25

Stockholm University
Department of Linguistics
Computational Linguistics

Towards automatic grammatical simplification of Swedish text

Anna Decker

Abstract

This thesis presents a pilot study which endeavors to explore a sample of a manually simplified text in order to determine if there exist consistent patterns within the simplifications which may be used in creating an automatic text simplification system. There are, as of yet, no automatic text simplifiers for Swedish, so current simplification, for whichever purpose, is done manually. This process is long and laborious, and it is imprecise – there is no guarantee that the simplified text is simple, other than the innate feeling of simplicity in the person who makes the simplification. If text simplification is automated, an audience who earlier had very limited access to reading material would be able to read a wider array of texts, and language learners could access simple versions of the canon of literature of the language they study. To determine how transformation rules for automatic simplification may be formed, this thesis investigates the grammatical properties of manual text simplification. More specifically, the differences in phrase structure between an original text written in standard Swedish, and a manually simplified version of the same text, are compared with the intention to extract and formalize these differences. The study showed that manual simplifications in the material were arbitrary and ad hoc. It was, however, possible to extract rules from them. The study resulted in 25 formalized simplification rules, designed to easily fit into any simplification program.

Master's Degree Thesis in Computational Linguistics
Supervisor: Lars Borin
March 2003

Abstrakt:

I denna uppsats presenteras en preliminär studie som har undersökt en manuellt förenklad text för att utröna ifall det finns regelbundenheter i förenklingarna som kan utnyttjas för att skapa ett automatiskt textförenklingssystem. Det finns inga automatiska textförenklingssystem för svenska, så textförenkling sker manuellt. Manuell textförenkling är en lång och arbetsam process och den är oprecis - det finns inga garantier för att den förenklade texten faktiskt är enklare, annat än den mänskliga förenklarens intuition för "enkelt språk". Om textförenklingen bleve automatiserad skulle målgrupper som tidigare haft väldigt begränsad tillgång till läsmaterial få tillgång till stora mängder text med stor tematisk spännvidd, och andraspråksinlärare skulle få tillgång till en stor mängd av förenklade versioner av litterära verk på målspråket. Att utveckla automatisk textförenkling är särskilt viktigt för ett litet språk som svenska, där marknaden är för liten för att kunna stödja en stor utgivning av manuellt förenklad text. För att avgöra hur transformationsregler för automatisk textförenkling skulle kunna utformas undersöker denna uppsats hur manuell grammatisk förenkling har gjorts. Skillnader i frasstrukturen mellan en orginaltext skriven på standardsvenska och en manuellt förenklad version av samma text jämförs, med avsikt att utvinna och formalisera dessa. De manuella förenklingarna i materialet visade sig vara flyktiga och otillräkneliga. Det var dock möjligt att utvinna regler ur dem. Den lingvistiska analysen av de grammatiska skillnaderna ledde till 25 förenklingsregler, som utformades för att lätt passa in i ett förenklingsprogram.

1 INTRODUCTION 7

2 BACKGROUND 8

2.1 SIMPLIFIED TEXT 8
2.1.1 *What is simplification?* 8
2.2 AUTOMATIC TEXT SIMPLIFICATION 8
2.2.1 *Simplification for NLP-purposes* 9
2.2.2 *PSET* 10
2.2.3 *Unresolved issues in current state-of-the-art text simplifiers* 11
2.3 SIMPLIFIED SWEDISH 11
2.4 READABILITY 12
2.4.1 *Readability research* 12
2.4.2 *Readability of Swedish text* 13
2.4.3 *Measuring readability* 13
2.4.4 *Critique of readability formulae* 15
2.5 SECOND LANGUAGE LEARNING 15
2.5.1 *Swedish as a second language* 16
2.5.2 *The second language learner* 16
2.6 SUMMARY 17

3 MATERIAL 18

3.1 INVANDRARTIDNINGEN 18
3.2 PÅ LÄTT SVENSKA 18
3.2.1 *Guidelines for simplification* 18
3.2.2 *Difficulty degrees* 20
3.2.3 *Word lists* 20
3.3 THE ARTICLES 20
3.3.1 *General statistics* 21
3.3.2 *Lix values for the articles* 21
3.3.3 *Lix values for the difficulty degrees* 23
3.4 SUMMARY 24

4 METHOD 25

4.1 ASSEMBLING THE MATERIAL 26
4.1.1 *Criteria for the selection of articles* 26
4.1.2 *Calculating lix* 26
4.1.3 *Extracting the simplifications* 26
4.2 ANALYZING THE PHRASES 27
4.2.1 *Establishing a typology* 27

5 LINGUISTIC ANALYSIS 28

5.1 ABOUT THE SIMPLIFIED PHRASES 28
5.2 TRANSFORMATIONS ON THE LEXICAL LEVEL 28
5.2.1 *Semantic relations* 28
5.3 TRANSFORMATIONS ON THE PHRASAL LEVEL 29
5.3.1 *Simplifications of the noun phrase* 29
5.3.2 *Pronouns* 32
5.3.3 *Simplification of the adjective phrase* 32
5.3.4 *Simplification of the adverb* 33
5.3.5 *Simplification of the prepositional phrase* 33
5.3.6 *Simplification of the verb phrase* 34
5.3.7 *The conjunction* 37

5.3.8 The subjunction 37
5.3.9 The anticipatory det 37
5.4 SIMPLIFICATION ABOVE PHRASE LEVEL 38

6 ANALYSIS 39

6.1 REVIEWING THE SIMPLIFICATIONS 39
6.1.1 Trends in simplification 39
6.1.2 Types of simplifications 39
6.2 THE SIMPLIFICATION RULES 40
6.2.1 Simplification rules for the noun phrase 40
6.2.2 Simplification rules for the adjective phrase 43
6.2.3 Simplification rules for the adverb 44
6.2.4 Simplification rules for the prepositional phrase 44
6.2.5 Simplification rules for the verb phrase 45
6.2.6 Simplification rules for the conjunction and subjunction 46
6.2.7 Simplification rules for the anticipatory det 46
6.3 RULES FOR AUTOMATIC SIMPLIFICATION 46
6.3.1 Rule order 46
6.3.2 Tentative simplification rules 47

7 DISCUSSION 48

7.1 FURTHER RESEARCH 49
7.1.1 Points of interest 49
7.2 CONCLUDING REMARKS 50

8 REFERENCES 51

9 APPENDIX 56

Table of Figures

Figure 2.1 The guidelines that *Centrum för lättläst* employs for writing easy texts. 12
Figure 2.2 The formula for calculating *lix*. 14
Figure 2.3 A general interpreter for results of the readability formula *lix*. 14
Figure 3.1 The original *På lätt svenska* version 19
Figure 3.2 Guidelines for writing and translating easy texts for L2 readers following Wiman (1998). 19
Figure 3.3 The lix values for the articles in *Invandrartidningen*. 22
Figure 3.4 The lix values for the articles in *På lätt svenska*. 22
Figure 3.5 The lix values for *På lätt svenska* and *Invandrartidningen* compared. 23
Figure 4.1 Example of the original *Invandrartidningen*. 25
Figure 6.1 Categories of simplifications found in the simplification of *Invandrartidningen*. 40
Figure 6.2 Simplification rules derived from the simplifications made of *Invandrartidningen*. 47

Table of Tables

Table 3.1 Statistics of the difficulty degrees of the articles in *På lätt svenska*. 21
Table 3.2 The data for computing *lix*, and lix values for both *Invandrartidningen* and *På lätt svenska*. 21
Table 3.3 The average lix values for the difficulty degrees in *På lätt svenska* and *Invandrartidningen*. 24
Table 5.1 Types of simplification of an NP consisting of a head noun. 29
Table 5.2 Types of simplifications of constituents preceding the NP head. 30
Table 5.3 Types of simplifications of constituents succeeding the NP head. 30
Table 5.4 Types of simplifications of the compound noun. 31
Table 5.5 Types of simplifications of the personal name. 31
Table 5.6 Types of simplifications of the numeral. 31
Table 5.7 Types of simplifications of the possessive. 32
Table 5.8 Types of simplifications of the adjective phrase. 32
Table 5.9 Types of simplifications of the adverb. 33
Table 5.10 Types of simplifications of the prepositional phrase. 33
Table 5.11 Simplifications of the auxiliary verb. 34
Table 5.12 Simplifications of the copula verb. 35
Table 5.13 Simplifications of the passive verb. 35
Table 5.14 Simplifications of the particle verb. 36
Table 5.15 Simplifications of verb tense. 36
Table 5.16 Simplifications of the conjunction. 37
Table 5.17 Simplifications of the subjunction. 37
Table 5.18 Simplification of the anticipatory it. 37

1 Introduction

There is little agreement among linguists on how to define text simplicity, or that simple texts exist at all. Some argue that it is the reduction of syntactic difficulty which makes something simpler, others champion the claim that language can be made simpler by paraphrasing or adding linguistic information, while still others suggest that simplification occurs when it is believed that the addressee is unable to decipher the standard variety of the language.

The bulk of the simplified texts available to the second language[1] (L2) learner today consist of school book texts and easy readers. Neither of these text types represent the language as a whole, as they mirror the individual author's instincts of what constitutes easy language. Moreover, most easy readers and simplified texts are written or adapted for the mentally disabled, aphasiacs or school children, and not for neurologically sound adults. This thesis focuses on learners of a second language, but many groups could benefit from simplified texts.

An automatized method for text simplification—and its extension, a computer program for automatic text simplification—would be useful tools for creating simplified texts. Other possible applications of text simplification might be in machine translation, information extraction and retrieval, and summarization, where a simplified text could circumvent a particular parsing problem which stems from complex syntactic structures. Text simplification could also be used to adapt texts for applications with limited space to display the text, for instance in mobile phones and PDA's, and for simplifying text on the internet in real time.

What little research on automatic text simplification exists today is, on the whole, sparse, most often tailored for the English language, and frequently geared towards very specific and limited purposes. Research into automatic text simplification for Swedish, among other languages, is lacking at the present time. It is particularly important that this line of research be further pursued for a language such as Swedish, as it has a small target audience and produces a low number of original simple texts. The development of a fully functioning automated text simplification system could drastically increase the number of simplified texts available for, among others, L2 learners.

This thesis endeavors to review and discuss text simplification of Swedish, it analyzes human-written simplifications from a weekly newspaper for immigrant readers, it offers a typology of these simplifications, and it presents a tentative set of formalized rules for text simplification of Swedish. To do this, the study for this thesis focuses on a small sample of a parallel corpus to investigate the feasibility of comparing a manually simplified text with a corresponding original text to extract simplification rules. The readability of the texts are calculated to establish that the simplifications are, indeed, grammatically simpler than the original texts. The rules discovered in this pilot study may be further developed for use in future automatic text simplification applications.

[1] It is customary to distinguish between *second language learning*, which takes place in the country of the language being learned, for instance immigrants learning Swedish in Sweden, and *foreign language learning*, which takes place in a (formal) setting away from the country of the language being learned, for instance students studying Swedish in a foreign country.

2 Background

To reach an understanding of what text simplification for second language learning of Swedish entails, we must first investigate text simplification and simplified text, and second language learning. An introduction to readability of texts is also provided.

2.1 Simplified text

Research on simplified Swedish texts is scarce. The majority of these studies have been conducted either on school book texts (e.g. Reichenberg, 2000; Ryhänen, 1987), or on texts written for language impaired readers or readers with a mental disability (*LL-stiftelsen*).

2.1.1 What is simplification?

Simple language is used in a variety of instances: in communication with L2 learners or language impaired persons, in child-directed speech, in telegrams, in sms-ing, and in the controlled languages used within, for example, the aerospace industry.

Lucas (1991) suggests five main types of simplification:

Lexico-semantic simplification	Simplification of the individual words and phrases that the simplifier defines as possible problems for the addressee.
Cultural simplification	Simplification explaining cultural presuppositions and inferences the writer makes which the addressee may find difficult or be unaware of.
Dialectal and idiolectal simplification	Simplification of features of the language which deviate from the standard variety of the language.
Grammatical simplification	Simplification of syntactic structures deemed to posit difficulties for the addressee.
Textual simplification	Simplification of textual properties such as deixis, ellipsis, and other cohesive features of the language.

Of these five types, the investigation in this thesis focuses on grammatical simplification. The grammatical simplifications in a text may be discovered in the comparison of an original text and an altered version of that text.

In this thesis, simplification of text is defined as: *simplification is the omission of certain grammatical features used in the source text, in the creation of the target text, while retaining the relative meaning of the source phrase.*

2.2 Automatic text simplification

Automatic text simplification is the process where simplification of text is automated and performed by a computer system or program. The amount of research conducted on automatic simplification, to date, has been insufficient; only a few large projects have been undertaken and focus on simplification of English.

The study of automatic text simplification is not old—the first research papers on the subject were published in the middle of the 1990s (Chandrasekar, 1994; Chandrasekar & Suresh, 1995; Chandrasekar, Doran, & Srinivas, 1996). In NLP terms, one can say that

simplification can be found somewhere inbetween summarization (McKeown et al., 1995; Knight & Marcu, 2000), abstracting (Endres-Miggemeyer et al., 1995), and aggregation (Dalianis, 1999; Reape & Mellish, 1999), with one important distinction – these areas all try to make a text shorter to some degree. This is not the case with simplification. In simplification, the aim is to make text easier. However, a simplified text may actually be longer than the original, as simplification rules may contradict one another; one rule reduces sentence length, while another upgrades a subordinate clause, making the resulting sentence longer.

2.2.1 *Simplification for NLP-purposes*

Some research of automatic simplification is related to the so called controlled languages, such as the AECMA Simplified English,[2] the Boeing Simplified English Checker,[3] and ScaniaSwedish (Almqvist & Sågvall Hein, 1996). These controlled languages are constructed to provide a vocabulary, and in some cases also a grammar, which follow a defined set of linguistic principles in order to facilitate communication and understanding of written language. This practice is found in industries with employees of many nationalities and in companies that conduct business globally. In many of these companies and industries work is under way to automate the process of creating simplified documentation.

Chandrasekar & Srinivas

A large portion of the research found in the field of automatic simplification for NLP has been carried out by Raman Chandrasekar and Bangalore Srinivas. In 1997, these gentlemen proposed to "develop a system to (semi-)automatically simplify text from any domain" (Chandrasekar & Srinivas, 1997:184) for the use in natural language technologies. Their motivation is to minimize parsing problems, which arise in complex syntactic structures and ambiguity in the input. Theirs is a two-stage system of simplification: analysis and transformation, which operates on one sentence at a time (Chandrasekar et al., 1996). The first module analyzes the input text to find dependencies between lexemes and their associated tree structures, dubbed supertags (Srinivas & Joshi, 1998).[4] The second module recognizes and extracts the components, i.e. the supertags, that can be simplified, and transforms the text into a simplified version. These simplification rules are applied recursively until there are no further possible simplifications.

In the continued efforts to develop their simplification system, Chandrasekar and Srinivas (1997) made an attempt to induce simplification rules. The training data consisted of an aligned corpus that links complex sentences with the corresponding simplified sentences. The data was analyzed, simplified, and finally generalized. This simplification technique focused on relative clauses, and turned out to be quite robust (Chandrasekar & Srinivas, 1997).

Reluctant Paraphrasing

By what he dubs 'Reluctant Paraphrasing', Dras (1999) is transforming sentences from one form to another, using syntactic rules for eliminating superfluous information. The goal of

---

[2] *Association Européenne des Constructeurs de Matérial Aérospatial*, in English *the European Association of Aerospace Industries*. `http://www.aecma.org/Publications/SEnglish/sengbrc.htm`

[3] `http://www.boeing.com/assocproducts/sechecker/se.html`

[4] The supertags are reminiscent of the tree notation widely used in linguistics, and do, in fact, allude to the constituent structure of the phrase the lexeme is a part of.

the system is to let the user adjust the level, to which s/he wants to transform a document, with the overriding goal being to paraphrase a long document, let us say, this thesis, into a shorter document, perhaps an eight-page article ready for submission to a text simplification conference. This system presupposes that the original document is spotless,[5] and that any change to it would be undesirable, and therefore be as minimal as possible.

The reluctant paraphrase system has three components. First, a set of paraphrase techniques, used to perform the actual paraphrase. Second, a set of text constraints that the paraphrase is obliged to obey, and, third, the effect of the paraphrase, which needs to be minimized.

The core mechanisms of reluctant paraphrasing are similar to those used in automatic simplification, but such simplifications of text go further. Automatic simplification does not only control the syntactic content to limit a text's length, but also controls the text's readability.

Other studies useful for simplification

Later work on simplification for NLP has been carried out by mister Advaith Siddhartan. He follows the path set out by Chandrasekar and Srinivas, but focuses more on the problems which arise on the discourse level of a simplification and the possible solutions of these problems (Siddhartan, 2002).

Daniel Marcu has conducted research in a similar vein, and resorts to using Mann and Thompson's RHETORICAL STRUCTURE THEORY (RST) (Mann & Thompson, 1988) to establish the clause structure. Marcu uses this theory to establish the rhetorical structure, which is akin to the discourse structure, of text (Marcu, 1997; 2000). By using the status of the various structures in RST, Marcu has found a way of eliminating the rhetorically less important phrases from the source text, and thereby simplifies it.

2.2.2 PSET

PSet (Practical Simplification of English Text) is a system for text simplification of English for language impaired readers, mainly aphasiacs (Canning et al., 2000). PSET takes as its input a newspaper published on the web, and runs it through a modular simplification system, which has two main modules: an analyzer module and a simplification module (Devlin et al., 1999). The analyzer determines the part-of-speech for the input text and tags it accordingly. Then, a morphological analyzer conducts inflectional analysis on the text. The analyzer module produces output, in the form of a shallow parse of the text, which then functions as input for the simplification module.

SYSTAR

The second module, which initializes after the analyzer, is the simplifier. The simplifier processes the text within its own two modules: first a syntactic simplifier, and then a lexical simplifier. The syntactic part of the module is called SYSTAR—SYntactic Simplification of Text for Aphasic Readers (Canning & Tait, 1999)—and handles anaphora resolution, syntactic simplification, and anaphora replacement. The syntactic simplification is divided into two stages, 1) unification pattern matching on the input, applied recursively until no further simplifications are possible, and 2) generation of a correct form of the simplified sentence (Canning & Tait, 1999). SYSTAR applies three syntactic rules: it transforms passive voice to active, it simplifies coordinated clauses into

[5] A presumption I do not entertain for this thesis.

separate sentences, and it upgrades subordinate clauses to main clauses by transforming them into separate sentences.

The anaphora replacement module resolves non-coreferring deictic pronouns, and replaces them with their referring expressions. The lexical simplifier replaces specialized and uncommon words with a more frequent synonym. The simplifier then selects the most frequent synonym, by cross-referencing all content words against WordNet[6] and the Oxford Psycholinguistic Database.[7] Post-processing, that is, generation of correct output, consists of morphological and orthographic adjustment.

Evaluation of SYSTAR has shown that it does retain both meaning and cohesion after simplification, but that it does not yet perform at a level where it can be made available to the public (Canning et al., 2000).

2.2.3 *Unresolved issues in current state-of-the-art text simplifiers*

Certain issues in today's text simplification systems have been identified, and are still unresolved (Canning et al., 2000; Canning & Tait, 1999; Chandrasekar & Srinivas, 1997). The order in which the simplified sentences are presented in the output must be established to maintain textual coherence in the output. There must exist rules for which referring expressions the simplifier should choose when resolving pronouns etc., and the issue of how to divine which gap-filling expressions are suitable for simplification of e.g. ellipsis needs to be solved.

2.3 Simplified Swedish

Simplified Swedish for language impaired readers has received much attention since the sixties, after a ruling in the parliament to found and subsidize a publishing company of easy reading texts, the predecessor to *LL-förlaget* (Easy Reader Publishing). This publisher evolved into *LL-stiftelsen* (The Easy Reading Foundation), which today is an organization which encompasses *LL-förlaget*, the biggest publisher of easy reader books in Swedish, and *Centrum för lättläst* (The Center for Easy Reading), a resource center for creating readable text. *LL-stiftelsen* is, however, targeting language impaired readers and readers with a mental disability, and not immigrant learners of the language. The guidelines for writing easy texts used at *LL-stiftelsen* (Wiman, 1998) are presented in Figure 2.1 below.

These guidelines include straight-forward rules for writing simple texts, as well as notes on what to be aware of when writing for language impaired readers. Additionally, there are rules about how to use typography and layout to create a readable document, for instance, it is recommend that writers start each new sentence on a new line, and use a type face of a larger size than normal.

[6] http://www.cogsci.princeton.edu/~wn

[7] ftp://ota.ox.ac.uk/pub/ota/public/dicts/1054/

- Write short.
  - Have something to say—say it—say nothing more.
  - Some readers have trouble distinguishing what is foreground and what is background information in the text.
- Write from the beginning to the end.
  - The actions in the text should be in chronological order without jumping back and forth.
- Write with simple words.
  - Do not use words that are hard to understand from either a morphological, etymological, or symbolic point of view.
  - Replace a hard word with a simpler more common synonym, and if you still need to use the hard word—explain it.
- Write without imagery.
  - Describe things as concretely as possible.
  - Avoid clichés.
  - Many readers interpret the text literally.
- Write suspiciously.
  - Metaphors and similes may have a concrete meaning as well as an abstract meaning.
- Write the same word.
  - Do not vary your vocabulary.
  - The reader may not understand coreferring expressions.
- Write without unnecessary numbers.
  - Many readers lack the intuitive sense of numbers and measures.
  - If you need to write numbers—explain and make concrete.
- Write without times.
  - Many readers just make the distinction between 'then' and 'now'.
- Write direct.
  - Make no implications.
- Write active.
  - 'who does what' instead of 'who did what to whom' or 'what was done by whom'.
- Write main clauses.
  - Use as few subordinate clauses as possible.

Figure 2.1 The guidelines that *Centrum för lättläst* employs for writing easy texts (my translation).

2.4 Readability

Readability is a word and a concept with a wide range of definitions. For the purposes of this thesis, readability is defined, following Björnsson (1968), as: *A measure of the lexical and syntactic complexity of a text, which makes it more or less accessible to the reader.* This definition presupposes that readability does not equate to "easy" or "hard" to read. Instead, each text with a certain syntactic complexity may be perceived as either "easy" or "hard" by each reader.[8]

2.4.1 *Readability research*

The first attempts to find out why certain texts are considered hard or easy to understand are found within early religious scholarship, but it was not until the late 19th century that research on readability developed into a full-fledged research area of its own. This

[8] Given that the text is coherent and well formed.

development occurred first and foremost in the United States, but also in Germany and Russia (Klare, 1963). This early research was initiated due to observations that school children had trouble reading texts with many unusual words. It had a pedagogical goal: to establish guidelines for a reading level corresponding to scholastic grade level. Eventually, other groups of professionals became interested in measuring and anticipating readability, e.g. journalists, publishers, and librarians.

The first readability formula for American English was published in the early 1920's, and there now exist some forty readability formulae for English, chiefly American (Klare, 1963).

Early in its development, readability research gravitated towards finding quick and stable methods for computing readability, which mostly resulted in counting easily identifiable linguistic elements rather than delving into the oblique reasons of the differences of the readability of texts. These 'surface'-methods can accurately predict how readable a text is to the average reader (Björnsson, 1968), but it was the emerge of psycholinguistics in the 1950's that plunged the readability research community into a whole new line of exploration; to gain further insight into why the 'hard' text is perceived as such. This research became largely based on the generative grammar of Chomsky and soon developed into research on the intelligibility of a text, rather than its readability.

2.4.2 *Readability of Swedish text*

In Sweden, readability research developed in the 1960's at *Stockholms skoldirektion* (the Stockholm County School Board) with the establishment of the nine-year elementary school in 1962. The first readability formula for Swedish was published in 1962/63 but the most widely used formula is lix, *läsbarhetsindex* (readability index), which was developed by C-H Björnsson and is presented in *Läsbarhet* (Björnsson, 1968). Lix was also the first actual attempt to apply methods developed by American readability research onto the Swedish language.

A number of studies have researched readability in Swedish texts. Most of these studies have been of texts written for school books (cf. Liberg, 2001; Reichenberg, 2000; Ryhänen, 1987), but there also exist studies of other types of text, for instance Platzack's studies (1974*a,b*) about the psycholinguistics underlying the grammatical surface structures of texts, and Gunnarsson's study (1982), which investigates the intelligibility of Swedish law texts. Other disciplines which draw upon readability studies are mass media and journalism, which focus primarily on the information content and structure of a text, above the paragraph level, rather than on grammar and lexicon (Oestreicher, 2000).

2.4.3 *Measuring readability*

There are many methods available to measure readability, but the most often utilized method is to use readability formulae. They are fast and easy to compute, and they return a tangible value of the text, usually in the form of a numeric. Readability formulae are normally calculated from visible and measurable factors in text, such as word and sentence parameters, but there have also been attempts to incorporate semantic complexity, lexical density and variation, and even discourse structure, into the readability formulae (Klare, 1963).

The most reliable readability formulae combine a word parameter, which is a measure of the lexical complexity of the text, with a sentence parameter, which reflects the syntactic complexity of the text. Most constructors of readability formulae agree that variables of the

formula should be:

stable	i.e. they should appear throughout the text.
easy to count	i.e. they should be easily distinguishable.
unambiguous	i.e. different persons should have to count them in the same way.

lix

Lix, *läsbarhetsindex* (readability index), was developed at the request of the Stockholm County School Board (Björnsson, 1968), which wanted to develop a tool for measuring readability. This tool would be used within the elementary school system to select appropriate literature for each grade level.

$$lix = \frac{\text{number of words}}{\text{number of sentences}} + \frac{\text{number of long words}}{\text{number of words}} * 100$$

Figure 2.2 The formula for calculating *lix*.

Lix is computed by counting the number of words, the number of long words, and the number of sentences in a text (see Figure 2.2). From these three variables, one computes the mean length of a sentence and adds to that the percentage of long words[9] in the text. The lix value usually falls within a range of 20–60, and is measured in integers.

The lix value itself does not convey the presumed difficulty of a text. For the lix value to be an effective tool, it needs interpretation, and for this purpose, lix interpreters have been developed. The most functional lix interpreter for Swedish is presented in Figure 2.3. In spite of its linguistically non-intuitive basis of calculation, lix is an accurate measure of readability. Björnsson (Björnsson, 1968; Björnsson & Hård af Segerstad, 1979) has calculated the lix values for a vast amount of Swedish prose and technical texts, and reports that his calculated lix values coincide with his test subjects' rankings of the same texts on the same scale, with an accuracy of 80–90% (Björnsson, 1968).

very easy text	25		
		30	children's and young reader's books
easy text	35		
		40	fiction
average text	45		
		50	non-fiction literature
hard text	55		
		60	technical literature
very hard text	65		

Figure 2.3 A general interpreter for results of the readability formula *lix*.

There exist several readability formulae for Swedish, other than lix. For a review of these, see Cedergren (1992).

[9] Long words are any words of more than six letters. Six letters was a good delineation point, since this is the point which maximizes the difference in percentage of long words between two texts with different readability (Björnsson, 1968:215). This decision produces the bizarre consequences that *svensk* (Swedish) is an easy word, but *Sverige* (Sweden) is hard, and that the surname *Person*, spelled with one *s*, is easy and *Persson* with two *s*'s is hard.

2.4.4 *Critique of readability formulae*

Heavy criticism has been leveled towards the usage, and the very notion, of readability formulae (Fulcher, 1997; Cedergren, 1992; Källgren, 1979; Platzack, 1974*b*). A readability formula does not give an assessment of how hard or easy a text is to read, but gives a rough estimate of the syntactic complexity of a text. This estimate is rough because a writer can construct a text with a low lix value, for instance, by writing only short words even if s/he uses a complex syntactic structure.

Many researchers share the view that text is too complex to fit into a scale, like the readability formulae do, and that the readability of a text cannot be reduced to a numeric value (Cedergren, 1992; Platzack, 1974*b*).

Another criticism of readability formulae (Källgren, 1979; Platzack, 1974*a,b*) is of their being shallow measurements of surface phenomena, and having little or nothing to do with the underlying, psycholinguistic factors that really constitute the text. From this point of view, the readability formulae are constructs which do not tell us anything about the actual readability of the text, i.e. the deep structure, but concentrate on the cosmetics, i.e. the surface structure.

In all fairness, Björnsson (1968) anticipated much of this criticism, and he stated that the readability formulae had to be used carefully as a tool, and not as a means in themselves. He also acknowledged that lix is a shallow formula which utilizes surface parameters, but, since lix correlates with perceived difficulty, it would still be usable as an estimate of text difficulty (Björnsson, 1968).

In the present day, a whole new type of criticism has seen the light—that the readability formulae are old, and that they might be dated. Should it not, with the knowledge collected during these past 40-something years, be possible to incoorporate something new into these formulae? As Fulcher writes:

> "It is time to ask whether tools which were developed 40–50 years ago can continue to serve the purpose of helping us choose texts wisely, or whether we need to develop a new awareness of the complexity of text and the nature of reading." (Fulcher, 1997:510)

2.5 Second language learning

Second language acquisition (SLA) is a vast research area, which spans many fields within linguistics, such as psycholinguistics, generative linguistics, sociolinguistics, cognitive linguistics, and phonetics and phonology. Other disciplines, related to linguistics, such as sociology, psychology, educational research, and language pedagogy, also research SLA. For additional information on SLA, see Ellis (1997), Gass (1997), Granberg (2001), Hammarberg & Viberg (1984), Larsen-Freeman & Long (1991), and Plaza Pust (2000).

Researchers often differentiate between child SLA and adult SLA, because of the different ways that both groups learn. One of the main questions facing modern SLA researchers has been whether or not adult L2 learners benefit from universal grammar (UG),[10] just as L1 learners and child L2 learners are thought to do. It is believed that L1 learners employ a modular way of learning, whereas adults are forced to develop a linear way. The differences in learning processes are contingent on the changes in the language

[10] The theory of universal grammar postulates that there exist linguistic constructs in all of the world's languages, which are universal, i.e. common to all. In learning an L1, certain constructs get 'switched' on or off, up to a certain age. The question is if L2 learners have access to the constructions that are 'off' in their first language, and if so, how these are accessed (Plaza Pust, 2000).

faculty of the brain, which are presumed to initiate not long before puberty. These changes are, theoretically, cementing the configuration of the UG, which the L1 learning process has established, effectively blocking any future adjustments of the UG. Before this change, language learners employ a modular learning strategy, effectively multi-tasking when encountering new structures. After the change, the ability to utilize this modular way of learning has been deactivated, and the learner is forced to resort to linear learning by processing each new structure, one at a time. At the same time, many components in child SLA and adult SLA are similar (Plaza Pust, 2000). Plaza Pust concludes that:

> "On the basis of a dynamic approach we realise that the similarities and the differences between child and adult language development and diachronic language change are in conformity with the self-similar or fractal nature which is characteristic of dynamic systems. In other words, each type of language development has its peculiarities but they also have something in common. UG delimits the range of variation but it does not prescribe the actual developmental path." (Plaza Pust, 2000:270)

2.5.1 *Swedish as a second language*

Most of the research on reading in a second language has, for obvious reasons, focused on reading in English (cf. Alderson & Urquhart, 1984), but a large body of literature on Swedish as a second language (SSL) does exist.

Research on Swedish as a second language started in the 1960's in connection with an influx of foreign labor, and a substantial body of literature on how immigrants learn Swedish has been produced (e.g. Granberg, 2001; Gunnarsson, 1982; Hammarberg & Viberg, 1984; Hyltenstam & Wassén, 1984; Kotsinas, 1982).

In 1999, Pienemann & Håkansson (1999) conducted a comprehensive study of learning sequences in Swedish as a second language. They compiled numerous earlier SSL-studies and formulated them into a coherent theory of SLA, in which they applied the so called processability theory[11] to Swedish morphology and syntax. Their studies resulted in the establishment of a processability hierarchy for Swedish. The processability hierarchy implies the sequence for processing certain syntactical and morphological procedures in a language.

To evaluate this Swedish processability hierarchy, Pienemann & Håkansson tested it against a number of empirical studies of SSL, conducted by prominent researchers of SSL, and Pienemann & Håkansson found that not one of these other studies contradicted their processability hierarchy.

2.5.2 *The second language learner*

SLA researchers have been attempting to construct generalized categories for investigation of learning similarities within and between these groups. Research on the individual learner has, however, been scarce. L2 learning is dependent upon a wide range of individual factors in learners, and in order to posit a general theory of SLA, one must include these factors (Granberg, 2001). Granberg writes:

> "There are no comprehensive theories about the influence of individual differences in second language learning, there is considerable confusion regarding the definition of constructs, and, above all, we know very little about

[11] The processability theory is essentially a hierarchy for processing language, which can predict the order in which linguistic structures are processed. Furthermore, it can be implemented in LFG (Lexico-Functional Grammar). For a full description, see Pienemann (1998)

> the individuality of each learner, especially the complex interaction of [individual differences], as well as longitudinal effects." (Granberg, 2001:45)

With this as a starting point, Granberg mentions a host of factors that must inevitably be weighed into a theory of SLA: personality, attitude, motivation, learning strategies, learning styles, age, and aptitude. He also mentions that different perspectives play a role in SLA, for instance, social, cultural, and emotional perspectives.

The immigrant second language learner

In light of the discussion in Granberg (2001), it is almost impossible to make any claims about immigrant second language learners as a group, because they are not a homogenous group of individuals. Some immigrants have received no formal education in their home countries and need to start with alphabetization, while other immigrants have earned academic degrees and are accustomed to learning languages. All immigrant L2 learners, however, exhibit one similarity throughout their development—they all require texts which present as much of the new language as the learner can absorb at that particular stage in their development.

2.6 Summary

There exists a significant void in the research area of text simplification, especially with regard to simplified texts for average adult L2 learners. Most easy readers and simplified texts are written or adapted for the mentally disabled, aphasiacs or school children, and not for neurologically sound adults.

Additionally, investigation into computer-generated text simplification is also lacking. It is especially important that this line of investigation be further pursued for a language like Swedish, as it has such a small target audience, and produces such a low number of original simple texts. The development of a fully-functioning automated text simplification system could drastically increase the number of simplified texts available for a variety of purposes and target-audiences. What little research on automatic text simplification exists today, is, on the whole, sparse, most often tailored for the English language, and frequently geared towards very specific and limited purposes.

The simplification systems that are being used today are a kind of transfer system, i.e. they take the input data and remodel it according to transformation rules. The transfer system seems to be a suitable format for building text simplifiers, as they only require rules for the structures that the writer wishes to transform, and thereby simplify. This means that the structures that should not be simplified, need no rule-writing, and may essentially be left alone. Moreover, the writer may add new simplification rules, and revise the rules already implemented, without disrupting the bulk of the text.

As a proven dependable source for measuring readability, defined as text complexity, readability formulae are widely used in many different settings. Readability formulae are utilized to measure the syntactic complexity of texts, and they roughly correspond to how readers rank texts on a readability scale. These tools are not always used for the purposes that they were intended for, and are sometimes viewed as absolute measures of texts, something which has resulted in criticism of the formulae. In spite of the criticism they have received, readability formulae continue to be utilized as a tool for determining text difficulty, and it is the ease with which they are calculated that their success may be attributed to.

3 Material

In this chapter, the material used for the study conducted in this thesis is described. The material constitutes a parallel corpus of newspaper text taken from *Invandrartidningen*, a weekly newspaper for immigrants.

3.1 Invandrartidningen

Invandrartidningen (The Immigrants' Paper), was a weekly newspaper for the immigrant population of Sweden.[12] It aimed at being a source of news for immigrants in Sweden, as well as a paper where this group could read and learn about Sweden and Swedish society. *Invandrartidningen* was published up until 1999 in seven languages,[13] including an easy Swedish version, *På lätt svenska*. Every issue was translated from an original text, written in standard Swedish. This original was never itself published, but served as the source text for the translations. It is the source text and the *På lätt svenska* version that together constitute the material for the investigation in this thesis.

3.2 På lätt svenska[14]

På lätt svenska (In Easy Swedish) was the simplified translation of the original Swedish version of *Invandrartidningen*. The simplifications were made by hand by the editorial staff of the paper, that is, by people who were not all trained translators. *På lätt svenska* is characterized by two features that the other translations do not have: the articles in *På lätt svenska* are ranked with a difficulty degree, and certain words are explained in a word list (see Figure 3.1).

3.2.1 Guidelines for simplification

The procedures for the simplification of *Invandrartidningen* were based on both the simple writing rules developed by *Centrum för lättläst* (see section 2.3 above), and rules-of-thumb the simplifiers of *Invandrartidningen* arrived at, through their ongoing work of translating the texts (J.B., p.c.).

In Figure 3.2, a modified version of *Centrum för lättläst*'s rules for simplification is presented; the criteria that were obviously developed for language impaired readers have been omitted, and rules specifically developed for L2 texts have been amended. The rules in Figure 3.2 are not applied in all instances to each of the texts, largely because of the lack of overall simplifying guidelines for the translators.

There are some interesting differences between *Centrum för lättläst*'s simplification rules in Figure 2.1, and the rules arrived at through empirical practice at *Invandrartidningen* (in Figure 3.2). The single most influential rule may be that of word order. The rules from *Centrum för lättläst* only indirectly address word order in the paragraphs about subordinate

[12] They lost their funding when the press subsidy was withdrawn, but were later reincarnated on the web as SESAM: `http://www.inv.se`

[13] Arabic, English, Farsi (Persian), Polish, Serbo-Croatian, Spanish, and easy Swedish.

[14] This chapter is based on my own observations of the material, the paper version of *På lätt svenska*, and on a telephone conversation with Jolin Boldt (2002-04-03), now editor-in-chief for SESAM and then associate of the editorial staff of Invandrartidningen.

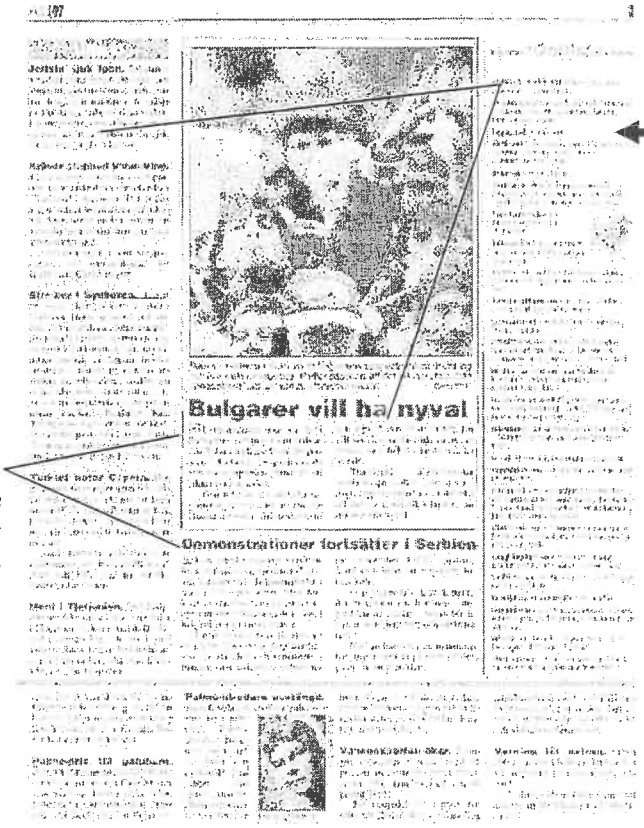

Figure 3.1 The original *På lätt svenska* version.

- Use straight word order.
  - Use as few subordinate clauses as possible.
  - Subject-verb inversion may be confusing to learners.
- Write from the beginning to the end.
  - The actions in the text should be in chronological order.
- Write sparse.
  - Learners may have trouble distinguishing what is foreground or background information in the text.
- Write with simple words.
  - Replace a hard word with a simpler, more common synonym.
  - If you still need to use the hard word—explain it in the text or a word list.
- Do not vary your vocabulary.
  - The reader may not understand coreferring expressions.
  - Use the present, or same, tense throughout the text.
- Describe things as concrete as possible
  - Learners often interpret the text literally.
    - Imagery, metaphors and similes may have a concrete meaning as well as an abstract meaning.
  - Make no implications.
- Write active.
  - The passive voice is harder to intuitively understand.
- With all that said, choose the vocabulary and the constructions that are necessary to write your text.

Figure 3.2 Guidelines for writing and translating easy texts for L2 readers following Wiman (1998).

clauses and passive voice, whereas at *Invandrartidningen*, the writers were asked to use the SVO (Subject-Verb-Object) word order to a higher degree. Another observation made by the staff at *Invandrartidningen*, is that when too many simple words are strung together, they can become hard to understand. It is for this reason that no guidelines for selecting the simplest word available were included in the list. Furthermore, certain subject matters and types of texts are not easily simplified, as they demand a degree of clarity in expression that cannot be compromised (J.B., p.c.).

The degree to which words and phrases are simplified or explained, is somewhat arbitrary, as the simplifications are entirely dependent upon the person who makes the translation.

3.2.2 *Difficulty degrees*

The difficulty degrees in *På lätt svenska* are a) very easy Swedish, b) easy Swedish, and c) not so easy Swedish. The three difficulty degrees are all easier than standard Swedish. Only some of the news articles featured in *På lätt svenska* were assigned a difficulty degree. The unmarked articles are primarily very short news items, but also include longer articles which appear to have been written exclusively for *På lätt svenska*, as they have no counterpart in Invandrartidningen.

The articles in *På lätt svenska* that have been assigned a difficulty degree are marked either A, B, or C, respectively. The articles which did not have a difficulty degree assigned to them, have been marked with an X, for the purposes of this thesis.

Furthermore, it appears that the editors of *På lätt svenska* had no specific difficulty degree requirements set for each issue.

3.2.3 *Word lists*

Word lists are dispersed throughout each issue of *På lätt svenska*. The words and definitions in these lists were selected by each article's translator. The types of words selected were primarily those which might aid the immigrant reader in following and navigating the general debate in the society. These were words such as *avgå* (resign), *dömas* (to be judged), *försvar* (defense), *medla* (mediate), and *opposition* (opposition) (*På lätt svenska*, 1/97:3), but also words that could be perceived as difficult or unusual, for instance *dissident* (dissident) and *ostron* (oyster) (*På lätt svenska*, 1/97:3). In *På lätt svenska*'s early days, only single words were explained in the word lists, but later on, expressions such as collocations, idioms, and phrasal verbs were explained as well (J.B., p.c.).
Single words that are found in the word list are marked with an asterisk (*) in the text, e.g. *avgå** in the sentence *Jeltsins politiska motståndare kräver att han ska avgå** (Jeltsin's political opponents demand that he resign*) (*På lätt svenska*, 1/97:3), while phrases are both underlined and marked with an asterisk, e.g. *De har slagits ihop* till Skåne län* (They have merged into* Skåne county) (*På lätt svenska*, 1/97:9).

3.3 The articles

The articles examined within this thesis have not been tagged for part-of-speech, nor have they been parsed. The extracted texts consist of approximately 5-6000 words each, and have been taken from 38 articles.

3.3.1 General statistics

As can be seen in Table 3.1, articles of difficulty degree B constitute 60.5% of the material. Articles of difficulty degree A represent 10.5% of the text, C represents 7.9%, and the X category forms a substantial part of the material, 21.1%.

Table 3.1 Statistics of the difficulty degrees of the articles in *På lätt svenska.*

Difficulty Degree	Number of Articles	Percentage of Articles
A	4	10,5%
B	23	60,5%
C	3	7,9%
X	8	21,1%
Total	38	100,0%

3.3.2 Lix values for the articles

To give an overview of the readability of the articles in both the original *Invandrartidningen* and the simplified *På lätt svenska* versions, necessary data to compute lix[15] is presented in Table 3.2.[16] When comparing the numbers of words, sentences, and long words for both versions of Invandrartidningen, it becomes clear that the original is longer, that is, has more words than *På lätt svenska – Invandrartidnignen* has 5970 words, and *På lätt svenska* has 5326 words. *Invandrartidningen* also has the higher number of long words, 1557 vs. 1223. The number of sentences, though, is higher in the simplified version, 532 vs. 486. These numbers predict that the original is indeed syntactically more difficult than the simplified version. The lix value of *På lätt svenska* is 33, and for *Invandrartidningen* lix is 38.

Table 3.2 The data for computing *lix*, and lix values for both *Invandrartidningen* and *På lätt svenska.*

Subcorpus	Number of words	Number of sentences	Number of long words	Words per Sentence	Percentage of Long Words	LIX
Invandrar-tidningen	5970	486	1557	12,28	26,08	38
På lätt svenska	5326	532	1223	10,01	22,96	33

When compared to the lix interpreter (Figure 2.3), the texts in both *Invandrartidningen* and *På lätt svenska* with lix 38 and 33, respectively, fall within the EASY TEXT-category; *På lätt svenska* in the lower range, and *Invandrartidningen* in the higher range of the category.

The lix values computed for the entire corpus in Table 3.2 are mean values, and when examining the lix values for each of the articles, a more diversified picture is shown. The lix values in *Invandrartidningen* (see Figure 3.3) are found in the interval 25–57, which corresponds to anything between VERY EASY TEXT and HARD TEXT. The majority (55%) of the lix values fall within the 35–45 range, which means EASY TEXT to AVERAGE TEXT.

[15] See section 2.4.3 for a presentation of *lix.*

[16] The raw data for Table 3.2 is presented in appendix A for *Invandrartidningen,* and appendix B for *På lätt svenska.*

There is also a cluster of very high lix values (18%), ranging from 50 to 57 which equals HARD TEXT, for *Invandrartidningen*. In four instances (11%), the lix value drops under 30.

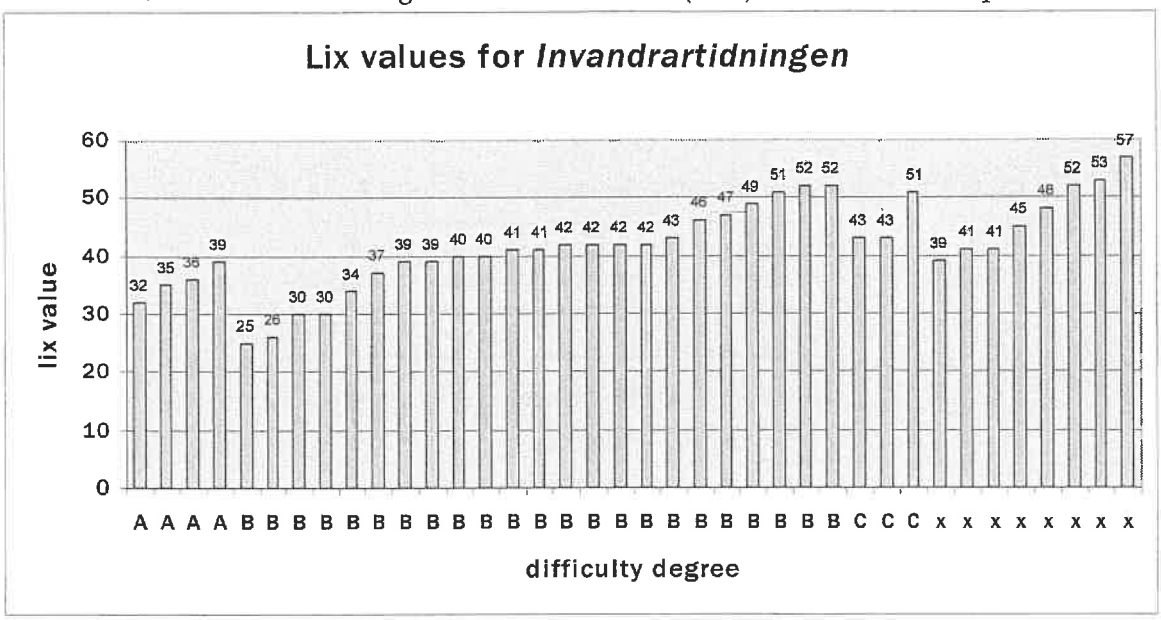

Figure 3.3 The lix values for the articles in *Invandrartidningen*.

Looking at the lix values for *På lätt svenska* (Figure 3.4), we find them more spread out across the field; between 11 and 50. This translates into very EASY TEXT to AVERAGE TEXT. The majority of the lix values for *På lätt svenska* fall within values of 30–40 (61%), which is centered round EASY TEXT in the lix interpreter. Only one article with lix 11 slides under the 20-limit, whereas the lix value for four articles (11%) crawls over 45.

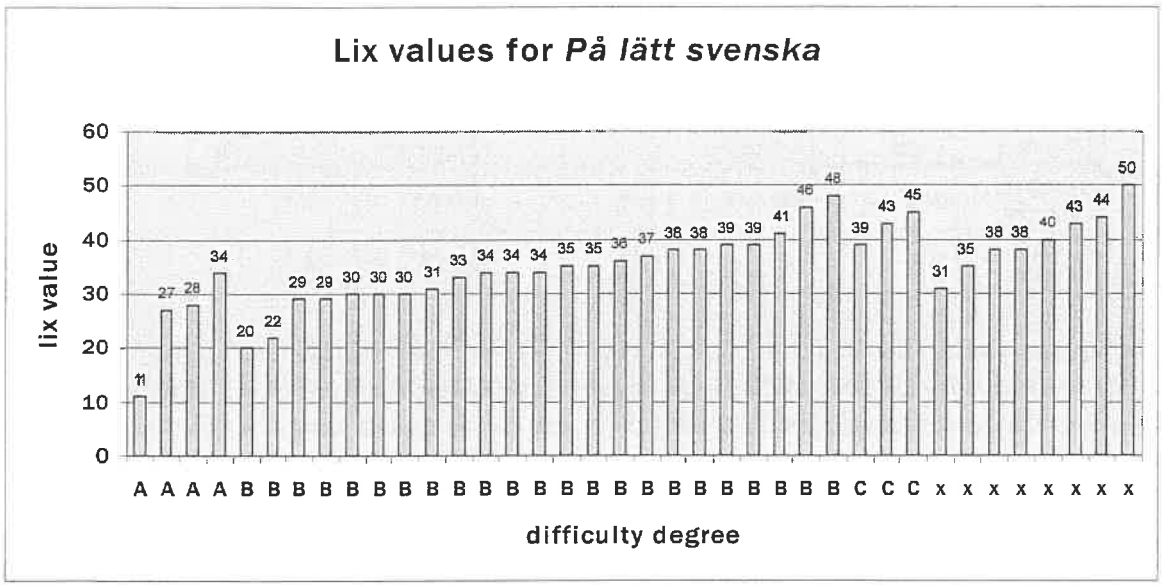

Figure 3.4 The lix values for the articles in *På lätt svenska*.

For both *Invandrartidnignen* and *På lätt svenska,* the difficulty degrees overlap with reference to their lix values. As is evident from Figures 3.3 and 3.4, the higher lix values of a lower difficulty degree are higher than the lower lix values of the following higher

difficulty degree. However, examining the difficulty degrees from A to X it is clear that they do have progressively higher lix values, the harder the difficulty degree.

When comparing *Invandrartidningen* with *På lätt svenska* with regards to their lix values, as is done in Figure 3.5, it is evident that the lix values of *På lätt svenska* are closely following those of *Invandrartidningen*, albeit at a lower value, with two exceptions—articles 13 and 36, which differ from *Invandrartidningen* by >10 lix. The reason for the difference is readily available from the data in appendix A and B, and it shows that article 13 is short and has a very large percentage of long words. For article 36, the explanation is that the *På lätt svenska* version has no long words at all. It also visualizes how *På lätt svenska* is on average simpler than *Invandrartidningen*.

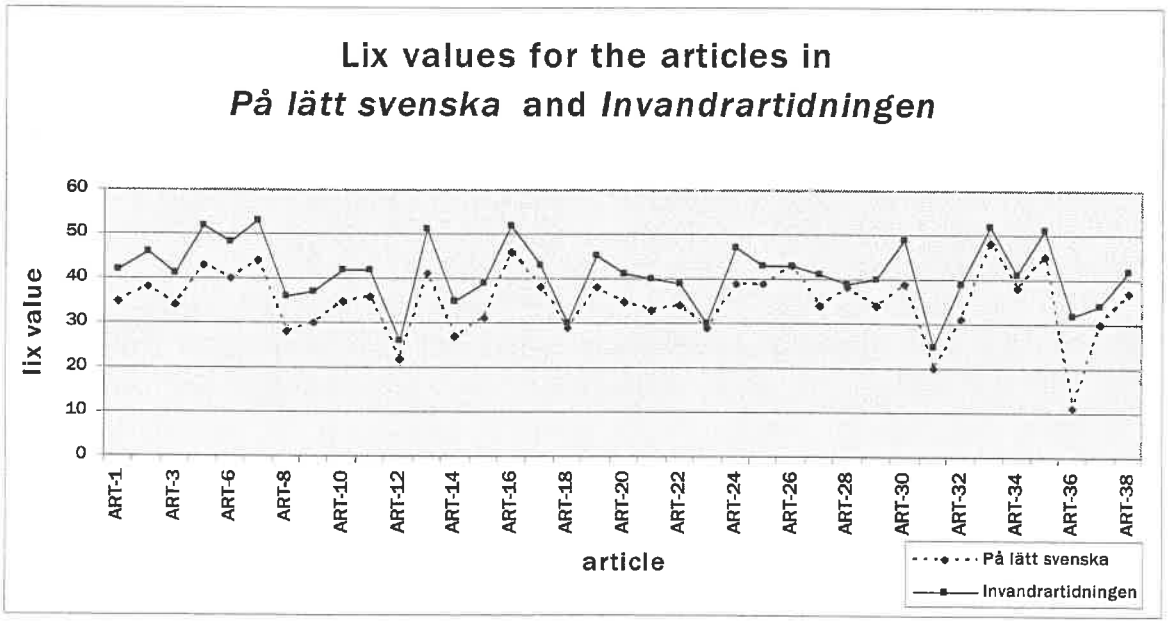

Figure 3.5 The lix values for *På lätt svenska* and *Invandrartidningen* compared.

3.3.3 *Lix values for the difficulty degrees*

The average lix values of the various difficulty degrees are presented in Table 3.3 and shows that the difficulty degrees in *På lätt svenska* are a valid predictor of difficulty.
The A category is the simplest, followed by the B category, and then the C category. Interestingly, the X category—the category which has no difficulty posed at the beginning—is the hardest category in Invandrartidningen, but not in *På lätt svenska*, where category C is the hardest. The texts of lower difficulty degrees, A and B, in *Invandrartidningen* are easier than the texts of harder difficulty degree, C and X, in *På lätt svenska*, even though *På lätt svenska* is said to be simpler than the original on the whole.

Table 3.3 The average lix values for the difficulty degrees in *På lätt svenska* and *Invandrartidningen.*

Difficulty Degree[17]	Average lix value for På lätt svenska	Average lix value for *Invandrartidningen*
A	25	36
B	34	40
C	42	46
X	40	47

3.4 Summary

The simplified version of *Invandrartidningen,* called *På lätt svenska,* is the work of several human minds, and as such, it presents little conformity to a norm. The guidelines for simplifying the original texts were adapted from a set of guidelines for simplifying texts for language impaired readers, and were amended as the work proceeded. The resulting praxis for simplification used for *På lätt svenska* has never been put into writing.

På lätt svenska has an overall lower difficulty degree than *Invandrartidningen,* and the difficulty degrees set in the simplified version agrees with the computed lix values of the texts. The lix values are also representative of the corresponding articles in *Invandrartidningen.*

[17] The articles in the original *Invandrartidningen* have no difficulty degrees, but correspond to an article in *På lätt svenska.* This is what the difficulty degrees stated in Table 3.3 refer to.

4 Method

Only one of the subcorpora, the original *Invandrartidningen,* seen in Figure 4.1, was machine readable at the beginning of this thesis study. Consequently, the first step in the investigation was to transcribe the plain newspaper text from the *På lätt svenska* version (see Figure 3.1) into machine readable text-files. Following this transcription, both subcorpora were automatically spell-checked, and the format was normalized.

```
Bättre ekonomi men fortsatt hög arbetslöshet.
- Vi går mot bättre tider, spår de ekonomiska prognosmakarna. Men
det tror vi inte på. Bara 21 procent av folket tror att Sveriges
ekonomi förbättras under 1997.
Experterna tror på ökad tillväxt, fortsatt låga räntor och mer
köpkraft för löntagarna. De flesta svenska storföretag väntar sig
större vinster 1997 på grund av effektivisering och ökad
försäljning. Bilförsäljningen - en säker indikator på
konjunkturförändringar - ökade med 14 procent under december 1996.
Antalet företagskonkurser gick ned med tre procent under 1996. Men
det som lägger sordin på vanligt folks optimism är att arbetslösheten
väntas förbli hög. Även i år kommer närmare 700 000 människor att
vara utan arbete i Sverige. Receptet för lägre arbetslöshet är enligt
OECD, de västliga industriländernas samarbetsorganisation, bl a
lägre skatter och större löneskillnader.

<!--page-->

Palmeutredare misstänkt för bedrägeri.
Hans Ölvebro avstängs från jobbet som spaningsledare för gruppen som
utreder mordet på Olof Palme. Ölvebro är misstänkt för
skattebedrägeri. Enligt åklagaren har han lurat staten på 98 000
kr. Ölvebros familj tvingades flytta från Stockholm efter
mordhot. Hans arbetsgivare betalade därför i flera år hyran för en
övernattningslägenhet och en tjänstebil. Åklagaren anser att
Ölvebro
borde tagit upp detta som skattepliktiga förmåner.

<!--page-->

Svensk vapenexport till diktaturer.
Nära en tredjedel av de vapen Sverige exporterar går till länder där
det förekommer omfattande kränkningar av de mänskliga
rättigheterna. Det hävdar Svenska freds- och skiljedomsföreningen
på
grundval av svensk statistik och uppgifter från Amnesty
international.
Vapenexporten har ökat kraftigt hävdar Svenska
Freds. Försvarsindustriföreningen påpekar att vapenexporten totalt
har
minskat sedan 1989.
```

Figure 4.1 Extract from the file containing the original *Invandrartidningen.*

4.1 Assembling the material

After the initial formatting, the parallel corpus was assembled. As the nature of this study was exploratory, and aimed to discover preliminary rules for the grammatical simplification of Swedish, the decision was made that the corpus should be limited to a moderate size, and that it should be prepared manually.[18] The articles that constituted the text sample were extracted in order of appearance, starting with issue 1/97, and then taken from consecutive issues until the two subcorpora consisted of approximately 5000 words each.[19]

The sample of news articles extracted for the investigation in this thesis contained text from 38 news articles from *Invandrartidningen* 1–5/1997.

4.1.1 Criteria for the selection of articles

In the early stages of extracting the sample, two criteria for selecting the articles were established. It was discovered that the two versions of *Invandrartidningen* did not always correspond with reference to their content; particularly *På lätt svenska* contained editorial material not found in *Invandrartidningen*. For this reason, the first criterium for the inclusion of an article within the text sample was that the article should appear in both subcorpora. The second criterium was that the article in *På lätt svenska* should not have been altered by additional editorial material in the simplification process.

Texts of particularly pronounced styles have been omitted from the selection. These texts were, for example: editorials, recipes, fairy tales, or letters from the readers. Moreover, these texts were, in most cases, written specifically for *På lätt svenska*. Neither were the difficulty degrees (see chapter 3.2.2) considered.

4.1.2 Calculating lix

After assembling the sample corpus, both subcorpora were manually proof-read, and as the articles were proof-read, the data for calculating lix—the number of words, number of sentences, and the number of long words—was collected and tabulated. This data was processed in Excel to easily calculate the mean sentence length, and the percentage of long words for each article, in order to calculate lix.[20]

4.1.3 Extracting the simplifications

During proof-reading, the sentences containing words which were explained in the word lists, were winnowed out. This step was necessary, as these sentences were not simplifications, but rather a deliberate choice of the human simplifier to retain a particular word, to explain it in detail in the word list, and to present it in its natural context. Many of these words appear in a syntactically complex context, and the whole context would probably have been simplified, if the word itself had been simplified.

Then the simplifications were systematically extracted. Pairs of simplifications—syntactical patters in the original, which were altered in the simplified version—were identified. If these were considered to constitute a simplification pair, as opposed to an alteration of a lexical component, then they were included in the text

[18] This decision was made because the task of automatically tagging and parsing the corpus, and evaluating the result, would have been unproportionally large for the purposes of this study.

[19] For exact numbers, see Table 3.2.

[20] Read more about calculating lix in chapter 2.4.3.

sample. A simplification pair is a pair of source and target phrase that have the same 'meaning', and where the target phrase is a simplification of the source phrase.

The simplifications appeared in various shapes; sometimes information was deleted, and sometimes information was added. In all cases where the source and target phrases were judged to have approximately the same semantic content, the transformation was labeled a simplification. On intuitive, as well as purely grammatical grounds, this may seem somewhat contradictory, but when examining the phrase in its original context it becomes clear that it is, in fact, a matter of simplification.

The pairs of simplified phrases were separated from the rest of the material in order to eliminate the context of the phrases. This was done because this study focuses on simplifications made on the phrase level and not on the sentence or textual levels.

The extraction phase resulted in 467 simplification pairs.

4.2 Analyzing the phrases

The linguistic analysis of the differences between the two versions was done subsequent to the extraction of the simplification pairs. The simplification pairs were first sorted into different simplification categories—NP-category, PP-category, Verb-category, etc.—according to the type of simplification that needed analyzing. Each category was scrutinized with the aim of distilling every single simplification pair. These individual simplification pairs were further analyzed in order to detect which alterations had been applied to make the simplifications.

4.2.1 *Establishing a typology*

Following the linguistic analysis, the simplifications were further examined to discern the simplifications that were frequent enough to constitute a preliminary typology.

To create a typology of the simplifications, the simplifications that occurred only once were deleted,[21] leaving those that occured two or more times. These types of simplification were formalized, and explained along with examples from the corpus. By this means, a set of formalized rules were developed, rules which consist of basic syntactic simplifications, and were suitable for implementation, in the form of simplification rules, into an automatic text simplification program.

[21] A single occurrance is not enough to establish a type.

5 Linguistic analysis

The linguistic analysis presents all simplifications found in the investigation of the articles.

5.1 About the simplified phrases

In categorizing the material, the description of Swedish found in Teleman (1974) and *Svenska akademiens grammatik* (Teleman, Hellberg, & Andersson, 1999; hereafter referred to as SAG) was followed.

Simplifications of four different patterns are found: lexeme-to-lexeme, lexeme-to-phrase, phrase-to-lexeme, and phrase-to-phrase. Of these four patterns, the lexeme-to-lexeme is the easiest to spot and to analyze. It also contains different types of simplification than the other categories, namely lexical relations. Due to this discovery, the presentation of the analysis is divided into two sections, the first section describing the lexeme-to-lexeme simplifications, and the second describing the simplifications of the remaining patterns.

5.2 Transformations on the lexical level

The transformations made on the lexical level are changes of one lexeme for another, and they are chiefly constituted by semantic relations. This type of simplification is not explored in depth in this thesis, since it is lexically, as opposed to syntactically, determined, and the analysis of the lexical changes is kept brief.[22]

5.2.1 Semantic relations

There are three categories of semantic relations represented in the simplification typology: synonymy, antonymy, hyponomy. The lexemes in this category are always simplified into another lexeme of the same part-of-speech, since the lexical relations are such that the two lexemes in the pair are syntactically completely interchangeable.

Synonymy
There are 54 occurrences of synonymy and they represent simplifications of most parts-of-speech. The original lexeme and the simplification rarely represent absolute synonyms, as a complete correspondence between interchangeable lexemes is somewhat redundant, and there exists a slight difference in stylistic meaning between the lexemes.

Antonymy
There are 5 occurrences of simplifications of one lexeme into its antonym. Since there are so few occurrences, it does not serve any purpose to differentiate between various kinds of antonymy.

---

[22] To achieve lexical simplification, a sophisticated lexicon of i.e. WordNet-type is needed. A Swedish WordNet, SwordNet, is currently developed at the University of Lund, and could be of interest once it is finished (cf. `http://www.ling.lu.se/projects/Swordnet`). Another solution would be to gain access to the dictionaries in Lexin at *Skolverket* (the National Agency for Education) (cf. `http://www-lexikon.nada.kth.se/skolverket/about-www-lexin-sv.shtml`).

Hyponymy

In the 22 occurrences of hyponymy, there are examples of both hyponymic (12 occurrences) and hypernymic (10 occurrences) relations between the original and the simplified lexemes.

5.3 Transformations on the phrasal level

There are two fundamentally different types of simplifications of phrases: a change to a grammatically simpler construction or a lexically bound simplification into another grammatical construction, due to an earlier simplification.

5.3.1 Simplifications of the noun phrase[23]

The noun phrase (NP)[24] is the phrase which occurs in most shapes, save the verb phrase. It appears that the type of simplifications of the noun phrase stem from the function they perform in the clause, rather than its phrase structure in itself.

Simplification of the NP consisting of the head noun

Table 5.1 Types of simplification of an NP consisting of a head noun.

np(n) (414)	Ø	13
	np(ap+n)	3
	np(pn, nn)	4
	np(poss+n)	5
	np(pron+n)	2
	pron	2
	subordinate clause	10

In Table 5.1, we see that the noun phrase consisting of just one noun is simplified into a variety of phrases. The most common simplification is the deletion of the noun phrase. This occurs most often when the NP is part of a superordinated phrase, for example, a deleted subordinate clause or a prepositional phrase (PP). A simplification by adding an adjective phrase (AP) occurs 3 times, and simplification by changing the noun into a person name (PN) or a compound noun (NN) occurs 4 times. There are 5 occurrences of an inserted possessive, and 2 of an inserted pronoun (PRON). The noun phrase is transformed into a subordinate clause a total of 10 times. Most of the times it occurs, the NP(N) is not simplified at all.

Simplification of constituents preceding the NP head

Table 5.2 shows the simplification of noun phrases with a preposed modifier. The noun phrase starting with a determiner plus a noun is simplified either by deleting the determiner (3 occasions), leaving a bare noun or by adding an adjective phrase (AP) to the noun phrase (3 occasions).

The simplifications of the noun phrase beginning with a determiner and an adjective phrase fall into three categories: one where the adjective phrase is deleted (2 occurrences),

[23] The noun (n) in these phrases may as well be a compound noun (nn) or a personal noun (pn), unless otherwise stated.

[24] The abbreviations of the phrase types are clarified in Appendix C.

a second where the determiner is deleted (4 occurrences), and a third in which both the determiner and the adjective phrase are deleted (3 occurrences), leaving only the noun (or compound noun or personal name) plus eventual postposed constituents such as prepositional phrases.

Table 5.2 Types of simplifications of constituents preceding the NP head.

np(det+n (36)	np(n)	3
	np(det+ap(adj)+n	3
np(det+ap+n (29)	np(det+n	2
	np(ap+n	4
	np(n)	3
np(ap+n (51)	np(det+n	2
	np(det+ap+n	5
	np(n	6
	np(n+subclause	3
	np(n+pp	4

The NP beginning with an adjective phrase gets simplified in many of the instances it occurs. The most common simplification is the deletion of the adjective (6 occurrences). The adjective is deleted more often than it is retained, and in the phrases where it is retained a determiner is inserted in 5 instances. In the deletion of the AP, a determiner is inserted (2 occurrences). The adjective is also simplified into constituents which follow the head (the noun), namely subordinate clauses (3 occurrences), and prepositional phrases (4 occurrences). Again, most occurrences of the phrase type are left unsimplified.

Simplification of constituents succeeding the NP head

Table 5.3 Types of simplifications of constituents succeeding the NP head.

np(n+pp (63)	main clause	2
	subclause	3
	infp(infm+v:inf	2
	active verb	4
n+subj(17)	∅	3

As can be noticed in Table 5.3, the simplifications of the noun phrase in which the head is followed by a prepositional phrase are of four types. This category seems to be a category where we find some of the nominalizations in the material. In 2 instances, the PP is simplified into a separate main clause—a new sentence. In other cases, the NP is changed into a subordinate clause (3 occurrences), and in yet other cases into an infinitive phrase (2 occurrences). In the last type of simplification for the constituents succeeding the NP head, the noun phrase is rewritten into an entirely new phrase structure, headed by an active verb (4 occurrences).

When this category is simplified, it is simplified to a very high degree, i.e. the distance between the source and the target phrase structure is large compared to the rest of the noun simplifications. Two of the simplification types fully retain the phrase structure of the noun phrase—the changes transforming the PP into either a subordinate clause or an infinitive phrase. Furthermore, the simplifications of the postposed prepositional phrase seem to be more dependent on a simplification of the PP than the head noun, which was retained in most instances.

The simplification of a noun followed by subjunction covers the instances where a subordinate clause is attached to the NP as an attribute. The only type of simplification of this phrase are deletions (3 occurrences).

Simplification of the compound noun

Table 5.4 Types of simplifications of the compound noun.

np(nn) (62)	np(ap(adj)+n)	2
	np(det+ap(adj)+n)	2
	np(n)	10
	np(n+pp(p+np(n)))	4
	np(pron+n)	2

The most common simplification (see Table 5.4) of the compound noun is into a simple noun (10 occurrences). Two types of simplifications inserts adjective phrases into the noun phrase; 2 occurrances of an adjective phrase, and 2 occurrences of insertion of both a determiner and an adjective phrase.

Another simplification inserts a prepositional phrase to a single noun (4 occurrences), and deletes the compound noun. The last simplification type for compound nouns is to a noun phrase containing a pronoun and a noun (2 occurrences).

Simplification of the personal name

Table 5.5 Types of simplifications of the personal name.

np(pn) (95)	np(n)	2
	np(n+pp)	4
	np(pn+IK+pn)	3

In Table 5.5, the simplifications of the personal name are described. The simplification of the personal name into a noun phrase containing a noun plus a prepositional phrase is the most frequent, even though it occurs just 4 times. The second most common simplification of the personal name is into two personal names separated by a comma, IK (3 instances). This is a type of simplification, where additional information added to the personal name in the form of a postposed attribute. In the last type, the personal name is simplified into a noun (2 occurrences).

The personal name is simplified in very few instances. The explanation for this may be found in the nature of the personal name—it cannot be easily paraphrased or omitted without losing vital information.

Simplification of the numeral

Table 5.6 Types of simplifications of the numeral.

num (48)	np(pron)	2

The numeral is a category which is not noticeably touched by simplification. Table 5.6 shows that the few instances the numeral is simplified, it is into a pronoun. There are also deletions in the material, but these are very few. Over all, the numeral is a fairly stable category in the simplifications of *Invandrartidningen*.

Simplification of the possessive

The category of the possessive contains all parts of speech, which can form the possessive, and no distinction between these is made in the analysis.

Table 5.7 Types of simplifications of the possessive.

poss (45)	pp	6
	verb phrase	7
	ø	6

The possessive always occurs inside a noun phrase, and the simplification of the possessive follows one of three patterns: first, it is simplified into a postposed prepositional phrase (6 occurrences). Second, it is transformed into a verb phrase (7 occurrences), something which is most common when the possessive noun is a nominalization. Third, the possessive can be deleted, together with eventual determiners, leaving a bare noun (6 instances).

The information loss in the simplification of the possessive does not seem to present a problem to the simplifiers. In this respect, the simplification of the possessive is comparable to that of the adjective phrase, although the AP is simplified to a higher degree.

5.3.2 *Pronouns*

Pronouns are deictic by nature, and are used anaphorically. Consequently, they are not simplified in the same way as the other categories; they are resolved rather than simplified. Therefore, the pronoun is not covered as a simplification, but left for further research.

5.3.3 *Simplification of the adjective phrase*

The adjective phrases[25] have two uses: as a modifier of a noun or as a predicative complement to a verb. In this material, we find adjective phrases with both functions. The adjective phrase appearing in a noun phrase is covered under the simplification of the noun phrase (see chapter 5.3.1 above). The AP forming a predicative complement will also be touched on in the chapter about verb simplification (see section 5.3.6 below)

Table 5.8 Types of simplifications of the adjective phrase.

ap(adj) (121)	ø	17
	pp(p+np(ap(adj)+n)	5
ap+n	v+ap	2
ap(adj+adj) (4)	ap(adj)	3
ap(adv+adj) (5)	ap(adj)	2
adj+konj+adj (3)	ap(adj)	3

The adjective phrase and its simplifications are elusive. Many times, the adjective phrase is retained, or simplified through reduction, but more times than not, it is simplified into a syntactically very different construction. Table 5.8 shows that the most consistent simplification of the adjective phrase modifying a noun head is deletion (17 deletions). The second most common simplification is the alteration of the adjective into a post-posed prepositional phrase (5 occurrences). An adjective modifying a noun is simplified into an

[25] In this thesis, the adjective category also includes participles.

adjective that is a complement to a verb in 2 instances. There is no evidence of the opposite. The adjective phrase functioning as a predicative complement remains largely untouched, but in many cases there is a change inside the adjective phrase. Other times the adjective is nominalized, and disappears. In a total of four instances of coordinated adjective phrases, these are simplified into either one of the adjectives (3 occurrences).

5.3.4 *Simplification of the adverb*

Table 5.9 Types of simplifications of the adverb.

adv (78)	ø	31
	np	3
	pp	3
	adv+adv	2
adv+adv	adv	2

In the majority of its occurrences, the adverb is deleted (see Table 5.9). In 3 occurrences each, the adverb is paraphrased into a noun phrase or a prepositional phrase, and in another 2 occurrences into two adverbs. A phrase with two adverbs is simplified into one adverb on 2 occasions.

The adverbs are often moved to a new position in the clause as a simplification. This type of simplification is not investigated further in this thesis.

5.3.5 *Simplification of the prepositional phrase*

The prepositional phrase is different from the phrases investigated thus far, in that the preposition very often is determined by a referent outside of the phrase, normally the referent preceding it (SAG:3). Simplification of the first referent, i.e. a structure outside the prepositional phrase, may require a different preposition, or a construction entirely unlike the original. This is the case in many of the simplifications of the prepositional phrase in *Invandrartidningen*.

Table 5.10 shows that 17 instances of the prepositional phrase are simplified into a subordinate clause, and in as many cases they are deleted. In 4 occurrences, the prepositional phrase is simplified into a noun phrase containing a pronoun and a noun.

Table 5.10 Types of simplifications of the prepositional phrase.

pp(p+np (182)	sub clause	17
	ø	17
	np(pron+n)	4
	np(n:poss	2
	adverb	6
	nominalizations	7
	verb construction	17
	passive > active	5

In two instances, the prepositional phrase denoting possession—the structure N+PP(P+NP)—is simplified into a noun phrase with a possessive attribute. Many prepositional phrases occurred after a nominalization, which in simplification is transformed into the corresponding verb. One of the most frequent alterations of the prepositional phrase (17 occurrences) is simplification into new verb constructions. This type of simplification does not depend on the prepositional phrases themselves, but rather

on the verb in the clause in which they occur. If the verb in a clause is simplified, sometimes the prepositional phrase in the same clause has to change or disappear in accordance with the context requirements of the new verb. In five cases, the verb changes from passive to active, and forces the prepositional phrase in the passive construction to undergo transformation.[26]

The prepositional phrases in which the preposition is succeeded by an infinitive phrase are few (8 occurrences) and undergo few simplifications, of which none occur two or more times.

5.3.6 *Simplification of the verb phrase*

The simplification of the verb is the type of simplification that causes the biggest transformations from the original to the simplified version. This both implies and verifies that the verb phrase is more diverse than any other phrase type. The constituents in a verb phrase, which are not verbs, are highly dependent on the main verb, something that is related to the verb's syntactical valency, and consequently, a simplification of the main verb alone may disrupt the structure of the entire verb phrase.

The verb phrase simplifications will not be analyzed and formalized into simplification rules like the other phrase types are, but will be discussed in order to illuminate certain issues for further research on the subject.

Verb chains cause many different simplification patterns, and it is difficult to extract any two simplification pairs containing verb chains that have the same phrase structure. In *Invandrartindningen*, there are examples of all kinds of verb chains in both the original and the simplified version.

Furthermore, the verb chains are difficult to formalize without a deeper linguistic analysis, because the subsequent verb constructions in the verb chain are regarded as verb phrases modifying the first (<second<third) verb (SAG:3). Thus, a simplification of an auxiliary high up in the hierarchy triggers transformations in the succeeding verbs in the verb chain, all the way down to the last verb.

In sentences containing more than one finite verb, for example, coordinated main clauses, ellipsis, or subordinated clauses, it is necessary to find a way to discern which verbs are subordinated to the other verbs, as the various verb phrases have different functions in the clause, and the simplifications often make changes in the hierarchy of these phrases.

The verbs have been analyzed and separated into the following categories: verb, auxiliary verb, copula verb, and passive verb. A particle verb has been identified. The verbs have also been marked for tense.

Simplification of the auxiliary verb

Table 5.11 Simplifications of the auxiliary verb.

vaux (92)		∅		23
vaux:pres (73)	+vpass:pres	v:pres		2
		vaux:pres	+v:inf	2
	+v:inf	v:pres		2
	+v:sup	vaux:pres	+v:inf	3

[26] This is the passive-to-active transformation, e.g. *X was made by Y* > *Y did X*.

In the majority of its occurrences, the auxiliary verb remains present after simplification. Also, verb chains of more than one auxiliary verb seem to get reduced more often, the longer they are. Table 5.11 illustrates how the auxiliary construction is simplified. The auxiliary is deleted in 23 instances, leaving the main verb. In 2 occurrences, an auxiliary followed by a verb in the present is simplified into a sole main verb, and in another 2 occurrences, the same structure is simplified into an auxiliary followed by an infinite verb. An auxiliary followed by an infinite verb is simplified into a verb in the present on 2 occasions. In 3 instances, an auxiliary plus a verb in the supine is transformed into an auxiliary plus an infinite verb.

Simplification of the copula verb

Table 5.12 illustrates how the copula verb is simplified. The single copula is most often simplified into a main verb (18 instances). In 2 occurrences, it acquires an auxiliary verb, and on 4 occasions it is transformed into the passive voice.

Table 5.12 Simplifications of the copula verb.

vkop (88)		verb		18
		v:aux	+v:inf	2
		vpass		4
vkop:pres (57)	+ap	v:pres		5
	+np	v:pres	+np	2
	+np	vkop:pres	+ap	2
vkop:pret (24)		v:pret		3

The copula is always closely connected to its complement, and the simplifications here suggest that rather than changing just the verb, the transformations also include the complement. In 5 occasions, the copula along with a predicative adjective phrase is simplified into a verb. In the case of a copula plus a predicative noun phrase, the copula, only, is transformed into a verb, while the noun phrase is left alone (2 occurrences). In 2 instances, a noun phrase is simplified into an adjective phrase, when functioning as predicative, while the copula remains unsimplified. On 3 occasions, a copula is simplified into a non-copula verb.

Tense seems to play only a minor role in these simplifications, as in almost all instances of simplifications for the copula, the simplifications do not change the tense.

Simplification of the passive verb

Table 5.13 Simplifications of the passive verb.

vpass:pres (31)	v:pres		12
	vaux:pres	+v:inf	6
	vkop:inf		2
	vkop:pres		5
vpass:pret (12)	vaux:pres	+v:sup	4
	vaux:pres	+v:inf	2
	vaux:pret	+v:inf	2
	v:pret		2
vpass:inf (7)	v:pres		2

Table 5.13 shows that the simplifications of the passive verb display a wide range of possible target structures. A large part of the simplifications of the passive verb is transformations to a compound tense with an auxiliary (14 occasions) or to a verb in the present tense (19 occasions). In the remaining 4 occurrences, the passive is transformed to a copula in the infinitive or to a past tense verb.

The passive also appears to be the verb form which is the most consistently simplified. This postulate implies that the human simplifiers believed the passive to present a particularly difficult problem for the readers if not simplified.

Simplification of the particle verb

Table 5.14 Simplifications of the particle verb.

v +vpart (16)	v		7
	vaux	+v:inf	3

The particle verb is simplified in many cases (see Table 5.14), but there are also many instances of particle verbs in the simplified version, which are not there in the corresponding original phrase. Since the particle verbs occur in as many instances in both subcorpora, 16 occurrences in the original and 16 in *På lätt svenska*, it would be presumptuous to draw any conclusions about the simplification of particle verbs from this material.

It seems safe, however, to deduce that an eventual simplification of, or into, a particle verb may not be governed by the particle verb in itself, but rather by some other construction in the phrase.

Tense

Table 5.15 Simplifications of verb tense.

present (290)	vaux:pres	+infinitive	15
		+supine	2
		v:pret	8
past (96)	present		5
	vaux:pret	+infinitive	3
	vaux:pres	+infinitive	4
		+supine	5
supine (32)	present		6
	past		3
	vaux	+infinitive	6
infinitive (109)	present		18
	past		3

Tense changes appear to form a special type of simplification, but not one that was widely employed by the human simplifiers of *Invandrartidningen*. As shown in Table 5.15, the present tense stays unsimplified in most of the instances, but is simplified to the past tense on 8 occasions. On 17 other occasions, the present tense is transformed into an auxiliary plus a verb in the infinitive (15 occasions) or a supine (2 occasions) verb. The past tense is transformed into the present tense in 5 cases, into an auxiliary in the past tense plus an infinitive in 3 cases, into an auxiliary in the present tense followed by an infinitive in 4 cases, and into an auxiliary in the present tense followed by a supine verb in 5 cases. The

supine, which appear exclusively in a compound tense together with an auxiliary, is simplified into a simple tense on 9 occasions, 6 simplifications into the present tense and 3 simplifications into the past tense. The remaining 6 implifications are of the supine verb into an infinitive verb. The infinitive tense is simplified into the present tense on 18 occasions, and into the past tense on 3 occasions.

5.3.7 *The conjunction*

Table 5.16 Simplifications of the conjunction.

konj (42)	Ø	12

The conjunction coordinates words and phrases of equal grammatical rank, and is not altered much in simplification. The most common way of simplifying this category is by deletion (see Table 5.16). Deleted conjunctions (12 occurrences) in the beginning of a clause are most often due to simplification of the main clauses, surrounding the conjunction, into separate sentences. The conjunctions which are deleted in the middle of a clause are in many cases deleted together with another constituent, such as a subordinate clause, or a constituent in a coordinated noun phrase, adjective phrase or prepositional phrase.

5.3.8 *The subjunction*

Table 5.17 Simplifications of the subjunction.

subj (53)	Ø	7
	new sentence	5
	pp	3
	ap	2
	conjunction	2

Table 5.17 shows that the subjunction[27] is most commonly deleted in simplification (7 occurrences), most often together with the subsequent clause. In 5 occurrences, the subjunction is deleted, while the subordinate clause following it is upgraded to a separate sentence. In other instances, the subjunction plus the subordinate clause following it are simplified into prepositional phrases (3 occurrences) or, in cases where it modifies a noun, into an adjective phrases (2 occurrences). In 2 occurrences, the subjunction is simplified to a conjunction, in which case the subordinate clause following it is upgraded to a main clause.

5.3.9 *The anticipatory it*

Table 5.18 Simplification of the anticipatory it.

formsubj	Ø	4
	np	10

[27] *Som* has been tagged as sunjunction, in line with e.g. Stroh-Wollin (2002), even in the cases where it can be argued that it should be analysed as a relative pronoun. In this material, there is just one occurrance of *som*, which I would have tagged as a pronoun had I made that choice, so this is not a big source for error. *För att* has been tagged as a compound subjunction consisting of two subjunctions, in line with Teleman (1974).

The simplification of the anticipatory subject *det* (it) (Table 5.18), captures the cleft sentence. *Det* is retained in most instances, but the simplifications that were made follow some patterns: the anticipatory subject is deleted (4 instances), and the construction is altered elsewhere, for instance with a different verb.

The second type of simplification (10 occurrences) presents the deletion of the anticipatory *det,* and the repositioning of a noun phrase, taken from elsewhere in the clause, to the position of the deleted *det*. This type of simplification covers the cleft subject.

Mostly, the cleft construction is left without simplification, or a simplification that affects the cleft takes place elewhere in the sentence.

5.4 Simplification above phrase level

It is important to remember that the simplifications in this chapter are simplifications within the phrase, and that the bulk of the material of *Invandrartidningen* is not affected by this type of simplification. Many simplifications occurred above the phrase level, and the most significant of these is a change in word order.

6 Analysis

Chapter 6 presents the results of the study of the two versions of *Invandrartidningen* through establishing a typology of simplifications suitable for implementation in a future text simplification program for Swedish.

6.1 Reviewing the simplifications

The analysis presented in chapter 5 suggests two supplementary notes to the definition of simplification adopted for this thesis (see chapter 2.1 above). The first addendum is that the the source and the target phrase of the simplification do not have to be of the same phrase type, i.e. the heads of the phrases do not need to have the same part-of-speech. The second addendum is that simplification may add or retract linguistic information to or from the original phrase, as well as restructuring extant information.

6.1.1 *Trends in simplification*

The investigation presented in chapter 5 shows that there exist no obvious tendencies towards a consistently employed model of simplification for the simplifications made of *Invandrartidningen*. Moreover, the consistent changes in phrase structure that were found and assumed to be simplifications, are scant and are applied in a haphazard manner. The explanation for this inconsistency in the simplifications is probably to be found in their nature; they were all applied by a number of people, not acting in unison, or in accordance with a predefined norm.

The inconsistencies in the production of the simplified version present a problem for predicting future simplifications by examining how earlier simplifications have been made. To predict simplifications is difficult because there is no guarantee that a phrase gets simplified in the same way as when it last appeared, or that the phrase is simplified at all. A simplification may also be employed in the reverse, that is, the phrase structures of the original source and target phrases change places, and form the inverted simplification. Furthermore, a phrase is not always simplified into the same phrase type, and is often altered into a different construction. Similarly, a first phrase type in the source phrase do not necessarily correspond to the first instance of the same phrase type in the source phrase.

6.1.2 *Types of simplifications*

The simplification types that were discovered in this study fall into four different categories of simplification, as shown in Figure 6.1 below. This classification of the simplifications provides four different subclasses of simplification: deletion, reduction, insertion, and paraphrase. Deletions and reductions always result in a loss of semantic meaning. In deletion, information is deleted in its entiry, while in reduction, parts of information are cut. The outcome of a deletion or a reduction is always a phrase of the same type as the source phrase, unless the phrase is deleted entirely. Insertion, on the other hand, adds supplemental information to the phrase, and may therefore be interpreted as actually increasing the semantic meaning. As insertion adds constituents within the source phrase, the result of insertion, too, is a phrase of the same kind as the source phrase. Paraphrase may either retract or add semantic information, just like deletion, reduction,

and insertion. As opposed to the first three simplification categories, paraphrase typically results in a new phrase structure.

deletion	a syntactic unit is deleted in its entirety
reduction	a syntactic unit is reduced by deleting parts of the unit
insertion	information is added to a syntactic unit
paraphrase	the grammatical information in a syntactic unit is restructured, without loss of semantic meaning. *Paraphrasing is essentially a case of deletion or reduction plus insertion, which results in a new phrase structure.*

Figure 6.1 Categories of simplifications found in the simplification of *Invandrartidningen*.

6.2 The simplification rules

The majority of the phrases of all categories of simplification remain unsimplified, a fact that suggests that perhaps it is not the individual simplifications but the totality of the changes in the text, which makes the text more readable. Since there are so few occurrences of each type of simplification, no effort has been spent on extracting the simplification types that are statistically significant—for each type of simplification, the simplifications which occur most often have been encoded below.

The simplifications are all presented with an example, and are briefly explained. Some examples contain more simplifications, than the simplification the particular rule it is presented under. In these cases, the simplification in question is made **bold**.

6.2.1 Simplification rules for the noun phrase

The noun phrase is the phrase, which has generated the most simplification rules. The simplification of the noun phrase appears to depend more on the function of the phrase, than on the form of the phrase.

First, the noun phrase, which consists of just a plain noun, is described.

NP(N1 + PP(P + NP(N2))) → NP2

(1) a. *resultatet av en folkomröstning*
'the outcome of a referendum'[28]
b. *en folkomröstning*
'a referendum'

This type of simplification deletes the first noun phrase in the phrase, along with the preposition of the prepositional phrase, promoting the second NP in the clause, that of the prepositional phrase.

[28] The translations are aimed at capturing the phrase structure of the Swedish as far as possible, not at providing idiomatic English equivalents of the sentences.

NP1 + VP:refl + PP(P + NP2(N)) → NP1(N + (som + AuxV + Vpass)) + VP:refl

(2) a. *tusentals **gömmer sig undan utvisning***
'thousands **hide from deportation**'
b. *flera tusen **som ska utvisas gömmer sig***
'several thousand **who will be deported are hiding**'

Here, a nominalized verb is simplified. In the original, it appears inside a post-modifying PP, but in the simplified version it is inserted in the initial NP, as the head of a post-posed passive subordinate clause.

NP(N1-s-N2) → NP(N1)

(3) a. *butikskedjor*
'chain stores'
b. *butiker*
'stores'

In this simplification, a compound noun is transformed to a simple noun. The second noun in the compound is deleted together with the *foge-s* (a linking morpheme *-s-*), leaving the first noun.

NP(N1:sg-N2) → NP(N2 + PP(P + NP(N1:pl))

(4) a. *bilförsäljningen*
'car sales'
b. *försäljningen av bilar*
'the selling of cars'

Another type of compound noun is simplified into a noun phrase consisting of a noun and a prepositional phrase post-modifier. The second noun in the compound forms the head of the new noun phrase, while the first noun in the compound forms the noun phrase in the prepositional phrase. In this process, the number of the N1 is changed from singular to plural.

The simplifications for the noun phrases with pre-posed constituents

The types of simplification described below all originate in an NP with constituents preceding the head.

NP(Det + N + subordinate clause) → NP(N + subordinate clause)

(5) a. ***de elever** som redan går på gymnasiet*
'the students who are already in high school'
b. ***elever** som redan går på gymnasiet*
'students who are already in high school'

This simplification deletes the determiner from the noun phrase, changing it from definite to indefinite.

NP(Det + AP + N:def) → NP(AP + N:indef)

(6) a. *de ekonomiska prognosmakarna*
'the economic forecasters'
b. *ekonomiska experter*
'economic experts'

This simplification deletes the initial determiner, while the head noun changes from definite to indefinite form. The head noun is replaced by a near-synonym in this simplification.

NP(Det + AP + N) → NP(N)

(7) a. *den stora arbetslösheten*
'the high unemployment'
b. *arbetslösheten* —
'the unemployment'

In this type, the determiner and adjective phrase of the original (7a) is deleted, leaving a simple head noun in the simplified phrase (7b).

NP(AP + N) → NP(N)

(8) a. ***konkreta minnen** av krusbär och saltgurka, svamp och påskbröd*
'**concrete memories** of gooseberry and pickled cucumber, mushrooms and pascha'
b. ***minnen** från barndomen*
'**memories** from the childhood'

Similar to simplification type (7) above, type (8) deletes the constituent preceding the head – the adjective phrase. Parenthetically, the noun phrase of the original prepositional phrase (8a) is significantly reduced and paraphrased in (8b), and even the preposition is changed.

NP(AP + N:indef) → NP(Det + AP + N:def)

(9) a. *nya jobb*
'new jobs'
b. *de nya jobben*
'the new jobs'

In this simplification type, a determiner is inserted, while the head noun is changed from indefinite to definite form.

NP(AP + NN:indef) → NP(NN:def + PP(P + NP(PN)))

(10) a. *svenska storföretag*
'Swedish large-scale companies'
b. *storföretagen i Sverige*
'large-scale companies in Sweden'

This simplification type covers the change of an adjective phrase, functioning as an appositive attribute of the noun, to a prepositional phrase, functioning as a predicative attribute. It is a case of paraphrase where the pre-posed adjective phrase is deleted and inserted again inside a post-posed prepopsitional phrase. In this simplification, the preposition must be determined. As in earlier cases of simplification, the definiteness of the noun changes from indefinite to definite.

The possessive behaves like the AP in simplification, as the following examples demonstrate.

NP(Det + POSS-N:indef + N:indef) → NP(N:def + PP(P + NP(N:def)))

(11) a. *ett lands kultur*
'a country's culture'
b. *kulturen i landet*
'the culture in a country'

This is a variant of simplification (10) above—a constituent preceding the noun is moved into a post-modifying prepositional phrase. In this case, the constituents are different; first, a determiner is deleted, then a possessive noun is demoted to the prepositional phrase . As in (6), (9), and (10), the phrase is changed from indefinite to definite form in the simplification. The possessive noun, too, is made definite.

NP(POSS-PN + N) → NP(N)

(12) a. *Invandrarverkets personal*
'the Immigration Board's personnel'
b. *personal*
'personnel'

In this simplification, a possessive, in this case a name of an institution, is deleted leaving the bare head noun.

The simplifications for the noun phrases with post-posed constituents

The NP with constituents placed after the head are explained here.

NP(N + PP) → NP(N)

(13) a. *eleverna i åttonde klass*
'the students in eighth grade'
b. *eleverna*
'the students'

In the simplest type of simplifying post-posed constituents, the prepositional phrase post-modifying the noun is deleted.

NP(N1 + PP(P + NP(N2 + PP))) → NP(N2 + (som + PRON + V) + PP)

(14) a. *saknaden efter människor i hemlandet*
'the longing for people in the home country'
b. *människor som hon saknar i hemlandet*
'people she misses in her home country'

In this simplification type, a nominalized verb inside a prepositional phrase is simplified into an active verb inside a subordinated clause, after an inserted *som* (that, which) and a generated 3rd person pronoun *hon* (she). The second noun phrase in the original is promoted to head of the whole noun phrase. The final prepositional phrase is moved up to the same level as the noun phrase.

6.2.2 *Simplification rules for the adjective phrase*

AP(Adj) → Ø

(15) a. ***kraftig** magsjuka*
'**severe** stomach illnes'
b. *magsjuka*
'stomach illness'

The first simplification type for the adjective phrase is the deletion of the adjective phrase.

AP(Adj1 + IK + Adj2) → AP(Adj1)

(16) a. *en **snygg, mönstrad** blus*
'a **nice, patterned** blouse'
b. *en **snygg** blus*
'a **nice** blouse'

This simplification manages the adjective phrase, in which two adjectives follow each other separated by a comma. The first adjective is retained, and both the comma and the second adjective are deleted.

AP(Adj1 + Konj + Adj2) → AP(Adj2)

(17) a. *__svår och dyr__ behandling*
'**difficult and expensive** treatment'
b. *__dyr__ behandling*
'**expensive** treatment'

This simplification is similar to example (16), but in this case, it is the second adjective that is kept, while the first is deleted. The conjunction is deleted as well.

The logic for determining which adjective is retained is rather oblique. In (16), it appears that it is the most frequent adjective that is retained, but in (17) the two adjectives appear to have approximately the same frequencies. In (17), it seems that the remaining adjective is the one most closely bound to the head of the phrase.

AP(Adv + Adj) → AP(Adj)

(18) a. *kaffet blir __allt dyrare__*
'the coffee is getting **more and more expensive**'
b. *__dyrare__ kaffe*
'**more expensive** coffee'

The adjective phrase, which contains both an adverb and an adjective, is reduced by the deletion of the adverb.

6.2.3 *Simplification rules for the adverb*

The adverb phrase is most often simplified by altering the adverb in the source phrase to a semantically similar adverb in the target phrase, or by relocating the adverb within the sentence. These transformations take place above the phrase level, and are not covered here.

The second most common simplification of the adverb is its deletion, and this type is described in (18), under the simplification of the adjective phrase.

6.2.4 *Simplification rules for the prepositional phrase*

The prepositional phrase behaves differently than the phrases of the open parts-of-speech.

For the simplifications of the prepositional phrases, as well as the noun phrases, it seems that it is their function, rather than their form that determines the simplification. Oftentimes, the preposition was simplified as a reaction to another simplification, most likely a verb.

PP(P + NP) → Ø

(19) a. *importerade ostron **från Irland***
'imported oysters **from Ireland**'
b. *ostron som är importerade*
'oysters that are imported'

The final prepositional phrase in (19) is deleted, reducing the noun phrase in which it occurred.

Paraphrase of the prepositional phrase

PP(P1 + NP(N + PP(P2 + NP1)) → som + V + PP(P1 + NP1))

(20) a. *En särskild arbetsgrupp finns **med ansvar för mångkultur***
'A specific working party exists **with responsibility for multiculture**'
b. *Det finns en särskild arbetsgrupp **som arbetar med det mångkulturella Stockholm***
'There exist a specific working party, **which works with the multicultural Stockholm**'

The prepositional phrase has been simplified into the full verb in the clause, with a shift in meaning; in (20) a, the semantic meaning in the clause was carried by the noun *ansvar* (responsibility), while the full verb in the simplification takes on the meaning – *arbetar* (works). *Som* (which, that) is inserted before the verb to create a full subordinate clause as a substitution for the prepositional phrase.

PP(P + NP) → NP + (som + V + PP(P + NP)))

(21) a. *barn i förskole- och skolåldern*
'children in pre-school and school age'
b. *lite större barn (som går i skola eller förskola)*
'slightly older childres (who are in school or pre-school)'

In this simplification, the prepositional phrase is changed into a subordinate clause with parentheses. The subjunction som (that, which) is inserted, as well as a full verb to break up the noun phrase in the original prepositional phrase. Moreover, the conjunction is changed from *och* (and) to *eller* (or).

PP(P + NP1) → NP1 + V + NP

(22) a. *asylskälen prövas i en domstolsprocess*
'reasons for seeking asylum are tried in a court procedure'
b. *domstol avgör asylärenden*
'court settles asylum cases'

In this simplification, the noun phrase in the simplified prepositional phrase is promoted to subject position, while the subject noun phrase in the original is demoted to object status. The verb in the clause is changed, and made active in the process. The change in the verb is what triggers the paraphrase of the prepositional phrase – the passive verb in the original takes a prepositional phrase complement, but not the active verb of the simplification.

6.2.5 Simplification rules for the verb phrase

As discussed in chapter 5.3.6 above, the verb simplifications need further analysis before consistent simplification rules based on the simplifications can be written.

6.2.6 *Simplification rules for the conjunction and subjunction*

Conjunctions and subjunctions behave similar to prepositional phrases when simplified, in that they themselves do not seem to trigger the simplification.

Konj + main clause → new sentence

(23) a. ***och*** *Stockholm står på tur 1998*
'**and** Stockholm is up 1998'
b. *1998 blir det Stockholm*
'1998 it will be Stockholm'

The most usual simplification of the conjunction is its deletion, exemplified in (23). In these cases, the phrase that follows it is upgraded to a separate sentence.

NP(N:indef + (Subj + subordinate clause)) → NP(N:def)

(24) a. *elever som behöver det*
'students who need it'
b. *eleverna*
'the students'

The simplification of the subordinate clause, which begins with a subjunction, is a deletion. In the simplified phrase, the preceding noun is changed to definite form.

6.2.7 *Simplification rules for the anticipatory det*

Formsubj + VKop + NP(N:poss + subclause) → NP + AuxV + (V + subclause)

(25) a. *det är polisens ansvar att*
'it is the police's responsibility to'
b. *polisen ska se till att*
'the police must see to'

Rule (25) captures the cleft subject, and replaces the anticipatory subject with the extraposed subject.

6.3 Rules for automatic simplification

The simplification rules in Figure 6.2 could be used as a starting point to create program code for simplification of Swedish. The focus in developing these suggestions for simplification rules has been to shape readable rules that are intuitively understood. Additionally, the rules should be kept general, so that they can be implemented on any material.

The grammatical changes of the simplifications are presented here in a grammatical notation, inspired by LFG, which can be readily converted into program code and implemented within a computer system for simplification of text.

6.3.1 *Rule order*

Indications of an order of simplification in *Invandrartidningen,* were not found—the simplifications seem to have been made at random.

It is, however, hypothetically feasible that a rule order could be extracted from research on the learning of languages in general, and Swedish in particular, carried out by for example Håkansson (1994) and Pienemann & Håkansson (1999).

6.3.2 *Tentative simplification rules*

The suggestions for simplification rules listed in Figure 6.2 are dependent on context, and are intended to be implemented in the order of their appearance in the code.

Constituents that can be realized as any of several different phrase types are marked by an X, and in some target phrases a whole word, *som*, is inserted. Morphological information is put within brackets, e.g. [defness=indef], and though it is not stated explicitly within the rules, it is understood that all tokens in a phrase should contain the same type of information, in order to provide congruence within the phrase. The possessive is marked as morphological information about case on the token where it occurs, e.g. n[case=poss].

Any rule that appeared in this type of program would alter the context for the subsequent rules. In effect, changing the rule order would influence the order in which the phrases in a text were simplified. This type of program would provide the rule writer with the option to manually control the rule order but would thus make the programmer's role similar to that of the human simplifier. An automatic simplification program would, however, apply these rules in the same order in all executions of the program.

np(n1+pp(p+np(n2)))	-->	np2
np1+vp+pp(p+np2(n))	-->	np1(n+som+aux+vpass)+vp
np(n1-s-n2)	-->	np(n1)
np(n1[number=sg]-n2)	-->	np(n2+pp(p+np(n1[number=pl])))
np(det+n+X)	-->	np(n+X)
np(det+ap+n[defness=def])	-->	np(ap+n[defness=indef])
np(det+ap+n)	-->	np(n)
np(ap+n)	-->	np(n)
np(ap+n[defness=indef])	-->	np(det+ap+n[defness=def])
np(ap+nn[defness=indef])	-->	np(nn[defness=def]+pp(p+np(pn)))
np(det+n[defness=indef,case=poss] +n[defness=indef])	-->	np(n[defness=def]+pp(p+np(n[defness=def]))
np(pn:poss+n)	-->	np(n)
np(n+pp)	-->	np(n)
np(n1+pp(p+np(n2)))	-->	np(n2+(som+pron+v))
np(n+subj+X)	-->	np(n)
ap(adj)	-->	ø
ap(adj1+ik+adj2)	-->	ap(adj1)
ap(adj1+konj+adj2)	-->	ap(adj2)
ap(adv+adj)	-->	ap(adj)
pp(p+np)	-->	ø
pp(p1+np(n+pp(p2+np1))	-->	som+v+pp(p1+np1))
pp(p+np)	-->	np+(som+v+pp(p+np)))
pp(p+np1)	-->	np1+v+np
konj+{main clause}	-->	new sentence
formsubj+vkop+np(n[case=poss]+subcl ause)	-->	np+aux+vp(v+subclause)

Figure 6.2 Simplification rules derived from the simplifications made of *Invandrartidningen*.

7 Discussion

Four categories of simplification were found to be present in the corpus: deletion, reduction, insertion, and paraphrase, but the majority of the simplifications in *Invandrartidningen* seem to have been applied at random. The human simplifiers were inconsistant, arbitrary, and ad hoc in their application of simplification rules. These characteristics of the simplifications are manifest in the low frequency of each simplification's occurrence. For instance, while it is understood that original source phrases are transformed into simplified target phrases, to form a simplification pair, on occasion, these human simplifiers inverted the simplification pairs and applied them in the reverse. Moreover, a given source phrase was not simplified each time it was encountered in the text—sometimes it was left unaltered. These inconsistencies in the human simplifiers' simplifications make it hard to draw any categorical conclusions about their simplification methods.

Readability

To measure the readability of the texts in the parallel corpus, the lix readability formula was used. Lix was found to be an adequate method for measuring the readability of the parallel texts used in this study, and for confirming the change in readability between the original and the simplified texts. The lix values also correspond with the difficulty degrees assigned to the articles in *På lätt svenska*. The simplification process lowered the lix values of the news articles by roughly the same amount across the four difficulty degree categories. It is still uncertain whether the phrase-internal simplifications in *På lätt svenska* directly produce the lower lix values.

Information loss and gain

The simplification pairs found in the corpus were semantically rarely one-to-one compatible with each other; in almost every instance of simplification, information and meaning was lost or gained. This information loss is incurred via deletion and reduction of all kinds of phrases, whereas the information gain is almost exclusively due to noun phrases. These noun phrases were often personal names or compound nouns, which were explained in *På lätt svenska* through appositions of inserted PP's, InfP's or NP's. This type of simplification is rather common in *På lätt svenska* and functions as a clarification of inferrences made in the original text.

To sidestep the problem of inferences, background information had been added to the simplified text to make the inferences less oblique. It is infeasible that a prototype simplification system could (or even should) be capable of handling this type of world knowledge in the early stages of its development. Nevertheless, a fully operational simplification system will ultimately require the means to process inferences.

The simplest solution to the problem of loss of meaning would be to accept it at face value. This is what would have to be done in an initial implementation of a simplification engine for Swedish. However, this acceptance of loss of meaning would result in a solution closer to a summarization system, than a simplification system.

One way to combat the information loss would be to develop a simplification system that could paraphrase the reductions and deletions with a separate sentence. The resulting sentences should consist of main clauses with simple word order.

Generation
When a source phrase is paraphrased, very often some of the information required to form a complete sentence is missing. This information is often referred to as gap-filling expressions and it must be generated. The generation of gap-filling expressions is problematic because these expressions need to fit into the context of the target phrase and the text surrounding it, while retaining the meaning of the source phrase. Gap-filling expressions may consist of just one word but might also consist of a whole new phrase. Finding these expressions is especially difficult for a purely syntactical system, working without lexical information. In any case, a system that aims to create well-formed simplifications must eventually include a module for controlling gap-filling expressions.

Form vs. function
This study investigated simplification within the phrase, with form and not function in mind. It was discovered that many phrases, for instance NP's and PP's (chapter 5.3.1 and 5.3.5 above) show different simplifications for different positions in the main clause. In light of this, it could be beneficial to conduct a study of the phrases extracted from this material, with regard to their function, rather than their form.

7.1 Further research

The simplification systems in existence today all apply their simplification rules recursively until no further simplifications of the material can be made. These systems produce resulting simplifications that are as far from the source as is possible. A system capable of applying different degrees of simplification to input texts could produce simplified texts for learners of any reading level.

Rule order could prove a fitting tool for calibrating the system's degree of simplification settings. Since research has already shown that language learners seem to follow the same order when learning certain grammatical structures (Håkansson, 1994), it could be useful to incorporate rule order, based on this research, into a simplification system. As an alternative, a tentative rule order for learning Swedish could be developed through the analysis of Swedish learner corpora, e.g. the ASU corpus (Hammarberg, 1997; Hammarberg & Viberg, 1984).[29]

7.1.1 Points of interest

- The four categories of simplification: deletion, reduction, insertion, and paraphrase, found in this material must be further investigated to ascertain if they show the same or different dynamics in their construction.
- Future studies of text simplification for Swedish should tag the *Invandrartidningen* parallel corpus for part-of-speech, in order to find and extract additional simplification rules from a much larger sample. It would be of enormous benefit to developers of simplification systems if Inductive Logic Programming (Muggleton & De Raedt, 1994) could be employed in this process.
- The preliminary simplification rules found in this thesis should be implemented and evaluated. In order to do this, the corpus needs to be tagged for part-of-speech, and the

[29] Learning texts must be adapted for the readers language level—they should not be too hard, neither should they be too easy, as simple texts can be as hard to understand as syntactically complex texts. A simplification system must reflect this.

simplification rules should be added to an existing system. For this purpose, Granska[30] would be of interest, as it is readily available online and easy to add rules to (Knutsson, 2001).

- Since lix proved to present a reliable estimation of readability of texts, it should be investigated further to determine if and how it can be integrated in a text simplification system.
- As the phrase-internal simplifications analyzed in this thesis are too few to have such a large impact on the readability values, it becomes apparent that something else influences the readability levels. The first place to turn to, then, is to simplifications above the phrase level, such as changes in word order. These are not investigated in this study, but require further research. It would also be of relevance to investigate the effect of simplifications on the paragraph level.
- The preservation of the original author's voice within the simplification is another issue to resolve, and is more important when simplifying fiction than when simplifying technical literature or newspapers.

7.2 Concluding remarks

This thesis presents a pilot study of how simplification of Swedish text was achieved by human simplifiers. This study extracted and formalized these text simplifications, outlined a typology for them, and drafted a rudimentary core of simplification rules for Swedish.

Automatic simplification of Swedish is still in its early stages, and is in dire need of basic research on both manual and automatic text simplification. Today's research on simplified Swedish has been focused on simple writing in school books, and on simplification for language impaired readers, but it is essential that text simplification be further developed with non-language impaired adults in mind.

In order to further develop automatic text simplification systems for the benefit of adult language learners, it would be advantageous to utilize existing research on language learning, and language learners' error typologies, to establish a framework for a solid simplification typology for Swedish.

[30] Granska is a grammar checker developed at NADA at KTH. Currently a project called CrossCheck is further developing it to provide writer's aid for L2 learners of Swedish (`http://www.nada.kth.se/theory/projects/xcheck/`).

8 References

Alderson, J. Charles, & Urquhart, A. H. (1984). *Reading in a Foreign Language*. Longman Inc., New York.

Almqvist, I. & A. Sågvall Hein (1996). Defining ScaniaSwedish - a Controlled Language for Truck Maintenance. In: *Proc. of the First International Workshop on Controlled Language Applications. 26–27 March 1996*. Katholieke Universiteit Leuven.

Björnsson, Carl Hugo (1968). *Läsbarhet*. Stockholm: Bokförlaget Liber AB.

Björnsson, Carl Hugo & Hård af Segerstad, Birgit (1979). *Lix på franska och tio andra språk: läsbarhetsprövning av franska skolböcker*. Stockholm: Pedagogiskt centrum, Stockholms skolförvaltning, 1979.

Canning, Yvonne & Tait, John (1999). Syntactic simplification of newspapertext for aphasic readers. In: *Proc. of the Customized Information Workshop held at SIGIR'99, Aug 19, 1999. Berkeley, CA, USA, pp. 6–11*

Canning, Yvonne, Tait John, Archibald, Jackie, & Crawley, Ros (2000). Cohesive regeneration of syntactically simplified newspaper text. *Proc. of the 1st workshop on ROMAND, Oct 19–20, 2000. Lausanne, Switzerland.*

Cedergren, Magnus (1992). Kvantitativa läsbarhetsanalyser som metod för datorstödd granskning. *Technical Report IPLab-55, TRITA-NA-P9217, NADA, KTH, 1992.*

Chandrasekar, R. and Suresh, T. (1995). Automatic text simplification. In *Proc. National Workshop on Software Tools for Text Analysis*. University of Hyderabad.

Chandrasekar, Raman (1994). *A Hybrid Approach to Machine Translation using Man-Machine Communication*. PhD thesis, Tata Institute of Fundamental Research/University of Bombay, Bombay.

Chandrasekar, Raman, Doran, Christine, & Srinivas, Bangalore (1996). Motivations and Methods for Text Simplification. *Proc. of COLING'96, Compenhagen, Denmark, poster paper, pp. 1041–1044.*

Chandrasekar, Raman, & Srinivas, Bangalore (1997). Automatic Induction of Rules for Text Simplification. *Knowledge-Based Systems*, 10: 183–190.

Dalianis, Hercules (1999). Aggregation in Natural Language Generation, *Journal of Computational Intelligence*, 15(4):384–414, November 1999.

Devlin, Siobhan, Tait, John, Canning, Yvonne, Carroll, John, Minnen, Guido, & Pearce, Darren (1999). The Application of Assistive Technology. In: C. Buhler & H. Knops (Eds.). *Assistive Technology on the Threshold of the New Millenium*, Assistive Technology Research Series vol. 6, 1999, Amsterdam: IOS Press.

Dras, Mark (1999). *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text.* PhD thesis, Macquarie University, Australia.

Muggleton, Stephen & De Raedt, Luc (1994). Inductive Logic Programming: Theory and Methods. *Journal of Logic Programming* 1994:(19, 20):629–679.

Ellis, Nick C. (Ed.). (1997). *Implicit and Explicit Learning of Languages.* London: Academic Press Ltd.

Endres-Niggemeyer, Brigitte, Maier, Elisabeth, & Sigel, Alexander (1995). How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing and Management,* 31(5): 631–674.

Fulcher, Glenn (1997). Text difficulty and accessibility: Reading formulae and expert judgement. *System,* 25(4): 497–513.

Gass, Susan M. (1997). *Input, Interaction, and the Second Language Learner.* Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Granberg, Nils (2001). *The Dynamics of Second Language Learning – A Longitudinal and Qualitative Study of an Adult's Learning Of Swedish.* Ph.D. Thesis. Dept. Of Contemporary Literature and Scandinavian Languages, Umeå University, Umeå.

Gunnarsson, Britt-Louise (1982). *Lagtexters begriplighet – En språkfunktionell studie av mebestämmandelagen.* Ph.D. Thesis, Dept. of Nordic Languages, Uppsala university, Uppsala.

Hammarberg, Björn (1997). ASU-korpusen, en longitudinell korpus av vuxna inlärares svenska. In: Håkansson, Gisela (Ed.). 1997. *Svenskans beskrivning,* 22. Lund University Press, Lund.

Hammarberg, Björn & Viberg, Åke (1984). Forskning kring svenska som målspråk. SSM Report 10. Stockholms universitet, Stockholm, Sweden.

Hyltenstam, Kenneth & Wassén, Kerstin (1984). *Svenska som andraspråk – en introduktion.* Studentlitteratur, Lund.

Håkansson, Gisela (1994). Rapid Profile – en snabbdiagnos av grammatisk nivå i inlärarspråk. In: Linnarud, Moira. *Språk – utvärdering – test. Rapport från ASLA:s höstsymposium,* Karlstad, 10–12 November, 1994.

Klare, George R. (1963). *The Measurement of Readability.* Ames, Ia.:The Iowa State University Press

Knight, Kevin & Marcu, Daniel (2000). Statistics Based Summarization – Step One: Sentence Compression. The 17th National Conference of the American Association for Artificial Intelligence AAAI'2000, pp. 703–710. Outstanding Paper Award. Austin, Texas, July 30-August 3, 2000.

Knutsson, Ola (2001). *Automatisk språkgranskning av svensk text.* Licenciate Thesis, IPLab 198, TRITA-NA-0105, Nada, KTH, Stockholm.

Kotsinas, Ulla-Britt (1982). Svenska svårt - Några invandrares svenska språk. PhD Thesis, Department of Scandinavian Languages, Stockholms univetsitet. Stockholm: Stockholms universitet.

Källgren, Gunnel (1979). *Innehåll i text.* Ord och Stil 11. Språkvårdssamfundets skrifter, Lund.

Larsen-Freeman, Diane & Long, Michael H. (1991). *An Introduction to Second Language Acquisition Research.* New York, NY: Longman, Inc.

Liberg, Caroline (2001). Läromedelstexter i ett andraspråksperspektiv - möjligheter och begränsningar. In: Nauclér, Kerstin (Ed.). 2001. *Symposium 2000 - ett andraspråksperspektiv på lärande.* Bulls Tryckeriaktiebolag, Halmstad.

Lucas, Michael A. (1991). Systematic grammatical simplification. *International Review of Applied Linguistics i Language Teaching,* 29(3): 241–249.

Mann, William C. & Thompson, Sandra A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text,* 8(3): 243–281.

Marcu, Daniel (1997). *The Rhetorical Parsing, Summarization, and Generation of Natural Language Texts.* PhD Thesis, Department of Computer Science, University of Toronto, December 1997

Marcu, Daniel (2000). The Rhetorical Parsing of Unrestricted Texts: A Surface-based Approach. *Computational Linguistics,* 26(3): 395–447.

McKeown, Kathleen, Robin, Jacques, & Kukich, Karen (1995). Generating concise natural language summaries. *Information Processing and Management,* 31(5): 703–733.

Oestreicher, Amelie (2000). *Bearbetning av tidningstext - En studie av texthantering vid sex svenska dagstidningar.* Skrifter utgivna av institutionen för nordiska språk vid Uppsala universitet 52, Uppsala.

Platzack, Christer (1974*a*). *Om läsbarhet.* In: Teleman, Ulf & Hultman, Tor G. *Språket i bruk.* Skrifter utgivna av Svensklärarföreningen 153, Lund.

Platzack, Christer (1974*b*). *Språket och läsbarheten.* Ph.D. thesis, Dept. of Nordic Languages, Lund university, Lund.

Plaza Pust, Carolina (2000). *Linguistic theory andadult second language acquisition: On the relation between the lexicon and the syntax.* European University Studies, Series XXI. Peter Lang–Europäischer Verlag der Weissenschaften, Frankfurt am Main.

Pienemann, Manfred (1998). Language Processing and Second Language Development: Processability Theory. Amsterdam: Benjamins.

Pienemann, Manfred & Gisela Håkansson (1999). A unified approach towards the development of Swedish as L2: a Processability account. *Studies of Second Language Acquisition* 21:383–420.

Reape, Mike & Mellish, Chris (1999). Just what *is* aggregation anyway? In : *European Workshop on Natural Language Generation, Toulouse, France. May 13–14 1999.*

Reichenberg, Monica (2000). *Röst och kausalitet i lärobokstexter – En studie av elevers förståelse av olika textversioner.* Göteborg Studies in Educational Sciences 149. Göteborg: ACTA UNIVERSITATIS GOTHOBURGIENSIS.

Ryhänen, Raija (1987). *Om språklig utvärdering av undervisningstexter i svenska som främmande språk.* Studies in Languages 9. Joensuu: University of Joensuu.

Siddharthan, Advaith (2002). An Architecture for a Text Simplification System. To appear in *Proceedings of the Language Engineering Conference 2002 (LEC 2002).* Location: `http://www.cl.cam.ac.uk/users/as372/cv.html`.

Srinivas, Bangalore, & Joshi, Arawind K. (1998). Supertagging: An Approach to Almost Parsing. *Computational Linguistics,* 22(3): 331–378.

Stroh-Wollin, Ulla (2002). *Som-satser med och utan som.* PhD Thesis. Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet, 58.

Teleman, Ulf (1974). *Manual för grammatisk beskrivning av talad och skriven svenska.* Studentlitteratur, Lund.

Teleman, Ulf, Hellberg, Staffan, & Andersson, Erik (1999). *Svenska akademiens grammatik.* Band 1–4. Nordsteds ordbok, Stockholm.

Wilkinson, John (1995). Aggregation in Natural Language Generation: Another Look. *Technical report, Computer Science Department, University of Waterloo, 1995.*

Wiman, Björn (1998). *Skriv lättläst.* Skrifter från Centrum för lättläst, Nr. 1. Stockholm, Centrum för Lättläst.

Digital sources

AECMA `http://www.aecma.org/Publications/SEnglish/sengbrc.htm`
(last accessed 2003-03-12)

Boeing Simplified English Checker
`http://www.boeing.com/assocproducts/sechecker/se.html`
(last accessed 2003-03-12)

Lexin `http://www-lexikon.nada.kth.se/skolverket/forord.shtml`
(last accessed 2003-03-12)

LL-stiftelsen & Centrum för lättläst
`http://www.llstiftelsen.se/stiftelsen/vadarll.html`
(last accessed 2003-03-12)

Oxford Psycholinguistic Database ftp:able version of the MRC Psycholinguistic Database, the original source of the Oxford database, is available at
`ftp://ota.ox.ac.uk/pub/ota/public/dicts/1054/`
(last accessed 2003-03-12)

PSET `http://osiris.sunderland.ac.uk/~pset/welcome.html`
(last accessed 2003-03-12)

SESAM `http://www.inv.se` (last accessed 2003-03-12)

SWordNet `http://www.ling.lu.se/projects/Swordnet`
(last accessed 2003-03-12)

WordNet `http://www.cogsci.princeton.edu/~wn` (last accessed 2003-03-12)

9 Appendix

Appendix A The raw data for the original *Invandrartidningen*.

Article	Difficulty Degree	Word Count	Sentence Count	Long Words	Words per Sentence	Percent Long Words	LIX
ART-1	B	122	10	36	12,20	29,51	42
ART-2	B	80	7	28	11,43	35,00	46
ART-3	B	105	8	29	13,13	27,62	41
ART-4	x	55	3	21	18,33	38,18	57
ART-5	x	51	4	20	12,75	39,22	52
ART-6	x	62	5	22	12,40	35,48	48
ART-7	x	39	2	13	19,50	33,33	53
ART-8	A	290	21	64	13,81	22,07	36
ART-9	B	416	35	104	11,89	25,00	37
ART-10	B	206	12	51	17,17	24,76	42
ART-11	B	32	3	10	10,67	31,25	42
ART-12	B	52	6	9	8,67	17,31	26
ART-13	B	55	5	22	11,00	40,00	51
ART-14	A	53	5	13	10,60	24,53	35
ART-15	x	306	28	87	10,93	28,43	39
ART-16	B	82	5	29	16,40	35,37	52
ART-17	B	59	4	16	14,75	27,12	42
ART-18	B	101	8	18	12,63	17,82	30
ART-19	x	39	2	10	19,50	25,64	45
ART-20	x	65	5	18	13,00	27,69	41
ART-21	B	193	14	50	13,79	25,91	40
ART-22	A	234	16	58	14,63	24,79	39
ART-23	B	188	17	36	11,06	19,15	30
ART-24	B	130	15	50	8,67	38,46	47
ART-25	C	404	37	128	10,92	31,68	43
ART-26	C	340	35	112	9,71	32,94	43
ART-27	B	228	16	60	14,25	26,32	41
ART-28	B	101	7	25	14,43	24,75	39
ART-29	B	78	6	21	13,00	26,92	40
ART-30	B	155	7	42	22,14	27,10	49
ART-31	B	758	70	106	10,83	13,98	25
ART-32	B	255	16	60	15,94	23,53	39
ART-33	B	108	8	42	13,50	38,89	52
ART-34	x	57	7	19	8,14	33,33	41
ART-35	C	128	9	47	14,22	36,72	51
ART-36	A	35	2	5	17,50	14,29	32
ART-37	B	211	18	47	11,72	22,27	34
ART-38	B	97	8	29	12,13	29,90	42
Total		5970	486	1557			
Mean					12,28	26,08	38
					Standard Deviation		7,50

Appendix B The raw data for *På lätt svenska.*

Article	Difficulty Degree	Word Count	Sentence Count	Long Words	Words per Sentence	Percent Long Words	LIX
ART-1	B	95	9	22	10,56	23,16	34
ART-2	B	65	8	26	8,13	40,00	48
ART-3	B	140	14	28	10,00	20,00	30
ART-4	x	51	3	17	17,00	33,33	50
ART-5	x	55	3	14	18,33	25,45	44
ART-6	x	51	5	15	10,20	29,41	40
ART-7	x	23	2	6	11,50	26,09	38
ART-8	A	262	30	50	8,73	19,08	28
ART-9	B	319	34	83	9,38	26,02	35
ART-10	B	186	13	41	14,31	22,04	36
ART-11	B	74	6	16	12,33	21,62	34
ART-12	B	62	5	15	12,40	24,19	37
ART-13	B	48	5	14	9,60	29,17	39
ART-14	A	52	6	13	8,67	25,00	34
ART-15	x	250	26	63	9,62	25,20	35
ART-16	B	146	12	50	12,17	34,25	46
ART-17	B	37	3	7	12,33	18,92	31
ART-18	B	45	5	9	9,00	20,00	29
ART-19	x	68	5	20	13,60	29,41	43
ART-20	x	53	6	12	8,83	22,64	31
ART-21	B	187	15	48	12,47	25,67	38
ART-22	A	183	16	28	11,44	15,30	27
ART-23	B	161	18	17	8,94	10,56	20
ART-24	B	145	16	43	9,06	29,66	39
ART-25	C	177	34	71	5,21	40,11	45
ART-26	C	233	30	73	7,77	31,33	39
ART-27	B	254	19	69	13,37	27,17	41
ART-28	B	117	10	31	11,70	26,50	38
ART-29	B	85	8	16	10,63	18,82	29
ART-30	B	68	5	13	13,60	19,12	33
ART-31	B	685	79	91	8,67	13,28	22
ART-32	B	247	19	54	13,00	21,86	35
ART-33	B	219	19	49	11,53	22,37	34
ART-34	x	39	4	11	9,75	28,21	38
ART-35	C	74	5	21	14,80	28,38	43
ART-36	A	21	2	0	10,50	0,00	11
ART-37	B	224	22	44	10,18	19,64	30
ART-38	B	125	11	23	11,36	18,40	30
Total		5326	532	1223			
Mean					10,01	22,96	33
					Standard Deviation		7,89

Appendix C The explanations for the abbreviations used in the thesis.

abbreviated form	full form
adj	adjective
adv	adverb
ap	adjective phrase
det	determiner
formsubj	expletive subject
IC	quotation mark
IK	comma
infm	infinite marker
IP	period
IT	dash
konj	conjunction
n	noun
neg	negation
nn	compound noun
np	noun phrase
num	numeral
p	preposition
pn	name
pp	preposition phrase
pron	pronoun
subj	subjunction
v	verb
vaux	auxiliary verb
vkop	copulative verb
vpart	verb particle
vpass	passive verb
extra tags	
:inf	infinite tense
:pres	present tense
:pret	past tense
:sup	supine tense
:poss	possessive