Pronominal Resolution in Automatic Text Summarisation

Martin Hassel   **+46-8-6044252**

Lördagsvägen 21, bv   m-hassel@dsv.su.se

123 60 Farsta   http://members.tripod.com/
                                          ~mhassel/curvit/writings.htm
Sweden

Master Thesis 20p

Abstract

Automatic text summarisation is the technique where a computer program summarises a text. One summarisation technique is to, on linguistical and statistical grounds, extract sentences that are central to the text and use them to form a shorter text, the summary. In this case, automatically summarised text can sometimes result in dangling anaphors (i.e. broken anaphoric references). This is due to the fact that the sentences are extracted without making any deeper linguistic analysis of the text. We will in this thesis show a novel method to resolve some types of pronouns and to summarise the text such that coherence and important information is preserved. The method is implemented as a text pre-processor, called PRM (Pronoun Resolution Module) written in Perl. PRM uses lists of likely focuses (Sidner 1984), here called *focus applicants*. These lists are by order of insertion sorted in order of likelihood. Choice of applicant for an anaphor is based upon salience (represented by the antecedents position in a list) and semantic likelihood (based on what list the antecedent is to be found in). The latter is determined by using semantic information in a noun lexicon. This method enables us to resolve anaphors non-linearly. The domain is Swedish HTML-tagged newspaper text. A live version of the automatic summariser SweSum and PRM can be found at http://www.nada.kth.se/~xmartin.

Table of Contents

1 Introduction

Automatic text summarisation is the technique where a computer program summarises a text. The program is given a text and returns a summary of the original text. This is done by reducing redundancy in the text and by extracting the essence of the text. The technique has been developed through research for more than 30 years and with the increased use of the Internet and electronic documents the need for fast and reliable summarisation has undergone a rapid growth. Today popular word processors come with tools for document summarisation and the demand for automatic text summarisation has arisen in many areas. Examples of such areas are:

- Summarisation of news articles (for journalists, media researchers, Business Intelligence etc).
- Summarisation of reports (for government officials, businessmen, investigators, researchers, etc).
- In search engines in order to receive summaries, or a summary, of document(s) found.
- Summarisation of e-mails and news flashes for SMS.
- Producing summaries of web pages to be synthetically read out through telephone.

1.1 Background

Many different approaches for text summarisation have been proposed (Luhn 1959, Edmundson 1969 and Salton 1989). According to Hovy and Lin (1997) there are two ways to view text summarisation; either as text extraction or as text abstraction. Text extraction means to extract pieces of an original text on a statistical basis or with heuristic methods and put together it to a new shorter text with the same information content. Text abstraction is to parse the original text using linguistic methods interpreting the text and finding new concepts to describe the text and then generate a new shorter text with the same information content. In this thesis we will concentrate on text/sentence extraction, the method of summarisation used by the automatic text summariser *SweSum* (Dalianis 1999).

When performing text extraction there is always a risk that anaphoric references are broken. This happens when the extracted pieces of text contain one, or more, anaphoric expression(s), such as for example a pronoun, which has its antecedent in one of the pieces of text that was not extracted. This is a serious problem since this not only breaks the cohesion of the text; it can even alter the meaning of the text. Below is an example of this in a summary of a review (see Appendix A3) of the motion picture "Analyze This". The summary has been made with SweSum, the original text and summary in Swedish can be found in Appendix A.

Analyze This
Director: Harold Ramis
Starring: Robert De Niro, Billy Crystal, Lisa Kudrow
Length: 1 hour, 45 min
…
One of many reasons to rejoice at *Analyze This* is that Robert De Niro here really makes use of his acting talents again. He accelerates emotionally from 0 to 100 in no time at all, only to cattishly slow down and park, calm and steadily. And he is quite irresistible. Here he has accomplished yet another intelligent comedy for all of us enjoying intelligence and comedy, preferably in combination.
SvD 99-10-08 (manually translated from the Swedish automatic summary; see Appendix A1)

Figure 1. Example of text summarised without pronominal resolution.

In the summary in Figure 1., the impression is given that Robert De Niro has "accomplished yet another intelligent comedy…" but when reading the original text it is apparent that it is Harold Ramis who has accomplished this feat.

1.2 Purpose

This thesis describes the development of an efficient and easily implemented technique for resolving a specific category of anaphoric references, pronominal references, in Swedish unrestricted text. The technique has been realised as a limited pre-processor for SweSum, a text summarisation tool, in order to evaluate the technique. The purpose of this thesis is mainly to describe the developed technique, issues in pronominal resolution and how these can be tackled. The evaluation of the technique is secondary and only serves as an indication of the techniques enhancement of SweSum. SweSum is an automatic text summariser, which uses text extraction, and its domain is Swedish HTML-tagged newspaper text.

The purpose of developing such a technique was not only to avoid dangling anaphors and consequently improve the coherence of the text, but also to increase the percent of summarised text.

1.3 Method

The technique for resolving pronominal references in Swedish unrestricted text is implemented in Perl 5 using lexicons of semantic and syntactic information about nouns, adjectives and auxiliary verbs. The resulting pronominal resolution module has been tested in a user centred field test where students have made subjective estimations on various textual summaries cohesiveness and encapsulation of the original texts information content. At the time of the field test the implementation of the pronominal resolution module only resolved the pronouns 'han' (he), 'honom' (him), 'hans' (his) and 'hon' (she), 'henne' (her), 'hennes' (her).

2   Pronominal Resolution

When humans perform anaphora resolution while reading we do not only use lexical or syntactic information. We do not simply rely on the scribblings that appear in front of on a piece of paper, the screen or any other media. We keep in mind the relations of both central and peripheral topics and when we don't remember, we go back and refresh our memory. We also make use of semantic information, such as something mentioned about someone or something, and knowledge about the world outside the context of the text, such as that a ripe strawberry is red. When implementing an efficient technique for anaphora resolution we cannot simply model the complete knowledge of the world into the computer so naturally some issues are bound to arise.

Issues in automatic anaphora resolution are:

- How much can be achieved with only syntactic information about the text?
- When do you have to add semantic information and knowledge about the world (e.g. heuristic rules) in order to understand the context?
- In order to be as easily expanded as utilised, in what form should this semantic information be realised (general architectural issues)?

2.1 Mending the Broken

As we saw in Figure 1, the extraction of key sentences to form a summary can lead to broken anaphoric references. In this specific example we could see three instances of the pronoun 'han' (he) being extract to the summary. When applying pronominal resolution, the resulting summary will appear like this:

Analyze This
Director: Harold Ramis
Starring: Robert De Niro, Billy Crystal, Lisa Kudrow
Length: 1 hour, 45 min
…
One of many reasons to rejoice at *Analyze This* is that Robert De Niro here really makes use of his acting talents again. **Robert** accelerates emotionally from 0 to 100 in no time at all, only to cattishly slow down and park, calm and steadily. And **Robert** is quite irresistible. Here **Harold** has accomplished yet another intelligent comedy for all of us enjoying intelligence and comedy, preferably in combination.
SvD 99-10-08 (manually translated from the Swedish automatic summary; see Appendix A2)

Figure 2. Example of text summarised with pronominal resolution

Here we clearly can see that it is in fact Harold that has "accomplished yet another intelligent comedy…". So how is this done?

2.2 Introducing the Pronoun

In Swedish pronouns are divided into two classes; animate and inanimate pronouns. These are in turn divided into masculine/feminine respectively uter/neuter. This gives us the following structure:
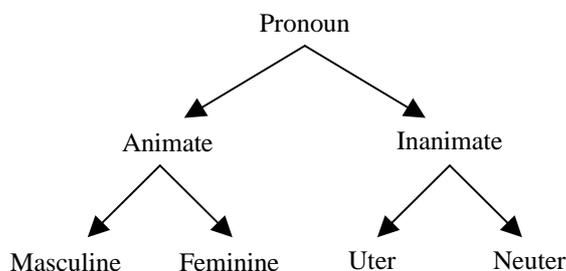

Figure 3. Pronoun structure in Swedish.

Which leads us to the following table:

Table 1. Pronoun categories in Swedish.

Category	Subject	Object	Possessive
Masculine	Han	Honom	Hans
Feminine	Hon	Henne	Hennes
Uter	Den	Den	Dess
Neuter	Det	Det	Dess

The plural form is the same for all four categories: *de* (s), *dem* (o) and *deras* (p).

The pronouns of the animate class are often said to be of natural gender since the referent in the real world 'naturally' belongs to either the masculine or the feminine category. There are though some exceptions, boats are for example sometimes referred to as 'she'. The pronouns of the inanimate class are in the same way often said to be of grammatical gender since there is nothing inherent in the real-world referent that makes it belong to either uter of neuter. Unfortunately there is nothing inherent in the *words* (the lexical entry) that make them belong to either of the categories.

2.3 Anaphoric Expressions and Cohesion

The ties that bind a text together are often referred to under the heading of *cohesion* (after Halliday & Hasan 1976). Several types of cohesive factor have been recognised. One of these is the coreference. Two types of coreferential relationship are recognised: *anaphoric* relations look backwards for their interpretation, and *cataphoric* relations look forwards (Crystal 1987). In this thesis we will focus on the anaphor but cataphors must be handled to some extent in order to be avoided (i.e. they must be identified as being cataphors). In order to resolve an anaphoric reference access to context is needed. Anaphors typically consist of a pronoun, which in itself do not carry any semantic information, but instead is interpreted by association with its antecedent's referent (Gustafson-Capková 1998). The pronoun can be seen as implicitly linked to its antecedent's counterpart in the real world *through the antecedent*, and the antecedent can be seen as being explicitly linked to its corresponding real-world entity. Graphically this relationship would look something like this:
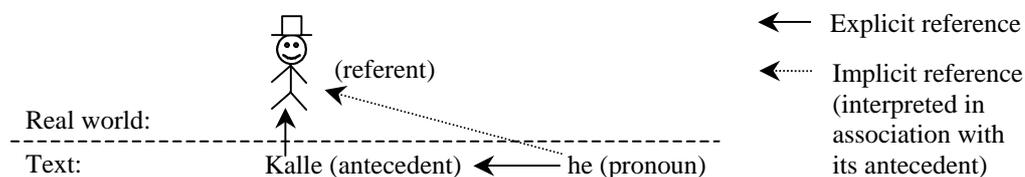
Figure 4. Pronoun->antecedent->referent relationship.

However, the two natural genders can be said to carry some semantic information. As we have already seen, a masculine pronoun is chosen if its antecedent's referent is of the male sex and conversely for the feminine pronoun and the female sex. Thus the pronouns of natural gender can be said to carry information about their antecedents referent, i.e. they can tell us *something* about the real-world referent, but they still cannot tell us *who* the real-world referent is all by themselves.

Usually two dimensions are used when classifying an anaphor: the type of antecedent and the relation between anaphor and antecedent. The type of the antecedent can be either an object (to put it simply, a noun phrase) or it can be a "situation relation" (Fraurud 1992) which is an occurrence or process (to put it simply, a verb phrase). The relation between the anaphoric expression and its antecedent can be either coreferential or non-coreferential. In the case of coreference the anaphoric expression has to agree in gender, number etc with its antecedent. The reason for this is the idea that the anaphoric expression simply is substituted with the antecedent when resolved (Lyons 1963). A special case of the anaphor is the 'situation anaphor' (Fraurud 1992) or 'discourse deixis' (Webber 1991). This type of anaphor refer to a situation or an event rather than an entity, for example:

"Kalle ramlade över bordet. Det gjorde verkligen ont."
"Kalle fell over the table. It really hurt."

In the above example, 'Det' (It) clearly doesn't refer to the table, but rather to the fact that 'Kalle fell over the table' or minimally to 'Kalle fell'. Situation anaphors present a very difficult problem and are beyond the scope of this thesis.

In the following example the anaphoric expression can be interpreted either as coreferential or non-coreferential depending on the scope chosen for the antecedent.

Kalle ramlade i vattnet. Det gjorde honom våt.
(Kalle fell into the water. This made him wet.)

- If 'Det' ('This') refers to 'vattnet' ('the water') the relation is said to be coreferential.
- If 'Det' ('This') refers to 'ramlade i vattnet' ('fell into the water') the relation is said to be non-coreferential situation reference.

A simple approach would be to, when possible, view all anaphoric relations as being coreferential. This would in this case be to disregard the ambiguity and simply assume that 'Det' (This) refers to 'vattnet' ('the water'). A useful term, coined by Sidner (1984), is to cospecify which comprises both coreferential and non-coreferential relations. This term is mainly used in computational linguistics where you more often than not have a data structure as referent.

2.4   Anaphoric Expressions and Automatic Text Summarisation/Extraction

The problem with anaphors when performing automatic text summarisation by extraction can be shown in the following example. Say that we have the following discourse fragment:

"Kalle kysste Lisa. Han har varit dödligt förälskad i Lisa de senaste 12 åren."
"Kalle kissed Lisa. He has been madly in love with Lisa for the past 12 years."

If the automatic text summarisation tool decides to extract only the second sentence of the two it will be hard, if not impossible, to tell whom 'Han' (He) is. In the worst of cases 'Han' (He) can appear to refer to some completely different male person (say if the sentence "Pelle bought a present" was extracted into the summary directly preceding "He has been madly in love with Lisa for the past 12 years."). If the anaphoric relation instead is resolved and 'Han' (He) is substituted with 'Kalle' the extract will be unambiguously intelligible.

2.5   Problems in Resolution of Anaphora

As mentioned above, there is nothing inherent in the words themselves that make them belong to either of the gender categories. For the computer (and probably for a human) this has to be seen as lexical and semantic information.

Pronouns do also not always refer linearly (in relation to each other) when pronouns of different categories are involved. For example:

Kalle älskade Lisa. Han älskade henne mer än något annat på jorden.
Kalle loved Lisa. He   loved   her   above all on Earth.

or:

Lisa   älskades av Kalle. Han älskade henne mer än något annat på jorden.
Lisa was loved by Kalle. He   loved   her   above all on Earth.

This calls for keeping track of *all* possible antecedents for all anaphor types you want to be able to resolve. In doing this we have to take into account the whole discourse so far since anaphors tend to not only refer intra-sentential but also inter-sentential. In rare cases, mainly with human referents, pronouns refer even over paragraphs (Fraurud 1992).

There is also the problem of identifying anaphors. Not every word that lexically looks like a pronoun is a pronoun. The lexical tokens 'den' and 'det' can for example be, depending on syntactic context, a pronoun ("Den var röd" – "It was red"), a determiner ("Den röda bilen" – "The red car") and it can also be a placeholder for the mandatory subject ("Det regnade" – "It rained").

2.6   Past Approaches

Quite a few approaches to the comprehension of anaphora have been suggested, for example; syntactic constraints on intra-sentential anaphora (Reinhart 1983), discourse structure (Grosz 1977), focusing (Sidner 1979, 1983) and Centering (Gordon, Grosz & Gilliom 1993, Grosz, Joshi & Weinstein 1995, Strube & Hahn 1996). Many of these approaches to anaphora resolution rely heavily on linguistic and domain knowledge. These approaches often require a strongly restricted discourse domain or rely on highly elaborate and yet unimplemented procedures in an effort to cover all cases of pronoun reference (Fraurud 1992).

Various alternatives have been proposed that make use of, for example, neural networks (Connoly, Burger & Day 1994), situation semantics frameworks (Tin & Akman 1995) or the principles of reasoning with uncertainty (Mitkov 1995). The problem with many of these approaches are that they are very process-heavy and not very practical as components for example a real-time summariser for the web.

There is a strong need for the development of robust and effective strategies to meet the demands of practical Natural Language Processing systems, and to enhance further the automatic processing of growing language resources (Mitkov 1998). Such computational

inexpensive strategies are also essential for Information Retrieval (for example a search engine presenting abstracts of the documents matching the query), Event Tracking systems (for example an agent that scans news media and e-mails an abstract of a news article given certain criteria) and many more real-time systems. They are also needed for less restrictive dialogue systems.

Two computational inexpensive, knowledge-poor algorithms have been presented by Ruslan Mitkov (1998) and Kari Fraurud (1992). We will describe and comment on these below.

Algorithm from Mitkov 1998:

```
1. Examine the current sentence and the two preceding sentences (if available).
   Look for noun phrases only to the left of the anaphor.
2. Select from the noun phrases identified only those which agree in gender and
   number with the pronominal anaphor and group them as a set of potential
   candidates.
3. Apply the antecedent indicators to each potential candidate and assign scores;
   the candidate with the highest aggregate score is proposed as antecedent. If
   two candidates have an equal score, the candidate with the higher score for
   immediate reference is proposed as antecedent. If immediate reference does not
   hold, propose the candidate with higher score for collocational pattern. If
   collocational pattern suggests a tie or does not hold, select the candidate
   with higher score for indicating verbs. If this indicator does not hold again,
   go for the most recent candidate.
```

This is a very attractive algorithm apart from that it only takes into account the current and (at most) two preceding sentences for each pronoun. Even if it is the usual case that the antecedent lies in the current or the nearest preceding sentences (for 90% of the pronouns in a empirical study done by Kari Fraurud (1992)) this is hardly always the case. To us the scoring system also seems unnecessary complex. A finite set of templates seems more linguistically sound.

Algorithm from Fraurud 1992:

```
1. Consider as a candidate antecedent every NP in the text that:
   1.1. precedes the pronoun in the text.
   1.2. agrees with the pronoun in number, animacy and gender.
   1.3. is not (co-referential with) the subject of the clause in which the
        pronoun occurs.
   1.4. is not in a clause where the pronoun is a subject.

2. Select as antecedent the NP in the set of candidate antecedents that:
   2.1. is the most recent candidate in the text, i.e. the NP whose head is
        closer to the pronoun than the head(s) of the other NP(s),
        unless:
   2.2. there is another candidate, which is the subject of the same clause as
        the most recent candidate, and the pronoun is not a semi-demonstrative;
        in that case select the subject NP.
```

Fraurud's approach is as we can see cleaner and more straightforward. This makes easier to understand and expand. It has also, contrary to Mitkovs algorithm, been successfully implemented for Swedish text, which makes it more interesting in our case.

3 The Pronominal Resolution Module

Our solution is, as we will se, closely related to Fraurud's algorithm. Contrary to Mitkov both Fraurud and we use the whole preceding text as discourse for an anaphor. Both Fraurud's and our algorithm also prefer the most recently named entity although we believe our approach to be more open. Our algorithm can easily be extended with more semantic knowledge and it includes an extendable system for taking care of special cases. This makes it more flexible but it also makes the framework a bit more complex.

The *Pronominal Resolution Module*, hereafter PRM, is written in Perl and works as a pre-processor to SweSum. The module can however be used in conjunction with basically any program that can run/use Perl scripts. Its usage with SweSum can graphically be represented as:
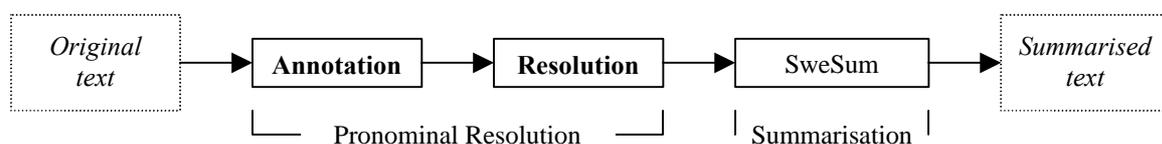

Figure 5. PRM's usage with SweSum.

PRM uses a lexicon of nouns that contains information about each entry's natural (han/hon - he/she) or grammatical (den/det - it/it) gender and has been constructed by remodelling extracts from the Stockholm Umeå Corpus[1]. This lexicon has also been replenished with 1000 names from Svenska Språknämnden, the Swedish Language Committee, (the 500/500 most common male/female first names 1979, web link now removed) and some additional names from Allén & Wåhlin (1995) as well as names encountered in texts used for testing. The current noun lexicon contains over 1500 gender specified first names.

3.1 Focus Applicants

PRM uses lists of likely focuses (Sidner 1984), here called *focus applicants*. By focus we hereby mean the person/item or group of persons/items that are most prominent at a specific point in the discourse. Every time a nominal phrase is recognised it is categorised and pushed onto the appropriate list for that category. Natural categories for these lists are, as mentioned above, the natural and grammatical genders as well as a category for enumerations and pluralised noun phrases (basically noun phrases referred to with 'de'/'dem'/'deras' – they/them/theirs).

These focus applicant lists are by order of insertion sorted in order of likelihood. This builds upon the supposition that the noun phrase most recently encountered is the most prominent noun phrase in that exact instant. What here have to do is to differentiate syntactic and semantic focus. These two focuses often, but do not have to, coincide. The syntactic focus is most often the last noun phrase of a sentence. For example, take the two sentences:

---

[1] The Stockholm Umeå Corpus is copyright (c) 1997 Dept of Linguistics, Umeå University, and Dept of Linguistics, Stockholm University and may be freely used in non-commercial research. The raw text of the corpus is annotated in the SGML-format for parts of speech (CD-ROM suc 1.0) and contains about 1.000.000 words and almost 300.000 word types.

1.1　　”Olle slog Nisse.”
　　　　”Olle hit Nisse.”

1.2　　”Han tyckte det gjorde ont.”
　　　　”He thought it hurt.”

The syntactic focus in sentence 1 is in this case 'Nisse' and most people would agree that the semantic focus in this case coincides and that 'Han' (He) refers to 'Nisse'. This can be changed by adding some semantic information to the discourse, for example:

2.1　　”Olle slog Nisse.”
　　　　”Olle hit Nisse.”

2.2　　”Han tyckte det gjorde ont *i handen*.”
　　　　”He thought it hurt *his hand*.”

In this case the semantic focus seems to shift to 'Olle' since it's common knowledge that, if nothing else is said, you usually hit someone with your hand. But what if we add still some more semantic information:

3.1　　”Olle slog Nisse *på handen*.”
　　　　”Olle hit Nisse *on his hand*.”

3.2　　”Han tyckte det gjorde ont i handen.”
　　　　”He thought it hurt his hand.”

Again the semantic focus seems to shift and we are back on 'Nisse' as both syntactic and semantic focus.

The example given in 2.1-2.2 is, however, not the common case. People tend to want to make themselves understood and thus usually clearly state their meaning. In these cases the syntactic and the semantic focus tend to coincide. Also, many semantic shifts are caused by syntactic constructions, for example active and passive phrase, and these can easily be handled with templates. Of course there are situations where a human can't resolve a anaphoric expression because of lack of information or a high degree of ambiguity. In these cases PRM tend to fall back on the syntactic focus. If there is no resolvable antecedent this poses no problem since the anaphoric expression will simply be left unresolved. If, on the other hand, there are several possible (syntactic and semantic) focuses it will choose the syntactic focus of the appropriate category (gender/number). This of course is not good but the problem is to recognise the ambiguity.

When, as we will see below, a suitable applicant is found in the text being processed, it is placed first in a suitable list. These lists can be seen as stacks of focus applicants and applicants are pushed upon appropriate stack when found. Focus applicant lists used by PRM so far are only two; one for each natural gender, but more lists/stacks can easily be added when needed.

PRM presupposes anaphora rather than cataphora. This also presupposes that the semantic and the syntactic focuses coincide but has an extendable template system to care for rare cases.

3.2   Templates

PRM uses templates for performing simple surface parsing. To exemplify this we can take the case of active and passive phrases containing pronouns with antecedents of the same gender.

1.   ”Kalle var arg på Pelle.”
       ”Kalle was angry with Pelle.”

2a.   ”Han slog honom.”
       ”He hit him.”

2b.   ”Han blev slagen av honom.”
       ”He was hit by him.”

As we can see the two phrases 2a and 2b carry fairly the same semantic information despite their difference in grammatical structure. The first one is said to be in active phrase and the latter in passive. These two phrases have quite distinct resolutions. 2a would yield "Pelle slog Kalle" – "Pelle hit Kalle" and 2b would yield "Kalle blev slagen av Pelle" - "Kalle was hit by Pelle". One way to accomplish this is to use two distinct templates, namely:

1.   [han][PHRASE][honom]
2.   [han][AUX VERB][PHRASE][av][honom]

Now all we have to do is to substitute [han] with the top most noun in the *male focus applicant list* in case 1 and the second top most name in case 2. The order is reversed with [honom].

3.3   Annotation in AHTML

Pronouns and abbreviations are marked with custom SGML (Standard Generalized Markup Language) tags. This is done during their respective annotation phase (phase 1 for abbreviations and phase 5.1 for anaphors, in the algorithm above). These tags serve as an extension to standard HTML (Hypertext Markup Language) and we have chosen to call this superset AHTML (after Anaphors and Abbreviations). Anaphors (pronouns) are marked with the tag pair *<!ANAPHOR REF="Referent" LINE="Line number">* and *</!ANAPHOR>*; and abbreviations are marked with the tag pair *<!ABBR>* and *</!ABBR>*.
For example, the annotation

*<!ANAPHOR REF="Robert" LINE="16">*han*</!ANAPHOR>*

represents that the pronoun "han" (he) has the antecedent "Robert" found in the sixteenth sentence (and hence line since the text is segmented into one sentence per line) in the discourse. The exclamation marks are used to distinguish the AHTML extension from standard HTML.

11

3.4 Algorithm Used

The pronominalisation algorithm consists of three distinct phases that act on three different levels; discourse, sentence and word level. At the discourse level we segment the text into sentences. This will be the discourse unit used by the algorithm. At the sentence level we identify and handle special cases like active and passive phrase and at the word level we identify and annotate the actual nouns and pronouns.

In order to perform both pronoun resolution and text summarisation we must distinguish a full stop (.) indicating a sentence delimiter from a period (.) used in abbreviations. This can be exemplified with the following discourse fragment:

"Han tog med sig sin sovsäck, tandborste, handduk, etc. till lägret."
"He took his sleeping bag, toothbrush, towel, etc. with him to the camp."

In order to properly segment sentences in even ill formatted text we found it necessary to annotate abbreviations also. These annotations are only internal and are removed in the output.

```
For each discourse (text):
  1)  Identify and annotate abbreviations.
      This step is necessary for sentence segmentation.
  2)  Identify and segment sentences.
For each sentence:
  3)  Use templates to identify special cases (i.e. active/passive phrase,
      cataphors, etc.).
      3.1)  If one or more pronouns/anaphoric expressions are found, annotate them
            using appropriate focus applicant lists.
  4)  Identify and segment words.
For each word in each sentence:
  5)  Search for pronoun/anaphoric expression (i.e. han (he), hon (she), den (it or
      the), det (it or the), företaget (the company), etc.).
      5.1)  If a pronoun/anaphor is found, annotate it in AHTML with the most
            likely antecedent and sentence number found in the corresponding focus
            applicant list (based on gender and/or any other supplied semantic
            information).
  6)  Search and compare with lexicon to see if it is a known noun.
      6.1)  If a noun is found, place it first in appropriate focus applicant list
            (according to category) along with information on in which sentence it
            was found.
  7)  Repeat 3) through 6) until end-of-discourse.
```

3.5 Example of a Pronominal Resolution

Choice of applicant for an anaphor is, as mentioned above, based upon salience, represented by the position in the list, and semantic likelihood (category), which is determined by using the semantic information (gender/number) in the lexicon. As an example we can take the following sentence segmented discourse fragment:

1. "I torsdags avgick Bo Persman som VD för Cyclop Enterprises."
    "Last Thursday Bo Persman resigned as Vice President for Cyclop Enterprises."

2. "Han anklagas för förskingring."
    "He is accused of embezzlement."

In accordance with the algorithm given above we first search sentence 1 for an anaphoric expression. Not having found one, we proceed with searching for known nouns. In this case we will find 'Bo' as in "Bo Persman" and it will be recognised as a male first name and thus will be placed in the *male focus applicant list* together with the number of the sentence it was found in, i.e. 1. The object pushed onto the list will thus look like this: [Bo, 1]. Now we continue with the next sentence, sentence two. This sentence is now, as sentence 1 was, first searched for an anaphoric expression. Here 'Han' (He) will be found and resolution will take part. The top most object from the male focus applicant list will be read and the anaphoric expression will be tagged using this information. The sentence will after tagging look like this:

"*<!ANAPHOR REF="Bo" LINE="1">*Han*</!ANAPHOR>* anklagas för förskingring."

This method enables us to resolve anaphors non-linearly since semantic information is used to differentiate antecedents. Of course this is not only applicable on proper names. Semantic information about nouns as 'man' (man), 'kvinna' (woman), 'flock' (herd) etc. can of course also be added.

4   Evaluation

In order to evaluate the PRM, SweSum was used in a user centred field test with and without the PRM. This was done within the framework of a 4-credit Human Language Technology course (2D1418 Språkteknologi) at NADA/KTH, Stockholm. Nine students were given the task to automatically summarise 10 texts of news articles and movie reviews with and without pronominal resolution. The purpose was to investigate how much a text can be summarised without losing coherence or important information and to validate if the summarisation could be improved with pronominal resolution.

The students carried out the test by first reading the text to be summarised and then gradually lowering the size of the summary giving SweSum the amount of the original text they would prefer in the summary. In a questionnaire they would note at what percentage they, subjectively, perceived that the coherence was broken and when important information in the original text was missing from the summary. This procedure was repeated with and without pronominal resolution for each of the 10 texts.

As can be seen in Appendix B, not all students completed the whole questionnaire leaving the field test inconclusive. Despite this one can conclude that most of the time the students have come to fairly the same conclusions. There are naturally exceptions, which only serve to exemplify the subjective nature of the test.

There are no corpora with manual extracts available for Swedish as for English. Therefore it is difficult to make a comparative evaluation of automatic summarisation in Swedish. We are planning on creating such corpora using the technique proposed by Marcu (1999). Since we had very few participants in our field test we decided to use median as a statistical measurement of our results. This because the median can somewhat compensate for diverging values. We first calculated the median of summarised text (given in percent) for each text (Table B5, Appendix B) and then we calculated the total median for all ten texts (Table 2 and Table B5), and for three movie reviews that were particularly pronoun dense (Table 3 and Table B6). From the field test we can see that the Swedish text summariser SweSum is as

good as the English state-of-the-art text summarisers. According to C-Y Lin (1999) about 30% summarisation gives an ideal summarised text for English.

As shown in Table 2 below (see also Table B5, Appendix B) our tests with SweSum show that a summary containing a median of 30% (based on all 10 texts) still contains all important information of the original text. The PRM does not seem to have any effect on this factor. The other factor measured was coherence. *Without* pronominal resolution, a summary containing a median of 24% (based on all 10 texts) of the sentences from the original text is still perceived as coherent. *With* pronominal resolution the summary is still perceived as coherent when only containing a median of 21% of the sentences from the original text.

Table 2. Results from the field test; all texts: no. 1 – 10.

	Important info. preserved (without pronominal resolution)	Important info. preserved (with pronominal resolution)	Coherence preserved (without pronominal resolution)	Coherence preserved (with pronominal resolution)
Total median:	30%	30%	24%	21%
Total average:	31%	34%	26%	29%

In the table below only texts 2, 3 and 10 have been used. These three texts have frequent occurrences of pronouns. We can here see that for texts containing a lot of pronouns the coherence of the summary is greatly improved with PRM from 25% without pronominal resolution to 15% with pronominal resolution.

Table 3. Results from the field test; texts no. 2, 3 and 10 containing many pronouns.

	Important info. preserved (without pronominal resolution)	Important info. preserved (with pronominal resolution)	Coherence preserved (without pronominal resolution)	Coherence preserved (with pronominal resolution)
Pron. Median:	30%	30%	25%	15%
Pron. average:	28%	31%	22%	18%

The PRM does not at this point seem to affect the threshold where important information no longer is being extracted to the summary. This may come to change as we add full support for grammatical gender, plurals, enumerations and set expressions (see Part 5, Conclusions and future improvements).

5 Future Improvements

Further improvements include partial resolution, that is, resolution solely of broken anaphors, the resolution of more types of anaphoric expressions and better recognition of names (i.e. by also using surnames).

5.1 Partial Resolution

In the summarised review of *Analyze This* (above), we can observe both anaphors with their antecedents within the summary and anaphors with their antecedents in the parts of text not extracted by SweSum. Clearly the resolution of all anaphoric references leads to redundant information and repetitive text and should therefore be avoided. As mentioned above, PRM

tags the anaphoric expressions with custom SGML tags that point out the antecedent and in which sentence it was found. An obvious use of this information would be to resolve only those anaphoric references that have their antecedents in the set of not extracted sentences.
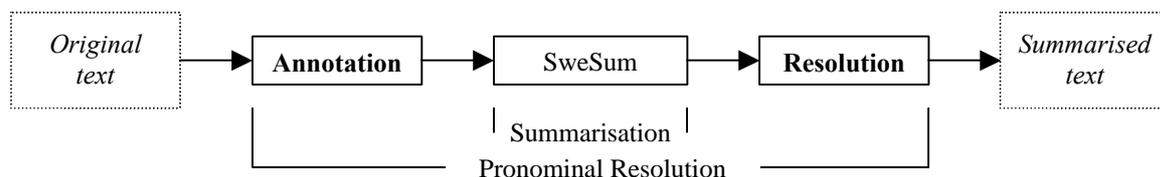
This can graphically be represented as:

Figure 6. PRM's future usage with SweSum.

Another tactic, which can be used by SweSum to further improve the coherence of the summary, would be to look at how far away the antecedent of a pronoun lies. If the antecedent is found in the nearest preceding sentence, this sentence and the sentence in which the pronoun was found can be treated as one unit instead of resolving the reference. This could possibly further the coherence of the summary, but tests still remain.

5.2 Plurals, Enumerations and Set Expressions

In order to handle the pronouns de/dem (they/them), PRM will have to be able to recognise plural forms; fåglarna (the birds), enumerations; Kalle, Pelle och Lisa (Kalle, Pelle and Lisa) and set expressions; fiskmåsflock (flock of seagulls). When recognised these will be treated as focus applicants and placed in the focus applicant list for plural pronouns.

5.3 Improved Recognition of Personal Names

One of PRM's weaker sides is its failure to recognise surnames. This can lead to some embarrassing misunderstandings as we well can see in the following discourse fragment:

"Stanley Kubricks sista film har Tom Cruise och Nicole Kidman i huvudrollerna. Tom Cruise spelar här Bill, läkare till yrket … Kubrick har som vanligt gjort en fantastisk film även om **han** tappar fart mot slutet." (*Swedish*)

"Stanley Kubrick's last film starrs Tom Cruise and Nicole Kidman in the leading parts. Tom Cruise here plays Bill, doctor by profession. … Kubrick has as usual made a fantastic film even though **he** loses speed towards the end." *(English translation of the text example above)*

With the current logic PRM will resolve the highlighted **han** (**he**) to Bill, the most recently (by first name) mentioned male person. Clearly it should resolve to Stanley (or Kubrick, or Stanley Kubrick). One way to amend this is to, when encountering a capitalised token, search through each of the two focus applicant lists for natural gender. Upon finding a match for a first name/surname combination this combination will be brought up front in the list.

5.4 Anaphoric and Non-Anaphoric Use of 'den'/'det'

When adding resolution of 'den'/'det' (it/it) we will need to add more templates in order to identify contexts in which 'den' and 'det' are used in the pronominal sense. 'Den' and 'det' also have non-anaphoric uses in Swedish. They can for example be used as determiners:

"Det röda äpplet och den gula bananen."
"The red apple and the yellow banana."

'Det' can also be used as a sort of placeholder for the in Swedish mandatory subject:

"Jag tittade ut genom fönstret och såg att det regnade"
"I looked out my window and saw that it rained."

One approach to handle expressions like these is to use flexible templates in the form of heuristic context-sensitive rules. This approach has rather successfully been applied on English scientific texts (Paice 1990). Similar rules should be devisable for Swedish. This may also be a satisfactory way of dealing with cataphors.

Another approach, in differentiating pronouns from determiners, is to use a part-of-speech tagger for Swedish unrestricted text developed at KTH (Carlberger & Kahn 1999).

5.5 More Considerations

Kari Fraurud (1992) examined randomly chosen passages of text from a corpus of stories, reports and articles. The sample consisted of 200 pronouns from each text type, a total of 600 pronouns. She found that 90% of the pronouns in the sample had their antecedent in the same or the immediately preceding sentences. In the total sample, the maximal linear distance between a pronoun and its antecedent was 15 sentences. This indicates that some restrictions on the scope should be introduced. One simple restriction would be to not let the scope cross headings in the text since these often are good indicators of topic change.

Worth pondering is also the use, albeit rare, of pronouns of natural gender when referring to non-human, and sometimes even non-sentient entities (boat – she, moon – he etc).

6   Conclusions

As expected, we found that the pronominal resolution module PRM improved the coherence for texts rich with pronouns. Our hope is that further improvements will push this limit a bit more and result in more readable and enjoyable summaries. PRM has proven to be stable and reliable in delivering results. It has also proven to be fast enough for real-time applications. Execution time can probably be shortened if the code is rewritten in C++. PRM is also so general that it should, with little changes to the module itself, easily port to other natural languages. A necessity in this case is to construct new or partially new lexicons for each target language (personal names, names of companies etc. can mostly be kept).

The current implementation of PRM is limited to resolving pronouns of the natural gender category. In order to make practical use of it in real-world applications it badly needs the future improvements outlined above. The development of PRM is now continued in the SeaSum project, a NADA-KTH, Euroseek joint project supported by Nutek. One of the aims of this project is to develop a Event Tracking system delivering summarised news articles by e-mail given certain search ('scout') criteria.

Acknowledgements

We would like to thank Johan Carlberger NADA, KTH, for his assistance with the acquisition of a Swedish morphological lexicon for SweSum. We would also like to thank the students of course 2D1418 for their willingness to participate in our field studies.

7 References

S. Allén & S. Wåhlin. 1995. *Förnamnsboken, de 10000 vanligaste förnamnen* (The Book of First Names, the 10000 most common first names; in Swedish). Stockholm Norstedt.

J. Carlberger and V. Kann. 1999. *Implementing an efficent part-of-speech tagger*. Software – Practice and Experinece, 29(9):815-832.

D. Connoly, J.D. Burger and D.S. Day. 1994. *A Machine learning approach to anaphoric reference*. Proceedings of the International Conference "New Methods in Language Processing", 255-261, Manchester, United Kingdom.

D. Crystal. 1995. *The Cambridge Encyclopedia of Language*. Cambridge University Press.

H. Dalianis. 1999. *SweSum - A Swedish Textsummarizer*. http://www.nada.kth.se/~hercules/Textsumsummary.html

H.P. Edmundson. 1969. *New Methods in Automatic Extraction*, Journal of the ACM 16(2) pp 264-285.

K. Fraurud. 1992. *Processing Noun Phrases in Natural Discourse.* Doctoral dissertation, Department of Linguistics, Stockholm University.

P.C. Gordon, B.J. Grosz, and L. Gilliom. 1993. *Pronouns, names, and the centering of attention in discourse*. Cognitive Science, 17(3), 311-348.

B.J. Grosz. 1977. The representation and use of focus in a system for understanding dialogs. In Proceedings of the International Joint Conference on Artificial Intelligence, 67-76.

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

S. Gustafson-Capková. 1998. *Diskursstruktur – hur kan sammanhang representeras?* (Discourse Structure – How can Cohesion in Text be Represented?; in Swedish), Handout - Datorlingvistiska metoder II, Dept of Linguistics, Stockholm University.

M.A.K. Halliday and R. Hasan. 1976. *Cohesion in English*, Longman, Harlow.

G. Källgren and E. Ejerhed. 1997. *CD-ROM Stockholm Umeå Corpus 1.0.* Dept of Linguistics, Umeå University, and Dept of Linguistics, Stockholm University.

C-Y Lin and E. Hovy. 1997. *Identify Topics by Position*, Proceedings of the 5th Conference on Applied Natural Language Processing, March.

C-Y Lin. 1999. *Training a Selection Function for Extraction,* Proceedings of the 8th International Conference on Information and Knowledge Management, 1999.

H.P. Luhn. 1959. *The Automatic Creation of Literature Abstracts*. IBM Journal of Research and Development pp 159-165.

J. Lyons. 1963. *Structural Semantics*. Blackwell Publishers, Oxford.

D. Marcu. 1999. *The automatic construction of large-scale corpora for summarization research*, In the Proceedings of International Conference on Research and Development in Information Retrieval SIGIR-99, pp.137-144.

R. Mitkov. 1995. *An uncertainty reasoning approach for anaphora resolution*. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95), 149-154, Seoul, Korea.

R. Mitkov. 1998. *Robust pronoun resolution with limited knowledge*. Proceedings of the 18.th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference, 869-875. Montreal, Canada.

C.D. Paice. 1990. *Constructing Literature Abstract by Computer: Techniques and Prospects*, Information & Management vol. 26, No. 1, pp 171-186.

T. Reinhart. 1983. *Anaphora and Semantic Interpretation*. The University of Chicago Press, Chicago, IL 60637

G. Salton, 1989. *Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer*. Addison Wesley Publishing Company.

C.L. Sidner. 1979. *Toward a computational theory of definite anaphora comprehension in English*. Technical Report AI-TR-537, MIT. International Joint Conference on Artificial Intelligence, pages 203-208, Vancouver, BC, Canada.

C.L. Sidner. 1984. *Focusing in the Comprehension of Definite Anaphora*. Berwich & Brady eds., Computational Models of Discourse, MIT Press, London.

M. Strube and U. Hahn. 1996. *Functional centering*. In ACL'96 - Proc. 34th Annual Meeting of the Association for Computational Linguistics, pages 270--277, 1996.

E. Tin & V. Akman. 1994. *Situated processing of pronominal anaphora.* Proceedings of the KONVENS'94 Conference, 369-378.

B.L. Webber. 1991. *Structure and Ostension in the Interpretation of Discourse Deixis*. Language and Cognitive Processes, 6 (2), p.107-135.

Appendices

Appendix A: Swedish Texts for the "Analyze This" Example

Appendix A1. 30% Summary without Pronominal Resolution

Analysera mera!
Regi: Harold Ramis
Medv: Robert De Niro, Billy Crystal, Lisa Kudrow
Längd: 1 tim, 45 min
…
Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. Han accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och han är tämligen oemotståndlig. Här har han åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.
SvD 99-10-08

Appendix A2. 30% Summary with Pronominal Resolution

Analysera mera!
Regi: Harold Ramis
Medv: Robert De Niro, Billy Crystal, Lisa Kudrow
Längd: 1 tim, 45 min
…
Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. **Robert** accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och **Robert** är tämligen oemotståndlig. Här har **Harold** åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.
SvD 99-10-08

Appendix A3. Complete Review of the Motion Picture "Analyze This".

Analysera mera

Analysera mera
(Analyze This)
Regi: Harold Ramis
Medv: Robert De Niro, Billy Crystal, Lisa Kudrow m fl
Längd: 1 tim, 45 min
Tillåten från 11 år

I många år nu har Robert De Niro ådet skandalösaste misshushållat med sin begåvning. I filmer som Casino och Heat har han visat upp ett slags tjurigt, utstuderat minimalistiskt antiagerande som kanske kunde vara intressant som övning i någon workshop någonstans, men som i filmisk praxis blir frustrerande: Ett hårt fruset ansiktsuttryck krusat av några enstaka ticks har, trots allt, en begränsad uttryckskraft även om det tillhör en av vår tids största stjärnor.

Också i Wag the Dog och Jackie Brown går han påsparlåga, medan han i andra sammanhang, som Lysande utsikter och Mary Shelleys Frankenstein, nöjer sig med hårt stiliserade, kuriösa krumelurer.

Ett av många skäl att glädjas å Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. Han är närvarande, avspänd och samtidigt skärpt, lekfull och samtidigt effektiv. Han accelererar emotionellt från 0 till 100 påingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Han är brutal och vulgär och stundtals upplöst av ångest och skuldkänslor. Och han är tämligen oemotståndlig.

Dessutom är han modig: han skojar filmen igenom med sitt eget varumärke, den kallhamrade mafioson. Hans rollfigur Paul Vitti är ännu en gangsterboss som utkämpar ett tvåfrontskrig mot dels den federala polisen, dels makthungriga rivaler. I filmens inledningsscen krossas en hel restauranginredning av kulkärvar adresserade till honom. Men inte heller en gangster är, hävdas det i den här filmen, mer än en människa. Paul Vitti blir rädd. Det hårda livet ger honom panikångest. Han har inte längre samma lust att misshandla motståndarsidans hejdukar med järnrör eller att lyckliggöra sin älskarinna. Så han skaffar sig, på sitt eget resoluta sätt (hot och kidnappning), en psykoterapeut.

Det är här som Billy Crystal kommer in i bilden. Även han har befunnit sig i en formsvacka ett tag, varit påagligt alldaglig och blek i filmer som Glöm Paris och Tvåpappor för mycket. Men i par med en De Niro på hög volym är han utsökt följsam. Dialogen dem emellan är en virtuos uppvisning i verbal dans. De Niro för och Crystal följer, men den senare får mot slutet chansen att briljera i ett eget solonummer, hårt granskad av en hoper hårda grabbar, som möjligen är det roligaste av allt.

Grundreceptet för Analysera mera är inte särskilt komplicerat: blanda maffiaklichéer med terapiklichéer och rör om. Filmen består i hög grad av upprepningar med minimala förskjutningar: De Niro hämtar Crystal under bisarra former och hotar sig till konsultation. Men om det är någon som kan göra stor konst av upprepningar med minimala förskjutningar, så är det regissören Harold Ramis, vars Måndag hela veckan (Groundhog Day) från 1993 i dag åtnjuter sin välförtjänta klassikerstatus.

Här har han åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination. Och vi är sannerligen inte bortskämda.

Jan Söderqvist

Appendix B: Results from the Field Test

In the tables below, **P1…P9** on the horizontal scale stands for person **1…10** in the test group. The numbers on the vertical scale each represent a text in the field test. The amounts given in percentage represent how much of the original text was extracted to the summary.

Table B1.
Without pronominal resolution important information is missing at:

Text\Person	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	37%	25%	40%	15%	35%	20%	30%	35%	49%
2	40%	25%	55%	45%	45%	20%	40%	25%	53%
3	30%	30%	45%	25%	50%	25%	20%	30%	55%
4	20%	15%	20%	15%	35%	35%	50%	25%	29%
5	60%	20%	10%	15%	20%	15%	20%	45%	15%
6	37%	30%	99%	35%	40%	40%	30%	25%	
7	44%	25%		30%	30%	25%	50%	50%	41%
8		30%	30%	30%	25%	35%	40%		
9		20%	15%	20%	25%	15%	30%		19%
10		15%	10%	10%	35%	10%	20%		

Table B2.
With pronominal resolution important information is missing at:

Text\Person	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	30%	30%	25%	15%	30%	20%	30%	35%	37%
2	40%	25%	15%	30%	25%	10%	35%	30%	79%
3	60%	30%	35%	50%	70%	25%	25%	65%	67%
4		99%	95%	25%	70%		100%	100%	59%
5	40%	20%	15%	15%	35%	15%	20%	45%	10%
6	29%	25%	99%	35%	40%	30%	20%	25%	
7	49%	25%		30%	30%	25%	40%	50%	31%
8		25%		20%	30%	25%	40%		
9		20%	5%	5%	20%	15%	40%		15%
10		15%	15%	5%	20%	10%	10%		

Table B3.
Without pronominal resolution coherence is broken at:

Text\Person	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	25%	20%	35%	30%	35%	25%	20%	20%	
2	20%		30%	55%	30%	60%	20%	15%	53%
3	27%	25%	25%	25%	70%	25%	15%	25%	55%
4	15%			15%	15%		10%	5%	
5	30%		20%	10%	20%	20%	15%	25%	
6	30%	10%	30%	35%	35%	10%	15%	15%	
7	20%	10%		25%	30%	40%	30%	35%	41%
8		20%	60%	25%	35%	35%	20%		
9		5%		20%	40%	10%	10%		44%
10		5%		10%	35%	10%	10%		

Table B4.
With pronominal resolution coherence is broken at:

Text\Person	P1	P2	P3	P4	P5	P6	P7	P8	P9
1	25%	20%	25%	25%	30%	15%	20%	20%	
2	20%		15%	10%	15%	25%	10%	15%	79%
3	32%	25%	35%	50%	75%	30%	20%	65%	100%
4		40%		25%	60%			100%	100%
5	20%		25%	10%	55%	20%	15%	25%	10%
6	14%	10%	45%	35%	35%	25%	10%	20%	
7	30%	10%		25%	35%	30%	20%	35%	31%
8		20%		20%	35%	45%	20%		
9		5%		8%	15%	5%	20%		60%
10		5%		2%	25%	5%	5%		

Table B5 below shows total median for each text in the test as well as total median and total average for all texts combined.

Table B5.
Results:

Text\Median	Important info. preserved (without pronominal resolution)	Important info. preserved (with pronominal resolution)	Coherence preserved (without pronominal resolution)	Coherence preserved (with pronominal resolution)
1	35%	30%	25%	23%
2	40%	30%	30%	15%
3	30%	50%	25%	35%
4	25%	95%	15%	60%
5	20%	20%	20%	20%
6	36%	30%	23%	23%
7	36%	31%	30%	30%
8	30%	25%	30%	20%
9	20%	15%	15%	12%
10	13%	13%	10%	5%
Total median:	30%	30%	24%	21%
Total average:	31%	34%	26%	29%

In Table B6 below only texts 2, 3 and 10 have been used. These three texts have frequent occurrences of pronouns.

Table B6.

	Important info. preserved (without pronominal resolution)	Important info. preserved (with pronominal resolution)	Coherence preserved (without pronominal resolution)	Coherence preserved (with pronominal resolution)
Pron. Median:	30%	30%	25%	15%
Pron. average:	28%	31%	22%	18%