory demands, see for instance [3,28]. A particular left anterior negativity has been further observed for function words, roughly between 400 and 700 ms, and it has been suggested that function words enhance working memory demands due to their anticipatory power to predict the next constituent [20]. The LAN might thus be affected by definiteness marking prior to integrative processes elicited by the entire noun phrase.

3 (In)Definiteness Marking

In the following, we turn to (in)definiteness marking in German. Burkhardt reported an ERP investigation that contrasted three types of *definite noun phrases* (corresponding to the first three examples in Table 1): direct anaphors (i.e. given entities that form an identity relation with an antecedent expression), indirect anaphors (i.e. entities that rely on the drawing of an inference to enter into a dependency with an anchor expression in discourse) and new entities (i.e. definite noun phrases that have no antecedent or anchor in discourse) [4]. Time-locked to the onset of the noun phrase, the results revealed N400-modulations as a function of the ease of dependency formation (guided by the strength of the lexical-semantic association): direct anaphor < indirect anaphor < new referent. In addition, a late positivity emerged for indirect anaphors and new referents; as alluded to in 2.2, the discourse integration of indirect anaphors is facilitated by an inferential link, however, this does not suffice for proper discourse integration, and an independent discourse representation must be created. In contrast, the new noun phrase cannot find a coherent link in discourse space, but a discourse representation still appears to be established (possibly in anticipation of further specification and integration as discourse unfolds; consider, for instance, backward anaphora).

Classical accounts of discourse representation have drawn a close correspondence between definiteness marking and anaphoric links on the one hand, and indefiniteness and new information on the other hand, e.g. [13]. Within the neurocognitive model of discourse comprehension, these accounts thus predict that

indefinite noun phrases introduce an independent discourse referent (reflected in a late positivity), yet are unaffected by contextual information (no N400-differences). In contrast, coherence-driven accounts have abandoned this strict correspondence and allow context to play a guiding role during discourse processing e.g. [2,24]. Accordingly, they predict immediate context-induced effects (N400-modulations) and possibly an additional late positivity relative to noun-onset. These predictions regarding indefinite noun phrases are tested in the following ERP study and are subsequently followed–up with three offline investigations.

In addition to the hypotheses associated with noun phrase integration, an immediate effect of definiteness marking time-locked to the determiner could be expected on the basis of the notion that the definite article serves as a pointer for referent identification and dependency formation. From an electrophysiological perspective, a LAN possibly reflects determiner-contingent working memory demands, and as pointed out above, LAN responses to function words have been related to their anticipatory potential. Following this line of reasoning, definite determiners might evoke a more enhanced negativity if their function is to address discourse and identify a discourse referent, while indefinite determiners should not exert a particular burden on working memory (but see the discussion below for conflicting findings from [1]).

3.1 ERP Study

The effects of definiteness marking were tested in a reading comprehension study in German during which ERPs were recorded. Of particular interest was the processing of indefinite noun phrases. In addition, this study sought to replicate the findings for definite noun phrases from [4]. Accordingly, indefinite and definite noun phrases occurred after three different context sentences, which are here labeled as Given, Inferred and New (on the basis of their impact on definite noun phrases).

Participants: Twenty-four native speakers of German (12 women) participated in this experiment. All were right-handed and reported normal or corrected-to-normal visual acuity and no history of neurological disorder. Their ages ranged from 21 to 29 years (Mean=25). They were paid for their participation. One participant had to be excluded from the analysis of the ERP data due to excessive artifacts in the recordings.

Materials: Stimuli were designed with the factors DEFINITENESS (2 levels: definite, indefinite) and CONTEXT (3 levels: given, inferred, new) yielding six conditions. Forty sets of these six conditions were constructed which consisted of a context and a target sentence each. Context sentences manipulated the contextual salience of the critical noun phrase, and target sentences varied the form of the determiner of the critical noun phrase - see Table 1 for examples. Stimuli were created on the basis of the inferential contexts, such that critical noun phrases and their anchors (in the inferred condition) represented highly associated noun pairs, as assessed in a previous questionnaire study (reported in [4]).

Table 1. Example stimulus set for the six critical conditions. Critical noun phrases are in bold.

Condition	Example	Translation
Definite Given	Peter besuchte neulich einen Redner in München. Er erzählte, dass **der Redner** sehr nett war.	*Peter has recently visited a speaker in Munich. He said that **the speaker** had been very nice.*
Definite Inferred	Peter besuchte neulich einen Vortrag in München. Er erzählte, dass **der Redner** sehr nett war.	*Peter has recently visited a lecture in Munich. He said that **the speaker** had been very nice.*
Definite New	Peter traf neulich Hannah in München. Er erzählte, dass **der Redner** sehr nett war.	*Peter has recently met Hannah in Munich. He said that **the speaker** had been very nice.*
Indefinite Given	Peter besuchte neulich einen Redner in München. Er erzählte, dass **ein Redner** sehr nett war.	*Peter has recently visited a speaker in Munich. He said that **a speaker** had been very nice.*
Indefinite Inferred	Peter besuchte neulich einen Vortrag in München. Er erzählte, dass **ein Redner** sehr nett war.	*Peter has recently visited a lecture in Munich. He said that **a speaker** had been very nice.*
Indefinite New	Peter traf neulich Hannah in München. Er erzählte, dass **ein Redner** sehr nett war.	*Peter has recently met Hannah in Munich. He said that **a speaker** had been very nice.*

All critical noun phrases therefore represented strongly associated, salient entities following the inference-inducing context. The contexts for the given and new conditions were created subsequently. Furthermore, the discourse referents that were introduced in the context sentences as potential antecedents or anchors carried an indefinite marker. The length of context-target sentence pairs was kept constant across all conditions and experimental items. A total of 100 additional sentence pairs were constructed that served as filler items in the final version of the experiment to divert attention from the critical material. This amounted to 340 items altogether. Verification questions were constructed for each stimulus, which probed the comprehension of either the context or target sentence; correct and incorrect responses were distributed evenly across all items. Two versions of different combinations of stimuli and verification questions were created and combined with three different randomizations each that were counterbalanced across all participants.

Procedure: Participants sat comfortably in a sound-attenuating booth and were instructed to read the stimuli material for comprehension. Stimuli were presented visually in the center of a computer screen in yellow letters against a blue background. Each trial started with the presentation of three asterisks to remind the participants to avoid blinking and moving around. Then the stimuli were presented word by word (400 ms each, followed by an inter-stimulus interval of 100 ms); the word-wise presentation was chosen to investigate whether an effect of definiteness could be observed by allowing more time for the determiner to be processed. After the presentation of the two successive sentences (i.e. context and target sentence), asterisks were presented on the computer screen for 500 ms, followed by the verification question, which was presented in its entirety. Participants were instructed to respond to this question as quickly and as accurately as possible, by pressing either a 'yes' or 'no' button on a response box. The purpose of this task was to assure that participants were properly reading the stimuli for comprehension. Generally, participants were instructed to blink only during the presentation of the verification question to limit ocular and other artifacts. Each experimental session started with two brief practice blocks of six stimuli each. The following experimental session consisted of 340 pseudo-randomized stimuli and was carried out in eight blocks with short breaks between blocks.

The EEG was recorded from 26 Ag/AgCl scalp electrodes mounted in an elastic cap *(Electro-Cap International)*, which conformed to the standard 10-20 system for electrode positioning (cf. [15]). The following positions were recorded: FPZ, FZ, CZ, and PZ for midline sites, and FP1/2, F3/4, F7/8, FC3/4, FT/8, C3/4, T/8, CP5/6, P3/4, P7/8, O1/2 for lateral sites. Recordings were referenced to the left mastoid and rereferenced offline to linked mastoids. Horizontal and vertical eye movements were monitored by means of two sets of additional electrode pairs, placed at the outer cantus of each eye and above and below the participant's left eye, to control for ocular artifacts. Electrode impedances were kept below 5 kΩ. Signals were recorded with a sampling rate of 250 Hz and all channels were amplified using a *Twente Medical Systems International* amplifier.

Data analysis: Analyses of the behavioral data for the verification task were computed for error percentages. Error percentages were averaged over incorrect and timed-out responses (i.e. responses that failed to be registered 4000 ms after the verification question was presented).

Average ERPs were time-locked to the onset of the determiner of the critical noun phrase and computed per condition per participant, prior to the calculation of the grand averages over all participants. In analogy to [4], a 200 ms baseline before the critical noun phrase was utilized during averaging. Trials that registered an incorrect response to the verification task or that contained ocular or other artifacts were excluded from averaging. For the statistical analysis of the ERP data, repeated-measures ANOVAs were performed with the factors CONTEXT and DEFINITENESS. Electrodes were grouped by location and entered the ANOVA as topographical factor: REGION OF INTEREST [ROI] left anterior (F3/F7/FC3/FT7), right anterior (F4/F8/FC4/FT8), left posterior (CP5/T7/P3/P7), and right posterior

(CP6/T8/P4/P8). Statistical analyses are based on the mean amplitude value per condition in predetermined time-windows. The analysis was carried out in a hierarchical manner. The data were corrected using the Huynh-Feldt procedure in order to control for potential type I errors due to violations of sphericity [14].

Based on previous studies and in particular the findings from [4], analyses were carried out for separate temporal windows corresponding to the N400 (300-500 ms) and late positivity components (550-700 ms) relative to the onset of the noun. To investigate potential immediate effects of the determiner, an additional analysis relative to the onset of the definite vs. indefinite determiners was conducted between 400 and 700 ms.

Results: In the behavioral task, participants performed at ceiling level, with a mean error rate of 2.58 for the critical items (Definite-Given: 3.17, Definite-Inferred: 1.62, Definite-New: 3.38, Indefinite-Given: 1.17, Indefinite-Inferred: 3.12, Indefinite-New: 3.00). This indicates that they were properly paying attention to the materials.

Fig.1 illustrates ERPs relative to the onset of the noun. The left panel presents the data for the definite noun phrases, which replicated previous investigations of referential processing [4]. The findings for the indefinites, illustrated in the right panel of Fig.1, cannot support the classical accounts on definiteness and indicate that lexical-semantic networks are activated to establish discourse coherence (reflected in the N400), even though the indefiniteness marking indicates that no dependency relations (in the narrow sense, i.e. direct or indirect anaphors) need to be established.

This was supported by statistical analyses, which revealed a reliable effect of CONTEXT in the N400-window (300-500 ms) [$F(2,44) = 4.18, p < .03$], which was reflected in modulations of the N400-amplitude with increasing amplitude for Given < Inferred < New. There was also a significant interaction of CONTEXT x ROI[$F(6,132) = 3.22, p < .01$] whose resolution revealed main effects of CONTEXT in the right anterior [$F(2,44) = 3.59, p < .04$], right posterior [$F(2,44) = 4.00, p < .04$] and left posterior ROIs [$F(2,44) = 8.98, p < .001$], substantiating a broad distribution of the negative deflection (which is a general characteristic of the N400-signature). The absence of a significant effect of DEFINITENESS indicates that processing costs occurred irrespective of definiteness marking in this time window.

Turning to the later time window, a significant late positivity effect was observed between 550-700 ms after the onset of the noun. In this later time window, there was a main effect of CONTEXT [$F(2,44) = 8.18, p < .002$] as well as interactions of CONTEXT X ROI [$F(6,132) = 7.67, p < .001$], DEFINITENESS X ROI [$F(3,66) = 4.48, p < .02$], and CONTEXT X DEFINITENESS [$F(2,44) = 4.16, p < .03$]. Resolution of this latter interaction upheld a main effect of CONTEXT [$F(2,44) = 7.90, p < .003$], and pairwise comparisons revealed a difference only between the two given conditions [$F(1,22) = 11.89, p < .003$] – reflected in the absence of a positivity for the Definite-Given condition – but no reliable differences between Definite/Indefinite-New and Definite/Indefinite-Inferred noun phrases [$Fs < 1$]. Fig.1 indicates that the late positive going

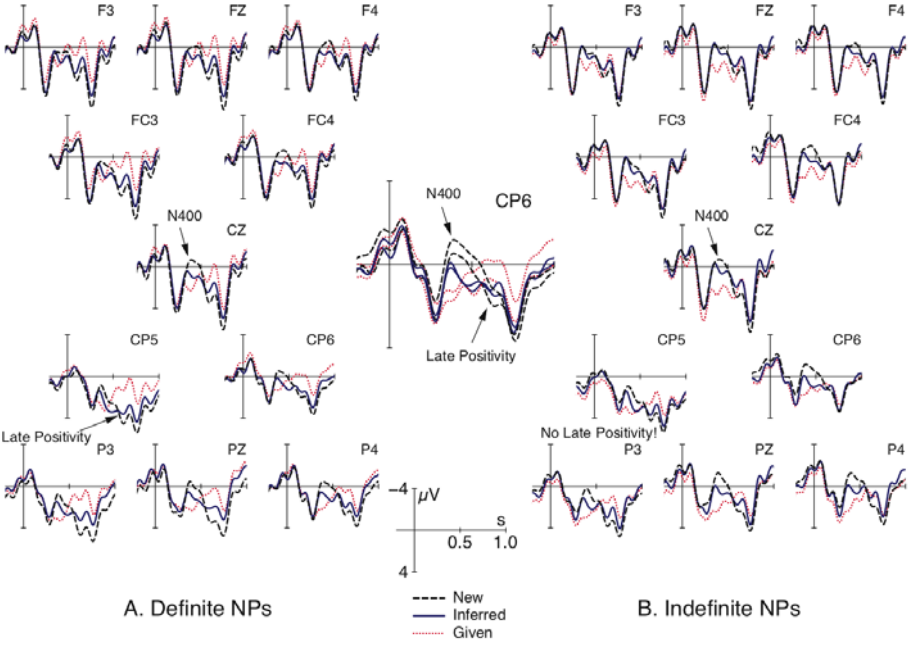

Fig. 1. Grand-average ERPs from 11 selected electrodes relative to noun-onset. Left and right panels present definite and indefinite noun phrases (NPs) respectively for given (dotted), inferred (solid) and new contexts (dashed). The enlarged electrode in the middle compares all six conditions. The vertical bar marks noun-onset; negative voltage is plotted upwards. The horizontal bar shows the time-axis from 200 ms before the noun until 1000 ms after.

deflection occurs for all conditions except for the Definite-Given condition (which is the only condition that allows coreference and thus does not require the introduction of an independent discourse referent). This divergence of the Definite-Given condition from all other conditions is depicted in the enlarged electrode (CP6) in the middle of Fig.1.

Additional analyses relative to the onset of the determiner revealed a statistically significant difference between definite and indefinite determiners between 400 and 700 ms. There was an interaction of DEFINITENESS X ROI [$F(3, 66) = 12.91, p < .001$], which was reliably resolved over left anterior [$F(1, 22) = 11.45, p < .003$] as well as left posterior sites [$F(1, 22) = 4.75, p < .05$]. While the negativity was significant over both left–lateralized regions, it had its maximum over anterior sites. However, contrary to the findings reported in [1], Fig.2 demonstrates that the definite determiner evoked a more pronounced negativity compared to the indefinite determiner.

Discussion: We have investigated the time-course of the processing of definite and indefinite noun phrases in discourse. The results reveal distinct electrophysiological correlates of discourse integration: the N400 is predominantly affected

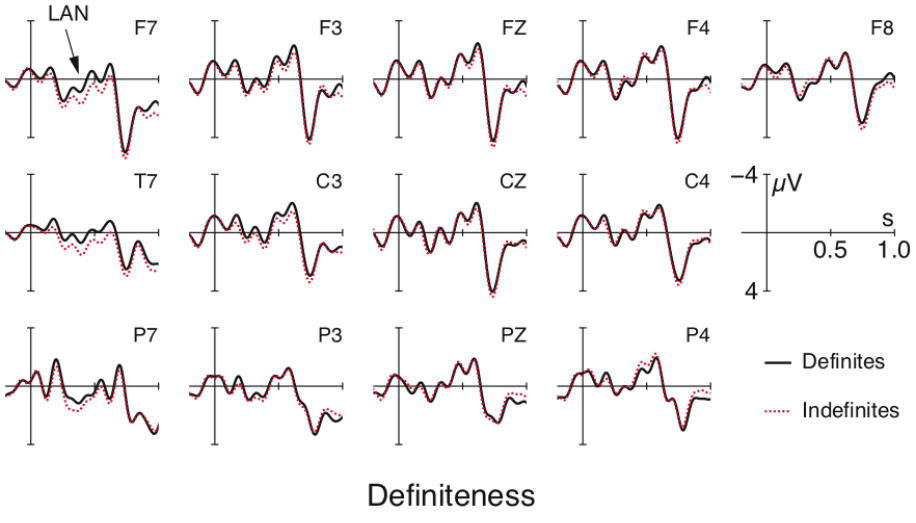

Fig. 2. Grand-average ERPs for definite (solid) vs. indefinite determiners (dotted) at selected electrode sites. The vertical axis marks the onset of the determiner and plots voltage in microvolts with negativity upwards. The horizontal bar shows the time-axis from 200 ms before till 1000 ms after the determiner.

by contextual cuing, while the late positivity indexes costs from the introduction of a new discourse entity, which is triggered by indefiniteness marking, but also by context information (i.e. absence of a matching referent in discourse). In addition, definite determiners exert early processing demands, reflected in a left anterior negativity.

The neurocognitive model outlined above can account for the findings in that initial attempts of linking up with discourse representation and dependency formation are reflected in the N400-signature. The presence of a discourse referent corresponding to the head noun facilitates linking, reflected in the least pronounced negativity for the two Given conditions; inference-based links recruit world knowledge and lexical-semantic associations, which exerts more demands; and the absence of supporting links and anchors in discourse yield the most enhanced negativity (see Fig. 1). Importantly, the linking attempts occur for both definite and indefinite entities and are governed by the fit of the head noun with the information provided by the context. This suggests that the underlying processes are guided by coherence constraints and that a strict correspondence between definiteness marking and integration processes cannot be maintained. The data indicate that morphological definiteness features are ignored during initial referential integration and that the parser seeks to establish discourse links. This is most compatible with coherence-driven accounts of discourse integration and could be modeled with respect to early applicable constraints like *Use Identity or Maximize Discourse Coherence* [2].

Crucially, indefiniteness information comes into play somewhat later, as evidenced by a late positivity for all three indefinite noun phrases, which is assumed to reflect processing costs from creating independent discourse representations. So even though associative links are formed initially, the functional information associated with indefiniteness (i.e. introducing a new entity) ultimately triggers additional processes. In particular, there is a clash between potential dependency links and the functional contribution of indefiniteness; for instance, in the Indefinite-Given condition, linking offers an anchor for an identity relation, but indefiniteness marking prevents this and forces the introduction of an independent representation. In the case of the definite noun phrases following Inferred and New contexts, which also elicited a late positivity, a general discourse constraint requires the introduction of an independent discourse referent in the absence of an identity relation, because definite noun phrases must have a corresponding discourse representation.

Finally, the current data revealed a left anterior negativity in response to the definite over the indefinite determiner. This effect was predicted on the basis of working memory accounts in relation to the definite's function of addressing discourse space. Contrary to the present findings, Anderson & Holcomb [1] report a LAN for indefinite determiners compared to definite determiners. However, since they tested definiteness effects in English, their data might be confounded by length differences (as for example reported in [21]), which is not the case for the German contrast tested here. Therefore, following Russell and others, the observed costs for definite determiners might be associated with the prompt to identify a unique referential entity for which the properties of the head noun are true. The ERP data substantiate this functional characterization of definiteness marking. Definite determiners thus signal the parser that a link must be established with an entity in discourse space, which results in enhanced working memory demands. As the data relative to noun-onset indicate, this pointer is only used in the later referential integration stage.

In order to further evaluate these conclusions, a series of offline paper-and-pencil studies was conducted. These studies investigated whether more general considerations of plausibility and predictability or even frequency of occurrence could also explain the ERP patterns.

3.2 Plausibility Questionnaire

Could the overall fit of context and target sentence – in other words discourse plausibility – account for the observed ERP effects? There are two different predictions with respect to the effect of plausibility on ERP patterns. First, a late positivity has been discussed with respect to considerations of well-formedness (cf. e.g. [17,22] for dispreferred structures), and this might be extended to the domain of discourse, where the evaluation of the fit of subsequent sentences may yield graded ratings. If this was the right kind of explanation for the observed positivity in the present experiment, all conditions except for the Definite-Given condition should be evaluated rather poorly in a plausibility rating task.

Second, semantic plausibility has been shown to affect the N400 component [19], and given the current ERP data, a rating study should then show no effect of definiteness, albeit a clear effect of context. To address these questions, participants were asked whether they felt that the second sentence was a good continuation of the first one.

Methods: 42 native speakers of German (22 women) participated in this questionnaire study (age range: $21 - 36$; M=26). Each participant rated five stimuli per conditions that were randomly selected from the material constructed for the ERP study. The 30 critical items were interspersed with 10 filler passages that represented poor continuations. Four different version of the questionnaire were created and distributed evenly across all participants, who were asked to evaluate the fit of the first and second sentence on a 7-point-scale ('1': the second sentence does not represent a good continuation of the first one; '7': the second sentence is a good continuation of the first one). An ANOVA was computed over mean response values.

Results: Mean scores are illustrated in Fig. 3: Definite-Given received the highest score ($M = 6.38$), followed by Definite-Inferred ($M = 5.96$), Indefinite-Inferred ($M = 4.90$), Indefinite-Given ($M = 4.41$), Indefinite-New ($M = 3.18$), Definite-New ($M = 2.71$). Statistical analysis revealed main effects of DEFINITENESS [$F(1, 41) = 33.83, p < .001$] and CONTEXT [$F(2, 82) = 125.21, p < .001$] and an interaction of DEFINITENESS X CONTEXT [$F(2, 82) = 23.46, p < .001$].

Discussion: The ratings indicate that Definite-Given is consider the most coherent condition, but Definite-Inferred is rated nearly as highly. This rules out a simple well-formedness account of the late positivity, which emerged for all but the Definite-Given condition. Similarly, the presence of an interaction and an effect of definiteness rule out a purely plausibility-based explanation of the N400.

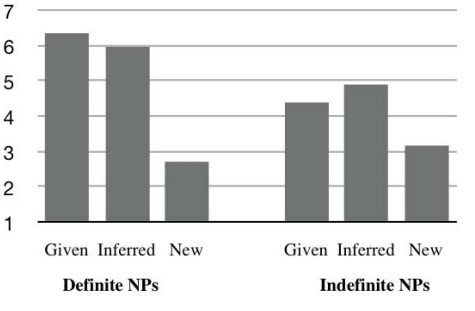

Fig. 3. Mean responses to the plausibility task for each condition. '7' stands for good fit of context and target sentence.

3.3 Completion Tasks

Another potential interpretation of the ERP data is related to the predictability or frequency of occurrence of certain noun phrases. The N400 is particularly susceptible to these criteria (cf. e.g. [19]). Thus, to determine whether the predictability of a certain continuation (especially, a particular referential representation) may explain the ERP effects, two completion studies were carried out that investigated how the probability of certain continuations changes over time. Previous research has demonstrated that completion studies provide probability measures that correlate with comprehension patterns (cf. e.g. [11]).

Methods: Sixteen participants filled out the first sentence completion form and forty-two the second (age range: 19-30; M=23.2). Each participant was instructed to complete a list of pseudo-randomized passages, which contained six stimuli each selected from the Given and Inferred conditions and 12 fillers. Two versions were distributed across participants. On the first form, the sentences were cut off after the complementizer *dass* ('that'); on the second form, completions started with a determiner. Responses were coded and assigned to different classes of continuations (e.g. pronominal, coreferential, inferential, neutral/no relation, ambiguous). Percentages are reported below for these continuation classes.

Results: Completions following the complementizer are summarized in Table 2 and revealed a strong preference for a pronominal continuation in the case of Given contexts (36% for the object antecedent – e.g. 'the speaker' in the example in Table 1 – and 8% for the subject antecedent – e.g. 'Peter') and neutral continuations (like 'the weather' or 'yesterday') following the Inferred contexts (31%). The Given contexts further elicited numerous complex anaphor continuation (e.g. 'this', 'this event'), which referred to the entire proposition introduced in the context sentence; interestingly, this type of continuation never occurred after the Inferred context. Indefinite noun phrase continuations were relatively rare.

When restricting the completions by adding a definite or indefinite determiner, the data for the definite determiner indicate a high predictability for a coreferential (direct) anaphor following a Given context sentence (which always contained an animate entity) amounting to 61% - evenly distributed over repetition of the same head noun (31%) and synonymous expressions (30%). The Inferred contexts (containing an inanimate event or object) registered a majority of inferentially linked indirect anaphors (45%) followed by coreferential anaphors (25% repetitions and 6% synonyms). When cued with an indefinite determiner, continuations following the Given and the Inferred context showed a trend for an indirectly related noun phrase (39% and 34% respectively), supporting the idea that coherence considerations guide processing, but both contexts also elicited many neutral continuations (33% in both cases). Table 3 presents these patterns for the four tested conditions. What seems to be odd with respect to the indefinite condition is the relatively high number of same head noun continuations; however, the count included reference to different tokens, which does not result in a coreferential link and therefore does not violate the constraints on indefinites.

Table 2. Percentages of continuations in first completion task (most frequent answers in bold)

	Given context	Inferred context
Pronoun		
– object antecedent	**36**	25
– subject antecedent	8	13
Direct anaphor (same head noun)	7	4
Complex anaphor (reference to proposition)	27	-
Indirect relation	21	27
Neutral continuation	-	**31**
Infelicitous continuation	-	-

Table 3. Percentages of continuations in the second completion task. The most frequent continuations following the determiner are marked in bold.

	Definite Determiner		Indefinite Determiner	
	Given context	Inferred context	Given context	Inferred context
Direct anaphor				
– same head noun	**31**	25	19	19
– synonymous	**30**	6	2	8
Indirect relation	19	**45**	**39**	**34**
Ambiguous/No referential relation	13	19	33	33
Infelicitous continuation	7	5	7	6

Discussion: These findings indicate that context-induced predictability alone cannot explain the observed ERP-patterns. Following the complementizer, pronominal and neutral continuations form the majority of responses, and definite and indefinite continuations appear to be equally less expected. Hence, the definiteness effect registered in response to the determiner (i.e. the LAN) is not guided by probability, but must be attributed to the inherent properties of the determiner. Likewise, the completion data following the determiner do not provide support for a probability-induced N400 effect, because there is no main effect of context. In general, the obtained referential preferences are in line with corpus data from direct and indirect anaphors and indicate that indirect relations occur frequently [9,23]. Furthermore, the completion data are in and off themselves interesting, because they reveal various factors related to referential processing, for instance with respect to prominence features that guide reference resolution (e.g. animate reference is preferred over reference to an inanimate entity, object antecedents are

preferentially chosen in these particular contexts). Yet for present purposes, we can only focus on the probability of a certain continuation that could affect N400-modulations. Such an interpretation cannot be supported by the findings.

4 Conclusion

Overall, the offline data do not provide a good alternative account of the observed ERP effects. Rather, the novel ERP data from indefinite noun phrases provide additional support for a model of discourse processing that proceeds from linking under coherence considerations to updating and the introduction of discourse representation structure. The former process is understood to be relational in nature and reflects the parser's immediate desire to establish links with the information already available in discourse. The latter mechanism targets the overall management and maintenance of discourse representation structure. In contrast to classical accounts of definiteness marking that advocate an immediate effect of definiteness onto discourse integration, the data demonstrate that processing costs are initially exerted when a definite determiner is encountered, but during reference resolution, definiteness features crucially affect later processing stages.

Acknowledgements. The ERP experiment was conducted during a stay at the Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany.

References

1. Anderson, J.E., Holcomb, P.J.: An Electrophysiological Investigation of the Effects of Coreference on Word Repetition and Synonymy. Brain Lang. 94, 200–216 (2005)
2. Asher, N., Lascarides, A.: Bridging. J. of Semantics 15, 83–113 (1983)
3. Burkhardt, P.: The Syntax-Discourse Interface: Representing and Interpreting Dependency. John Benjamins, Amsterdam (2005)
4. Burkhardt, P.: Inferential Bridging Relations Reveal Distinct Neural Mechanisms: Evidence From Event-Related Brain Potentials. Brain Lang. 98, 159–168 (2006)
5. Burkhardt, P.: The P600 Reflects Cost of New Information in Discourse Memory. Neuroreport 18, 1851–1854 (2007)
6. Burkhardt, P., Roehm, D.: Differential Effects of Saliency: An Event-Related Brain Potential Study. Neurosci. Lett. 413, 115–120 (2007)
7. Coulson, S., Van Petten, C.: Conceptual Integration and Metaphor: An Event-Related Potential Study. Mem. Cogn. 30, 958–968 (2002)
8. Donnellan, K.S.: Reference and Definite Descriptions. The Philosophical Review 77, 281–304 (1966)
9. Fraurud, K.: Definiteness and the Processing of Nps in Natural Language. J. of Semantics 7, 395–433 (1990)
10. Frege, G.: Über Sinn und Bedeutung. Zeitschrift für Philosophie und philosophische Kritik NF 100, 25–50 (1892)
11. Gennari, S.P., MacDonald, M.C.: Linking Production and Comprehension Processes: The Case of Relative Clauses. Cognition 111, 1–23 (2009)

12. Hawkins, J.A.: Definiteness and Indefiniteness. Humanities Press, Atlantic Highland (1978)
13. Heim, I.: The Semantics of Definite and Indefinite Noun Phrases. Unpublished Ph.D. dissertation, University of Massachusetts, Amherst (1982)
14. Huynh, H., Feldt, L.S.: Conditions Under Which Mean Square Ratios Repeated Measurements Designs Have Exact F Distributions. J. of the American Statistical Assocation 65, 1582–1589 (1970)
15. Jasper, H.H.: The Ten Twenty Electrode System of the International Federation. Electroencephalogr Clin Neurophysiol 10, 371–375 (1958)
16. Kaan, E., Dallas, A.C., Barkley, C.M.: Processing Bare Quantifiers in Discourse. Brain Res. 1146, 199–209 (2007)
17. Kaan, E., Swaab, T.Y.: Repair, Revision, and Complexity in Syntactic Analysis: An Electrophysiological Differentiation. J. Cogn. Neurosci. 15, 98–110 (2003)
18. King, J.W., Kutas, M.: Who Did What and When? Using Word- and Clause-Level ERPs to Monitor Working Memory Usage in Reading. J. Cogn. Neurosci. 7, 376–395 (1995)
19. Kutas, M., Federmeier, K.D.: Electrophysiology Reveals Semantic Memory Use in Language Comprehension. Trends Cogn. Sci. 4, 463–470 (2000)
20. Neville, H.J., Mills, D.L., Lawson, D.S.: Fractionating Language: Different Neural Subsystems With Different Sensitive Periods. Cereb. Cortex 2, 244–258 (1992)
21. Osterhout, L., Allen, M., McLaughlin, J.: Words in the Brain: Lexical Determinants of Word-Induced Brain Activity. J. Neurolinguist 15, 171–187 (2002)
22. Osterhout, L., Holcomb, P.J., Swinney, D.A.: Brain Potentials Elicited By Garden-Path Sentences - Evidence of the Application of Verb Information During Parsing. J. Exp. Psychol. Learn Mem. Cogn. 20, 786–803 (1994)
23. Poesio, M., Vieira, R.: A Corpus-Based Investigation of Definite Description Use. Computational Linguistics 24, 183–216 (1998)
24. Prince, E.F.: Toward a Taxonomy of Given-New Information. In: Cole, P. (ed.) Radical Pragmatics, pp. 223–255. Academic, New York (1981)
25. Russell, B.: On Denoting. Mind 14, 479–493 (1905)
26. Schumacher, P.B.: The Hepatitis Called.: Electrophysiological Evidence for Enriched Composition. In: Meibauer, J., Steinbach, M. (eds.) Experimental Pragmatics/Semantics. John Benjamins, Amsterdam (2009)
27. Strawson, P.F.: On Referring. Mind 59, 320–344 (1950)
28. Streb, J., Rösler, F., Hennighausen, E.: Event-Related Responses to Pronoun and Proper Name Anaphors in Parallel and Nonparallel Discourse Structures. Brain Lang. 70, 273–286 (1999)

Pronoun Resolution to Commanders and Recessors: A View from Event-Related Brain Potentials

José Augusto Leitão[1], António Branco[2], Maria Mercedes Piñango[3], and Luís Pires[1]

[1] University of Coimbra, Faculdade de Psicologia, Rua do Colégio Novo, 3001-802 Coimbra, Portugal
[2] University of Lisbon, Faculdade de Ciências, Cidade Universitária, 1749-016 Lisboa, Portugal
[3] Yale University, Psycho/Neurolinguistics Lab, 370 Temple St, New Haven, CT 06520-8366, USA
jleitao@fpce.uc.pt, Antonio.Branco@di.fc.ul.pt, maria.pinango@yale.edu, sirleopires@gmail.com

Abstract. We present results from an online experiment designed to probe the cognitive underpinnings of intra-sentential pronoun resolution. Event-related brain potentials were used to test the hypothesis that the processing of anaphoric links established between pronouns and non commanding antecedents demands more cognitive resources than the processing of anaphoric links to commanding antecedents. The experimental results obtained show, among others, a major N400-like effect elicited by the pronouns resolved to the non-commanding antecedent. This enhanced negativity suggests that, as hypothesized, resolving a pronoun to a non commanding antecedent is a more resource demanding process than resolving it to an antecedent in a commanding position. Our results can be interpreted within a theoretical framework for anaphor resolution that distinguishes two processing routes: a more resource-demanding discourse-based route and a less taxing syntax-only route.

Keywords: Pronoun resolution, Intra-sentential anaphora, Cognitive processing, Binding theory, Command relation, Event-related potentials.

1 Introduction

Anaphora is a prominent research subject in cognitive science because it brings to fore one of the most startling properties of language: whereas most content words in a language determine their meaning by having a concept associated with it (e.g., the word *boy*), some words don't behave in this way. Instead, their meaning must be determined by another expression in the context. Such is the case of pronominals: in a context formed by the single sentence *The boy said the barber shaved him* the meaning of the word *him* is completely dependent on the meaning of its antecedent *the boy*. The only independent information

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 107–120, 2009.

this pronoun conveys is of gender (masculine) and number (singular). Given this situation, the question arises as to what the mechanisms are that allow the pronominal to be interpreted. In other words, what are the mechanisms that allow to determine its antecedent, i.e. what is the process by which this anaphoric expression is "resolved".

The full availability of pronominals (defined in terms of this referential dependency) is attested across the languages of the world. This ubiquitous presence seems then the manifestation of an organizing principle of the language system and of its connection to the rest of the cognitive system.

The inherently dynamic nature of anaphora resolution has inspired information processing-based models from a multiplicity of disciplines including artificial intelligence and psycholinguistics. Supported by current computational technology, the models designed and implemented in the area of computational natural language processing underlie anaphora resolution modules that aim at optimal performance when running on free input text. In accordance with the nature and goals of computer science as supporting discipline, these are algorithmic models, relying on fully specified and articulated information processing procedures (e.g. [2,9]) . By contrast, the models from psychological approaches are usually not specified at the algorithmic level and contain, therefore, substantially less computational detail (e.g. [5]). This is to be expected since their construction is driven by the higher level goal of capturing generalizations and generating testable hypotheses based on behavioral evidence obtained from human subjects. This difference in objectives and methods has led to a multifaced understanding of anaphora resolution which has been most beneficial to our understanding of this domain. Many experimental findings have been integrated on the anaphora resolution models subscribed by computational linguistics. And in the reverse direction, increasingly articulated models of anaphora resolution can inspire exploratory experimentation aimed at testing increasingly non trivial empirical hypotheses.

In a broad outline, the anaphora resolution process that emerges as common to all or most of these models is that the anaphoric links between anaphors and their antecedents can be viewed as being established as a result of the interplay of a number of constraints (e.g. [1,3]). These constraints can be conceptually split into two categories: filters (hard constraints) and preferences (soft constraints). For a given anaphor, the relevant hard constraints circumscribe the set of its admissible antecedents by filtering out the non admissible ones from the set of all the possible antecedent candidates around in the context. The relevant preferences, in turn, concur to favor, though not to determine, the selection of the actual antecedent (or actual antecedents, in cases of some occurrences of plural anaphors) against which the anaphor happens to be eventually interpreted.

Against this background, the present work investigates the operation of a particular subset of preferences, namely, those that rely on the use of structural information as criteria to single out the preferred antecedents. We further focus our study upon the operation of these preferences as they apply during the processing of intra-sentential anaphoric links i.e., cases where the anaphor and its

antecedent candidates occur in the same sentence. Preferences operate in tandem with hard constraints. The hard constraints relevant for the processing of intra-sentential anaphora are dubbed binding principles in grammatical studies. In this work, we will be concerned with binding principle B that holds for the anaphoric expressions classified as pronouns. Thus, we set out to study the operation of preferences defined upon structural relations – namely the *command* relation - holding between pronouns and intra-sentential antecedent candidates.

The remainder of the paper is organized as follows. Section 2 presents in more detail the constraints on intra-sentential anaphoric links, with a focus on binding principle B, followed by a review of previous work on the linguistic, psycholinguistic and computational manifestations of this principle and by the full articulation of our present approach, including our hypotheses and corresponding predictions.

Section 3 describes our study including methods and results. Section 4 discusses the hypotheses in the light of the results obtained and concludes the paper.

2 Background and Hypotheses

2.1 Grammatical Studies

The subset of grammatical constraints impinging on intra-sentential anaphoric links, known as binding principles, are defined in terms of two auxiliary relations: the command relation and the locality relation. We introduce each in turn.

The command relation is established over predicate-argument structures. When an expression is an argument of a given predicator, it is said to be commanded by its less oblique co-arguments — i.e. by the other less oblique arguments of that predicator. For instance, a Direct Object of a given predicator is commanded by the Subject of that predicate, an Indirect Object by the Subject and the Direct Object, etc. In the example sentence *The brother of Mary offered a book to John*, the Indirect Object *to John* is commanded by the Direct Object *a book* and by the Subject *the brother of Mary*, both arguments of the predicator *offered*; even though *to John* is the most oblique of the arguments selected by the predicator *offered*, it is not commanded by *of Mary* as this expression is not an argument of that predicator.

Moreover, the command relation is established recursively along the embedding of successive argument selection relations. For instance, a Direct Object α is commanded by the Subject of its predicator and, if the corresponding predication domain is the argument of another, upwards predicator,α is commanded also by all the arguments of this upwards predicator that commands its predication domain. In the example sentence *Tom promised to Peter that the brother of Mary will offer a book to John*, the Indirect Object *to John* of the embedded clause is commanded also by the Subject *Tom* of the predicator *promised* as this predicator selects as its Direct Object the predication domain *the brother of Mary will offer a book to John*, where *to John* occurs; *to John* is not commanded

by *to Peter* as this expression is not less oblique than the predication domain where *to John* occurs.

The locality relation is also established over predicate-argument structures. Two expressions are local to each other when they are co-arguments with respect to a given predicator. For instance, the Direct and Indirect Objects of a given predicator are local to each other. In the example sentence *Tom promised to Peter that the brother of Mary will offer a book to John*, the Direct Object *a book* of the predicator *will offer* is local with respect to the Subject *the brother of Mary* and the Indirect Object *to John* of that predicate; it is not local with respect to the Subject *Tom* or the Indirect Object *to Peter* of the upwards predicator *promised*.

The grammatical constraints on intra-sentential anaphoric links of interest here are defined in terms of these two auxiliary relations. Given its aim and length, it is not in the scope of the present paper to offer a thorough discussion of these constraints. They will be partly introduced by way of two key examples.

(1) a. Peter$_i$ said that [[[John's]$_k$ brother]$_j$ shaved himself $_{*i/j/*k}$].
b. Peter$_i$ said that [[[John's]$_k$ brother]$_j$ shaved him$_{i/*j/k}$].

For an expression to qualify as an admissible antecedent of a reflexive anaphor like *himself*, that expression has to be one of its local commanders in case the reflexive is commanded (principle A).[1] This is exemplified with the contrasts in (1a). The Subject *John's brother* is the only expression in that sentence that is both local with respect to *himself* and commands it. The Subject *Peter* of the upwards predicator *said* is not local with respect to it. And the argument *John* of the relational noun *brother* does not command it. Only the first is an admissible antecedent of *himself* and this anaphoric link, between *himself* and *John's brother*, is an instance of the empirical generalization that principle A seeks to capture.

While this is a first example illustrating the role of the command and the locality relations in the definition of observed constraints on anaphoric links, the focus of the present paper is on the specific constraint on intra-sentential anaphoric links for pronouns, an empirical generalization that is sought to be captured by principle B. An instance of this generalization is exemplified with the contrasts in (1b).

For an expression to qualify as an admissible antecedent of a pronominal anaphor like *him*, that expression cannot be one of its local commanders (principle B). The Subject *John's brother* is the only expression in that sentence that

[1] For a reflexive that is not locally commanded, the constraint expressed by principle A does not hold and the reflexive can establish anaphoric links with antecedent candidates that are not its local commanders. In such cases, the anaphoric expression has been characterized as a logophor [12]. In other theoretical settings, this behavior is explained on the basis that the reflexive is in a so-called exempt position, i.e. in a position where it is exempt from the discipline of principle A [11]. In either case, it is expected that the resolution mechanisms for these elements are different from those for pronouns with commanding antecedents.

is both local with respect to *him* and commands it. Therefore while *Peter* and *John* are admissible antecedents of *him*, *John's brother* is the only antecedent candidate in that sentence that does not qualify as an admissible candidate of the pronoun.

2.2 Previous Behavioral Work

As illustrated in the brief discussion above, the grammatical notions of command and locality have emerged chiefly to characterize the observed constraints on intra-sentential anaphoric links between anaphoric expressions and their admissible antecedents.

Concomitantly, the command relation has been claimed to play a role also in terms of characterizing preferences impinging on the selection of the actual antecedent. The set of admissible antecedents can, in fact, be split into two disjoint subsets: One of the subsets contains the admissible antecedents that command the anaphoric expression at stake, and the other subset the ones that do not command it.[2] It has been argued that these two subsets are not of equal standing with respect to the processing of anaphoric expressions with intra-sentential antecedents. This instance of anaphor resolution is allegedly subject to a preference that favors resolution against elements of one of these two subsets – the commanders – to the detriment of the elements of the other group – the recessors.

Some authors have proposed to explain these performance effects by resorting to syntactic and/or discursive factors possibly impinging on the cognitive processing of anaphora (e.g., [6,12]). Namely, the preference for commanders would be due to the fact that resolution to antecedents holding such status could obtain by syntactic means alone, while resolution to recessors would require processing at the discourse level. Koornneef et al. [8] expand on this rationale, by identifying two alternative fine-grained explanations for the preference towards commanders: either anaphor resolution involving the discourse level is intrinsically more resource demanding than resolution by syntactic means alone, or the experimental effects observed are in fact due to the processing time-course, in which commanders would be available before non-commanding antecedent candidates. The hypothesis most extensively explored in this connection has been the "increased processing load" hypothesis. We review below a couple of contributions that have assessed this hypothesis by resorting to behavioral experimentation.

Piñango and Burkhardt [10] studied possible differences in the processing of anaphoric links to commanders (2a) vs. links to recessors (2b) holding between reflexives and their antecedents. In one of the experiments, test materials like the following were used:

[2] The candidate antecedents that do not command the relevant anaphoric expression occur in a position of the predicate-argument structure of the sentence that can be viewed as a recess with respect to the position of that anaphor (when climbing up that structure from the anaphor's position). For the sake of brevity, non commanding candidate antecedents (of a given anaphor) will be refered to in the remainder of this paper as recessors (of that anaphor).

(2) a. [The driver$_i$ who caused a crash blamed himself$_i$].
b. The therapist$_i$ rolled a ball [around himself$_i$]

In (2b), the reflexive *himself* is the sole argument of a predicate, namely the semantically-loaded preposition *around*. As this reflexive is not commanded, the constraint captured in principle A does not apply and it can establish anaphoric links with commanders or with recessors. In the present case, the reflexive is resolved against a recessor, namely *the therapist*.

In contrast, in (2a) *himself* is commanded and the constraint captured in principle A is in force. In this example, the reflexive is anaphorically linked to a commander, namely *the driver*.

For their experiment, these authors resorted to the cross-modal lexical decision interference paradigm. The lexical decision task consisted in pressing a button if the string displayed was a word. They recorded the reaction time to a visual probe appearing immediately after the occurrence of the anaphor in the sentence being listened.

The result was in line with the hypothesis as the reaction time for the condition concerning anaphors resolving to recessors was "statistically significantly higher" than the condition for anaphors resolving to commanders. This experiment was later replicated with materials involving the Dutch *zich*, leading to similar results [4].

Another experimental assessment of the hypothesis at stake was undertaken by [8]. This study resorted to the eye-tracking experimental methodology. It tested possible differences in the processing of pronouns anaphorically resolved to commanders (3a) or to recessors (3b), by resorting to materials illustrated by the following excerpts:

(3) a. [Every worker who just like Paul was running out of energy]$_i$ thought it was very nice that he$_i$ could go home early this afternoon.
b. [Every worker who knew that Paul$_i$ was running out of energy] thought it was very nice that he$_i$ could go home early this afternoon.

It was found that "readers refixated the critical region (i.e. containing the pronoun) and the preceding region longer in the [recessor] condition than in the [commander] condition". Hence, the results reported in this study can also be interpreted as being in line with the above hypothesis.

2.3 Previous Electrophysiological Work

A number of studies, with a specific focus on anaphora resolution, have resorted to evidence based on indicators of neural activity obtained through event-related brain potentials (ERP). However, to the best of our knowledge, the present is the first attempt at measuring antecedent selection using this methodology. Moreover, this is also the first time that this kind of research is carried out in

European Portuguese, which differs in interesting ways from the better studied Germanic counterparts such as English, German and Dutch.

Some studies have been concerned with ambiguous anaphors or the contrast between different types of anaphoric expressions. For instance, Streb et al. [14] found a 270-400 ms frontal negativity and a 510-600 ms parietal negativity elicited by pronouns in contrast to definite descriptions that are resolved to the same extra-sentential antecedent. Van Berkum et al. [15,16] identified a sustained frontal negativity, emerging at about 300-400 ms, elicited by ambiguous pronouns in contrast to non-ambiguous ones. The authors dubbed this effect Nref and suggest that it specifically indexes ambiguity, possibly reflecting the additional neuronal activity required to simultaneously keep two competing referential interpretations in working memory.

Harris et al. [7], concerned with a specific constraint on intra-sentential anaphoric links, identified a P600 effect elicited by a violation of principle A while resolving reflexives.

Still other studies have focused on preferences for anaphora resolution. Streb et al. [13] studied the recency preference and brought to light a N400 effect elicited by pronouns resolved to more distant inter-sentential antecedents than pronouns resolved to more recent ones. Streb et al. [14], in turn, found a 510-630 ms parietal negativity elicited by pronouns resolved to inter-sentential antecedents in a non parallel grammatical function, in contrast to pronouns resolved to antecedents in parallel grammatical functions. The authors interpret this enhanced negativity as a member of the N400 family. This indexing increased the processing demands for resolution of antecedents in a non parallel grammatical function.

2.4 Hypotheses

In this paper, we test the hypotheses that intra-sentential anaphor resolution to recessors differs from resolution to commanders (i) in terms of computational cost, with resolution to a recessor being more costly than resolution to a commander; and (ii) in terms of processing time course, with commanders being made available to the processor before recessors.

We expect (i) to entail a N400-like effect, as described by [14], elicited by pronouns resolved to recessors, and (ii) to entail a Nref effect, as described [15,16], elicited by pronouns resolved to recessors.

Hypothesis (i) stands straightforwardly from [14], taking the amplitude of the N400-like effect evoked by pronouns as an index of computational cost.

The rationale for hypothesis (ii) is as follows: In our material, in the cases where resolution is made to recessors, this resolution is forced by gender agreement – only the non-commanding antecedent candidate agrees in gender with the pronoun. If the morphological information relevant to determine gender agreement can be accessed in parallel for both the commanding and non-commanding antecedent candidates, resolution to the commander should be blocked, given that an antecedent with the suitable gender inflection value is available to the processor. However, if the commander bearing the gender mismatch is momentarily the only alternative

available, the processor should pursue the possibility of resolving the pronoun to that antecedent, repairing the gender mismatch at a later stage, in the P600 window. As a candidate with the suitable gender value is eventually made available to the processor, repairing the gender mismatch is rendered unnecessary: a referential ambiguity should emerge instead, indexed by a transient Nref effect, as the ambiguity is subsequently resolved on the basis of gender information.

The contribution of this study relies not only on the novelty of the hypotheses — which focus on intra-sentential pronoun interpretation —, but also on the methodology used — event-related response potentials (ERPs) —, and on being the first one which reports on observation from Portuguese, a Romance Language.

3 Experiment

3.1 Methods

Participants: Eighteen students (five female) at Coimbra University participated in the experiment for partial fulfillment of course requirements. All participants were right handed monolingual native speakers of Portuguese, with normal or corrected-to-normal vision. Their age ranged from 18 to 25 years (mean age: 20.75; SD = 2.34). Data from six subjects were excluded from further analysis due to insufficient number of valid trials.

Materials: Two conditions were tested in this study: pronouns resolved intra-sententially to commanding antecedents (Antec-Comm) and to non-commanding antecedents (Antec-Recess).

Forty eight pairs of items were designed, differing in the factor Antecedent Command Status (Antec-Comm/Antec-Recess). Gender agreement was used to disambiguate the intended resolution either to the commanding antecedent (4a) or to the non-commanding antecedent (4b).

(4) a. [O mordomo-MASC de [a condessa-FEM]]-MASC$_i$ discutiu com [a criada]-FEM a quem ele-MASC$_i$ tinha emprestado algum dinheiro. (Antec-Comm)
The butler$_{male}$ of the countess quarreled with the servant$_{female}$ to whom he had lent some money.

b. [A empregada-FEM de [o talhante-MASC$_i$]]-FEM discutiu com [a cliente]-FEM a quem ele-MASC$_i$ tinha vendido carne estragada.
The employee$_{female}$ of the butcher quarreled with the client$_{female}$ to whom he had sold spoiled meat.

Forty eight different matrix verbs were used to create the experimental sentences. Each verb occurred once in both conditions, yielding 48 pairs of sentences sharing the same matrix verb. Gender inflection values of the pronominal element

and of its intended antecedent were counterbalanced across pairs. Except for the intended antecedent, all other nominal phrases occurring before the pronominal bear a gender value that is opposite to the gender value of that pronominal. The pronominal element was immediately followed by an auxiliary verb, counterbalanced across pairs. The remainder sentential material, occurring after the time-windows of interest for the ERP analysis, was tailored in order to maximize the pragmatic acceptability of each individual sentence. In addition to the 96 experimental stimuli, 144 filler sentences were created. Items were pseudo-randomized and counterbalanced. Three different orderings were used across subjects to control for sequence effects.

Procedure: Participants were seated comfortably in front of a 19" computer screen, at a distance of approximately 100 cm, and presented with the task instructions, followed by a block of 9 practice trials. They were asked to process the sentences for comprehension and instructed to move as little as possible.

The experimental stimuli were presented visually, word by word, in the center of the computer screen. A fixation cross, appearing for 500 ms, served as a reminder for the participants to stop blinking. Each word was displayed for 300 ms, followed by a 300 ms blank screen interval. The final word of the experimental sentences was presented together with a period sign. Following the final word, three dots were displayed in the center of the screen, signaling to the participant that she was free to blink until the next fixation cross would appear. In order to foster the participant's commitment to the sentence comprehension task, 800 ms after each sentence, a force-choice question was presented. The next trial began 2500 ms after collection of the participant's answer to the comprehension question.

Answers were collected by means of two response switches, one held in the participant's right hand, the other in her left hand. The comprehension question was displayed on the top-middle section of the screen, together with two answer options, one on the left lower corner of the screen, the other on the right lower corner. Three different types of questions were used, asking the participant to decide (i) which of the two entities referred before the main verb was the agent of the action conveyed by the main clause; (ii) which entity was the possessor in the genitive construction occurring before the main verb; (iii) which of the two entities referred before the main verb was the agent of the action conveyed by the relative clause. The participants were instructed to press the switch that they were holding in the hand directly in front of the correct option.

Short breaks were introduced approximately every 8 minutes.

EEG recording: Electroencephalogram recordings were collected from 64 Ag/AgCl scalp active electrodes mounted in an electrode cap conforming to the 10-20 system for electrode positioning. Vertical eye movements and blinks were monitored via a supra- to sub-orbital bipolar montage. A right-to-left canthal bipolar montage was used to monitor for horizontal eye movements. Electrode offsets were kept within the interval 25 μV to -25 μV. The signals were recorded continuously with a digitization rate of 512 Hz and referenced to the average of all electrodes.

3.2 Results

Behavioral data: All participants performed at near ceiling level.

ERP data: Data were band-pass filtered offline to 0.5-40 Hz and screened for eye-movements, muscle artifacts, and electrode drifting. A total of 18% trials were rejected due to artifact contamination. Blink artifacts were removed using an independent component analysis filter algorithm. Data from six subjects were excluded from further analysis because one of the experimental conditions had less than 25 acceptable trials. ERPs were time-locked to the pronominal element and computed using the waveforms from all the trials of the remainder twelve participants: epochs raging from 150 ms pre stimulus to 1500 ms post stimulus were extracted, baseline corrected using the pre stimulus period, and averaged per condition.

ERP data from 61 electrodes were analyzed for the LAN (250-450 ms), N400 (400-600 ms) and P600 (550-800 ms) time windows, by means of repeated-measures ANOVAs. Separate ANOVAS were performed for lateral and central scalp regions. The electrodes were grouped into nine regions, on the basis of their topographical distribution. The lateral ANOVAs were conducted with the factors Gradient – anterior, medial and posterior - and Hemisphere – left and right, corresponding to six regions: anterior left (AF7 AF3 F7 F5 F3), medial left (FT7 FC5 FC3 T7 C5 C3 TP7 CP5 CP3), posterior left (P7 P5 P3 PO7 PO3), anterior right (AF8 AF4 F8 F6 F4), medial right (FT8 FC6 FC4 T8 C6 C4 TP8 CP6 CP4) and posterior right (P8 P6 P4 PO8 PO4). The central ANOVAs

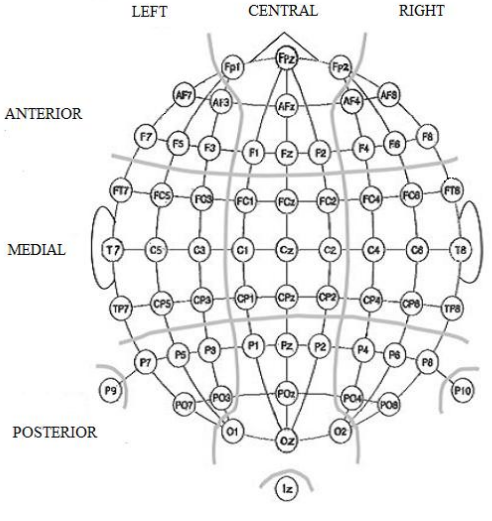

Fig. 1. Electrode groupings corresponding to the regions of interest used in the lateral ANOVAs (anterior left, medial left, posterior left, anterior right, medial right, posterior right) and central ANOVAs (anterior central, medial central, posterior central)

were conducted with the factor Gradient, corresponding to three regions: anterior central (FP1 FPz FP2 AFz F1 Fz F2), medial central (FC1 FCz FC2 C1 Cz C2 CP1 CPz CP2), and posterior central (P1 Pz P2 POz O1 Oz O2). Electrodes P9, lz and P10 were excluded from the analysis.

We report below the results for the two analyses that yielded significant effects or trends involving the Antecedent Command Status variable (Antc. Status). Huynh-Feldt correction was used whenever there was more than one degree of freedom in the numerator. Follow-up pairwise comparisons were computed using Bonferroni adjustment for multiple comparisons.

The central ANOVA for the 400-600 ms window showed a significant main effect for the variable Antec. Status ($F(1, 11)=11.93$, $p<0.01$, $MSE=0.062$). Inspection of the estimated marginal means for this variable reveals a more pronounced negativity when the pronoun is bound to a non-commanding antecedent. A significant polynomial quadratic trend occurs for the Gradient x Antec. Status interaction ($F(1, 11)=12.80$, $p<0.01$, $MSE=0.043$). Pairwise comparisons for the Gradient x Antec. Status interaction show a significant effect for the Antec. Status variable only for the medial central region.

The lateral ANOVA for the 250-450 ms window showed a significant main effect for the variable Hemisphere ($F(1, 11)=11.93$, $p<.05$, $MSE=0.62$), a significant interaction Gradient x Hemisphere ($F(2, 22)=4.91$, $p<.05$, $MSE=0.339$), and a marginally significant main effect for the variable Antec. Status ($F(1, 11)=3.89$, $p<0.1$, $MSE=0.09$). Inspection of estimated marginal means for Hemisphere reveals a more pronounced negativity over the left hemisphere. Follow-up pairwise comparisons for the Gradient x Hemisphere interaction show that this lateralized negativity only holds for the medial and posterior regions; the anterior region bears a negativity that spreads to the right hemisphere. The marginally significant main effect for Antec. Status suggests that pronoun resolution with a non-commanding antecedent elicits a more pronounced overall negativity, which conforms to the previously described spatial distribution pattern.

4 Discussion

As mentioned above, the contribution of this study relies on the novelty of the hypotheses, concerning anaphora resolution against commanders vs. recessors, which focus on intra-sentential pronoun interpretation, and not on reflexives as in previous related works. It relies also on the methodology, which for the first time explores ERPs to investigate the issues at stake. And last but not least, it relies on being the first one to report on observation from Portuguese, thus extending this type of inquiry to language materials from Romance Languages.

In line with [14], we interpret the medial relative negativity found in the 400-600 ms window as an N400-like effect, signaling the effects of the experimental manipulation upon the formation of the pronoun-antecedent dependency. This enhanced negativity for pronouns resolved to non-commanding antecedents suggests that, in line with hypothesis (i), resolving a pronoun to a recessor is a more resource demanding process than resolving it to a commander.

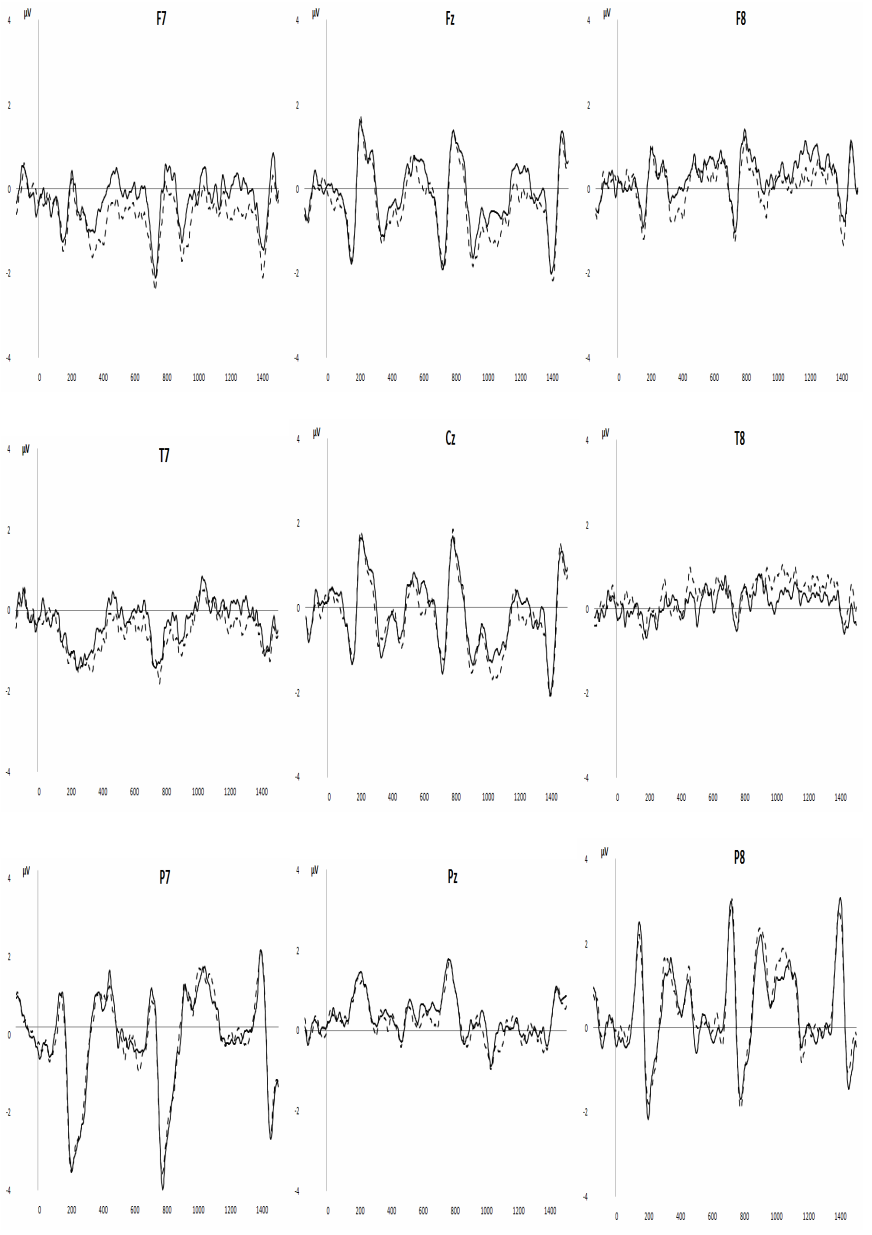

Fig. 2. Grand average ERPs (n=12) measured to the onset of the critical pronoun resolved with a commanding antecedent (solid line) and pronoun resolved to a non-commanding antecedent (dashed line). Waveforms are plotted from a 150 ms pre stimulus baseline to 1500 ms post stimulus. Pronouns resolved to a non-commanding antecedent elicit a fronto-lateral dominant negativity in the 250-450 ms window, as illustrated at F7 and F8, and a central negativity in the 400-600 ms window, as illustrated at Cz. Negative is plotted down.

The most apt explanation for this difference is that whereas the selection of commanders as antecedents is done through syntactic means alone, the selection of recessors as antecedents requires the implementation of the selection through an additional informational layer, namely, discourse representation. The level of complexity is thus determined by how many layers of information (syntax only vs syntax+discourse) are required to ultimately establish the antecedent selection.

Hypothesis (ii) also gathered confirmatory evidence. The marginally significant effect found in the 250-450 ms window for the Antecedent Command Status variable consists of a wide-spread relative negativity elicited by pronouns resolved to recessors. This enhanced negativity is distributed according to an overall pattern of Hemisphere and Gradient effects, characterized by (i) a similar gradient pattern over each hemisphere, with anterior and medial regions more negative than the posterior region, (ii) a left-lateralized negativity for the medial and posterior regions, (iii) a bi-hemispheric negativity for the anterior region. The spreading of the anterior negativity to the right hemisphere is mainly due to the contribution of the Antec-Recess condition. The relative negativity elicited by pronouns resolved to recessors is, therefore, indicative of a (short-lived) Nref-like effect. Van Berkum et al. [15,16] describe the Nref as a bilaterally and globally distributed negativity, frontally dominant, elicited by anaphors with two admissible antecedent candidates in contrast to anaphors with a single admissible antecedent candidate.

We interpret the Nref-like pattern found in our experiment as indicating that the commanding antecedent is made available to the resolution process before the recessor, and momentarily entertained as the sole available alternative. The processor therefore pursues the possibility of resolving the pronoun to that antecedent even when it mismatches the pronoun in gender. A referential ambiguity emerges when a recessor with a suitable gender value is eventually made available to the processor, indexed by Nref effect. The Nref negativity is not a sustained one, unlike what is more frequently observed in manipulations that evoke the Nref effect. This is to be expected, since in this instance, the ambiguity is readily resolved on the basis of gender information.

Altogether, these findings nicely converge with a large body of work based on behavioral evidence showing that establishing this kind of discourse-based dependency (resolution to recessor) is more computationally demanding than establishing dependencies based on syntactic mechanisms alone (resolution to commander).

References

1. Asher, N., Wada, H.: A Computational Account of Syntactic, Semantic and Discourse Principles for Anaphora Resolution. Journal of Semantics 15, 83–113 (1988)
2. Branco, A.: Binding Machines. Computational Linguistics 28, 1–18 (2002)
3. Branco, A.: Anaphoric Constraints and Dualities in the Semantics of Nominals. Journal of Logic, Language and Information 14, 149–171 (2005)
4. Burkhardt, P., Piñango, M.M., Ruijgendijk, E., Avrutin, S.: Reference Assignments in Dutch: Evidence for the syntax-discourse divide, Lingua (accepted for publication, 2009)

5. Garnham, A.: Reference and Anaphora. In: Garrod, Pickering (eds.) Language Processing. Psychology Press (1999)
6. Grodzinsky, Y., Reinhart, T.: The Innateness of Binding and Coreference. Linguistic Inquiry 24, 60–101 (1993)
7. Harris, T., Wexler, K., Holcomb, P.: An ERP Investigation of Binding and Coreference. Brain and Language 75, 313–346 (2000)
8. Koornneef, W., Wijnen, F., Reuland, E.: Towards a Modular Approach to Anaphor Resolution. In: Ambiguity in Anaphora Workshop Proceedings, ESSLL 2006, pp. 65–72 (2006)
9. Mitkov, R.: Anaphora Resolution. Longman (2002)
10. Piñango, M.M., Burkhardt, P.: Pronominal Interpretation and the Syntax-Discourse Interface: Real-time Comprehension and Neurological Properties. In: Branco, McEnery, Mitkov (eds.) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, pp. 221–237 (2002)
11. Pollard, C.J., Sag, I.A.: Head-driven Phrase Structure Grammar. CSLI (1994)
12. Reuland, E.: Primitives of Binding. Linguistic Inquiry 32, 439–492 (2001)
13. Streb, J., Henninghausen, E., Rösler, F.: Different Anaphoric Expressions are Investigated by Event Related Brain Potentials. Journal of Psycholinguistic Research 33, 175–201 (2004)
14. Streb, J., Rösler, F., Hennighausen, E.: Event-related Responses to Pronoun and Proper Name Anaphors in Parallel and Nonparallel Discourse Structures. Brain and Language 70, 273–286 (1999)
15. Van Berkum, J.J.A., Brown, C.M., Hagoort, P., Zwitserlood, P.: Event-Related Brain Potentials Reflect Discourse-Referential Ambiguity in Spoken-Language Comprehension. Psychophysiology 40, 235–248 (2003)
16. Van Berkum, J.J.A., Koornneef, A.W., Otten, M., Nieuwland, M.S.: Establishing Reference in Language Comprehension: An Electrophysiological Perspective. Brain Research 1146, 158–171 (2007)

Effects of Anaphoric Dependencies and Semantic Representations on Pronoun Interpretation

Elsi Kaiser

University of Southern California, Department of Linguistics,
Los Angeles, California 90089, USA
emkaiser@usc.edu

Abstract. We describe three experiments that use priming methodology to investigate the nature of the abstract mental representations activated during pronoun resolution, in order to contribute to our understanding of how structural representations and semantic coherence representations influence pronoun interpretation. The results of Experiment 1 suggest that there exists a level of abstract anaphoric dependency representations. Experiments 2 and 3 indicate that general coherence representations are activated during pronoun interpretation and thus provide direct evidence for the existence of shared abstract representations between (non-pronominal) coherence-related inferencing and pronoun resolution processes. Moreover, Experiment 3 has implications for our understanding of the connections between linguistic and non-linguistic cognitive processes.

Keywords: Discourse coherence, Pronoun interpretation, Psycholinguistics, Eye-tracking, Priming, Anaphoric dependencies.

1 Introduction

Pronouns are central for communication, but the processes and representations underlying pronoun resolution are not yet fully understood. Competing accounts disagree regarding the contribution of syntactic factors, semantic factors and factors influencing referent prominence more generally. For example, coherence-based accounts (e.g. [6,7,8,9,12]) attribute a more central role to intersentential semantic relations (e.g. *X is a result of Y, W is similar to Z*) than to syntactic representations, in contrast to more structurally-oriented approaches. We describe three experiments that use priming methodology to investigate the nature of the abstract mental representations activated during pronoun resolution, in order to contribute to our understanding of how structural representations and semantic coherence representations influence pronoun interpretation. More broadly, this research aims to further our understanding of how humans perceive the semantic relations between objects, events and situations.

Priming is a well-known phenomenon according to which prior exposure to a stimulus influences (often facilitates) subsequent processing of a similar stimulus (or the same stimulus). For example, existing work has shown robust effects of syntactic priming in production [2,11], indicating that producing a particular syntactic structure facilitates subsequent production of the same structure.

S. Lalitha Devi, A. Branco, and R. Mitkov (Eds.): DAARC 2009, LNAI 5847, pp. 121–129, 2009.

Recent work suggests that syntactic priming also exists in comprehension (e.g. [5,16]).

In the experiments discussed here, we build on the logic of existing research and use priming as a tool to test whether two processes make use of the same (or overlapping) underlying representations. Although existing work has found evidence for priming of abstract syntactic representations, the abstract representations activated during pronoun interpretation have not yet been directly investigated. We explored two related questions: (i) *Experiment 1:* Does pronoun interpretation result in the activation of abstract anaphoric dependency representations? (ii) *Experiments 2 and 3:* Does pronoun interpretation result in the activation of abstract, possibly domain-general, coherence representations?

2 Experiment 1: Abstract Referential Dependencies

This experiment tested for the possibility of priming on the level of anaphoric dependency relations. Does processing a certain coreferential configuration facilitate subsequent comprehension of the same configuration? In other words, if a comprehender has recently constructed a certain kind of anaphoric dependency (e.g. interpreted an object pronoun as referring to the subject of the preceding clause), is s/he biased to construct the same kind of dependency again when faced with an ambiguous pronoun? If we find evidence for such priming, this would provide evidence for the existence of a distinct level of anaphoric dependency representations.

To investigate these issues, we conducted a comprehension-priming experiment. Participants (n=24) read sentences on a computer screen. Nonsense verbs were used to avoid effects of verb semantics. The prime sentences (ex. (1)) used gender cues to force a subject interpretation (ex. (1a)) or an object interpretation (ex. (1b)) of the pronoun in object position (Note that the syntactic structures of (1a) and (1b) are the same, and thus, any differences between subject primes and object primes cannot be attributed to syntactic priming). Neutral primes ended in intransitives (ex. (1c)). All sentences (targets, primes, fillers) were followed by questions. There was no noun or verb overlap between primes and targets.

(1a) William swooked Betty and Kevin brucked him. [Subject Prime]
(1b) William swooked Betty and Kevin brucked her. [Object Prime]
(1c) William swooked Betty and Kevin brucked. [Neutral Prime]
(2) Target: Stephen tulvered Peter and Diane churbited him.
[Question: Diane churbited ___. Stephen Peter]

The critical target sentences contained ambiguous object-pronouns (ex. (2)), whose interpretation (preceding subject or object?) was probed by the subsequent question, to test whether the preceding prime influences pronoun interpretation. Targets were preceded by subject primes, object primes or neutral primes, as we wanted to analyze whether the anaphoric dependency in the prime influences participants' interpretation of the ambiguous pronoun in the target. We measured and analyzed participants' responses to the questions and the speed of these responses.

2.1 Predictions

We expected targets to show a baseline object preference, given that we were testing object pronouns and that existing research has repeatedly found effects of structural parallelism (e.g. [13,3]), i.e., pronouns prefer antecedents in parallel structural positions. However, if it is the case that activation/construction of a particular kind of anaphoric dependency makes participants more likely to activate/construct that same kind of dependency again, then we predict that the object preference should be weakened by subject primes (e.g. 1a). In other words, subject responses should be more likely to occur or easier to process after subject primes than after neutral or object primes.

2.2 Responses

Overall, there were more object responses (89%) than subject responses (10%) as expected given structural parallelism. Crucially, participants' responses were modulated by preceding primes: There were roughly twice as many subject-responses after subject primes as after object primes or neutral primes (p's<.05).

The effect of the primes can be seen visually in Fig. 1 below, which shows the *object advantage score* for each of the three conditions, derived by subtracting the proportion of subject responses from the proportion of object responses. As Fig. 1 shows, although all three conditions have an overall object preference (as shown by the relatively high object advantage score in all three conditions), the object advantage score is lower after subject primes than after object primes and neutral primes.

2.3 Speed of Responses

Object-responses were approximately twice as fast as subject-responses. There was a numerical effect of prime on response speed: Subject-responses were faster after subject primes (<1500ms) than object primes (>1600ms) or neutral primes (approximately 2000ms).

2.4 Discussion

The results of Experiment 1 suggest that if a comprehender has recently processed/constructed a particular kind of anaphoric dependency, s/he is more likely to construct the same kind of dependency again when faced with an ambiguous pronoun.

In other words, it seems that anaphoric dependencies can be primed, even in the absence of noun/verb overlap between targets and fillers. This finding fits well with the idea that there exists a distinct level of anaphoric dependency representations, whose activation can linger and influence the interpretation of subsequent pronouns.

However, we should also keep in mind the question of how anaphoric dependency representation relate to coherence representations, because, as discussed

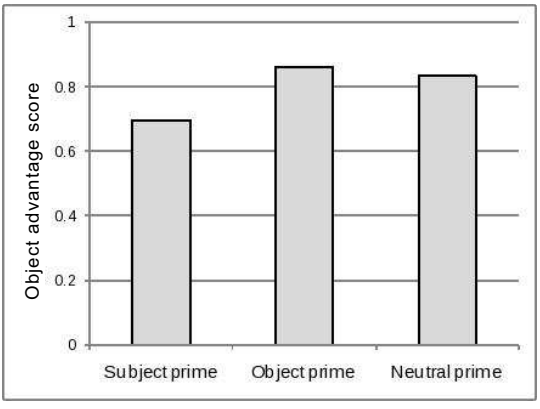

Fig. 1. The object advantage score for target sentences, shown as a function of prime type (The object advantage score is a measure of how strongly the object was preferred over the subject, and it was computed by subtracting the proportion of subject responses from the proportion of object responses)

in Section 3 below, certain anaphoric dependency patterns tend to be correlated with certain coherence relations. We will consider the implications of Experiment 1 again in Section 4, after presenting Experiments 2 and 3.

3 Experiments 2 and 3: Coherence Relation Priming

Experiments 2 and 3 tested whether the process of pronoun interpretation also activates abstract coherence representations. According to coherence-based approaches, pronoun resolution is largely a by-product of general inferencing about inter-clausal relations, and a growing body of research agrees that a successful account of pronoun interpretation needs to take into account the coherence relation between the pronoun-containing clause and the preceding clause. For example, in ex. (3), a subject interpretation of 'him' is more likely with a Cause-effect relation (3a) than with a Resemblance relation (3b) (e.g. [7,8,9,17]):

(3) Phil tickled Stan, and Liz poked him.
  (a) Phil tickled Stan, and [as a result] Liz poked him$_{Phil}$
  (b) Phil tickled Stan, and [similarly] Liz poked him$_{Stan}$

However, the nature of these coherence relations is not yet well understood. Experiments 2 and 3 aim to contribute to our understanding of the linguistic and cognitive properties of these representations. We address two main questions:

First, *Experiment 2* tested whether these representations are specific to reference resolution. According to [6], coherence relations are not restricted to the domain of pronoun resolution – Hobbs claims that inferences about coherence relations exist independently of pronoun interpretation. However, so far we have no

direct evidence as to whether the representations activated during pronoun interpretation are the same as the representations activated during (non-pronominal) coherence establishment. Experiment 2 tackles this question.

Second, *Experiment 3* investigated whether these coherence representations are specific to the linguistic level of representation. In other words, how domain-general are these representations? Are they restricted to the linguistic domain, or potentially shared between different cognitive domains? Given that relations such as cause-effect and similarity also exist in other domains, e.g. vision, one might well expect coherence relations to be domain-general.

3.1 Design and Methods

Using priming and visual-world eye-tracking, we tested whether processing a particular coherence relation influences interpretation of a subsequent ambiguous pronoun. We used *linguistic primes* (Experiment 2) and *visuo-spatial primes* (Experiment 3) of three types: (i) Cause-effect, (ii) Resemblance, and (iii) Neutral/Baseline. The Neutral primes are best regarded as a baseline, because they were designed to evoke other kinds of coherence relations that, crucially, are neither Cause-effect nor Resemblance.

The **linguistic primes** were visually-presented two-clause sentences (see ex. (4) for an example of a cause-effect prime), whereas the **non-linguistic/visuo-spatial primes** were silent video clips of moving geometric shapes of various colors (see ex. (5), Fig. 2 for an example of a cause-effect video clip).[1] The linguistic primes contained no subject or object pronouns, in order to prevent anaphoric dependency priming from occurring. Furthermore, both linguistic and visuo-spatial primes were normed beforehand to ensure that the intended coherence relation was clear.

(4) *Sample linguistic prime*
[Cause-effect condition]: The patient pressed the red emergency button near the bed and a nurse quickly ran into the room.

(5) *Description of sample video prime*
[Cause-effect condition]: A triangle knocks into a circle which falls off a ledge.[2]

Two tasks were used to ensure that participants attended to the primes. In Experiment 2, with linguistic primes, participants were shown a prime sentence on the computer screen and instructed to read it aloud and indicate whether they had seen it earlier during the experiment. In Experiment 3, with visuo-spatial primes, participants were instructed to watch the video and afterwards to use the mouse to trace the paths of the objects.

[1] In Experiment 3, the video prime component was played twice on critical trials (as well as some filler trials), to increase detectability of potential priming effects.

[2] No linguistic information accompanied the actual videos. Furthermore, the arrows in the example image are there simply for purposes of illustration; they were not present in the actual videos.

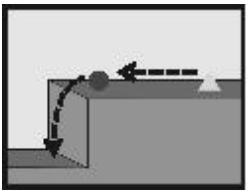

Fig. 2. Sample cause-effect prime (The arrows were not present in the actual videos)

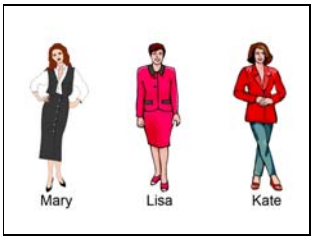

Fig. 3. Sample target visual display

The **critical targets** (which were the same in Experiments 2 and 3) were pictures of three same-sex characters (see Fig. 3), accompanied by an auditory sentence with an ambiguous object pronoun (ex. (6)). As in Experiment 1, we used nonsense verbs to eliminate any potential effects of verb semantics, and participants had been familiarized with the characters' names before the main experiment began. Each experiment contained 15 critical trials (prime+target sequences).

In addition, both experiments contained a large number of filler trials of two types: Some fillers resembled the targets whereas others resembled the primes. Fillers and targets were interspersed such that although critical trials consisted of a prime+target sequence, this patterning was not apparent to participants.

(6) Mary linded Lisa and Kate hepped her.

Eighteen native English speakers participated in each of the two experiments. Participants saw a scene and heard a sentence while their eye-movements were recorded. On target-type trials, the participants' task was to use the mouse to click on the last-mentioned person or thing.

Eye-movements in the visual-world paradigm are well-suited for investigating reference resolution, because existing research has shown that-eye movements to objects or pictures in a display are closely time-locked to the potential referents that a listener considers as language unfolds over time ([4,14], for a review see [15]). Thus, we can use looks to the different characters to gain insights into what participants are considering as potential referents for the ambiguous pronouns.

3.2 Predictions

If the abstract representations involved in pronoun resolution are connected to those in (pronoun-independent) coherence-related inferencing, we predict that lingering activation from primes can bias interpretation of ambiguous pronouns, resulting in more consideration of the subject after Cause-effect primes than after Resemblance or Neutral primes. (Because the pronouns are in object position, we expect an overall/eventual object preference, e.g. [13]) Crucially, using both linguistic primes (Experiment 2) and visuo-spatial primes (Experiment 3) allowed us to test whether the effects of coherence relation activation are specific to linguistic input, or whether they are more domain-general.

3.3 Results

Mouse click results: Participants' mouse clicks showed a strong object preference in all conditions.

Eye-movement results: After both linguistic and visuo-spatial Resemblance primes, eye-movements showed an early, persistent object preference (p's<.05) that emerged within 200 ms of pronoun onset. Neutral primes also resulted in an object preference, although it reached significance later than with Resemblance primes. In contrast, Cause-effect primes resulted in initial competition between subject and object. In the Cause-effect condition of Experiment 2 (linguistic primes), the object preference did not reach significance until the 600-800 ms time slice after pronoun-onset (p<.01). In the Cause-effect condition of Experiment 3 (video primes), the emergence of the object preference was also delayed relative to Resemblance primes and Neutral primes: With Cause-effect primes, the object preference did not reach significance (p<.01) until the 400-600 ms time slice. Thus, participants' eye-movement patterns show that Cause-effect primes resulted in relatively more consideration of the subject early on, in both Experiment 2 and Experiment 3. In sum, we found priming effects both with linguistic primes and with visuo-spatial primes.

4 Conclusions

As a whole, the outcomes of Experiment 1 (anaphoric dependency priming) and Experiments 2 and 3 (coherence representation priming) contribute to our understanding of what representations are activated during the pronoun interpretation process. The results of Experiment 1 indicate that priming exists in the domain of reference resolution: Processing a subject interpretation appears to make a subsequent subject interpretation more likely to occur and/or easier to process. These findings fit well with the idea that there exists a level of abstract coreference representations or procedures, whose activation can linger and thereby bias the interpretation of a subsequent ambiguous pronoun.

Furthermore, given existing research showing that certain anaphoric dependencies are associated with certain kinds of coherence relations (recall the subject

vs. object effects associated with Cause-effect relations and Resemblance relations in ex. (3)), we are faced with the interesting question of how (and whether) the anaphoric dependency priming of Experiment 1 is related to the priming of coherence relations. In other words, could the effects observed in Experiment 1 actually be due to activation originating from representations of coherence relations, which may in turn be activating particular anaphoric dependencies? Or perhaps a combination of anaphoric dependency priming and coherence relation priming? Although the design of Experiment 1 does not allow us to offer a definite answer to this question at this stage, the results of Experiments 2 and 3 certainly fit well with these possibilities.

The results of Experiments 2 and 3 show that pronoun interpretation can be primed by coherence relations in preceding linguistic input as well as preceding visual input, even when primes and targets are connected only on the level of abstract coherence relations, and even when they are not in the same modality. This shows that general coherence representations are activated during pronoun interpretation, and thus provides direct evidence for the existence of shared abstract representations between (non-pronominal) coherence-related inferencing and pronoun resolution processes.

The finding that visuo-spatial primes have an effect on pronoun interpretation suggests that the abstract coherence representations may in fact be *domain-general*, i.e., shared between linguistic and non-linguistic domains. This is an important question that deserves further study. In particular, we are faced with the question of whether the participants in Experiment 3 were perhaps re-coding the visuo-spatial information into linguistic information. To investigate this issue, an articulatory suppression experiment is currently underway. When participants are asked to do an articulatory suppression task (e.g., say the syllable 'the' repeatedly), the phonological component of working memory is engaged. This prevents subvocal rehearsal and as a result, participants' ability to re-code the visuo-spatial information into linguistic form is expected to be significantly impaired (see e.g. [10,1]). Thus, the outcomes of this experiment will help us to find out whether the abstract representations underlying coherence relations are domain general or specific to language.

In sum, the results of the three experiments discussed here shed light on the cognitive processes and representations activated during the process of pronoun resolution – in particular, they provide evidence for the existence of a distinct, non-pronoun-specific level of coherence representations – and also have implications for our understanding of the relationship between linguistic and non-linguistic representations.

Acknowledgements. Special thanks to Edward Holsinger for help with programming, stimulus creation and data collection. I would also like to thank three anonymous reviewers for their useful comments and insightful feedback. This research was partially supported by the funding from the Advancing Scholarship in the Humanities and Social Sciences Fund at the University of Southern California.

References

1. Baddeley, A.D.: Working memory. Oxford University Press, New York (1986)
2. Bock, K.: Syntactic persistence in language production. Cognitive Psychology 18, 355–387 (1986)
3. Chambers, G.C., Smyth, R.: Structural parallelism and discourse coherence: A test of Centering Theory. Journal of Memory and Language 39, 593–608 (1998)
4. Cooper, R.M.: The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory and language processing. Cognitive Psychology 6, 84–107 (1974)
5. Fedorenko, E., Gibson, E.: Syntactic priming in comprehension. Poster presented at the 18th CUNY Conference on Human Sentence Processing, Tucson, Arizona (2005)
6. Hobbs, J.R.: Coherence and coreference. Cognitive Science 3, 67–90 (1979)
7. Kehler, A.: Coherence, Reference, and the Theory of Grammar. CSLI Publications, Stanford (2002)
8. Kehler, A., Kertz, L., Rohde, H., Elman, J.: Coherence and Coreference Revisited. Journal of Semantics (Special Issue on Processing Meaning) 2, 1–44 (2008)
9. Kertz, L., Kehler, A., Elman, J.: Grammatical and coherence-based factors in pronoun interpretation. In: Proceedings of the 28th Annual Conference of the Cognitive Science Society, pp. 1605–1610 (2006)
10. Murray, D.J.: Articulation and acoustic confusability in short-term memory. Journal of Experimental Psychology 78, 679–684 (1965)
11. Pickering, M.J., Branigan, H.P.: The representation of verbs: Evidence from syntactic persistence in written language production. Journal of Memory & Language 39, 633–651 (1998)
12. Rohde, H., Kehler, A., Elman, J.: Pronoun Interpretation as a Side Effect of Discourse Coherence. In: Proceedings of the 29th Annual Conference of the Cognitive Science Society, pp. 617–622 (2007)
13. Smyth, R.: Grammatical determinants of ambiguous pronoun resolution. Journal of Psycholinguistic Research 23, 197–229 (1994)
14. Tanenhaus, M.K., Spivey-Knowlton, M.K., Eberhard, K.M., Sedivy, J.E.: Integration of visual and linguistic information in spoken language comprehension. Science 268, 632–634 (1995)
15. Tanenhaus, M., Trueswell, J.: Eye movements and spoken language comprehension. In: Traxler, M., Gernsbacher, M.A. (eds.) Handbook of Psycholinguistics, pp. 863–900. Academic Press, Amsterdam (2006)
16. Traxler, M., Pickering, M.: Syntactic priming in comprehension: Evidence from eye-movements. Talk presented at the 18th CUNY Conference on Human Sentence Processing, Tucson, Arizona (2005)
17. Wolf, F., Gibson, E., Desmet, T.: Discourse coherence and pronoun resolution. Language and Cognitive Processes 19, 665–675 (2004)

Author Index