UNIVERSITY OF TARTU

FACULTY OF MATHEMATICS AND COMPUTER SCIENCE

Institute of Computer Science

Speciality of Information Technology

Pilleriin Mutso

Knowledge-poor Anaphora Resolution System for Estonian

Master’s Thesis (20 cp.)

Supervisor: Kaili Müürisep

Senior researcher of Language Technology

Author: …………………………………… “…..“ may 2008

Supervisor: ………………………………. “…..“ may 2008

Professor: ……………….……….……… “…..“ ……... 2008

TARTU 2008

Contents

Introduction

The aim of the current thesis is to develop a pronominal anaphora resolution tool for the Estonian language. The problem of anaphora resolution is well known in the field of natural language processing, it is faced for example in information extraction, translation, summarization and question answering. Many solutions to this topic have been implemented and proposed throughout the decades; however a remarkable number of questions and issues still remain open and therefore anaphora resolution has remained a very actual topic of discussion in the field of computer linguistics. The actuality of the problem shows by the fact that every couple of years Discourse Anaphora and Anaphor Resolution Colloquium (DAARC) is held. The most recent conference, 6th, was held in Portugal (DAARC 07) and a number of automatic anaphora resolution systems for languages like Czech (Linh, Žabokrtský 07), Turkish (Küçük 05) and Norwegian (Holen 06) were also presented among other topics.

However, the majority of the algorithms that resolve anaphora were initially designed for English texts only. Therefore it may often be rather difficult to make these algorithms function as effectively on texts in other languages due to grammatical differences. The current thesis introduces a program that attempts to solve third person personal pronouns in Estonian newspaper and scientific texts. For Estonian no anaphora resolution systems had been implemented up to this point and in general this field of natural language processing has rarely been researched in Estonian, especially when the field of computer linguistics is considered. From purely filological point of view Renate Pajusalu has done research on deicsis in Estonian (Pajusalu 99) and on the usage of the third person personals and demonstratives in Estonian (Pajusalu 97). She has also discussed the anaphoric pronouns in spoken Estonian (Laury, Pajusalu 05). Still, a step to the side where artificial intelligence is integrated had not been made.

As Ruslan Mitkov's knowledge-poor implementation (Miktov 98) has been successfully adapted to a number of languages (like English, Polish and Arabic), a thought was to use it for the first attempts on resolving anaphora automatically in Estonian. The following thesis gives an overview of how it was done and which were the achieved results.

This thesis consists of eight chapters:

- Chapter 1 is introductive – it gives an overview of the aim of the thesis and describes the problem and the definitions in general. An overview of the history of knowledge-poor anaphora resolution is presented.
- Chapter 2 gives an overview of the pronouns in the Estonian language.
- Chapter 3 introduces the knowledge-poor algorithm implemented by Ruslan Mitkov (Mitkov 98) as this is the basis of the pronominal anaphora resolution system for Estonian.
- Chapter 4 concentrates on the details of the implementation – how the tool for Estonian texts has been implemented, which input and output has been used and gives a detailed description of the algorithm and the corpora used.
- Chapter 5 covers the testing process and the obtained results.
- Chapter 6 discusses the results and proposes the solutions to unanswered questions that could be implemented in the future.
- Chapter 7 summarizes what has been done in the current paper.

The source code of the implemented program and the used test and training corpora files are accessible from a CD that comes together with the thesis.

1. Anaphora Resolution Overview

This chapter gives a general introduction to the anaphora resolution. At first the definition and types of anaphor are presented. Then algorithms and approaches to anaphora resolution are discussed. The basics of anaphora resolution process are given and the most frequent issues and problems occurring in resolution are discussed. Finally an overview of the history of knowledge-poor anaphora resolution is presented.

1.1 Definition and categories

According to the Britannica Online Encyclopedia (Britannica) the term *“anaphora”* means *“a carrying up or back”* in Greek. In written or oral text it can be thought of as *“pointing back or referring to something or someone mentioned earlier”*. The entity to which it refers is called its antecedent. The process of determining the antecedent of an anaphor is called anaphora resolution (Mitkov 99). Two simple examples ((1) and (2)) of anaphora are given below:

(1) **Poiss** oli väsinud. **Ta** läks magama.
The **boy** was tired. **He** went to sleep.

In that sentence *“ta“ (“he”)* is an anaphor referring to the antecedent *“poiss“ (“boy”).*

For resolving some cases of anaphors, semantic information is needed. In example (3) the words *“medicine”* and *“ill”* must be associated with each other in order to resolve the anaphor correctly. It cannot be done without world knowledge.

(2) Tom andis **Bobile** rohtu, sest **ta** oli haige.
Tom gave **Bob** medicine, because **he** was ill.

Anaphora can be divided into different types according to the categories (Mitkov 99, Hirst 81) below:

- Pronominal – the most widespread form of anaphora: *“he”, “she”, “they”* etc.
- Lexical noun phrase – widely used in newspaper articles, represented by definite

noun phrases, synonyms or proper names: *“the 26 year old singer”, “the Iron Lady”* etc.

- Ordinal – *“first”, “last”, “former”, “latter”* etc.
- One-anaphora – *“I take the red book, you take the blue one.”*
- Adverb – *“My sister went to Tallinn and stayed there for two days”*. The anaphor *“there”* refers to *“Tallinn”*.
- Verb phrase – *“She chose the third door. The decision was wrong”*. Here *“the decision”* refers to the act of choosing the third door in the first sentence.
- Zero-anaphora – the pronoun is left out: *“John was tired and went to bed”*

Also, sometimes anaphora is also classified by its location in sentences. There are two types (Mitkov 99):

- Intersentential anaphor - refers to the antecedent in a different sentence
- Intrasentential anaphor - refers to the antecedent in the same sentence

There is another type of anaphora that refers forwards, not backwards. This type of referral is called cataphora and is defined as follows: *“Cataphora is the coreference of one expression with another expression which follows it. The following expression provides the information necessary for interpretation of the preceding one. This is often understood as an expression “referring” forward to another expression.”* (SIL)

1.2 Common approaches to anaphora resolution

Various approaches have been used for anaphora resolution throughout the decades. According to Mitkov (Mitkov 99) these approaches can be classified to the following two groups:

1) Traditional approach: integrates knowledge sources/indicators that discount unlikely candidates until a minimal set of plausible candidates is obtained and then makes use of center or focus, or other preference.
2) Alternative approach: computes the most likely candidate on the basis of statistical or artificial intelligence techniques/models.

Both of these approaches use a number of methods for evaluating the likeliness and unlikeliness of the candidates. For example syntactic and semantic analysis, constraints and preferences, centering and statistics (using a statistical Bayesian engine to suggest the most probable center on the basis of a new piece of evidence) or patterns may be used. As Mitkov has already claimed (Mitkov 99), many of these methods are based on world (or domain) knowledge, which means that a remarkable amount of resources (like time, computational power, manual input of people) are needed. As a result a new, knowledge-poor approach (also called robust or rule-based approach) has come forth.

In knowledge-poor approach semantic, linguistic or discourse knowledge is used minimally or completely discarded. This kind of approach may use eliminative or preferential techniques for finding the most suitable candidate, but most commonly a combination of those two is used. For example, the current Estonian anaphora resolution system also uses both of these techniques:

- Eliminative technique: operates on a list of possible antecedent candidates and eliminates them one by one if they do not match certain criteria (for example, gender, number, case, semantic consistency etc). In the end only one candidate remains and is proposed as the correct solution.
- Preferential technique: processes a list of possible antecedent candidates and awards them with scores based on certain criteria (for example, syntactic or semantic parallelism, frequency of mention, grammatical role, distance from the anaphor etc). In the end the candidate with the highest score is proposed as the correct antecedent.

An overview of some of the most well-known knowledge-poor anaphora resolution systems is given in section 1.5.

1.3 Resolution process

Most of the anaphora resolution systems deal with pronoun and/or noun phrase resolution, because the other tasks (like resolving verb phrases, clauses, sentences or even paragraphs) are too complicated. Most commonly a scope of 2-4 sentences is considered when finding a suitable referent for the anaphor, however there can be cases where the correct antecedent is further away than 4 sentences. The texts in which anaphora is resolved can be from various genres, for example newspaper texts, technical manuals, fiction, scientific texts etc.

In general the anaphora resolution process can be divided into the following smaller tasks (Mitkov 99):

1) At first the anaphora must be allocated in a sentence in the text.
2) Then all the possible antecedent candidates of the anaphora are found in the defined range of preceding sentences.
3) Then the best candidate of the antecedents' list is found by calculating scores or eliminating inappropriate candidates.
4) Finally, the last remaining candidate or the candidate with the highest score (depending on the technique used) is proposed as the correct solution for the anaphora.

1.4 Frequent issues in anaphora resolution

According to Mitkov (Mitkov 01) the following problems are faced when implementing tools for anaphora resolution:

- Computational power - semantic algorithms require more computational power and resources than the robust ones.
- The range of human assistance - majority of the anaphora resolution tools still require human input in some stages of the resolution process. If the process is done without any assistance, then the efficiency of the system drops as full automatic resolution is prone to mistakes.
- Language dependency - the implemented tools are not always language independent and that may require a remarkable amount of restructuring when it is adapted to another language.
- The preprocessed input may contain a number of errors - before a resolution tool can start its work, a lot of processing must be done on the raw text. It involves such issues as morphological and syntactical analysis, proper names recognition, extraction of noun phrases etc. As many of these processes require human intervention, they are prone to errors and that as result reduces the quality of the whole resolution process.
- Lack of annotated corpora – there are not many corpora annotated with anaphoric or coreferential links that would be widely available.
- In English the pleonastic pronoun *"it"* is considered a serious issue. The problem is

that *"it"* in the sentences *"It is snowing" or "It is six o'clock"* does not refer to anything, but is just an expression (similarly in German: *"Es schneit"*). This kind of anaphor is called pleonastic (Lappin & Leass 94). Similar use of pronouns can also be found in Estonian, but they occur very rarely compared to English. Some of them are for example the phrases like *"Nii ta on" ("So it is"), "Nüüd ta vihmale jääbki" ("Now it will remain raining"), "Nii nad väidavad" ("So they claim")*

1.5 Knowledge-poor anaphora resolution systems

The problem of anaphora resolution dates back to the time of 1960s, where first attempts on anaphora resolution were made. The implemented tools were not independent, but they were used as part of question answering systems. In 1964 a high school algebra problem answering system STUDENT was created by Daniel Bobrow (Bobrow 1964). It contained limited heuristics and could solve anaphora to a certain extent. In 1972 Terry Winograd implemented a system called SHRDLU. It understood the directions that were given to it and could pick up and move blocks. When processing the the given directions it was also able to solve anaphorical references (SHRDLU).

Hobbs 1976

The naïve, syntax-based algorithm implemented by Jerry Hobbs (Hobbs 76) works on the surface parse trees of sentences. These trees describe the grammatical structures of sentences where subjects, verbs, objects, adverbs etc. are marked. The algorithm looks for a noun phrase of the correct gender and number. The nodes of the tree are processed in an optimal left-to-right order in such way that the noun phrase upon which it terminates is regarded as the probable antecedent of the pronoun at which the algorithm starts. The implementation was tested on an archaeology book, a novel and newspaper articles and it produced the correct result in 88.3% of the cases. However, it is worth mentioning that in more than half of the cases there was only one plausible antecedent. From the anaphora that had more than one possible antecedent it managed to resolve 81.8%.

Hobbs' algorithm has remained one of the most influential implementations in the history of anaphora resolution as it is often used as a benchmark when evaluating the success of the newly implemented system. Baldwin (Baldwin 96), Mitkov (Mitkov et al. 02) and

Lappin and Leass (Lappin & Leass 94) have all used Hobbs' algorithm as comparison.

RAP (Lappin & Leass 94)

Shalom Lappin and Herbert Leass presented an algorithm for identifying the noun phrase antecedents of third person pronouns and lexical anaphors (reflexives and reciprocals). The implementation is referred to as RAP – Resolution of Anaphora Procedure. RAP has been implemented for English and German slot grammars and written in Prolog. It does not use semantic or real-world knowledge, but salience measures derived from the syntactic structure of sentences. Syntactic and morphological filters are applied to lists of pronoun-noun phrase pairs to reduce the number of possible antecedents for the pronoun. Weights are assigned to candidates based on their grammatical role, parallelism of grammatical roles, frequency of mention, proximity and sentence recency. The following preferences are used in RAP:

- Subject is preferred (i.e. by assigning higher weights) over non-subject NPs
- Direct objects are preferred over other complements
- Arguments of a verb are preferred over adjuncts and objects of prepositional phrase adjuncts of the verb
- Head nouns are preferred over complements of head nouns.

If the remaining candidates have equal weights, then the noun phrase that is closer to the anaphor is selected as the correct antecedent. Also, intrasentential antencedents are preferred to intersentential candidates. The program was trained and tested on computer manuals. For testing a set of 345 sentences randomly selected from a corpus of 48 computer manuals containing 1.25 million words was used. The program managed to identify 86% of the anaphors correctly.

Christopher Kennedy and Branimir Boguraev (Kennedy & Boguraev 96)

In 1996 they implemented an algorithm that was a modified and extended version of RAP (Lappin & Leass 94). The algorithm does not require in-depth syntactic parsing of text, but works on the output of POS-tagger which has been annotated with syntactic functions and position ID-s of each token.

The resolution procedure involves moving through the text sentence by sentence and interpreting the discourse referents in each sentence from left to right. There are two possible interpretations of a discourse referent: either it is taken to introduce a new participant in the discourse, or it is taken to refer to a previously interpreted discourse referent. The coreference between words is represented as an equivalence class. Coreference is determined by first eliminating those discourse referents to whichan anaphoric expression cannot possibly refer, then selecting the optimal antecedent from the candidates that remain, where optimality is determined by a salience measure. Then the morphological and syntactic filters are applied, after which a set of discourse referents remains. This set is processed based on the following criteria: cataphora is penalized and locality and parallelism features are boosted.

Pronoun resolution accuracy of 75% was achieved when the implementation was tested on a corpus consisting of press releases, news and magazine articles.

Breck Baldwin's CogNIAC (Baldwin 96)

The implemented system uses such features as part-of-speech tagging, simple noun phrase recognition, basic semantic category information like gender, number, and in one configuration, full parse trees. The rules followed are:

1) Unique in discourse: if there is a single possible antecedent i in the read-in portion of the entire discourse, then pick i as the antecedent
2) Reflexive: pick the nearest possible antecedent in the read-in portion of current sentence if the anaphora is a reflexive pronoun
3) Unique in current and prior: if there is a single possible antecedent i in the prior sentence and the read-in portion of the current sentence, then pick i as the antecedent
4) Possessive pronoun: if the anaphor is a possessive pronoun and there is a single exact string match i of the possessive in the prior sentence, then pick i as the antecedent
5) Unique current sentence: if there is a single possible antecedent i the read-in portion of the current sentence, then pick i as the antecedent
6) If the subject of the prior sentence contains a single possible antecedent i, and the anaphor is the subject of the current sentence, then pick i as the antecedent

The method of resolving pronouns within CogNIAC works as follows: Pronouns are resolved left-to-right in the text. For each pronoun, the rules are applied in the presented order. For a given rule, if an antecedent is found, then the appropriate annotations are made to the text and no more rules are tried for that pronoun, otherwise the next rule is tried. If no rules resolve the pronoun then it is left unresolved. The system is tested on a number of categories and the results are in the range of 75% - 89%.

Ruslan Mitkov’s approaches

Mitkov has implemented a robust tool for resolving anaphora (Mitkov 1998). It operates on text tagged by a POS-tagger and filters out incorrect candidates by applying syntactic constraints (like gender and number agreement) on them. Finally, it applies antecedent indicators to the remaining candidates by assigning scores to each candidate based on the indicator. The indicators are related to salience, for example definiteness/indefiniteness, givenness, indicating verbs, lexical reiteration, section headings and non-prepositional noun phrases are considered. Initially the tool was developed for English only, but later it has been adapted for Polish and Arabic. Mitkov found that the approach could be adapted with minimum modification to both languages and moreover, even if used without any modification, it still delivered acceptable success rates. Evaluation shows a success rate of 89.7% for English, 93.3% for Polish and 95.8% for Arabic. This approach was developed further as MARS – Mitkov’s Anaphora Resolution System (Mitkov et al. 02). Both of these Mitkov’s approaches are described in more detail in Chapter 3.

2. Estonian Pronouns

This chapter describes briefly, how the personal pronouns in Estonian are used and which types of pronouns there are in general (Erelt 03).

As there is no grammatical gender in Estonian, there is no difference in the pronouns used for referring to males and females. However, there are 14 cases in Estonian, meaning that each one of those pronouns can have 14 different word forms. These word forms are formed of the stem of the word and the case ending, which is different in almost every declination.

Below the eight types of pronouns that exist in Estonian are listed. A longer description is given about the personal pronouns as they are relevant for the current thesis.

Personal pronouns

There are six personal pronouns given in the Table 2.2.1 below. The forms written in the brackets are short forms of these pronouns.

Singular		Plural	
mina (ma)	I	meie (me)	We
sina (sa)	you	teie (te)	You
Tema (ta)	he/she	nemad (nad)	They

Table 2.2.1. Personal pronouns in Estonian

All the personal pronouns in Estonian have two forms – long and short. In general the short forms are used when there is no need to emphasize the subject, however sometimes they can also be in the stressed position. The long forms are frequently used in both positions. According to Renate Pajusalu (Pajusalu 1997) the short form *"ta"* refers to the most salient entity in the sentence, whereas the long form *"tema"* is used when the referent is contrasted to some other referent in the text. The second person plural *"teie/te"* is also used when addressing a person formally.

In general different pronouns are used for referring to animate and non-animate objects: *“ta”/“tema” (“(s)he”)* are used when talking about living creatures and the demonstrative pronouns *“see”/”too” (“this”/”that”)* are used for referring to all the other objects. However, sometimes the short form *“ta”* is also used for referring to non-animate objects. As an example (3) the following sentence from the Estonian newspaper “Eesti Päevaleht” is given:

(3) *Waigeli seekordne ettepanek nimetada uus* ***raha*** *lihtsalt euroks paistis leidvat üksmeelse toetuse, sest* ***ta*** *vastab tähtsamatele eurorahale esitatavatele nõuetele:* ***ta*** *ei oma negatiivseid ajaloolisi seoseid, on lühike , kergelt hääldatav ja identne igas riigis ning sisaldab viidet Euroopale .*

Waigel’s present proposal to name the new ***currency*** *simply euro seemed to find a unanimous support, because* ***it*** *meets the important requirements set for the euro currency:* ***it*** *does not have negative historical relations, is short, easily pronounciable, identical in every country and contains a referral to Europe.*

Here the non-animate object *“raha” (“money”, “currency”)* is referred to with the pronoun *“ta”* that is mostly used when pointing back to living objects.

Demonstrative pronouns

The most common demonstratives in Estonian are *“see” (“this”)* and *“too” (“that”).* Usually they refer to inanimate objects, but sometimes they can be used when referring to persons. This happen in two cases:

1) The referral has negative emotion, for example *“Mis tollel nüüd häda on?” (“What’s wrong with that one?”).*
2) There are two objects mentioned in the previous text, one is inanimate, the other is animate. In that case the inanimate object is referred to with a demonstrative and the animate object with a pronoun as illustrated in example (4):

(4) **Anne** ostis endale uue **kleidi**. **Ta** näeb hea välja. **See** näeb hea välja.
Anne bought herself a new **dress**. **She** looks good. **It** looks good.

3) There are two persons mentioned in one sentence. When a demonstrative is used, it refers to the person mentioned last in the sentence. The example (2) given earlier in this thesis can be easily resoved in the following case (2a):

(2a) Tom andis **Bobile** rohtu, sest **see** oli haige.
Tom gave **Bob** medicine, because **that** was ill.

The rest six types of the pronouns are listed here:

- Reflexive pronouns
- Possessive pronouns
- Reciprocal pronouns
- Interrogative-relative pronouns
- Determinative pronouns
- Indefinite pronouns

As can be seen, the list of possible pronouns in Estonian is rather long and therefore it is unrealistic to wish that the anaphora resolution system presented in the current thesis would be able to cover all these categories. The other reason for choosing only the third person personal pronouns for resolving is that these pronouns are used more frequently compared to the use of other pronouns. Statistical analysis showed that the three most frequent pronoun types that occurred in the test and training corpora used in the current research were

1) Demonstrative pronouns
2) Personal pronouns
3) Interrogative-relative pronouns

3. Mitkov's knowledge poor approach

The current chapter describes the Mitkov’s robust approach (Mitkov 98) in detail as it is used as a basis for implementing anaphora resolution system for Estonian.

3.1 Basics

Ruslan Mitkov presented a robust, knowledge-poor approach (Mitkov 98) for resolving pronominal anaphora in technical manuals. His work was a continuation of the latest trends in the search for an algorithm that would be computationally cheap, fast and reliable in terms of efficiency. It also proved that the basic set of antecedent indicators can work well not only for English, but also for other languages like Polish and Arabic. His research shows that it is possible to resolve anaphors quite successfully without thorough linguistic knowledge.

The input of the Mitkov’s program is neither parsed nor analysed, but only a part-of-speech tagger combined with noun phrase rules is used. Preference rules or so-called antecedent indicators are applied to possible candidates. The candidate with the highest score is proposed as a correct antecedent. During the work of the program the following steps are made:

1) The input is taken from the output of the POS-tagger
2) The noun phrases that precede the anaphor within the distance of two sentences are identified
3) The number and gender agreement with the anaphor is checked
4) The antecedent indicators are applied to the identified noun phrases to find the most appropriate antecedent
5) The noun phrase with the highest score is proposed as the correct antecedent of the anaphor

If two or more noun phrases have an equal score, then the following rules are considered:

1) The candidate with the higher score for immediate reference is preferred

2) In case immediate reference has not been identified, the candidate with the best collocation pattern score is selected
3) In case it still does not help to solve the disambiguity, the candidate with the higher score for indicating verbs is selected
4) In case the problem is still unsolved, the most recent from the candidates is selected

3.2 Indicators used

The indicators can be boosting or impeding, i.e. some indicators increase scores, some decrease scores. Candidate noun phrases are assigned a score in the scale of -1, 0, 1 or 2 for each indicator. The indicators have been identified empirically and are based on salience, structural matches, referential distance and preference of terms. Most of the indicators are genre-independent, i.e. they can be used in multiple texts from different genres. Below the used indicators are described in more detail.

Definiteness

Definite noun phrases are more likely antecedents of the pronominal anaphors than the indefinite ones. The noun phrase is regarded as definite if the head noun is modified by a definite article or by demonstrative or possessive pronouns. The indefinite noun phrase candidates are assigned a score of -1 whereas the definite ones score 0. If there are no articles, demonstrative or possessive pronouns in the processed sentence, then this indicator is not applied.

Givenness

Noun phrases in previous sentences representing the “given information“ are considered to be good candidates for antecedents and they score +1, whereas the candidates not representing the given information score 0. The given information is regarded as the first noun phrase in a non-imperative sentence.

Indicating verbs

If a verb is a member of the following set of verbs: {discuss, present, illustrate, identify, summarize, examine, describe, define, show, check, develop, review, report, outline, consider, investigate, explore, assess, analyse, synthesise, study, survey, deal, cover}, then the first noun phrase following it scores +1, otherwise 0. Empirical evidence suggests that because of the salience of the noun phrases which follow them, the verbs listed above are particularly good indicators.

Lexical reiteration

Lexically reiterated items are very likely the candidates for antecedent. A noun phrase scores +2 if it is repeated within the same paragraph twice or more, +1 if repeated once and 0 if it is not repeated. Lexically reiterated items include repeated synonymous noun phrases which may often be preceded by definite articles or demonstratives. Also, a sequence of noun phrases with the same head counts as lexical reiteration (e.g. *„toner bottle“, „bottle of toner“, „the bottle“*).

Section heading preference

If a noun phrase occurs in the heading of the section, part of which is the current sentence, then it is considered as the preferred candidate and it scores +1, otherwise 0.

Non-prepositional noun phrases

A non-prepositional noun phrase is preferred (scores 0) over a noun phrase which is part of a prepositional phrase (penalized with -1). This preference can be explained in terms of salience from the point of view of the Centering Theory (Grosz et al. 95). The latter proposes the ranking “subject, direct object, indirect object“ (Brennan et al. 1987) and noun phrases which are parts of prepositional phrases are usually indirect objects.

Collocation pattern preference

This preference is given to candidates that have an identical collocation pattern with the pronoun. In that case they are rewarded with +2 points. The collocation patterns that the pronoun must match are the following:

- <noun phrase> , <pronoun> <verb>

- <verb>, <noun phrase> <pronoun>

The example (5) illustrates the collocation pattern case:

(5) Press the **key**$_i$ down and turn the volume up... Press **it**$_i$ again.

Immediate reference

In technical manuals the immediate reference indicator can often be useful for identifying the correct antecedent. In constructions of the form

... you V1 NP ... con (you) V2 it (con (you)) V3 it) where con $\in$ {and/or/before/after...},

the noun phrase immediately after V1 is a very likely candidate for antecedent of the pronoun *"it"* immediately following V2 and is therefore given preference by having assigned a score of +2. If the noun phrase does not follow the pattern, it is assigned 0 points. This preference can be viewed as a modification of the collocation preference. It is also frequent with imperative constructions.

Referential distance

In complex sentences, noun phrases in the previous clause are the best candidate for the antecedent of an anaphor in the subsequent clause, followed by noun phrases in the previous sentence, then by nouns situated 2 sentences further back and finally nouns 3 sentences further back. The assigned scores for matching these criteria are +2, +1, 0 and -1, correspondingly. For anaphora in simple sentences, noun phrases in the previous sentence are the best candidate for antecedent, followed by noun phrases situated 2 sentences further back and finally nouns 3 sentences further back. The assigned scores for matching these conditions are +1, 0 and -1, correspondingly.

Term preference

If a noun phrase represents a term in the field that the text is about then it is more likely to be the antecedent (scores +1). The noun phrase which is not among the predefined terms list scores 0.

3.3 Evaluation

For practical reasons, the approach presented does not incorporate syntactic and semantic information (other than a list of domain terms) and it is not realistic to expect its performance to be as good as an approach which makes use of syntactic and semantic knowledge in terms of constraints and preferences. The evaluation however shows that much less is lost than might be feared and the results are in fact comparable to syntax-based methods (Lappin & Leass 94). Mitkov suggests that such good result has been achieved because various indicators were used and no indicator was given absolute preference.

Two evaluations were carried out on technical manuals in English. The achieved success rates (the number of correctly resolved pronouns divided to all pronouns attempted to be resolved) were 95.8% and 83.6%. There were cases where the program failed to resolve the anaphor when

- The anaphor and its antecedent were in the same sentence, but the preference was given to a candidate in the previous sentence
- The sentence where the anaphor was located had a more complex syntactic structure

The described approach was also adapted to languages like Polish and Arabic with the success rates of 93.3% and 95.2%, correspondingly.

3.4 MARS

A few years later Mitkov's knowledge-poor approach was improved in the way that it became fully automatic. The new version was named MARS – Mitkov's Anaphora Resolution System (Mitkov et al. 02) and it introduced the following new features:

1) Boost pronoun indicator: the pronouns itself are also considered as possible antecedent candidates. One reason for that is the fact that pronouns indicate salience, as they represent noun phrases. The other reason is that one pronoun can refer to another pronoun that can refer to a noun phrase i.e. the anaphora can be solved transitively. All the pronouns are assigned a score of +1.
2) Syntactic parallelism: the noun phrases that represent the same syntactic role as the anaphor are awarded with a score of +1. The syntactic roles of the candidates were

obtained by using new pre-processing software: FDG Parser.

3) Frequent candidates: the three candidate noun phrases that occur most frequently in the sets of competing candidates of all pronouns in the text are assigned a score of +1.
4) Givenness indicator was changed in the way that subject nouns score +2, objects score +1, indirect objects score 0 and the nouns for which the parser is not able to assign a function are penalized by -1.

The success rate achieved by MARS was 61.55% which was considered encouraging because the research made in (Palomar et al. 01) has shown that the approaches operating without any semantic knowledge usually do not achieve a success rate higher than 75%.

4. Anaphora Resolution System for Estonian

This chapter gives a thorough description of the anaphora resolution system implemented for Estonian. At first an overview of the used corpora is given. Then the architecture of the developed system is explained with examples of the program output. Finally the indicators used in the algorithm are described in detail.

The resolution system for Estonian texts has been implemented in Java programming language and it attempts to solve third person personal pronouns (both in singular and plural forms) in newspaper and scientific articles. The first and second person pronouns are not considered in the current solution as resolving them requires knowledge that might be outside the text and outside the scope of the implemented system, i.e. semantic knowledge. Also, the first and second person pronouns are less frequent in newspaper and scientific texts compared to the third person pronouns.

The current version of the program operates with the morphological and syntactical information of text, but not with semantic data. The algorithm used for implementing the system is based on the knowledge poor approach described by Mitkov (Mitkov 98) and its successor MARS-system (Mitkov et al. 02). Both of these approaches have been covered in Chapter 3 of this thesis. Although the main idea of this algorithm remains the same, a number of language and genre-specific modifications were made to the algorithm due to different characteristics of the Estonian language and different character of target texts. All these new and modified features are described in more detail in section 4.1 of the current thesis.

4.1 The training and test data

The program is trained and tested on corpora of texts that are morphologically analyzed and manually disambiguated. The training and test corpora have been created from the following files:

1) The files in the morphologically disambiguated corpus (MDC) of Estonian newspaper texts, fiction, scientific texts and legal texts

2) The morphologically and syntactically disambiguated corpus of Estonian newspaper texts, fiction and scientific texts. The contents of these files partly overlap with the contents of the MDC files

The main reason for using two different types of files was to get more resources for training and testing – the problem with Estonian language resources is that they are minimal. Another reason was to test using syntactic annotation during anaphora resolution. The statistical information of the data in the training and test corpora is shown in the tables 4.1.1 and 4.1.2 below. The last column displays the number of third person personal pronouns in the corpora.

Data	Words	Pronouns	3.p.p.pronouns
File 1: pm99	12215	875	130
File 2: ml98	10601	754	117
File 3: epl98	10437	670	126
File 4: sl	2966	262	37
File 5: arip	2835	130	28
File 6: horisont	5232	475	93
File 7: pm97	8208	607	79
File 8: epl	4979	354	36
	57473	**4127**	**646**

Table 4.1.1 Training corpus data

Data	Words	Pronouns	3.p.p.pronouns
File 1: epl99	4390	277	62
File 2: pm99	8267	576	95
File 3: ml98	8496	642	74
File 4: epl98	5733	432	59
File 5: ilu_00001	2052	263	66
File 6: ilu_00011	1985	244	60
File 7: horisont	13487	1244	203
File 8: ee	12734	1153	176
File 9: pm97	6823	366	61
	63967	**5197**	**856**

Table 4.1.2. Test corpus data

The training corpus consisted of 57473 words out of which 646 were the pronouns of interest. The test corpus consisted of 63967 words, out of which 856 were the pronouns of interest. The number of all pronouns in the training corpus was 4127 and in test corpus 5197. Although the ratio of third person personal pronouns to the total number of pronouns might seem very small, one has to keep in mind that there are 8 types of different pronouns in Estonian altogether and they were all represented in the corpus as well, but the percentage compared to the personal pronouns was lower. The three most frequent types of pronouns in the training and test corpora were:

- Demonstrative pronouns (~18% of all the pronouns)
- Third person personal pronouns (~16% of all the pronouns)
- Interrogative-relative pronouns (~12% of all the pronouns)

As can be seen, these three types of pronouns make a bit less than a half (46 %) of all the pronouns occurring in the corpora.

The training corpus consists mostly of texts from Estonian newspapers and a small part of texts from a scientific magazine “Horisont”. During the testing also mostly newspaper texts were used and a smaller number of texts from “Horisont”. Additionally fictional texts were involved in the test corpus – the files 5 and 6. Although the initial idea of the system was to attempt to resolve the third person personal pronouns in newspaper texts, it was tested on a small number of fiction texts, just to evaluate the future perspectives - how the program is able to deal with such genre. The results achieved have been covered in Chapter 5 of this thesis.

4.2 Annotation of the corpora

As the corpora used for training and testing consists of two types of files, the ones that are only morphologically annotated and the ones that include both morphological and syntactical data, an overview of the annotation of the both types of files is presented below.

Morphologically annotated files

In morphologically annotated files every sentence is placed between the <s> and </s> tags. All the words and punctuation marks are placed on separate lines, followed by the morphological information. At first the word form as it was found in the text is presented.

It is followed by (separated by 4 spaces) the stem of the word and the ending of the word separated by “+” sign.

The data between the “//” signs contains the morphological category of the given word. The most important category data for resolving anaphora are the word type, number and declination, but in addition to that the category data includes a number of other markers. All the existing categories and the corresponding markers are described in detail in the documentation of the MDC and can be retrieved from the webpage (MDC). The alphabet characters that are specific to the Estonian language are presented as HTML entities. The description of all the entities can be retrieved from the webpage of the Research Group of Computer Linguistics of the University of Tartu (RGCLUT).

Hereby an example (6) of a noun record in the file is discussed. The word token is “reporteri” which is a genitive form of the word “reporter” (“correspondent”). The initial word form (“reporteri”) is followed by the stem of the word and declination ending (“reporter+0”). The morphological category which is displayed between the “//” markers indicates that the word is a noun (“_S_”), a common noun (“com”), in singular (“sg”) and in genitive case (“gen”).

(6) Reporteri   reporter+0 //_S_ com sg gen //

Below a longer excerpt of a morphologically annotated text is shown:

```
<s>
Reporteri    reporter+0 //_S_ com sg gen //
uudishimu    uudis_himu+0 //_S_ com sg nom //
&uuml;letab    &uuml;leta+b //_V_ main indic pres ps3 sg ps af //
n&uuml;&uuml;d    n&uuml;&uuml;d+0 //_D_ //
juba    juba+0 //_D_ //
igasugused    iga_sugune+d //_P_ pl nom //
piirid    piir+d //_S_ com pl nom //
ja    ja+0 //_J_ crd //
ta    tema+0 //_P_ sg nom //
pinnib    pinni+b //_V_ main indic pres ps3 sg ps af //
,    , //_Z_ Com //
kas    kas+0 //_D_ //
&otilde;petajaid    &otilde;petaja+id //_S_ com pl part //
```

```
ikka    ikka+0 //_D_ //
j&auml;tkub    j&auml;tku+b //_V_ main indic pres ps3 sg ps af //
.    . //_Z_ Fst //
</s>
```

Morphologically and syntactically annotated files

The morphologically and syntactically annotated files are quite similar to the files described above; however there are some minor differences. Here the sentences are separated by $<s> and $</s> tags followed by a line of four “#” characters. The words and punctuation marks and their morphological and syntactical data are represented on separate lines. The word stem and the ending and the morphological category of the word are displayed exactly in the same way as in the morphologically annotated files. However, an addition here is the syntactical information. For the anaphora resolution only the subject (“@SUBJ”) and object (“@OBJ”) tags were used.

An excerpt of a morphologically and syntactically annotated file is given below:

```
$<s>
    ####
&ldquo;
    &ldquo; //_Z_ Oqu //  **CLB
Neis
    see+s //_P_ pl in #cap //  @ADVL
oli
    ole+i //_V_ main indic impf ps3 sg ps af #FinV #Intr //  @+FMV
ka
    ka+0 //_D_ //  @ADVL
algul
    algul+0 //_D_ //  @ADVL
korralik
    korralik+0 //_A_ pos sg nom //  @AN>
põllumajandustoodangu
    põllu_majandus_toodang+0 //_S_ com sg gen //  @NN>
müük
    müük+0 //_S_ com sg nom //  @SUBJ
$,
    $, //_Z_ Com //
&rdquo;
    &rdquo; //_Z_ Cqu //  **CLB
```

```
meenutas
    meenuta+s //_V_ main indic impf ps3 sg ps af #FinV //  @+FMV
ta
    tema+0 //_P_ sg nom //  @SUBJ
$.
    $. //_Z_ Fst //
$</s>
    ####
```

4.3 Architecture of the implemented system

Hereby the structure of the anaphora resolution system is presented. A diagram of the general idea of the implementation is presented on Figure 4.3.1:

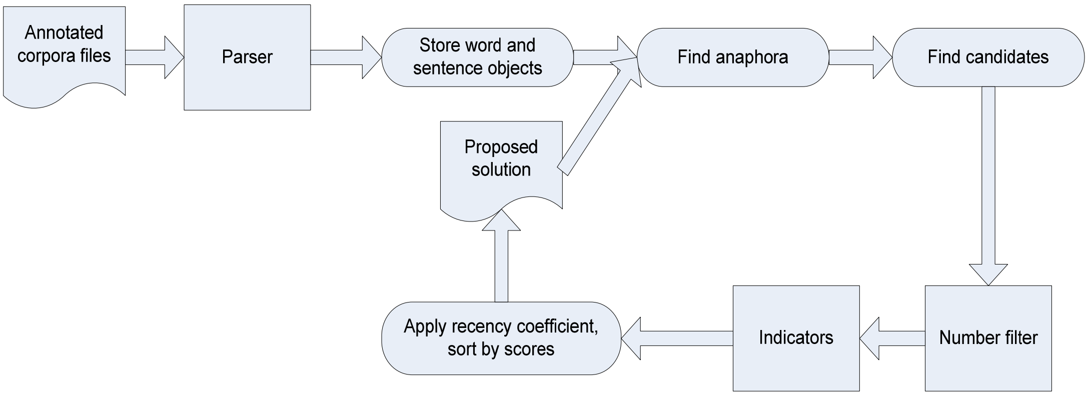

Figure 4.3.1 Architecture of the implemented program

The main aim of the program is to output the correct antecedent for the given anaphor. For example (7) in the following sentence

(7) Meri kuulas hapu näoga **Rüütli** selja taga **tema** eestikeelset pressikonverentsi Valge Maja trepil.

Meri sourly listened, standing behind **Rüütel's** back **his** press conference in Estonian on the stairs of the White House.

the program must locate the pronoun *"tema" ("(s)he")* first and then reach to the correct antecedent of *"Rüütli" ("Rüütel's").*

Four Java classes have been implemented to create the pronoun resolution system for Estonian:

- Resolve.java – it is the class where the main method is located
- Resolver.java – in this class the main work of the resolution is done, it contains the parsers, object creations, number filter and indicators
- Word.java – this class represents a word object and all the data that is stored with a word object
- Sentence.java – this class represents a sentence object

The source code written was about 950 rows long in total. The source code can be found on a CD that comes together with this thesis.

At first the input text file is read in line by line and parsed according to the file type. As there are two different annotations, two parsers were used in order to retrieve the data correctly. In the current version of the program the parsers are both located in the Resolver class, but it might be a good idea to store them as separate files (classes) in the future. All the words and sentences taken from the input text file are stored as objects. The morphological information about every word as well as the location info (word number in text, word number in sentence, sentence number) and the scoring info (the score the word is given for different indicators) is stored together with the word itself.

For sentences an independent object is created – an object that contains the word objects. Sentence type info is stored together with the sentence. The sentences can be of three different types: normal, complex and headings.

When the text file has been read into the memory, the program starts to search for the anaphora to resolve. It looks only for the third person personal pronouns. These pronouns are identified by the morphological information that was previously saved for every word object, i.e. if the word type is "_P_" then it is a pronoun. As the current version of the program does not consider all the pronouns in the text, the third person personal pronouns are filtered out by the criterion that the stem of the word contains the pattern "tema" ("he"/"she"). In that way the plural forms are also retrieved as the plural forms of the word "tema" are annotated as shown in the example (8) below:

(8)
```
ta    tema+0 //_P_ sg nom //
```

```
tema    tema+0 //_P_ sg nom //

nad    tema+d //_P_ pl nom //

nemad    tema+d //_P_ pl nom //
```

This filter is able to find all the declination forms as the stem of the word always remains the same, as shown in example (9) for the word „nende” („their”):

(9) `nende    tema+de //_P_ pl gen //`

If a pronoun has been found in the text, it is saved as an anaphora and the program continues to look for the appropriate candidate nouns for the anaphora. The candidates are again found on the basis of their morphological data: the program looks for nouns (words of type “_S_”) and pronouns (words of type “_P_”). Range for finding the possible antecedents is 3 sentences: the sentence, where the anaphor was found, the previous sentence and the sentence before the previous sentence. Hereby it should be discussed whether the 3 sentence scope is wide enough for finding the antecedents. A statistical survey of the test corpus showed that only in about 1.5% of the all cases that the program attempted to resolve the correct antecedent was not in the scope of 3 sentences. There were a number of cases where the final antecedent itself was not inside the 3 sentence range, however one of these sentences still contained a pronominal referral to the correct antecedent, i.e. it was possible to use the previous pronouns as hints to the correct solution. This topic is further discussed in section 4.4 “Filters and indicators”. By widening the range of sentences by one or two would have involved more resources and time, but as the statistical analysis showed it would not have been of much benefit.

In the sentence where the anaphora itself is located only the words preceding the anaphora are considered (i.e. the cataphora is not identified). The matching candidates are saved as word objects and their morphological and syntactical data is stored with them as well. The resolution process continues by processing the vector of candidate words by a number of resolution indicator modules to calculate the best antecedent for the anaphor. The exact filtering and indicator calculation process is described in section 4.4.

After the list of candidate words has been processed by all the indicators, they are multiplied with the recency constant:

- The nouns located in the previous sentence are multiplied by 0.75
- The nouns in the sentence before the previous sentence are multiplied by 0.5

These exact coefficients were successfully implemented in the Norwegian Anaphora Resolution system (Holen 06) and they proved to be efficient enough in the current system as well, because changing them did not give any better effect than the initial combination did.

In the last step the candidates are sorted by their total score. The candidate with the highest score is proposed as an antecedent. If there is a tie between two or more nouns, then the noun that is closer to the anaphor in the text is preferred. The algorithm proceeds to the place where the previous anaphora was found and continues to look for the next anaphora to be resolved.

Hereby an example (10) of how the program works is given. The English translation is not part of the initial output, but was added for better understanding. The numbers mark the word's number in the texts and function as a unique identifier for the word.

(10)

```
5. Lahendada anafoor 'Ta' (191) lauses:
```

Resolve the anaphora 'Ta' in the sentence:

```
Ta (191) käis (192) kirikus (193) ja (194) tundis (195)
vajadust (196) pihtida (197) . (198)
```

He went church$_{\textit{(inessive)}}$ *and felt need*$_{\textit{(partitive)}}$ *confess*$_{\textit{(infinitive).}}$

He went to church and felt a need to confess.

```
Eelmine lause: Goebbelsi (182) noorusaastate (183) kohta
(184) ei (185) ole (186) midagi (187) halba (188) väidetud
(189) . (190)
```

Goebbels$_{\textit{(genitive)}}$ *youth-years*$_{\textit{(genitive)}}$ *about not is nothing bad claimed.*

Nothing bad has been claimed about Goebbels' youth years.

```
Üleelmine lause: Aasta (168) hiljem (169) kaitses (170) ta
(171) samas (172) dissertatsiooni (173) ja (174) sai (175)
filosoofiadoktori (176) kraadi (177) saksa (178) kirjanduse
(179) alal (180) . (181)
```

Year later defended he dissertation$_{\textit{(genitive)}}$ *and received philosophy-doctor* *degree*$_{(genitive)}$ *German literature field*$_{(addessive)}$

A year later he defended the dissertation and received a Doctor of Philosophy degree in the field of German literature.

```
Anafoori lahendikandidaadid on:
```

The anaphora resolution candidates are:

```
Goebbelsi (182): 3.75 (Fre:2.0 Dec:0.0 Name:2.0 Pat:0.0
Dist:1.0 Ind:0.0 Quo:0.0 Pro: 0.0)
Aasta (168): 2.5 (Fre:2.0 Dec:3.0 Name:0.0 Pat:0.0 Dist:0.0
Ind:0.0 Quo:0.0 Pro: 0.0)
ta (171): 2.5 (Fre:2.0 Dec:3.0 Name:0.0 Pat:0.0 Dist:0.0
Ind:0.0 Quo:0.0 Pro: 0.0)
alal (180): 0.5 (Fre:1.0 Dec:0.0 Name:0.0 Pat:0.0 Dist:0.0
Ind:0.0 Quo:0.0 Pro: 0.0)
filosoofiadoktori (176): 0.0 (Fre:0.0 Dec:0.0 Name:0.0
Pat:0.0 Dist:0.0 Ind:0.0 Quo:0.0 Pro: 0.0)
kraadi (177): 0.0 (Fre:0.0 Dec:0.0 Name:0.0 Pat:0.0 Dist:0.0
Ind:0.0 Quo:0.0 Pro: 0.0)
kirjanduse (179): 0.0 (Fre:0.0 Dec:0.0 Name:0.0 Pat:0.0
Dist:0.0 Ind:0.0 Quo:0.0 Pro: 0.0)
dissertatsiooni (173): 0.0 (Fre:0.0 Dec:0.0 Name:0.0 Pat:0.0
Dist:0.0 Ind:0.0 Quo:0.0 Pro: 0.0)
```

At first the program presents the sentence where the anaphora was found and indicates the ID of the anaphora to be resolved. Then the two previous sentences are printed (if they exist). The noun and pronoun candidates from these three sentences are processed by the number filter and salience indicators (see section 4.4). The total score for each of the candidates is multiplied by the recency coefficient, which is 0.75 for the second sentence and 0.5 for the third sentence. The coefficient for the anaphora sentence is 1, i.e. the score of the candidates that are located in the same sentence as the anaphora remains the same. Finally a ranked list is made of all the candidates and printed to the screen. For every candidate their ID, the final score and the scores from each indicator are printed (Frequency, Declination, Name, Pattern, Distance (also called Recency), Indicative, Quotes, Pronoun boosting). The first candidate in the list is proposed as the suitable antecedent for the anaphora.

As can be seen in example 5 the program finds the correct antecedent “Goebbelsi” which ends up with a score of 3.75 points from the following indicators: 2 points by Frequency indicator, 2 points by Name indicator, 2 points by Recency (Distance) indicator. The sum of 5 points is multiplied by the recency coefficient of 0.75 as the candidate is located in the previous sentence of the anaphora.

4.4 Filters and Indicators

The list of candidate words is processed by the number filter and by 8 different indicators described below in detail.

Number agreement filter

The number filter is the only eliminative module in the program (i.e. here some of the candidates may be dropped from the initial list); there is no gender filter as it is common for languages like English or German as there is no grammatical gender in Estonian. The number filter checks if the anaphor and the candidate noun match in number. In case of a mismatch the inappropriate candidate is dropped from the list. If no noun with a matching number is found, then the initial set of candidate words is returned without processing. This kind of situation can mean two things:

1) The antecedent was out of the range of the scope of the three sentence limit set initially
2) The antecedent’s number is different from the pronoun number.

The latter case is more likely to happen than the first one. There are a number of cases where a noun in singular is referred to with a pronoun in plural form (“nemad” or “nad” – “they”), for example the words that describe a group of people: parliament, government, party, a name of a sports team or a company etc. Also the referent of a plural pronoun may contain more than one word, even a whole list of words. Resolving that kind of anaphora has proved to be one of the bigger problems that occurred when attempting to resolve the Estonian anaphora. As the current system only outputs up to two words (these were only names) as the correct resolution, the cases with a higher number of antecedents have been omitted when evaluating the test results. These cases are described in more detail in Chapter 5 of the current thesis.

During the training phase the 8 indicators described below were implemented, tested and analyzed. No candidates were eliminated when applying the indicators, but the candidates were assigned positive or negative scores based on different criteria. The score values were implemented similarly to the Mitkov's algorithm, many of the scores remained the same and the scores for new indicators were assigned based on the same scale and their effectiveness was tested, analyzed and changed when necessary to obtain the best results. The lowest possible score assigned is -3 and the highest is +2.

The goal of the training phase was to achieve such combination of indicators and indicator scores that it would produce as much correctly resolved anaphora as possible. The cases where the system failed were analyzed and the indicators were improved in the way that would make the performance better.

Givenness indicator

Initially the givenness indicator was implemented similarly to the way it was described in Mitkov's original approach (Mitkov 98). According to his algorithm the first noun phrase in the previous sentence is awarded with +1 point if the anaphora is located in a simple sentence. If the anaphora is in a complex sentence, the first noun phrase of the previous clause is assigned a score of +1. This feature was implemented in the Estonian system with minor changes – as the clauses are not annotated in the used corpora, the first noun phrase of the complex sentence was assigned +1 if the anaphora itself was also located in the same complex sentence. However, this indicator was finally removed from the Estonian system as it did not prove to be efficient enough. The main reason for that is that Estonian is a language of free word order, meaning that the salient information is not always located in the first part of the sentence, but it can also be in the very last words of the sentence.

However, givenness is still considered in the current implementation, but in a different form. The declination indicator described further down is actually an indicator that considers the salience or givenness of the words, but not by the location of the word, but by its declination.

Frequency indicator (lexical reiteration)

The nouns and pronouns that appeared frequently in the text were awarded points according to the following criteria:

- The candidate was assigned a score of +1 if it occurred in the section twice (section is considered to be a range from one heading to another)
- The candidate was given a score of +2 if it occurred in the section more than twice

As the words in Estonian can be in different declinations, the stem of the word was considered when comparing the words instead of the initial word form.

In the original Mitkov's algorithm the frequency of a word was calculated in the range of one paragraph, but as the paragraphs were not steadily marked in the Estonian corpora, the range of text from one heading to another was used.

Declination indicator

As morphological annotation provides us with the words' declination info, this kind of knowledge can be used for finding the salient entities. The nominative case refers to a possible subject; a smaller likeliness is that it refers to an object or an adverbial. The partitive case mostly refers to an object. This indicator was added to the system to replace the syntactic indicator in these files of the corpora that were not provided with syntactical annotation. The idea of the indicator is the following:

- The candidates in nominative case are awarded +2 points
- The candidates in partitive case are awarded +1 point
- The candidates that are in the same declination as the anaphor are awarded +1 point

Name indicator

This indicator was added to the system due to the fact that the names tend to be one of the most salient entities in newspaper texts and in fiction. The names list was obtained during the file parsing by filtering out all the proper nouns that occur next to each other in the text file. Most likely that kind of combination represents a first and a last name of a person, but it can also represent the name of a company, country, sports team, institution, organization etc.

- The candidate that appears in the list of names is awarded +2 points

Pattern indicator

The pattern indicator was implemented identically to the one described by Mitkov: the pattern <NOUN><VERB> or <VERB><NOUN> was searched.

- If the anaphora was preceded by a verb, then the candidate that was also preceded by a verb was assigned a score of +1
- If the anaphora was followed by a verb, then the candidate that was also followed by a verb was assigned a score of +1

These kinds of patterns are often used in newspaper articles when thoughts or statements of people are expressed. An example (11) is given below:

(11) *”Hommikusel visiidil on tal haige ära kuulamiseks ainult veerand tundi, “; räägib Aleksei.*

”During the morning visit (s)he has only a quarter of an hour to listen to the patient, “; says Aleksei.

Here the pattern <VERB><NOUN> is matched by the construction “räägib Aleksei”.

Referential distance indicator

The distance filter was implemented similarly to the one described by Mitkov. The difference here is that as no clauses have been identified in the Estonian corpora, all the candidates in the same sentence are awarded with points if the anaphor is in a complex sentence.

- If the anaphor is located in a complex sentence, the candidates in the same sentence are awarded +2 points.
- Irrespective of the type of the sentence of the anaphora all the candidates in the previous sentence are awarded +1 points.

Section heading indicator

Section heading indicator boosts the candidates that appear in the section heading by assigning them +1 points. However, it did not prove to be efficient enough and was omitted.

Indicative verbs indicator

The general idea of Mitkov’s algorithm was used, but with some modifications. A list of verbs and phrases that tend to occur frequently in newspaper texts before salient entities, most likely names of persons, was composed. The list is as follows: ("ütles", "ütleb", "väitis", "lisas", "märkis", "lausus", "kinnitas", "möönis", "sõnas", "soovitas", "hinnagul", "sõnul") / (“said”, “says”, “claimed”, “added”, “noted”, “uttered/said”, “declared”, “admitted”, “uttered/remarked”, “recommended”, “in the opinion of”, “according to the words of”).

- The candidate that is preceded by one of the phrases in the list is awarded +2 points
- The candidate that is followed by one of the phrases in the list is awarded +2 points

Syntactic parallelism indicator

The syntactic parallelism indicator was at first implemented in the way it was implemented in Mitkov’s MARS. According to that system the following scores were assigned:

- If the candidate is marked as subject and the anaphora is marked as subject, the candidate is assigned as score of +1
- If the candidate is marked as object and the anaphora is marked as object, the candidate is assigned as score of +1

However, in the final version of the implemented program the syntactic indicator was omitted as it did not prove to be efficient enough. When it was left out, the efficiency of the resolution increased by about 0.5%.

Quotation indicator

This indicator was added to the system due to the fact that newspaper texts contain a remarkable number of direct speech. If the anaphora is located inside the direct speech (i.e. inside the quotation marks), it is very likely that the antecedent is also located inside the direct speech. Respectively, if the anaphora is not located in the quotation marks, the probability for the antecedent to be outside the quotations is also more likely. These two cases can be illustrated with examples (12) and (13) from the newspaper text taken from the training corpora.

(12) ”Kõigepealt las see **asi** juhtub, siis analüüsime, miks ja kuidas **ta** juhtus ja alles seejärel võtame vastu otsused, ” lausus Vähi.

”At first let this **thing** happen, then we analyze, why and how **it** happened and only then we will make the decisions, ” said Vähi.

(13) Vastates küsimusele, kas valitsusliit ei lagune ka juhul, kui selgub mõne valitsusliidu poliitiku osalus SIA jälitustegevuses, ütles **Vähi**, et ei hakka kunagi mõtlema sellele, ”mis juhtub siis, kui miski asi juhtub”. “Siseminister Savisaar on praegu peaministri ülesannetes, ” ütles **ta**.

Answering to the question if the coalition will not fall apart even if it will come out that some of its politicians are part of SIA’s detective work **Vähi** said that he will never start to think of ”what happens if something happens”. ”Minister of the Interior Savisaar is in the duties of prime minister right now,” **he** said.

The candidates that are part of direct speech are penalized with -3 points if the anaphora itself is not part of direct speech. If the anaphora itself is located between the quotation marks, the penalty is not implemented.

Boost pronoun indicator

According to Mitkov (Mitkov et al. 02) the pronouns in texts carry salient information and they can be used as hints for reaching the correct antecedent. The pronouns can create chains that may lead to the correct antecedent step by step.

- The pronoun candidates in the current and previous sentences are regarded to be salient and they are awarded +1 point.

Indicators summary

To sum up the described indicators section, the list of the indicators implemented in the Estonian anaphora resolution system is once more presented:

- Frequency indicator (lexical reiteration)

- Pattern indicator
- Referential distance indicator
- Indicative verbs indicator
- Boost pronoun indicator
- Name indicator
- Declination indicator
- Quotes indicator

Three of the indicators that were initially implemented and later left out due to lack of efficiency:

- Givenness indicator
- Syntactic parallelism indicator
- Section heading preference

The following indicators that were presented by Mitkov were not tested at all on various reasons:

- Definiteness indicator - there are no articles in Estonian.
- Term preference indicator – as the resolution is tested on newspaper texts and scientific texts that cover a very wide field of topics, it is not possible to create a final list of terms that might be salient in the topics discussed in these texts.
- Non-prepositional noun phrase indicator – in Estonian the number of prepositions is very small (postpositions are used more frequently) and they are used not nearly as often as in English.
- Immediate reference indicator – it was not implemented, because it was too genre-specific.

5. Testing

The current chapter presents the results that were achieved when the Estonian anaphora resolution algorithm was applied to training and test corpora. The general cases where the anaphora was resolved wrongly are described and analysed.

5.1 Overview

The current chapter describes which results were obtained when the program was run on test corpus. The statistical data about the test corpus in presented in section 4.1 in Table 4.2. Compared to the training corpus some fiction texts were added to the corpus to see how well the system manages to resolve the anaphora that does not belong to the target genre.

The texts files used for testing (and training) were not manually checked for errors; however, the errors that occurred during the resolution were fixed as they prevented the system to work properly. The most common errors were that some of the tags were located on the wrong line (the next line) or there were redundant whitespaces in the middle of morphological information tags. In some places the sentence end tags were missing. Another problem was that the files were partly inconsistent – sometimes the quotes were marked as “” and sometimes as *“ldquo”* and *“rdquo”*. The paragraph tags were also inconsistent in the corpora files - some files contained these tags, some files did not. Due to that reason the paragraph tags were overlooked during parsing.

The test and training results are shown in tables 5.2 and 5.3.

- In the training phase the resolution system was run on 8 files. The file 6 was from the scientific magazine “Horisont” and the rest of the files were from newspapers.
- In the testing phase the resolution system was run on nine files: the files 1-4 and 8-9 were newspaper articles, the files 5-6 were fiction texts and the file 7 was from ”Horisont”.

In order to evaluate the success rate of the program it has to be calculated by using the following formula (Mitkov 98):

Success rate$_{\text{All}}$ = correctly resolved anaphora / total number of anaphora resolved

However, the program is not able to solve a number of cases due to the implemented algorithm. These cases are discussed in detail in section 5.1. For fair evaluation of the results the pronouns that were impossible for the program to resolve were counted out of the anaphora. Hence a new formula for finding the percentage of correctly resolved anaphora would be the following:

Success rate$_{\text{In scope}}$ = correctly resolved anaphora / total number of anaphora in scope resolved

Considering these two formulas the following results were obtained (Table 5.1):

	Success rate (all)	Success rate (in scope)
Training corpus	0.6904	0.7357
Test corpus	0.6414	0.7369

Table 5.1. Training and test corpus overall results

The results are described in more detail in tables 5.2 and 5.3. The last column shows the Success rate$_{\text{In scope}}$ for every file in the corpus. The results vary from 65.2% to 88.9% in the training corpus and from 64% to 81.8% in the test corpus. The analysis of the achieved results is done in Chapter 6.

Data	3.p.p. Pronouns	Correctly resolved	Incorrectly resolved	Out of scope	Percentage correct
File 1: pm99	130	91	30	9	75.2
File 2: ml98	117	73	39	5	65.2
File 3: epl98	126	94	28	4	77.0
File 4: sl	37	27	6	4	81.8
File 5: arip	28	19	9	0	67.9
File 6: horisont	93	54	21	18	72.0
File 7: pm97	79	50	21	8	70.4
File 8: epl	36	32	4	0	88.9
	646	**440**	**158**	**48**	**73.6**

Table 5.2. Training corpus results

Data	3.p.p. pronouns	Correctly resolved	Incorrectly resolved	Out of scope	Percentage correct
File 1: epl99	62	31	15	16	67.4
File 2: pm99	95	48	27	20	64.0
File 3: ml98	74	46	16	12	74.2
File 4: epl98	59	34	11	14	75.6
File 5: ilu_00001	66	46	14	6	76.7
File 6: ilu_00011	60	40	15	5	72.7
File 7: horisont	203	146	50	7	74.5
File 8: ee	176	113	38	25	74.8
File 9: pm97	61	45	10	6	81.8
	856	**549**	**196**	**111**	**73.7**

Table 5.3. Test corpus results

5.2 Error analysis

The cases where the program did not produce the correct answer can be divided into two bigger categories by the character of the errors:

1) The cases where the wrong antecedent is picked, because it is impossible for the system to resolve the anaphora due to the algorithm itself
2) The cases where a wrong antecedent is picked due to the domination of a wrong indicator because of sentence structure and due to the wide range of variations of the language (i.e. more rules are needed to solve certain cases)

Non-resolvable anaphora

During the training of the program and analysis the following cases evolved that the program could not resolve:

1) The correct antecedent is split between two words or the antecedent was a whole list of words. The program is only able to find an antecedent that consists up to two words, but they have to be located next to each other in the text (i.e. full names of

persons). An example (14) is given where the algorithm fails to find the two antecedents:

(14) Varemgi on **Statoil** ja **Neste** tõstnud kütuse hinda koos, mõne päeva pärast on **nende** eeskuju järginud ka Shell.

Previously too **Statoil** and **Neste** have raised the price of the oil at the same time, some days later **their** example was followed by Shell.

2) The correct antecedent is in singular, but the anaphora referring to it is in plural. There are a number of cases where the pronoun in plural refers to a noun in singular. Currently all the mismatching candidates are removed, so the correct antecedent might be eliminated as well. The program is only able to reach the correct solution if there are no noun candidates in the plural form. If no corresponding nouns of the matching number are found by the number filter, it means that the antecedent is probably of other number and the candidates list is processed further in its initial form. The example (15) given below illustrates the mismatch of an anaphora and its antecedent.

(15) **Soome** näiteks teatas, et **nende** esindaja Erkki Liikanen võiks jätkata.

Finland for example announced that **their** representative Erkki Liikanen could continue.

The number mismatch is also very common with the constructions where something is counted, for example “12 meest” (“12 men”). Here the noun is in singular and in partitive case. As it is referred to with the pronoun in plural, the correct antecedent is eliminated in the number filter because of number mismatch and a wrong answer is given as a result.

3) The pronoun refers forwards not backwards, i.e. it was cataphora instead of anaphora. The program only processes the antecedents that precede the anaphora so it is impossible for the program to find the correct answer to the following problem (16). Secondly, the referred pronoun is not pronominal, but it is an interrogative-relative pronoun that is not handled by the current version of the program.

(16) Eutanaasiavastased on teleesinemise vastu, sest nende hinnangul võib seesuurendada enesetappe **nende** seas, **kes** pole surmahaiged.

The people who are anti-euthanasia are opposing TV-performances, because in their opinion it may increase suicides among **those who** are not deathly ill.

4) The pronoun refers to a pronoun that is not a personal pronoun. As the implemented program only looks for the nouns and third person personal pronouns, then the example (17) below cannot be resolved by the program as the correct antecedent of the word “nad” is “paljud” which is an indefinite pronoun.

(17) **Paljud** on esimest korda vastutusrikkal töökohal ja see ei tohiks tähendada, et **nad** hakkama ei saa.

For **many** (people) it is the first time on a responsible positions and it should not mean that **they** will not manage.

5) The correct antecedent is not in the range of 3 sentences. The example (18) illustrates this case. As the algorithm only looks for antecedent candidates in the current and two previous sentences, then it cannot find the referents that are located further back in the text. However, the percentage of the cases where the correct referent was not in the 3 sentence scope was very low – only 1.5% and therefore the range was not widened. Also, if a candidate is located that far back in the text, its final score is decreased remarkably as it is multiplied by the recency coefficient. So even if the sentences further back are considered, the candidates in them would most likely be dominated by the candidates in more recent sentences.

(18) “Teatasin klubi juhatuse esimehele Jaak Kiikerile, et kui raha juurde ei leita, olen lihtsalt sunnitud vanemate talusse tagasi pöörduma. Tartus viibitud aja eest lubati maksta, seni pole ma saanud sentigi.” “Andsime **talle** hotellitoa”; pareeris Jaak Kiiker.

“I notified the club’s chairman of the board Jaak Kiiker, that if no more money is found, I am forced to return to my parents’ household. I was promised to be paid for the time I spent in Tartu, up till now I haven't gotten even a cent.“ “We gave **him/her** a hotel room”; claimed Jaak Kiiker.

Here the unresolved pronoun is “talle” (“to him/her”), which probably refers to the speaker who expresses his or her opinion in direct speech. However, it is not possible to detect from the two preceding sentences who the speaker is.

6) The correct antecedent is not marked as a noun in the corpus. There were a couple of cases where the correct referent was an abbreviation of a name, for example “TTÜ”, which type was marked in the corpus as “_Y_” and due to that the program did not consider it as a possible candidate. However, the total number of such cases in the whole corpora was too small (only 3 cases) to create a separate rule for them.

7) There is no antecedent at all, because the anaphora is idiomatic (“nii ta on”)

Incorrectly resolved anaphora

One of the issues during the training of the program was the steps to be taken in case of a tie between two or more antecedents. Analysis showed that the most efficient way would be to select the candidate that is located closest to the anaphora compared to other candidates. However, there are still a number of cases where the most recent antecedent is not the correct one. It is something that cannot be resolved with a simple rule, but needs a deeper analysis of the language structure and semantics. The cases were there was a tie between the first candidates did not appear very frequently – in about 9% of all the cases.

Secondly, there were cases where due to the frequency indicator the wrong candidate got higher points than the correct candidate. An example (19) is presented.

(19) Ka siis kahtlustasid tuletõrjujad süütamist. Neljapäeva õhtuks saadi põleng kustutatud, kuid juba reede hommikul kell 6.30 süttisid liiprid(2745) uuesti. **Pritsimeeste** (2748) kannatuste karikas hakkas aga täis saama eile, mil **neil** taas tuli samal aadressil välja sõita.

Also back then the fire fighters suspected arson. By Thursday evening the fire was extinguished, but already on Friday morning at 6:30 the sleepers ignited again. **Firemen’s** patience started to become to an end yesterday when **they** had to drive to the same place again.

```
liiprid (2745): 3.75 (Fre:2.0 Dec:2.0 Name:0.0 Pat:0.0
Dist:1.0 Ind:0.0 Quo:0.0 Pro: 0.0)
Pritsimeeste (2748): 3.0 (Fre:1.0 Dec:0.0 Name:0.0 Pat:0.0
Dist:2.0 Ind:0.0 Quo:0.0 Pro: 0.0)
```

The word “liiprid” (“sleepers”) obtains the highest score, because it appears more times in the text (the frequency range is from one heading to the next) than the second word “Pritsimeeste”. Here the correct antecedent is actually a referent itself – the words “pritsimees” and “tuletõrjuja” are synonyms in Estonian, but using the synonym here reduces the frequency count and as a result the false candidate is preferred over the right one. It is not a rare case in newspaper articles – synonyms, hyponyms and hyperonyms occur very frequently in this kind of texts, these cases of anaphora are called lexical noun phrase anaphora and for resolving them semantic analysis is necessary. For example, the Estonian WordNet TEKsaurus successfully presents the relation between “tuletõrjuja” and “pritsimees”.

Thirdly, the declination indicator may produce wrong answers. It is true, that in Estonian the nominative case indicates that the word is salient and very often a subject. But the subject is not always the referred antecedent. The program described in the current thesis prefers nominative and partitive case which often indicate to subject and object, correspondingly, but there are still 12 other declinations that can also be in the role of antecedents. The examples (20), (21) and (22) describe the cases where the score from preferring nominative case wrongly dominated over the correct candidate in partitive, allative and adessive cases. These were by no means all the occurring cases – there were also sentences where the correct antecedent was for example in genitive, komitative and essive declination.

(20) Näitus(4223) analüüsib inimest(4225) ja seda, kuidas **ta** struktureerib oma maailma, millistest elementidest ta selle kokku paneb, milliseid sümboleid ja kategooriaid selle ülesehitamiseks kasutab.

The exhibition analyses human and that, how (s)he structures his/her own world, from which elements (s)he compiles it, which symbols and categories (s)he uses for constructing it.

```
Näitus (4223): 8.0 (Fre:2.0 Dec:3.0 Name:0.0 Pat:1.0 Dist:2.0
Ind:0.0 Quo:0.0 Pro: 0.0)
inimest (4225): 3.0 (Fre:0.0 Dec:1.0 Name:0.0 Pat:0.0
Dist:2.0 Ind:0.0 Quo:0.0 Pro: 0.0)
```

Here the word “näitus” (“exhibition”) beats the correct antecedent “inimest” (“human”, partitive) in a number of categories like Frequency, Declination and Pattern. It is true that the exhibition is the most salient entity in the article; however it is not the correct antecedent in the current case.

In example (21) the word “Naatsaret” (“Nazareth”) dominates over the word “Maarja” (“Mary”), which is assigned zero points in the Declination category as it is in allative case.

(21) Muudest ajalooallikatest on teada, et just Naatsaret(1442) oli see paik, kuhu kogunesid juudi preestrid, kes valmistusid reisima Jeruusalemma, et teenida seal nädal aega Templis. Niisiis oli juudi kultuuritraditsioonide mõju **Maarjale**(1496) ja **tema** perele üsna suur.

It is known from other historical sources that Nazareth was the place where Jewish priests, who were preparing for a journey to Jerusalem, gathered to serve in the Temple for a week. So the impact of the Jewish culture traditions on **Maarja** and **her** family was rather big.

```
Naatsaret (1442): 3.75 (Fre:2.0 Dec:2.0 Name:0.0 Pat:0.0
Dist:1.0 Ind:0.0 Quo:0.0 Pro: 0.0)
Maarjale (1469): 3.0 (Fre:1.0 Dec:0.0 Name:2.0 Pat:0.0
Dist:0.0 Ind:0.0 Quo:0.0 Pro: 0.0)
```

In example (22) the correct antecedent “naisel” (“woman”, addessive) is dominated over a number of other candidates in various categories and also the correct antecedents is not awarded any points in by the Declination indicator, because it is in “wrong” declination.

(22) Ometigi ei hüljanud Maarja(1719) poega, viibis sündmuspaigal ja kannatas koos jünger Johannesega, kellel Jeesus(1731) palus oma ema eest hoolitseda. See oli tollases ühiskonnas väga oluline, sest **naisel**(1746) ei olnud peaaegu mingisugust iseseisvat positsiooni (1752), kui **tema** kõrval ei

olnud isa, meest või poega.

Though, Maarja did not abandon her son, was on the scene and suffered together with the follower Johannes, whom Jesus had asked to take care of his mother. It was very important in the society of that time, because **woman** had almost no independent position if there was no father, husband or son next to **her**.

```
Jeesus (1731): 5.25 (Fre:2.0 Dec:2.0 Name:2.0 Pat:0.0
Dist:1.0 Ind:0.0 Quo:0.0 Pro: 0.0)
Maarja (1719): 5.25 (Fre:2.0 Dec:2.0 Name:2.0 Pat:0.0
Dist:1.0 Ind:0.0 Quo:0.0 Pro: 0.0)
positsiooni (1752): 4.0 (Fre:1.0 Dec:1.0 Name:0.0 Pat:0.0
Dist:2.0 Ind:0.0 Quo:0.0 Pro: 0.0)
...
naisel (1746): 2.0 (Fre:0.0 Dec:0.0 Name:0.0 Pat:0.0 Dist:2.0
Ind:0.0 Quo:0.0 Pro: 0.0)
```

In addition to the problems already described some more issues occurred during the testing phase. For example, there were cases where the correct antecedent was too far, in the third sentence from the anaphora and therefore multiplied by 0.5 which remarkably reduced its score. Then there were cases where names dominated too much and the actual correct solution was not the name, but some other word that received much lower score than the name.

To sum up the topic, the list of the mentioned sources of errors is once more given:

- Selecting the more recent antecedent in case of a tie does not work
- The frequency indicator boosts the wrong candidate
- The declination indicator boosts the wrong candidate
- The correct antecedent is too far and therefore only half of its actual points are counted
- The name indicator boosts the wrong candidate

As can be seen, there are exceptions in almost every category. At first it might seem that the number of exceptions in one category is low, but if all these cases from different indicators are added together, it already makes up a bigger percentage of erroneous cases. However, changing the scoring of some indicators or adding new rules does not always help, because it seems rather impossible to try and map all the rules about a language, which is a dynamic system, not static. In many of these cases analysis of the word semantics would be helpful. Deeper analysis about the possible solution to those errors is presented in section 6.2.

.

6. Evaluation

The current chapter compares the achieved results to the results of other similar anaphora resolution systems and discusses how the program could be improved.

6.1 Comparison to other systems

Hereby it is useful to compare the Estonian anaphora resolution system to other similar systems that have been implemented earlier. Comparing the systems that are so different - in terms of the target language and the genre of the text in which the anaphora is resolved - might not be a fair case, however it still gives some kind of overview of the field.

As can be seen in the table 5.2.1 below, that kind of systems have been created for many different languages The success rates for the systems that operate on technical manuals seems to be higher than other genres like fiction and newspapers. It is understandable as the technical manuals do not contain very complicated language constructions compared to fiction and newspaper articles. Also, the technical manuals mostly contain only one type of pronominal anaphora that refer to non-animate objects.

System	Language	Genre	Result %	Algorithm
Mitkov 98	English, Polish, Arabic	Technical manuals	89.7 – 95.2	Rule-based (morphology)
MARS (Mitkov et al. 02)	English	Technical manuals	61.55	Rule-based (syntax, morphology)
RAP (Lappin & Leass 94)	English	Technical manuals	86	Rule-based (syntax, morphology)
ARN (Holen 06)	Norwegian	Fiction, newspaper articles	70.5	Based on MARS and RAP
Filippova 05	German	Newspapers	71.6 – 84.2	Rule-based (syntax, morphology)
Trouilleux 02	French	Newspapers	74.8	Rule-based (syntax)
Linh, Žabokrtsky 06	Czech	Newspapers	74.5	Based on RAP and Mitkov 98

Tanev & Mitkov 02	Bulgarian	Tourist guides, technical manuals	72.6 – 75.7	Rule-based (syntax)
Kücük 05	Turkish	Children’s fiction	73.6 – 85.2	Based on MARS

Table 5.2.1. Comparison of different anaphora resolution systems

The anaphora resolution system presented in the current thesis has achieved the results that are comparable to the systems for French (Trouilleux 02), German (Filippova 05), Norwegian (Holen 06) and Czech (Linh, Žabokrtsky 06). The achieved result is not bad at all compared to other systems.

One has to keep in mind that Estonian is an agglutinative language with 14 different declinations which means that every anaphora can occur in text in 14 different forms. Also it is the language of free word order which makes it more difficult to find the salient entities as they can be located basically anywhere in the text compared to the languages like French, English or German. Surprisingly the fact that the pronouns do not denote the gender of the persons whom they are referring was not an issue at all. The cases where there were female and male names both among the candidates did appear maximally in 2% of the cases of the whole corpora. Also, it has to be kept in mind that even if a pronoun gives hints about the gender of the searched antecedent, the same information about noun candidates must be known as well. Hence just knowing the gender of a pronoun will not make the solution process easier.

6.2 Possible solutions and future work

However, there is lot to improve. The erroneous cases and possible solutions to them were discussed in the section 5.2 of the current thesis. There cases where the program could not solve the anaphora can be roughly divided into two categories:

1) The cases where the wrong antecedent is picked, because it is impossible for the system to resolve the anaphora due to the algorithm itself
2) The cases where a wrong antecedent is picked due to the domination of a wrong indicator because of sentence structure and due to the wide range of variations of the language (i.e. more rules are needed to solve certain cases)

To make the non-resolvable anaphora resolvable, more constraints and rules should be implemented and applied to the candidates and to the anaphora themselves. Some of them have also been discussed in (Evans 02):

- For successfully resolving split antecedents the scope of antecedents' search should be widened and/or some more patterns should be added. Wider scope considers not only the candidate noun itself, but also the surroundings of the noun as it may be part of a more complex expression. Patterns might be useful to identify <noun> AND <noun> or similar constructions.
- To relieve the problem with number mismatch between the anaphora and the correct antecedent the number filter should be improved. Currently it eliminates all the candidates that do not match numerically. Penalizing, not removing the candidates of wrong number may help to improve the resolution.
- The cataphora issue can be overlooked at the moment, as it was not the target of the system to resolve forwards referring entities.
- The problem with the pronouns that refer to entities that are neither nouns nor third person personal pronouns can be relieved by widening the range when searching possible candidates. At the moment only nouns and third person personal pronouns are considered as possible candidates, but some more pronoun types can be added to the criteria.

The rest of the error sources like abbreviations as antecedents, too small search range or pleonastic pronouns can be overlooked at the moment, because all these cases occurred only a few times in the corpora and it is not cost-effective to implement additional rules and constraints for resolving them, at least for the current target genre of text (newspapers). Almost all of the errors described above were also described in the follow-up of MARS (Evans 02) – plural disagreement, cataphoric anaphora, idiomatic anaphora, discontinous antecedent consisting of a number of smaller NPs, annotation errors. Evans' solution for improving the systems' performance is suggested as follows:

- Correct the annotation errors
- Modify the enforcement of agreement constraints (gender and number)
- Extend the algorithm's search scope
- Implement the classification of pronouns

Secondly, the possible solutions for improving the wrongly resolved anaphora are proposed:

- Frequency indicator boosts wrong candidate – semantic information could be used to bind synonyms to each other and increase their frequency count
- Declination indicator boosts wrong candidate – more rules involving more cases could be implemented, however, this topic probably needs a deeper research about the declinations
- The name indicator boosts the wrong candidate – the names of persons and names of inanimate objects should be differentiated, but this is not an easy thing to do. Creating lists with names of countries etc. could help a little; however, it is not possible to make a final list of all the names in the world.

7. Summary

The presented Master's Thesis gives an introduction the world of anaphora resolution and gives a detailed overview of the tool implemented for the resolution of third person pronouns in Estonian.

Knowledge-poor anaphora resolution systems have been implemented for a number of languages. However, for Estonian this kind of application had never been made. The basis of the implemented program is Mitkov's knowledge-poor approach that was created for resolving pronominal anaphora in technical manuals in English. The algorithm does not use semantic or deep syntactic knowledge, but operates on the output of a POS-tagger. Later this approach was successfully adapted to Polish and Arabic by Mitkov himself. His example has been followed by a number of successful implementations for other languages by different authors who also built their systems on Mitkov's approach. It was encouraging enough to try and implement a similar system for Estonian, however, with a small number of modifications.

The system for Estonian has been implemented in Java and it uses morphologically and syntactically annotated corpora containing newspaper and scientific articles. At first it searches the third person personal pronouns ("tema/ta" ("(s)he"), "nemad/nad" ("they")) in the text. If a matching pronoun is found, then it locates the possible antecedent candidates that precede the anaphor in 3 sentence range. After that number filter and the antecedent indicators (criteria based on what every candidate is assigned bonus points or penalized by negative points) are applied to the list of candidates. Compared to the original approach some new genre-specific indicators have been added and some of the indicators used in the initial approach have been left out due to the character of the Estonian language or due to the genre-difference. Finally, the candidate with the highest score is proposed as the correct antecedent.

The achieved success rate of the system is as high as 73.6 %. It can be considered satisfactory, as this result is comparable to similar systems on similar target texts implemented for French, German, Czech and Norwegian. A surprising find was that preferring syntactic categories like subject and object did not increase the performance of the system; furthermore, it decreased the results by 0.5 %.

8. Teadmistevaene anafooride lahendamine eestikeelsetes tekstides

Magistritöö (20 AP)

Pilleriin Mutso

Resümee

Käesolev magistritöö tutvustab ühte aktuaalset probleemi tänapäeva loomuliku keele töötluse alal, milleks on anafooride lahendamine. Anafoori võib defineerida kui viidet tekstis (ka kõnes) eelpool mainitule. Üheks tuntuimaks anafooride liigiks on asesõnad, eriti just isikulised asesõnad. Antud magistritöö eesmärgiks ongi luua isikuliste asesõnade lahendaja eestikeelsete ajaleheartiklite jaoks.

Teadmistevaeseid, reeglitel põhinevaid anafooride lahendamise rakendusi on loodud mitmetele keeltele, aga eesti keele jaoks analoogset süsteemi pole varem tehtud. Käsitletava programmi aluseks võetakse Ruslan Mitkovi teadmistevaene algoritm, mis tehti eesmärgiga lahendada isikulisi asesõnu inglisekeelsetes tehnilistes manuaalides. See algoritm ei kasuta semantilist ega põhjalikku süntaktilist infot, vaid töötab tekstil, milles on annoteeritud vaid sõnade morfoloogilised kategooriad. Mainitud algoritmi rakendas Mitkov ise edukalt ka poola ja araabia keelte peal. Tema eeskuju on hiljem järginud mitmed autorid, kes on püüdnud sama algoritmi erinevate keelte jaoks kohandada ja rakendada ning on saavutanud rahuldavaid tulemusi. Need asjaolud julgustasid tegema analoogset rakendust, siiski küll väikeste muudatustega, ka eesti keelele.

Loodud programm on kirjutatud programmeerimiskeeles Java ja seda rakendatakse morfoloogiliselt ja süntaktiliselt märgendatud korpuse peal, mis koosneb eestikeelsetest ajaleheartiklitest ja teaduslikest artiklitest. Tekstist otsitakse asesõnu, mis vastaksid etteantud tingimusele ehk oleksid järgmised 3. isiku isikulised asesõnad: “tema”, “ta”, “nemad”, “nad”. Kui tekstist leitakse vastav asesõna, siis järgmiseks sammuks on üles otsida tema võimalikud lahendikandidaadid – need nimi- ja asesõnad, mis asuvad tekstis anafoorist eespool. Lahendikandidaate otsitakse 3-lauselises skoobis – anafoori sisaldavast lausest, eelmisest ja üleelmisest lausest. Peale kõikide lahendikandidaatide kaardistamist rakendatakse neile arvufiltrit ja kaheksat indikaatorit. Indikaatoriteks on erinevad

kriteeriumid, mille põhjal antakse igale kandidaadile plusspunkte või võetakse punkte maha. Näiteks antakse kandidaadile boonuspunkte, kui tegemist on nimega või tekstis sagedasti esineva sõnaga. Võrreldes originaalalgoritmiga on eestikeelses programmis mõned indikaatorid välja jäetud ja mõned uued lisatud, tulenevalt keele eripärast ja tekstide žanri eripärast. Viimase sammuna pakub programm välja kõrgeima punktisumma saavutanud kandidaadi kui sobivaima lahendiks anafoorile.

Tulemuste analüüs näitab, et programm suudab ära lahendada 73.6% kõikidest nendest anafooridest, mis on talle lahendamiseks jõukohased. Selle tulemusega võib rahule jääda, sest sarnaseid tulemusi saavutasid ka teiste keelte, prantsuse, saksa, norra ja tšehhi keele jaoks implementeeritud isikuliste anafooride lahendamise programmid. Üllatavaks tulemuseks oli see, et süntaktilise info (alus, sihitis) rakendamine ei parandanud programmi efektiivsust, vaid vastupidi – vähendas seda umbes 0.5% võrra.

Reference

(Bobrow 64): Daniel Bobrow. Natural Language Input for a Computer Problem Solving System. 1964.

(Brennan et al. 87): Susan Brennan, Marylin Friedman, Carl Pollard. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting of the ACL (ACL'87), 155-162. Stanford, CA,USA*, 1987.

(Britannica): Britannica Online Encyclopedia. http://www.britannica.com/ (last visited 25.05.2008)

(DAARC 07): 6th Discourse Anaphora and Anaphor Resolution Colloquium. March 29-30, 2007, Lagos (Algarve), Portugal, hosted by the University of Lisbon, Faculty of Sciences. http://daarc2007.di.fc.ul.pt/ (last visited 24.05.2008).

(Erelt 03): Estonian language / [Estonian Academy of Sciences]; edited by Mati Erelt. Tallinn, Estonian Academy Publishers, 2003.

(Evans 02): Richard Evans. Refined salience weighting and error analysis in anaphora resolution In *Proceedings of International Symposium on Reference Resolution for Natural Language Processing 2002, Alicante, Spain*.

(Filippova 05): Katja Filippova. A Memory-Based Learning Approach to Pronominal Anaphora Resolution in German Newspaper Texts. Master’s Thesis in Computational Linguistics, University of Tübingen, 2005.

(Grosz et al. 95): Barbara Grosz, Joshi Aravind, Scott Weinstein. 1995. "Centering: a framework for modelling the local coherence of discourse". Computational Linguistics, 21(2), 44-50.

(Hirst 81): Graeme Hirst. Anaphora in natural language understanding. Berlin Springer Verlag, 1981.

(Hobbs 76): Jerry Hobbs. Pronoun Resolution. City University of New York, 1976.

(Holen 06): Gordana Ilic Holen. Automatic Anaphora Resolution for Norwegian (ARN). Thesis submitted for the degree of Candidata Philologiae, University of Oslo, 2006.

(K&B 96): Christopher Kennedy, Branimir Boguraev. Anaphora for Everyone: Pronominal Anaphora Resolution without a Parser. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING'96), 113-118, Copenhagen, Denmark*. 1996

(Küçük 05): Dilek Küçük. A knowledge-poor pronoun resolution system for Turkish. A thesis submitted to the Graduate School of Natural and Applied Sciences of Middle East Technical University, 2005.

(Lappin&Leass 94): Shalom Lappin, Herbert Leass. An algorithm for Pronominal Anaphora Resolution. In *Computational Linguistics 20(4), 535–561*. 1994.

(Linh, Žabokrtský 07): Nguy Giang Linh, Zdenek Žabokrtský. Rule-based approach to pronominal anaphora resolution applied on the Prague Dependency Treebank 2.0 data. In *Proceedings of DAARC 2007 (6th Discourse Anaphora and Anaphor Resolution Colloquium).*

(MDC) Morphologically disambiguated corpus.
http://www.cl.ut.ee/korpused/morfkorpus/index.php?lang=en (last visited 26.05.08)

(Mitkov 98): Ruslan Mitkov. Robust pronoun resolution with limited knowledge. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING'98)/ACL'98 Conference,* pages 869–875. 1998.

(Mitkov 99): Ruslan Mitkov. Anaphora resolution: the state of the art. Working paper (based on the *COLING'98/ACL'98 tutorial on anaphora resolution*), University of Wolverhampton, Wolverhampton. 1999.

(Mitkov 01) Ruslan Mitkov. Outstanding issues in anaphora resolution. In *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*. Springer-Verlag. 2001.

(Mitkov et al. 02): Ruslan Mitkov, Richard Evans, Constantin Orasan. A new, fully automatic version of Mitkov's knowledge-poor pronoun resolution method. In *Proceedings of CICLing-2000, Mexico City, Mexico*. 2002.

(Palomar et al. 01): Manuel Palomar, Lidia Moreno, Jesus Peral, Rafael Munoz, Antonio Ferrandez, Patricio Martinez-Barco, Maximiliano Saiz-Noeda. An algorithm for anaphora resolution in Spanish texts. Computational Linguistics, 27(4), 545-567. 2001.

(RGCLUT) HTML entities. http://www.cl.ut.ee/abi/olemid.html.en (last visited 26.05.08)

(SHRDLU): http://hci.stanford.edu/~winograd/shrdlu/ (last visited 25.05.2008)

(SIL) http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/ (last visited 26.05.08)

(Tanev, Mitkov 02): Hristo Tanev, Ruslan Mitkov. Shallow language processing architecture for Bulgarian. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*. 2002.

(Trouilleux 02): Francois Trouilleux. A Rule-based Pronoun Resolution System for French. In *Proceedings of the 4th Discourse Anaphora and Anaphora Resolution Colloquium (DAARC'02).*

(Pajusalu 97): Renate Pajusalu. Eesti pronoomeneid I. Ühiskeele see, too ja tema/ta. - Keel ja Kirjandus 1997, lk 24-30 ja 106-115.

(Pajusalu 99): Renate Pajusalu. Deiktikud eesti keeles. Dissertationes Philologiae Estonicae Universitatis Tartuensis 8. Tartu, 1999.

(Pajusalu, Laury 05): Renate Pajusalu, Ritva Laury. Anaphoric pronouns in Spoken Estonian: crossing the paradigms. Studia Fennica, SKS, Helsinki, lk 107-134. 2005