Katarina Heimann Mühlenbock

I see what you mean

Data linguistica

<http://www.svenska.gu.se/publikationer/data-linguistica/>

Editor: Lars Borin

Språkbanken
Department of Swedish
University of Gothenburg

24 • 2013

Katarina Heimann Mühlenbock

I see what you mean

Assessing readability for specific target groups

Gothenburg 2013

Data linguistica 24
ISBN 978-91-87850-50-9
ISSN 0347-948X
GUPEA <http://hdl.handle.net/2077/32472>

Printed in Sweden by
Ineko AB Göteborg 2013

Typeset in LaTeX 2$_\varepsilon$ by the author

Cover design by Kjell Edgren, Informat.se

Front cover illustration:
detail from "I see what you're saying", 2002
by Eileen Cowin ©

Author photo on back cover by Rudolf Rydstedt

ABSTRACT

This thesis aims to identify linguistic factors that affect readability and text comprehension, viewed as a function of text complexity. Features at various linguistic levels suggested in existing literature are evaluated, including the Swedish readability formula LIX. Natural language processing methods and resources are employed to investigate characteristics that go beyond traditional superficial measures.

A comparable corpus of easy-to-read and ordinary texts from three genres is investigated, and it is shown how features present at various levels of representation differ quantitatively across text types and genres. The findings are confirmed in significance tests as well as principal component analysis. Three machine learning algorithms are employed and evaluated in order to build a statistical model for text classification. The results demonstrate that a proposed language model for Swedish (SVIT), utilizing a combination of linguistic features, actually predicts text complexity and genre with a higher accuracy than LIX.

It is suggested that the SVIT language model should be adopted to assess surface language properties, vocabulary load, sentence structure, idea density levels as well as the personal interest of different texts. Specific target groups of readers may then be provided with materials tailored to their level of proficiency.

SAMMANFATTNING

I den här avhandlingen undersöks lingvistiska faktorer som påverkar texters komplexitet och därmed också deras läsbarhet. Idag ställs stora krav på individen när det gäller förmåga att orientera sig i samhället och att självständigt fatta viktiga beslut. De flesta samhällstjänster bygger numera på elektronisk kommunikation, vilket kräver en relativt god läsförmåga. Man har dock funnit att en stor andel vuxna inte kan tillgodogöra sig den typ av text som i avhandlingen beskrivs som "ordinär", utan har behov av "förenklad" text. Avhandlingen syftar till att identifiera de språkliga särdrag som kan förmodas inverka på olika målgruppers förståelse av en text.

I Sverige har man sedan 1968 förlitat sig på LIX som ett mått på läsbarhet. Med aktuella språkteknologiska metoder och digitala språkresurser har dock möjligheten ökat att göra mer korrekta läsbarhetsanalyser. I avhandlingen används en jämförbar korpus med ordinär och förenklad text från tre olika genrer för att identifiera språkliga särdrag på olika nivåer. Ytstruktur, vokabulärtyngd, meningsstruktur, idétäthet och intressegrad undersöks kvantitativt och statistiska metoder används för att säkerställa skillnader mellan ordinär och förenklad text.

De deskriptiva statistiska resultaten undersöks vidare genom automatisk textklassificering. De mest signifikanta särdragen integreras därvid i en vektormodell, där tre olika algoritmer för maskininlärning utvärderas. Man finner att en implementering av SVM (support vector machines) ger bäst resultat. Resultatet är en språkmodell för svenska (SVIT), som visar sig predicera textkomplexitet och textgenre med högre noggrannhet än LIX. I avhandlingen föreslås att SVIT kan användas för att bedöma textegenskaper på de nämnda nivåerna. Beroende på den specifika målgruppens språkliga förutsättningar och individuella önskemål i form av textgenre och tema kan personer med nedsatt läsförmåga därmed förses med lämpliga texter.

ACKNOWLEDGEMENTS

The journey from a first dawning thought to a final thesis has been long and eventful, and a large number of persons have helped me along the way. First of all I want to express my deepest gratitude to my supervisors Sofie Johansson Kokkinakis, Lars Borin and Jerker Järborg. Sofie has, in addition to her devout friendship, believed in my project no matter what and offered her continuous support and guidance. Lars Borin gave me invaluable skilled and expert input during the process. Jerker Järborg introduced me to the meaning of meaning and believed in my ability to do the job.

I am deeply grateful to Elisabet Engdahl who, apart from being an excellent graduate advisor, also made her support available in times of need. Benjamin Lyngfelt took over her responsibilities and led me with steady hand through the final stages of the dissertation. Åsa Wengelin made an excellent review of the thesis, and provided inestimable final comments.

The admittance to the National Graduate School of Language Technology (GSLT) allowed me both the financial freedom and the scientific means to finish. I am very thankful to all the supervisors, graduate students and staff of GSLT for supplying a generous, well-organized and friendly research environment. The graduate students at the Department of Swedish welcomed me promptly in the group and made me feel very comfortable.

Some people within and outside the area of computational linguistics made this journey extra rewarding. Researchers and technical staff at Språkbanken have helped out in a number of ways, sharing solutions and a never-ending faith in the importance of language resources and Thursday's coffee-breaks. I want to direct special thanks to Maria Toporowska Gronostaj who has been a compatible room-mate and an inexhaustible source of grammar knowledge throughout the years. My warmest thanks also go to Rudolf Rydstedt who assisted with technical tips and tricks, in addition to relieving chats and first-aid emergencies during the final thesis writing. Dana Dannélls, Emma Sköldberg, Dimitrios Kokkinakis, Elena Volodina, Judy Ribeck, Karin Friberg

Heppin, Karin Warmenius, Leif-Jöran Olsson, Markus Forsberg, Martin Kaså, Susanne Lindstrand, Taraka Rama, Yvonne Adesam and Yvonne Cederholm have contributed to a warm and friendly atmosphere during the years. Pernilla Danielsson jumped off the language technology train, but has remained a close and inspiring friend. Arne Jönsson, Henrik Danielsson and Johan Falkenjack at Linköping university provided essential scientific support in their respective fields. Arne invited me into his readability research group, Henrik straightened out some statistical questionmarks, and Johan explained hyperplane concepts in an understandable manner.

My former boss Paul Uvebrant at Queen Silvia's Children Hospital offered me the time needed to fulfill the work by approving a temporary leave from my position as head of DART - Centre for augmentative and alternative communication and assistive technology. Anna Carlstrand shouldered the burden as my substitute in a highly competent and responsible manner. My present boss Goran Delic has been generous and considerate during these last months of split attention. DART staff Britt Claesson, Eva Holmqvist, Gunilla Thunberg, Ingrid Mattson Müller, Jan Övrevik, Lage Persson, Margret Buchholz, Maria Olsson, Mats Lundälv, Mia Tengel Jöborn, Sandra Derbring and Ulrika Ferm paved my way into the field of assistive technology and cognitive disabilities. Without their enthusiasm, energy and versatile support, I would never have found the specific direction and goal of my thesis.

My thanks also go to all other friends who have stood by me through fail and foul. Your encouraging words and blessings have made the way a lot easier.

My love go to my parents, who I surprised by learning to read at the age of four, and who I have continued to surprise by persisting in my reading ambitions. My father introduced me to the world of books, and my mother into project management. Thank you for always being supportive and encouraging.

My by now grown-up children have promoted my work at a distance by thriving and making all family gatherings such joyful and pleasant ones. David and Katie checked my English in a very competent manner. Thank you!

And finally Mikael – without you, my journey would certainly not have reached its happy ending!

Katarina Heimann Mühlenbock
Gothenburg, March 2013

CONTENTS

1 INTRODUCTION

The ultimate goal of reading is to understand the thoughts of others. These thoughts can be more or less readily packaged, and the ease of accessing the content does not only depend upon its size and shape, but also on the recipient's ability to untie the laces. Seamless and fluent reading is no guarantee for a person's capacity to really understand a text, although it certainly is of great benefit. Many people find it difficult to orient themselves in an abundance of text at hand, and for persons with reading difficulties the problem becomes circular: In order to know what text to choose or reject, you must first understand it.

For reading to be rewarding, it requires a suitable match between reader and text. Readability metrics are superficial judgments of how easy a text is to understand, and are the fruits of readability research conducted internationally over the past 100 years. Swedish readability metrics has long been limited to the LIX formula, which is a general rule-of-thumb for an estimation of sentence and word lengths in a text. Empirical readability research suggests a range of other characteristics that might contribute to complexity and hence to comprehensibility of text materials. In the field of computational linguistics, a wide variety of resources and tools are developed for the purpose of supplementing written text with informative linguistic clues. The present thesis aims at identifying linguistic features that might replace or replenish the shallow factors in LIX by combining results from linguistics and computational linguistics. The study is corpus-based, which means that authentic texts have been consulted for identification of appropriate features. Statistical analyses have then been carried out in order to confirm or reject hypotheses about the relationship between these features and the degree of complexity across text genres and types. Finally, good results from text classification experiments have supported the theories of readability being a function of a wide range of features, observable at different text levels.

1.1 Literacy – an essential prerequisite

Historically, reading skill has a very long tradition in Sweden. Already at the end of the 17th century a canon imposed on the clerk to 'with diligence and fidelity pursue the instruction of children' *med all flit och trohet driva barnaläran*. The parish priest kept track of the efficiency of the tuition during his yearly 'household examinations' *husförhör*. All persons over the age of 15 were examined in the knowledge of their religion and the ability to read and recite the Cathechism. The priest made notes in the clerical surveys, later on consulted for confirmation and marriage. Anyone not able to read was not confirmed, and the confirmation was a prerequisite for marriage. This does certainly not imply that all married parishioners were literate in today's sense. In the 17th century, literacy was regarded as the ability to more or less fluently spell out the articles of the Lutheran Cathechism. An approval or fail was most probably dependent on the examiner, i.e. the priest and his arbitrariness, and most manifestations of reading full and proper were certainly coupled to the auditive memory and a reciting by heart. Today we regard literacy as a human world-wide right and vital for anyone living and functioning in the information society.

1.2 Readability

The reader's own comprehension of a text depends on a variety of factors unique to each person. First of all, and most obviously, the reading level of the individual must match the materials in question. The vocabulary used and the syntactical structure must correspond to the reading stage of the individual. The decoding skills must be developed to a certain degree of fluency in order to master the challenge of reading unknown words. Another prerequisite for unhampered reading is prior knowledge of topics and phenomena addressed in the text.

In a world-wide perspective, readability research has primarily been directed towards the difficulty of style of written English. A wide range of metrics for leveling texts have been established in order to meet the requirements of official and instructory publishing. For Swedish, readability research has mainly been a topic of interest for pedagogues and teachers, although a growing demand of simplified texts has arisen along with the increasing immigration and an enhanced focus on information accessibility.

The pioneer of Swedish readability research is Björnsson (1968), who conceived the LIX formula as a method to estimate lexical and syntactical difficulty of texts. The purpose of this thesis is to go beyond the superficial metrics of LIX and to suggest more sophisticated means to assess the suitability of texts for individuals with specific needs. To this end, a combination of different features at the vocabulary, syntactical and conceptual levels will be investigated and suggested.

1.3 Outline of the thesis

The thesis is organized in the following way:
Chapter 2: Background starts with an overview of factors involved in the reading process. A rough outline of the characteristics of different reading difficulties is given, followed by a discussion of atypical readers' different needs profiles. Some words are also said about neutral techniques for human reading evaluation. Levels of text analysis are suggested, as well as key concepts in the study of textual properties. The notion of easy-to-read is introduced and various facets of simplified language are exemplified, followed by an overview of different aspects of readability and a summary of common readability formulas. A multi-level partition of linguistic features is proposed and the principles behind the overall framework of feature levelings adopted in the thesis are described. A short introduction to text classification is provided. The last part of the chapter is dedicated to a discussion on the issue of matching texts to specific target groups of readers.
Chapter 3: Material describes the text corpora, lexica and computer tools employed. The notion of a monolingual comparable corpus is presented. The LäSBarT corpus, which is compiled as a subtask within the thesis project, is presented more extensively. Another focal point is a Swedish base vocabulary word list, SweVoc, also produced within the frame of the present work.
Chapter 4: Method starts with a description of the design of the study and the descriptive statistical methods used. The language feature model SVIT, based on a multi-level partition of textual properties, is introduced. The adopted algorithms for text classification are described, followed by an account of the evaluation procedure.
Chapter 5: Descriptive analysis provides an overview of the results from statistical analyses of feature similarities and significant differences in texts from different types and genres.
Chapter 6: Document classification is devoted to the presentation of re-

sults from classification experiments made on written corpus materials across genres and types. The experiments concern the performance of three different algorithms for text classification, evaluated as the difference in accuracy between a base model and the multi-level SVIT model.
Chapter 7: Concluding results provide combined results from descriptive statistical analyses and document classification. The impact of salient features is discussed, and correspondences between the original hypothesis about readability as a combination of multi-level linguistic features and actual findings in corpora are presented. Details are given about the feature selection outcome, performed on the basis of statistical significance testings and principal component analyses. Finally, the conclusive results, in terms of an enriched readability assessment model, are presented.
Chapter 8: Discussion and conclusions completes the thesis by summarizing its results, contributions to the field, and implications for further research.

2 BACKGROUND

The primary task for this thesis is to investigate factors assumed to influence the complexity and implicitly the *readability* of various texts. Determining readability involves different components that can be viewed from the qualitative, quantitative or reader-task oriented perspective, and the aim is to integrate these perspectives into a single readability model. For this reason an overview of concepts connected to the terms *reading*, *reader* and *text* will be given. The work is restricted to the analysis of texts primarily directed towards persons with cognitive disabilities, but no authentic user studies confirming or rejecting the results from analysis have been made. The first part of the background chapter will therefore be dedicated to a description of the finds from various human reading evaluation studies presented by other researchers. This overview will serve as a scientific basis for selection of textual features suitable to integrate into a language model.

Another goal is to implement and evaluate a text classifier able to decide on texts appropriate for a hypothetical target group of readers. A background to text classification will hence be provided. Readability regarded as value scales correlating with levels of difficulty will be put forward in the section presenting the most common readability formulas and text complexity measures.

The study is also intended to demonstrate how natural language processing methods can be used for text analysis, and how different computer-based language resources can be adopted for a comprehensive investigation of text complexity.

2.1 Reading

Reading is essentially the cognitive process of understandig visual codes for spoken language. Throughout history, a variety of symbolic writing systems have been invented, including ideographic, logographic,

syllabic and alphabetic systems. A very general description of each of these systems will be given below.

- At the most abstract level we find the ideograms, which represent ideas rather than words and morphemes. A person with severe language problems, such as lacking phonemic awareness, knowledge of sight words, phonics and other reading skills, can rely on some symbolic system at hand. These systems are part of the field of augmentative and alternative communication (AAC). AAC denotes all communication that is not speech, but is used to enhance or replace speech. Special augmentative aids, such as picture and symbol communication boards and electronic devices, are low and high technical solutions available for transmission of these symbols.

- Logographic systems consist of a set of logograms, which are visual symbols representing a word or morpheme. A logogram is not linked to the actual pronunciation of a specific word, which is why several languages can use the same grapheme. An example of a logographic system is the Bliss language created by Charles Bliss (1949) as an effort to bridge the gap between different cultures. Sight word reading is a logographical process that takes place when a word is immediately recognized as a whole and does not require phonological analysis for identification.

- Syllabic systems refer to sets of written symbols for consonants, vowels or syllables. Japanese is the best-known example of a language using syllabic writing as one of its writing systems.

- In the alphabetic systems, characters or combinations of characters are the symbols used to represent the speech sounds of a language. Alphabets represent phonemes with more or less transparency depending on the language. Alphabetic reading is the subject of the present thesis.

In the Latin-based writing system of standard contemporary Swedish, the alphabetic characters include the upper and lower case forms of twenty-nine letters. Nine vowels and twenty consonants (in the most recent SAOL), individually or in combination, represent approximately twenty-seven phonemes in Swedish (Elert 1997). In addition to this the graphic system contains punctuation marks and a few other symbols such as those for numerals. Swedish is not very consistent in the correspondences of spelling to sounds. It is to be found somewhere at the

middle of a continuum between English, which is very inconsistent in grapheme-phoneme correspondences, and Finnish which is highly regular (Aro 2004). The basic challenge for a beginning reader is to map the graphical representations to the language sounds in order to retrieve the intended words. With the increasing literacy comes the capacity to read sequences of words forming phrases, sentences, paragraphs and entire texts.

Although most children learn to talk and successively learn to read without any major conscious effort, the path from written symbols on paper to a mental representation in the brain is regarded as one of the most complicated motor skills that we acquire in developing from toddlers to school children. From an evolutionary perspective, the human brain has existed for approximately 60,000 years, while written representations of words has been in use for only 5,000 years. There are countless theories and explanatory models for illustrating the reading process. The remaining part of this section will concentrate on a few that have direct bearing upon the overall perspective of this thesis.

Reading acquisition research has a long history as part of experimental psychology, leading to various hypotheses about the nature of and relationship between the different modules involved. The bottom-up reading model accentuates a single-direction, part-to-whole processing of text, that gives little emphasis to the influences of the reader's world knowledge, contextual information, and other higher-order processing strategies. The top-down model, on the other hand, advocates a view where the process proceeds from whole to part when the reader identifies characters and words in order to confirm a previous assumption about the meaning of the text. In-between these views lays the interactive model which recognizes the collaboration of different processes simultaneously throughout the reading process.

A convincing standpoint has been taken by Hoover and Gough (1990), Gough and Tunmer (1986), and Juel (1988). They argue that what distinguishes reading is that the reader is exercising abilities involving patterns of higher mental processes that may be developed; persons that could not read have also used these processes. These abilities would respond to graphic rather than acoustic signals. According to this view only two components are involved, *decoding* and linguistic *comprehension*, and the underlying assumption is that this complexity can be made simple by dividing it into two parts of equal importance. A further assumption is that this can be expressed as a mathematic equation where decoding (D) and listening comprehension (C) are the factors that when multiplied produce reading comprehension (R) as a result. As opposed

to the additive case, i.e. where *R* is regarded as the sum of the *D* and *C* factors, the multiplicative case yields zero if one of the individual constants equals zero. An implication of this reasoning is that each skill is necessary but not sufficient on its own. Even if it is well established that reading comprehension is some function of decoding and listening comprehension, this *simple view of reading* makes the stronger prediction that the effect of either skill on reading ability depends on the reader's level of competence in the other skill (Gough and Tunmer 1986; Hoover and Gough 1990; Tunmer and Hoover 1992). This view will be fundamental for the coming reasoning about readability and reading difficulties.

What Halliday (1985) called language strata, has been reformulated by Goodman and Goodman (2009) into a leveling of three cuing systems, or levels, that readers use in making sense of print. By using these cues at the same time, a reader is supposed to comprehend written language. The basic, observable level, is the *signal level*, which includes the phonology, the orthography and the phonic relationships between them in alphabetically written language. The *lexico-grammatical level* comprises both the vocabulary and the grammar of the language, while the *semantic level* obviously contributes with the knowledge necessary to convey meaning to a certain text. Making sense of print involves a set of psycholinguistic strategies for using cues from these levels simultaneously, according to the authors.

In the model of Wren (2001) language comprehension and decoding is conceptually illustrated as two cooperating areas, both comprising separate elements and also interacting at different levels, ranging from relatively low level for phonological decoding to high level for inference generation based on background knowledge. Wren's reading model is illustrated as a pyramid, where background knowledge, phonology, syntax and semantics are integrated into the language comprehension area. The decoding area, i.e. recognition of written representations of words, is constituted by different cognitive elements such as word decoding, which at base level is supposed to act through concepts about print. This module is, for readers of alphabetic writing systems, built by letter knowledge and knowledge of the alphabetic principle. Another basic element of the decoding area is phonological awareness; a central concept in explaining variation in early reading acquisition (Jorm and Share 1983). The two areas diverge at a higher level, where linguistic knowledge, cipher knowledge and lexical knowledge interact into the second highest level, which is language comprehension and decoding. At the top of this pyramid we find the reading com-

prehension level. While Goodman and Goodman (2009) describe a process that is circular and incremental, Wren's pyramid concept seems to illustrate a process where different abilities are used as static building blocks. It is beyond the scope of this thesis to dive deeper into the question of whether there exists a single explanatory model for reading comprehension. Suffice it to say that the field has been profoundly investigated in an abundance of studies on humans in oral test situations, and more recently in neurocognitive experiments. While this work is dedicated to the matter of finding suitable literature for persons having some reading performance deficiencies, the earlier mentioned theory of Gough and Tunmer (1986) will be kept as a general framework for the description of reading component skills. Although it has the reputation of a *simple view of reading*, it includes all components that are generally regarded as crucial for reading performance.

From a developmental perspective, oral language is the foundation on which literacy initially builds, and the listening comprehension rests on the ability to derive meaning from spoken language. The syllable is the primary linguistic processing unit, and each syllable making up a word can be decomposed into onsets, rimes, and phonemes in a hierarchical fashion. Developmentally, spoken language precedes printed language, on the individual as well as the evolutionary level. Each language has its own specific rules for the syllabic structure. The common view is that syllables have a linguistic organization between vowels and consonants in linear order, following the phonological rules of the specific language (Colé, Magnan and Grainger 1999). For Swedish, the typical pattern is an initial consonant cluster, followed by a vocal, then a final consonant cluster. Syllable counts reflect word length based on phonological principles, but they also require a preprocessing of the textual representations. Lexicographical syllabification can serve two different purposes, either as an indicator of the orthographical hyphenation, i.e. where to break at wordwrap, or as a marker of the internal structure of a word (Svensén 2004). The latter case is to be regarded as a morphological rather than phonological marker. Researchers have found it plausible that syllables do play a role in visual word recognition. There is evidence for the reality of syllables in mental representations of words (Yap and Balota 2009). Empirical evidence from different languages concerning phonological development and reading development in children has shown that the development of reading depends on phonological awareness. It has been shown that distinctive reading strategies emerge for different languages due to variances in both syllabic structure and grain size of lexical representations by

which phonology is represented by the orthography (Goswami 2008). Words are composed of sequences of phonemes, and the phonemes are grouped together into individual words. Children acquire more than 14,000 words between the ages of 1 and 6 years (Dollaghan 1994), and the phonological awareness is crucial for the ability to detect and manipulate the component sounds that compose these words. In addition to the letter-to-sound rules, there are several aspects that affect the development of phonological representations of different words. The phonological neighborhood density (Goswami 2008) is one of these factors. It is the count of similar-sounding words to a particular target word.

Turning to the linguistic form of words, the easiest, and most obvious way to make some statement about a text is to perform a simple word frequency calculation. In reading, one of the most robust findings in the word recognition literature, is that frequency influences the efficiency with which units are processed. Numerous experimental studies have shown that the lexical latencies decrease as the whole word frequencies in print increase. To mention a few, Just and Carpenter (1980) demonstrated greater cognitive loads while readers were accessing infrequent words. Later on Juhasz and Rayner (2003) showed in eye-tracking studies that both word frequency and familiarity showed an early but lasting influence on eye fixation durations. Effects of whole word surface frequency are interpreted to reflect processing at the level of the whole word (lexical processing), while effects of stem or lemma frequency provide a means to measure sublexical processing efforts.

The *dual-route model* of word recognition assumes that written language processing is accomplished by two distinct but interactive procedures that are referred to as the *lexical* and *non-lexical routes*.

It is not possible to discuss word frequency without mentioning the early findings of Zipf. With the amount of data available at the time, Zipf (1932) observed that the distribution of word frequencies in English is an inverse power law with the exponent very close to 1, if the words are aligned according to their ranks. That is, if the most frequently occurring word appears in a text with the frequency P(1), the next most frequently occurring word in the same text has the frequency P(2), and the rank-r word has the frequency P(r), the frequency distribution can be written as

$$P(r) = \frac{C}{r^{\alpha}} \tag{1}$$

where C ≈ 0.1 and $\alpha \approx 1$, or more simply, the most frequent word will occur approximately twice as often as the second most frequent word,

three times as often as the third most frequent word, etc. Furthermore, Zipf attempted to explain a variety of human traits and behavioral patterns in this way, including for instance the population ranks of cities, structure of music, and income distribution. The underlying notion was that humans act in ways that require them to make minimal effort (Zipf 1949). In fact, Baayen and Lieber (1996) investigated the relation between meaning, lexical productivity, and frequency of use, and showed that differences in semantic structure was reflected in probability density functions estimated for word frequency distributions.

In authentic reading assessment tests, non-words are often used because in contrast to real words they are equally unfamiliar to all subjects. It has been found that reading familiar words differs from reading non-words in two ways. First, word reading is faster and more accurate than reading of non-words. Second, effects of word length are reduced for real words, particularly when they are presented in the right visual field in familiar formats (Grigorenko and Naples 2007).

When it comes to visual word recognition, it has been shown in experiments that lexical decision, perceptual identification, and semantic categorization tasks can be performed successfully on the basis of orthographic and/or semantic information alone. When a person is faced with a task involving control of orthography, the manipulation of phonological variables have been shown to have a large impact (Colé, Magnan and Grainger 1999).

The cognitive process by which a person verbally produces or confirms semantic information about an object or the image of an object is under constant re-evaluation. Theories built upon different dimensions of categorization (Rosch 1978) have later on been followed by models where network simulations are used to defend a pure connectionistic view (Rogers and McClelland 2004). Regardless of the theory one adheres to, principles involving the presence of a semantic base categorization seem to be mutually agreed upon. The *lexical base level* has been defined as the hierarchical level where the maximal degree of information (informativeness) and the maximal degree of distinction (distinctiveness) coincide (Murphy and Lassaline 1997). The inflected word forms in categories that are too general are per definition less informative, while more specific categories are informative, but not particularly distinctive because they are abstruse.

It also seems certain that children can name many objects at the correct base level before they can name them on a more general or specific level, which could mean that children learn base level categorization first in language development (Brown 1958; Chapman and Mervis

1989). Similarly, researchers have found that people affected by progressing dementia keep the base level categories longest during the course of the disease (Hodges, Graham and Patterson 1995).

2.2 The reader

Fish (1970) introduced the theory of *affective stylistics* which was built on principles of readers' emotive responses to texts. By having a reader-oriented perspective, the author creates a text "assisted" by a hypothetical reader. An implication of this view is that the content of a text has to be presented in different manners depending on the individual reader and his/her purpose of reading.

An individual's reading skill level rests on many different reading components. Assessment techniques of *reading skills* have traditionally been limited to verbal tests, but more recently neurophysiological evidences from brain activity measurements and eye-tracking finds have shed new light on old theories. The advantages of these techniques is that they are neutral. Neuroimaging studies have in fact shown that different cognitive processes are activated depending on the reading task. Reading of sentences involve other processes than single-word-reading, even after eliminating the contribution from word-level processes inherent to the task (Cutting et al. 2006). This means that the task of reading a sentence is not compositionally proportionate to the task of reading separate words, storing them in the working memory and analyzing them according to syntactical clues.

In general electroencephalogram (EEG) technique, electrodes attached to the scalp allow researchers to measure the brain's electrical activity. Several experiments for different languages show (Zaidel, Hill and Weems 2008) that lexical variables had physiological correlates, observed as EEG gamma signal changes as a function of lexicality (wordness), semantic (word frequency), orthographic (word regularity), and phonological (nonword pronounceability) variables. Another method to trace brain activity and to identify the localization of processes in reading is by using functional Magnetic Resonance Imaging (fMRI) (Richards 2001). Although fMRI has been widely used as a technique for applications in mapping motor, visual and auditory systems, it has a major drawback which is to be found in the time resolution of the method. As stated earlier, the reading process is based on the information processing system and on the stages of activation from perception to processing. In order to optimally trace the different stages of activation on-line

during reading, the time units of the brain sample must be very small. The disadvantage of fMRI in this respect is that it allows sampling only within relatively large frames of time measurement. Thus, the neurophysiological technology recently adopted in reading research capable of overcoming some of these resolution limitations is ERP. The basic idea behind the ERP methodology is that different stimuli of interest cause different brain waves. These differences can be used just like any other dependent measure in research on language processing, similar to behavioral measures of text comprehension rates and reading time. Many new finds in ERP studies give valuable information about human parsing, such as the process of mappings of form onto meaning (Friederici et al. 2006), comprehension of simple transitive sentences (Bornkessel-Schlesewsky and Schlesewsky 2008), and application of grammatical principles during human parsing (Bornkessel, Schlesewsky and Friederici 2002).

Much attention has been paid to studies of eye movements in reading and information processing tasks during the last 30 years. One of the most exhaustive overviews in this field is presented by Rayner (1998). Most eye tracking studies aim to identify and analyze patterns of visual attention of individuals, when performing specific tasks. In these studies eye movements are typically analyzed in terms of fixations and saccades. During each saccade visual sharpness is suppressed, so we can only perceive and interpret something clearly during fixations. The light sensitive surface of the eye, the retina, is not equally sensitive everywhere. A limited part of the visual field in the eye, called the foveal area, registers details clearly, while the much larger, peripheral area of the visual field is better adapted to low light vision. During each fixation individuals place the foveal area on the feature which is most interesting to extract information about. There are several techniques to detect and track the movements of the eye, the most commonly used is Pupil Centre Corneal Reflection (PCCR). Basically, it uses a light source to illuminate the eye causing highly visible reflections, and a camera to capture an image of the eye showing these reflections. Advanced image processing algorithms and a physiological model of the eye are then used to calculate the position of the eye and the point of gaze.

Generally, reading skill is closely connected to short- and long-term memory processes. Reading difficulties may be caused by insufficient working memory capacity or poorly organized long-term memory. The relationship between working memory, or particular components of it, and aspect of oral language development has been subject to different research studies. Baddeley (1990) claimed working memory to support

language processing in two ways, the first acting as storage for information as language is being processed. The second way would be to support information processing in supplying working space for the necessary linguistic operations. The concrete effect of the working memory capacity on language skill would then be an influence on *vocabulary acquisition* and *comprehension of language*. Working memory may support phonological learning, which in turn benefits vocabulary acquisition. Acquiring a new word involves both a long-term semantic construction of the underlying concept and its association to a particular phonological sequence, that is a possible word in the language (Rondal and Edwards 1997). The storage capacity of working memory would play a limiting role in the buffering of strings of incoming words for a time, pending the construction of more durable representations of the structure and meaning of the sentences. An ample storage space would then be an important asset for language comprehension.

The simple view of reading provides an account of the different forms of *reading difficulties* (Gough and Tunmer 1986; Tunmer and Hoover 1992). Depending on the magnitude of the two factors *D* and *C* mentioned earlier, a schematic categorization of different forms of reading difficulties can be illustrated as in figure 2.1. The model predicts that a person that can understand a text when it is read aloud, but is unable to decode its written representation, might be afflicted by some degree of *dyslexia*. On the other hand, a person who is a skilled decoder of printed text but unable to comprehend the same message in spoken form might have some form of *hyperlexia*. The lower left-hand square of the figure denotes persons that have problems within each of the two preceding areas.

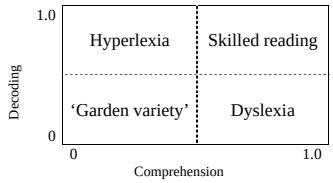


Figure 2.1: Categorization of different forms of reading difficulties. From Tunmer and Hoover (1992)

From the atypical reader's perspective, one unique adaptation of a text into some kind of easy-to-read format is no guarantee for its accessibility. Persons with intellectual disabilities, and those suffering from autism, aphasia, or dyslexia, people who are deaf from childhood, the

elderly and second-language learners all have their specific needs in terms of reading materials. In an ideal world, a reader should be able to access texts tailored to compensate for his or her individual linguistic deficits. As will be further discussed later on, a person who has dyslexia has quite different supportive needs than a second-language learner immigrant or a visually impaired person. Natural language processing (NLP) technology brings potential to adapt textual information to the needs of specific readers.

The disability movement exponents express different ideas regarding the value of identifying persons belonging to certain groups. The concept of a "group" is here to be interpreted in its metaphorical sense, where we assign a set of people certain common properties, namely that they exhibit reading difficulties. These difficulties may in turn have different etiologies, where a medical diagnosis or ethnical background gives rise to additional grouping. By way of example, we will envisage a hypothetical target group of readers consisting of persons characterized by mild intellectual disability.

In clinical terms, the diagnosis *mental retardation* (MR) (World Health Organization 2008), generally assigned to 2-3 % of the population, is divided into six grades of severity, with regard to social functioning, adaptability and intellectual capacity. Persons diagnosed with the mildest form acquire language with some delay, most achieve the ability to use speech for everyday purposes and to hold conversations. The main difficulties are usually seen in school work, and many have particular problems in reading and writing. Persons with mild mental retardation (diagnose code F70 in ICD-10 (World Health Organization 2008)), i.e. IQ scores 55-70, account for 65 to 75 % of all cases with MR, which means a prevalence of 1.5 % of the population nationwide (World Health Organization 2011). This can be regarded as a relatively high prevalence for a chronic condition. Down's syndrome has long served as the major reference for moderate and severe MR conditions, although various syndromes related to MR may have specific language profiles. Similar to the normal population individual variations evidently exist across syndromes at similar levels of MR, and also within a syndrome. The present work will nevertheless address persons with mild MR and able to read as a specific group of persons with some general language difficulties in common, which makes them eligible to be included into a "group", although with large internal variations. In table 2.1 (from Rondal and Edwards (1997)) three syndromic profiles for speech and language are presented.

Language aspect	**Down's**	**Williams**	Syndomes Fragile-X
Phonetico-phonological	−−	+	−−
Lexical	-	++	+
Thematic semantic	+	+	?
Morphosyntactic	−−	+ (comprenhension?)	-
Pragmatic	+	−−	-
Discursive	−−	+	-

Table 2.1: Three MR syndromic profiles for speech and language. Key: +(+): relative strength; -(-): relative weakness; ?: insufficient data available.

Some literacy impairments seem to distinguish people with mild MR from other low-literacy adults. Although IQ is irrelevant to the definition of reading disability *per se* (Siegel 1989), it seems that IQ score is correlated with reading in subjects with mild mental retardation (Cohen et al. 2001, 2006). Limitations in verbal short-term memory in combination with slower speed of semantic encoding results in loss of units from the working memory before they are processed (Feng, Elhadad and Huenerfauth 2009). They are also often limited in their choice of reading materials, due to a mismatch between their interests and their literacy, which in turn has a negative impact on their reading-skill practice. In a study conducted by Feng, Elhadad and Huenerfauth (2009) participants were asked about their preferences regarding reading materials. The majority mentioned news and information that would be relevant to their daily lives. A Swedish study was carried out in order to evaluate the easy-to-read newspaper *8 SIDOR* (Göransson 1985). Forty subscribers, diagnosed with MR, were interviewed in order to have their opinions regarding the general quality of the newspaper and personal preferences regarding the content. The conclusions in this report were that the reading interests of the interviewed persons largely correspond to that of the "ordinary" reader.

The present study will rest on results from statistical analyses of sentences and pseudodocuments, which are not directly portable into theories of how a reader would process isolated words. Nevertheless, since different target groups of readers experience dissimilar reading difficulties, it is likely that an NLP approach considering characteristics at various textual levels depending on the intended reader audience would

be successful. In order to pave the way for NLP solutions tackling a wider range of reading problems, some things will also be said about linguistic features related to single words and sentences.

2.3 The text

The term *text* will be used throughout the thesis as a cover term for natural written language of any length, and texts will be studied from particular situations of use. Normally, one would start by viewing the text from a holistic perspective, i.e. the broadest possible context through which the complexities, interconnections, and interdependencies of a text can be comprehended. Structural cohesion is one important factor to consider within the framework of text theory, based on more or less clearly pronounced correlations between objective and text or text and efficacy (Melin and Lange 2000). These researchers also argue that the only textual property that has repeatedly been tested scientifically is readability, and in their opinion it is clear which syntactical relationships affect and complicate the reading process. Studies have shown that a reader's understanding of a text increases if the text in some way is given voice (Reichenberg 2000). Text comprehension will also be further enhanced if aspects such as cohesion (Siddharthan 2006) and clear causal relations (Reichenberg 2000) are taken into consideration. Other textual features can emphasize aspects of a text's content or structure without adding to the content. Such features, i.e. explicitly or implicitly marked signals, comprise discourse markers, titles, headings, summaries and typographical cues. Such signaling makes sentences longer and readability scores soar, but eases readability for readers employing the structure strategy and looking for such signals. In general terms, one would say that the reader makes use of all these aspects when he or she "reads between the lines". Even though discourse markers have a significant impact on readability, they are not explicitly annotated in the corpora and will thus not be specifically addressed in this study. Moreover, texts studied at the discourse level would demand them to be analyzed from beginning to end and not in chunks of equal size which is the case for the present material. Emphasis will instead be put on quantitative linguistic features signaling text complexity at other levels, as well as text genre and type properties.

Text varieties and the difference among them constantly affect peoples daily lives (Biber and Conrad 2009). The earlier mentioned easy-to-read format is a text type characterized by simple vocabulary, short-

ened sentences and reduced linguistic complexity. Lundberg and Reichenberg (2008) found Swedish easy-to-read texts to present some common characteristics, i.e. the texts were generally short, long and short sentences alternated, there were few foreign words, long nouns and passives. Stylistically, the texts were characterized by clear causal relationships and the sentences were linked by connectives. However, as is the case for many other central terms in connection to text research, no general consensus concerning the use of *easy-to-read* exist. In what follows, texts labelled as easy-to-read (ETR) will be referred to being of an *easy-to-read type* as opposed to texts of *ordinary type*. Texts will also be studied from the perspective of *genre*. In literary studies the concept of genre denotes varieties of literature that employ different textual conventions. The present study will consistently use the genre perspective for fiction, daily news and information texts.

As pointed out by Biber and Conrad (2009), general consensus also lacks concerning the use of the terms *register*, *genre*, and *style*. The distinction between register and genre made by Biber is that *genre* perspective emphasizes the conventional features of whole texts, while *register* variation emphasizes variation in the use of linguistic features. The term *style* has been used for a wide range of concepts. In a general perspective, as applied in literary studies, it is a way of describing characteristic modes of using language. In order to avoid confusion, the term *genre* will be used according to the definition of Biber and Conrad (2009): "The genre denotes varieties of literature that employ different textual conventions". For the sake of simplicity, we stick to the term *type* in order to distinguish between easy-to-read and ordinary texts.

The Swedish terms "Lättläst", "Klarspråk" and "Klartext" have achieved a more or less established status as trademarks for different concepts within the same range of efforts to achieve textual clarity. Although the terms are meant to distinguish between separate initiatives or works promoting readability, they are not very transparent for the non-expert. *Lättläst* 'Easy-to-read', is broadly controlled natural language (CNL), a subset of natural languages obtained by restricting the grammar and vocabulary in order to reduce or eliminate ambiguity and complexity. The term *Klarspråk* 'Plain Swedish Language', denotes official texts written in a neat, simple and understandable language, and is promoted by the Swedish Language Council. *Klartext* 'Plain text' is the title of a Swedish radio show, broadcasting news in a simple and understandable fashion.

Two text types will be investigated from a complexity perspective. The first type consists of texts in the easy-to-read format, and is ex-

pected to be least complex at crucial language levels. The second type are ordinary texts retrieved from a representative corpus of Swedish texts, assumed to be more linguistically complex. We will dedicate a separate chapter to a description of the characteristics of each of these text two types.

2.3.1 Text classification

Classification is the task of assigning objects to one of several predefined categories. Within the literary domain, a vast amount of *text classification* methods have been developed and used for decades. The simplest bag-of-word model, where a text document is converted into a vector of word counts, is often used for text representation when no prior knowledge is available with regard to specific classification tasks. For a complex document, it results in a high dimensional vector space, where many features are irrelevant. In order to reduce the computational cost and produce a classifier with good generalizability, feature reduction is normally performed as a primary step, usually by means of statistical feature selection. Compared with traditional, or hard classification, soft classification provides more information about the probabilities that one attribute set belongs to a specific class. An approach of using a soft classifier trained on ETR texts and ordinary texts is described by Sjöholm (2012) and Falkenjack and Heimann Mühlenbock (2012). The results show that almost all documents in the test set had slightly different probabilities of belonging to either class. However, in order to confirm the accuracy of this approach, appropriate training materials previously ranked by human readers according to degree of readability is needed.

Computational analysis tools have been used for tasks such as authorship attribution and stylistic analysis of topics, styles and text genres. Automatic text classification methods provide other approaches to these and other text analysis problems. Two popular algorithms, the Naïve Bayes and support vector machines (SVMs) have been found to work well, and a number of studies have tested these and other methods for topic classification tasks on benchmark data sets. Classifiers are mostly evaluated by the measure of classification accuracy. High classification accuracy provides evidence that some patterns have been inferred to separate the classes. Studies performed outside the literary domain indicate that SVMs generally perform better than Naïve Bayes classifiers (Joachims 1998). Yu (2008) reported high accuracy in liter-

ary text classification for both algorithms, but also that the Naïve Bayes classifier outperformed the SVM classifier due to different feature selection ranges. This in turn caused a divergence in the choice of relevant characteristic of the target classes. Furthermore, it was recommended that the choice of classification method and feature selection procedure should be carefully considered. It also emphasized that empirical experience on classification methods obtained from one domain is not directly portable into a new domain.

In the present study hard classification is performed, defined as the task of learning a classification model that maps each attribute set X to one class label Y. It serves as a descriptive model for two specific purposes; the first being to explain which features define a text to be ETR, and the second to explain which features distinguish text genres. It might be that individual classification algorithms perform differently depending on the text genre and/or text type analyzed. Thus, the optimal classification algorithm for each classification task will also be presented.

2.4 Readability

There are almost as many definitions of readability as there are experts to define it. The major point of disagreement seems to be to which extent the human reader is to be included in the model. In the categorization made by Klare (1963) the definitions are made up by three major groups:

1. To indicate legibility of either handwriting or typography
2. To indicate ease of reading due to the interest-value or the pleasantness of writing
3. To indicate ease of understanding or comprehension due to the style of writing

One definition of the concept *readability* is expressed in the large lexical database WordNet (Miller 1995; Fellbaum 1998a, b): *The quality of written language that makes it easy to read and understand*, i.e. it might be interpreted as an intersection of Klare's third and second category. An earlier and more wordy definition is proposed by J. Chall, cited in Dale and Tyler (1934): "The sum total (including all the interactions) of all those elements within a given piece of printed materials that affect

the success a group of readers have with it. The success is the extent to which they understand it, read it at an optimal speed, and find it interesting." Some issues often related to ease or difficulty of reading are connected to layout and design of written materials. These are considered to fall outside the scope of this thesis, as they are principally taken to promote the *legibility* of a text, which in turn is secondary to readability.

Quantitative measures of readability are easier to perform computationally than qualitative, as they are purely descriptive, not interpretative. Readability measures are devised to form a link between the quantitative textual surface properties and the qualitative characteristics. The question whether these links are valid interpretations of real facts or not can be answered either through human studies or by a comparison between different materials already qualitatively and quantitatively evaluated. Traditional readability formulas utilize similar forms of quantitative analysis to assess the reading level of a text, but fail to consider factors such as the skill or interests of the specific reader.

The Swedish researcher Platzack (1974) considers readability to be a meaningful property only within texts conveying information, since these texts are expected to *provide the interested reader with maximal information against minimal effort* (author's translation) (Platzack 1974: 17). Platzack refers to Cassirer (1970), who argues that a characteristics of nonfiction as opposed to fiction is the possibility to separate language meaning and language form.

Readability, according to Platzack, is a function producing a measurable output in terms of effort (E). The input, or arguments, of the function are: *Content* (C), *Typography* (T), *Language* (L), *Reader* (R), and *Understanding* (U), and a pseudo formula is constructed in this way:

$$E = f(C, T, L, R, U) \tag{2}$$

The effort (E) is measured in terms of reading speed. "If two linguistically different but otherwise identical versions of a text are read and equally understood by two similar groups of trial subjects, the version which on average was read the fastest is also to be judged as read with the least effort" (author's translation) (Platzack 1974: 22). Experiences of a text differ depending on the reader's prior knowledge, which obviously affects the content (C) factor. Other points made by Platzack is that the typographical factors (T) mentioned concern fonts and line length, and that the reader factor (R) refers to a person's reading skill or ability rather than to the individual himself. The understanding factor (U) is finally to be checked by questionnaires related to content.

Although this formula is meant to be exhaustive, Platzack also admits that effort is highly correlated to the reader's interest and frame of reference.

A qualitative approach for Swedish has also been taken by for instance Falk (2003), in guidelines addressing professionals writing for the easy-to-read audience. It is, however, not clear whether these rules of thumb have emerged through intuition or not. Sandberg, Spånning-Westerlund and Wejderot (2005) have reported interesting finds from a project involving persons with different types of reading difficulties, although made on a small-scale basis.

2.4.1 Quantitative readability measures

When turning to readability indices, the questions whether some materials is easy or difficult to read are put to the materials itself and the answers sought in an analysis of it. The tricky part is to decide which questions to put, and how to analyze the answers, i.e. to define a criterion. Another challenge is to choose the most representative materials. During early readability research, factors studied usually arrived from intuition, personal experience and surveys of opinion. One condition for definition of a readability factor is that it must be easily operationalized and possible to combine into a formula. Most readability formulas aim to calculate some measure of syntactic complexity and semantic difficulty by way of surface features. Normally, syntactic complexity is sought in the sentence length and letter or syllable count or word frequencies to mirror the semantic complexity. A readability formula is mostly a regression equation, based on counting and weighting of the most significant internal factors. The degree of relationship between the factors is normally expressed by a coefficient of correlation.

Research on readability started in the 1920's and had its peak during 1930 to 1960. Studies were mainly carried out in the US on American English (Lively and Pressey 1923; Vogel and Washburne 1928; Lewerentz 1929; Morriss and Holversen 1938; Dale and Chall 1948; Flesch 1948), predominantly performed as quantitative associational studies on shallow linguistic features. Still, it is necessary to bear in mind that computations carried out on large data-sets were not easily performed and that even calculations such as mechanical counting demanded a high degree of manual labor. These manual calculations are obviously easier to perform on enumerable and unambiguous units.

A very thorough investigation was carried out by Gray and Leary (1935), who examined more than 200 style elements and the relationships between them. By combining variables that were highly predictive but not related to each other they created readability formulas with five variables and a high correlation to reading-difficulty scores previously assigned by informants.

In the 1970's a simplified view suggested that most variables used in readability formulas contained a semantic (meaning) measure, such as difficulty of vocabulary, and a syntactic (sentence structure) measure, such as average sentence length. By virtue of this, Björnsson (1968) presented his readability formula *LIX* 'Readability Index' for Swedish, based on only sentence and word length. An overview of the variables considered by predominant readability researchers is presented in table 2.2.

2.4.2 Readability indices and formulas

This section is dedicated to a description and analysis of international and Swedish readability indices and formulas that have had the highest impact over the years. In the following, the term *readability index* will be used for denoting readability as a numerical scale by means of which *readability levels* can be compared to each other. A *readability formula* is a set of symbols denoting the most significant internal factors – the *readability variables* – and a coefficient of correlation, expressed as a mathematical rule. It is usually a regression equation based on a counting and weighting of the most significant elements in predicting a criterion.

Seven different readability formulas will be described, where the only variations seem to regard whether they operate on character or syllable counts. From this on, the following notation will be used:

	Word length	In word list		Hard words	Image bearing sensory words	Personal words		Prepositions	Different technical words	Different words	Prepositional phrases	Sentence length	Simple sentences – empirical eval	Context	Ideas
	letters syllables	Dolch Dale Thorndike				nouns	pronouns								
Thorndike (1921)		■													
Lively and Pressey (1923)		■		■						■					
Vogel and Washburne (1928)		■						■		■			■		
Patty and Painter (1931)		■								■					
Dale and Tyler (1934)	■			■	■	■			■		■	■			
Gray and Leary (1935)	■			■		■				■	■	■			■
Morriss and Holversen (1938)														■	■
Yoakam (1939)		■													
Flesch (1943, 1948)	■					■						■			
Lorge (1944)				■			■				■	■			
Dale and Chall (1948)		■		■								■			
Dolch (1949)	■		■							■		■			
Farr, Jenkins and Paterson (1951)	■											■			
Gunning (1952)	■											■			
Spache (1953)												■			
Björnsson (1968)	■											■			

Table 2.2: Summary by authors of readability factors for English and Swedish

- S_n = Number of sentences. Sentences end with punctuation marks, i.e. period, exclamation point, and question marks.
- C_n = Number of characters. We regard the letters of the Swedish alphabet as characters. Numerals and punctuation marks are annotated, but excluded in the lexical analyses, unless otherwise is stated.
- W_n = Number of word tokens.
- Sl_n = Number of phonological syllables.
- CoW_n = Number of word tokens ≥ 3 syllables (in formula (7) referred to as "complex words")
- LW_n = Number of word tokens > 6 characters
- XLW_n = Number of word tokens > 13 characters
- UW_n = Number of unique word tokens
- ULM_n = Number of unique lemmas

Flesch reading ease

It is beyond dispute that the ideas put forth by Rudolf Flesch have had the most long-lived impact on American readability research. In his dissertation, Flesch (1943) published his first readability formula. It was designed to measure adults' reading materials, and used variables such as affixes for estimating the word length, and personal pronouns and names for identifying personal references.

In his next work, Flesch (1948) published a new formula composed by two separate parts. He advocated a return to phonics, so the first part in the new Reading Ease Formula took into account the number of syllables and the number of sentences for 100-word samples. Flesch found that the correlation between syllable count and the affix count in the earlier formula was $r = .87$, and considered the two measures equivalent. The correlation coefficient used for the formula was taken from the earlier findings by Lorge (1944). The second part of this new formula reconnected to the earlier version by using personal words and sentences as predictors of human interest.

The specific mathematical formula is given in (3):

$$RE = 206.835 - (1.015 \times \frac{W_n}{S_n}) - (84.6 \times \frac{Sl_n}{W_n}) \tag{3}$$

Reading Ease Score	Style Description	Estimated Reading Grade	Estimated Percent of US Adults (1949)
0 to 30	Very Difficult	College graduate	4.5
30 to 40	Difficult	13^{th} to 16^{th} grade	33
50 to 60	Fairly Difficult	10^{th} to 12^{th} grade	54
60 to 70	Standard	8^{th} to 9^{th} grade	83
70 to 80	Fairly Easy	7^{th} grade	88
80 to 90	Easy	6^{th} grade	91
90 to 100	Very Easy	5^{th} grade	93

Table 2.3: Description of Flesch's Reading Ease Scale

The output, i.e., Reading Ease is a number ranging from 0 to 100. The higher the number, the easier the text is to read. For reference values, see table 2.3.

The Flesch–Kincaid formula

Another contribution to American readability research was made by Rudolf Flesch in collaboration with J. Peter Kincaid (Kincaid et al. 1975). The "Flesch-Kincaid Grade Level Formula" translates the 0-100 score to a U.S. school grade level, or the number of years of education generally required to understand a text. It is calculated with the formula presented in (4):

$$FKF = 0.39 \times (\frac{W_n}{S_n}) + 11.8 \times (\frac{Sl_n}{W_n}) - 15.59 \tag{4}$$

A score of 6.2 would indicate that the text is expected to be understandable by an average US student in 6th grade.

Automated readability index

The Automated readability index (ARI), was developed in 1966, presented in Senter and Smith (1967). It operates on the average number of characters, words and sentences. Like the formula above, the ARI produces an approximate representation of the US grade level needed to comprehend the text. The formula adopted is presented in (5):

$$ARI = 4.71 \times (\frac{C_n}{W_n}) + 0.5 \times (\frac{W_n}{S_n}) - 21.43 \quad (5)$$

Coleman–Liau formula

Like the ARI but unlike most of the other American indices, this formula (Coleman and Liau 1975) relies on characters instead of syllables per word. The formula for Coleman–Liau calculation is given as (6):

$$CLF = (0.0589 \times \frac{C_n}{W_n} \times 100) - 0.3 \times (\frac{S_n}{W_n} \times 100) - 15.8 \quad (6)$$

The output approximates the US grade levels, in the same manner as the indices above.

Gunning Fog index

The Gunning Fog index was developed by Robert Gunning (Gunning 1952). The formula operates on chunks of text of approx. 100 words in a sequence. After calculation of average sentence length, the number of "complex" words is established. According to this formula (7), a complex word (CW) is regarded as a word with three or more syllables, common suffixes excluded.

$$GF = 0.4 \times ((\frac{W_n}{S_n}) + 100 \times (\frac{CoW_n}{W_n})) \quad (7)$$

Even this formula is designed to measure the readability of texts aimed at different US school levels.

SMOG index

The SMOG grading, invented by McLaughlin (1969) also operates on text chunks. For proper measuring, the chunks should consist of at least 30 sentences where the "complex" words, consisting of at least 3 syllables, are counted. The variables are then inserted into formula (8)

$$SMOG = 1.043 \times \sqrt{CoW_n \times \frac{30}{S_n}} + 3.1291 \quad (8)$$

The SMOG grading estimates the years of US education needed to understand a piece of writing.

LIX

The LIX readability formula was presented by Björnsson in 1968. Like most of the formulas for English, it operates on word length and sentence length, and gives an output between 0 and 100, where 100 is regarded to indicate the most difficult text. The instructions for LIX measurements on words by Björnsson (1968), is that *text, numbers and abbreviations* shall be counted but not *punctuation marks*. The formula, which provides the average of words per sentence and the relative frequency of long words in a text, is shown in (9):

$$LIX = \frac{W_n}{S_n} + \frac{LW_n \times 100}{W_n} \tag{9}$$

The reference values for LIX are given in table 2.4.

LIX score	**Style description**	**Text type**
< 25	Very easy	Children's literature
25 – 30	Easy	Young Adults' literature
30 – 40	Standard	Fiction and daily news
40 – 50	Fairly difficult	Informative texts and non-fiction
50 – 60	Difficult	Specialist texts
> 60	Very difficult	Scientific texts

Table 2.4: Description of LIX scale

2.4.3 Multilevel readability analyses

Different theoretical frameworks have been developed for the purpose of assigning textual features to specific levels of language and discourse. Graesser, McNamara and Kulikowich (2011) mention five different levels of language and discourse proposed in several multilevel theoretical readability frameworks. These are *words*, *syntax*, *textbase*, *situation model*, and *genre and rhetorical structure*. The word level embraces different measures ranging from simple word frequency counts to measurements of vocabulary variation and diversity. The syntactical level includes assignments of part-of-speech categories, and aspects of phrases and syntactic tree structures. The text base level considers the explicit ideas in the text, where the meaning rather than the surface coding is present. The situation model concerns subject matter content of the

text and inferences that are activated by the explicit text. Finally, the genre structure refers to the category of text (Biber and Conrad 2009). Dell'Orletta, Montemagni and Venturi (2011) propose a four-fold partition that closely follows the automatic preprocessing of a text before readability analysis. These are *raw text features*, *lexical features*, *morpho-syntactic features*, and *syntactic features*.

The categorization used in this study was originally proposed by Chall (1958), who believed that

> Only four types of elements seem to be significant for a readability criterion, namely vocabulary load, sentence structure, idea density and human interest. (Chall 1958: 40).

With this division, the quantitative, qualitative and reader-task oriented aspects seem to fuse into one integrated view. Many objectives might be raised against this distribution, the primary and most difficult to meet being the fact that vocabulary is used to express ideas, and that no distinction can be made between vocabulary load and idea density. As pointed out already by Halliday (1985), almost every constituent enters into more than one structural configuration at a number of levels simultaneously, and has more than one function at a time, and above all: linguistics should deal with meaning. Without challenging Halliday's view, we would like to stick to another standard and regard vocabulary load as a measure of lexical diversities and frequencies in the text sources, while the idea density refers to the meaning behind lexical units that might exceed word boundaries.

A division into **vocabulary load**, **sentence structure**, **idea density**, and **human interest** will thus be maintained as a conceptual framework for features operating at different language levels. Two more levels will be presented and discussed in addition to the previous four. The first one comprises aspects observed at the **surface** of texts. We have already presented seven readability formulas that all operate at the surface level. The second additional level relates to the individual's **personal interest**. With the aspect of personal interest in mind, it will be possible to extend the reasoning on whether it might be possible to produce NLP solutions for matching texts to specific target groups of readers or even to one individual. Linguistic measures proposed by other researchers to reflect specific issues related to readability will be introduced along the description of language levels, and assigned an appropriate language level.

Vocabulary load

Vocabulary plays a central role in all modes of communication, and the relationship between word knowledge and text understanding have been demonstrated empirically in many classical works. At the individual level, we denote vocabulary as the set of words within a specific language that are familiar to that person. There is a long-standing recognition that vocabulary knowledge has a heavy impact on reading comprehension. The first American study was provided in a series of works by E.L. Thorndike, among them Thorndike (1917). All quantitative investigations consider vocabulary to some degree related to difficulty by word length, either as number of letters or syllables. An overview presented in table 2.2 reveals that the majority refer to lexical properties.

One method to investigate vocabulary impact is to measure its *diversity*, which can be seen as the vocabulary range (Lively and Pressey 1923), vocabulary burden (Stone 1938) or average repetition. These studies suggested that materials with fewer different words were easier than materials with a higher percentage of different words. Lively and Pressey (1923) argued that diversity seems to be the best predictor of difficulty when the criterion is restricted to children's materials, at the lower levels of difficulty and for poor readers. A language use characterized by a high degree of variation is normally described in positive terms and considered as highly creative. For the non-fluent reader, a limited lexical variation enhances the ease of text processing and production.

Word length seems to correlate with frequency, which in turn affects the lexical variation in texts. As reported by New et al. (2006), the effect on reading speed has been examined with different techniques and found to be inconsistent, ranging from inhibitory to null effect. In a re-examination, New and colleagues studied the influence of stimulus length on lexical decision and naming latency data for a large set of words in the English Lexicon Project. They found that English words ranging from 3–5 and from 9–13 letters caused prolonged lexical decision times, while words between 6 and 8 characters showed shorter times. One factor likely to play a role is physiologically conditioned: eye tracking studies have shown that in reading, most saccades are from 5 to 11 character spaces long, which an average of 8 spaces. The implication is that words with a length of 6–9 letters have the highest chance of being processed after a single fixation.

The **type-token ratio (TTR)** refers to the richness of the vocabulary and is given as the ratio between the number of word types and tokens in a text (Chotlos 1944). Its efficacy as a general measure of language diversity has been discussed extensively over the years (McKee, Malvern and Richards 2000; McCarthy and Jarvis 2007). Attempts to eliminate its sensitivity to sample size have been made, for instance through logarithmic transformation (Honoré 1979). Another objection is that the common units for measuring the TTR are graphical tokens without lemma and lexeme information, which means that a high degree of inflected forms, regardless of their lemma form, contribute to a high TTR. Still, it might be valuable as a quick and simple means to calculate language diversity, in for instance children's speech.

Word length based on phonological measures correlate in general highly with word length in characters, number of orthographic neighbors (LND) and frequency, at least for English (New et al. 2006). Swedish readability research has not paid very much interest to the question whether the syllable as a linguistic unit has any impact on the readability of words. Björnsson (1968) stated that a Swedish syllable is composed by in average 2.8 characters, that a disyllabic word increases into a length of 5.6 characters, and to a length 8.4 characters for trisyllabic words. The internal distribution of word types with regard to number of syllables and number of characters, according to Björnsson is given in table 2.5. According to these calculations, about 90 % of the polysyllabic (≥ 3) words were supposed to be long in terms of characters (> 6), and 90 % of the disyllabic words short (< 7), the latter mainly consisting of simple disyllabic words such as *bada* 'bathe', *pappa* 'daddy', and *hunden* 'the dog'. Long disyllabic words, on the other hand, frequently contain consonant clusters such as *klister* 'glue', *törstig* 'thirsty', and *skrivning* 'exam'. Björnsson ends by concluding that:

> There are strong reasons to believe that these latter words as a group are to be considered as more difficult than the former ones.

The word variation index, **OVIX** (Hultman and Westman 1977) is another way to capture lexical richness. As for type/token ratio, it displays the ratio of unique word tokens to the total number of word tokens in a text, but is constructed in order to approximately compensate for differences in sample size. It is calculated according to formula (10).

Word types	No. of characters				
	1–3	4–6	7–9	10–12	Sum
Monosyllabic	46	8	–	–	54
2-syllabic	–	29	4	–	33
Polysyllabic	–	1	7	5	13
Sum	46	38	11	5	100

Table 2.5: Internal distribution of Swedish words with regard to syllables and characters

$$OVIX = \frac{log(W_n)}{log(2 - \frac{log(UW_n)}{log(W_n)})} \tag{10}$$

The algorithm is similar to that of Honoré (1979) in that it takes the growth of unique words into account. Hultman and Westman (1977) consider an OVIX value below 60 to denote a low lexical variation, while values above 70 would indicate a high degree of variation. Although OVIX seems to constitute a more elaborated way to calculate lexical richness, it is performed on word tokens only without any information of lemma or even lexeme properties. This means that a text with several lexical representations of the same concept gets a higher OVIX than it would have had considering lemma frequencies instead.

Behind the concept of lemma lays an observation of the nature of language. Languages do not invent unique signs for every conceivable nuance of meaning, but generalize and re-use signs in sets of related words. Relatedness is the starting point for the study of the lemma, and consequently, the **lemma variation index**[1] would be a better way to represent lexical variation, as it reduces the set of lexical units.

For obtaining the lemma variation index the ratio of unique lemmas to the total number of word tokens in a text is calculated. In order to compensate for differences in sample size, we use the logarithmic values in the same way as the OVIX formula, see (11)

$$LVIX = \frac{log(W_n)}{log(2 - \frac{log(ULM_n)}{log(W_n)})} \tag{11}$$

The count of a word's **lexical neighbors** seems to be a variable that affects the early stages of visual and auditory identification in language

[1] further on referred to as LVIX

comprehension. An orthographic neighbor is defined by for instance Coltheart et al. (1977), as any word token that can be created by changing one letter of the stimulus word while preserving letter positions (e.g. *hatt* and *kant* are orthographic neighbors of *katt*). Similarly, two words are said to be phonological neighbors if they have the same number of phonemes and differ by one phoneme substitution.

Two neighborhood variables are of interest in readability research:

- the number of neighbors, i.e. the **lexical neigborhood size** or **neigborhood density** (LND), which is the number of strings (or acoustic words) at Hamming distance 1. This metric has been found to be related to different reading tasks, such as lexical decision, naming, and semantic categorization.
- the number of neighbors with higher frequency **lexical neighborhood frequency** (LNF).

It has been called in question whether the neighborhood factors have facilitating or inhibiting effects on reading. Grainger et al. (1989) and Grainger and Segui (1990) presented studies suggesting that the time to recognize a visually presented word increased significantly when the stimulus word was orthographically similar to at least one other higher frequency word in the reference language. A later review (Perea and Rosa 2000) suggested that both lexical inhibition and facilitative lexical-sublexical feedback play a mayor role in identifying words, and that the number of higher frequency neighbors is inhibitory in reading. Other researchers (Yates, Friend and Ploetz 2008), found that phonological neighborhood facilitated reading in eye movement data, which was evidenced by shorter fixations for words with large neighborhoods. It has also been assumed that facilitation is generally obtained only in lexical decision reading tasks, and changing the task might imply a reversal of the effect. The suspicion that orthographic neighborhood effects were progressively modulated by reading skill was opposed by Duabeitia and Vidal-Abarca (2008), who found that children, like adults, show clear neighborhood effects, and that these effects did not seem to depend on reading expertise. It is plausible that a high lexical neighbor density of words belonging to the same lemma facilitates word recognition in reading. Frequency influences the efficiency with which units are processed, and these units can be defined with respect to whole words or their morphemic constituents. There are several measures which index a correspondence between a sequence of letters and mean-

ing based on the number or on the frequency of the word that share a particular morpheme. In addition, measures tend to be intercorrelated despite if they are token based, as the LNF measure, or type based, as the LND, which considers only the number of different related forms. The stem morpheme frequency seems to influence word recognition as decision latencies decrease when base frequency increases (Taft 1979).

Another view of vocabulary is to consider its *difficulty*, which originally was performed by setting up specific guiding principles in the form of word lists, which defined words as easy or hard, familiar or unfamiliar. The word lists used for English are for instance Thorndike's basic word lists (Thorndike 1921; Thorndike and Lorge 1944), or the Dale-Chall word list (Dale and Chall 1948). The words within these particular lists were considered easy, while those not contained in the lists were considered hard. Vogel and Washburne (1928), Dale and Tyler (1934), Gray and Leary (1935), Lorge (1944), Dale and Chall (1948), Dolch (1949), Thorndike (1921), Thorndike and Lorge (1944), and Spache (1953) classified words within particular lists as easy, those not contained in the list as hard. Patty and Painter (1931), Yoakam (1939) and Forbes and Cottle (1953) assigned differential weights to words according to their commonness in a specific word list. Other methods are to estimate the amount of technical words (Dale and Tyler 1934), or the quite unconventional method of looking at the amount of words with certain initial letters (Lewerentz 1929). Reference lists were also made with words known to children in various school grades (Dolch 1936), or graded according to conceptual difficulty as a measure of the abstractness of its vocabulary. It is important to point out that the word lists mentioned so far mostly lacked lexical information, i.e. the entries were represented only at the graphemic level, and that they were constructed on fairly subjective grounds.

A relatively small number of words are used in every-day communication. This is why an individual with small, relatively fixed vocabulary of words can cope with an extremely rich and open-ended world. Hirsh and Nation (1992) found that a 5,000-word vocabulary was necessary for adequate coverage in pleasure reading. It is noteworthy that the key concept underlying this estimation is the *word family*, which is a group of words that share the same stem, such as *love*, *lovely*, *lover*, *loveliness*. This makes the word list shorter, in that one entire word family is regarded as one entry in the list. It is suggested that morphemes function for adult readers as perceptual units that influence word recognition. Nagy and Scott (2000) found that for derived words, the number of words in a word family and the frequency of words in that family af-

fect adults' speed of recognition of the base word. The implication of this is, that the larger a word family, the greater the likelihood that the base word will facilitate recognition of words, even if these words are new to the individual. Reichle and Perfetti (2003) suggested that the likelihood that morphemes in a word play a role in word identification depends on exposure to those morphemes in different word contexts. Out of a genre perspective, Yildirim, Yildiz and Ates (2011) found that vocabulary made more contribution to expository text comprehension than narrative text comprehension.

By way of example, the Swedish particle verb *reda ut* can be viewed as a member of the expanded word family of *reda*, containing additional verbs such as *reda* 'order', *inreda* 'furnish', *utreda* 'investigate', *bereda* 'process', *tillreda* 'prepare', and nouns as for instance *redning* 'thickening', *inredning* 'furniture', *utredning* 'investigation', *beredning* 'processing', and *tillredning* 'preparation'.

The vocabulary rate refers to the internal composition of the vocabulary of a text, compared to a reference word list. To this might also be added the out-of-vocabulary rate, i.e. the occurrences of out-of-vocabulary words. In this study, the *SweVoc word list*, described in section 3.2.4 was used as reference list.

Some specifics for Swedish are frequently mentioned in guides on how to design ETR texts, namely the use of particle verbs and long compounds.

The **particle verb** combinations can be either enclitic or free. A free adverbial particle can change the meaning of the verb completely, as for instance in *hoppa av (hästen)* 'to jump off (the horse)' as opposed to *hoppa (av glädje)* 'to jump (for joy)', where *av* is a preposition. The phrasal verbs imply at least two complications for a reader. First, as illustrated in the previous examples, it can be difficult to tell a particle verb from a verb with PP complement. Secondly, there exist two types of phrasal verbs, which are not easy to differentiate between. In transparent particle verb combinations, the meaning is determined by the meaning of its parts, as for instance *slå ut (fönsterrutan)* 'break (the window)'. Regarding the idiomatic combinations, as for instance *(blomman) slår ut* '(the flower) blooms', the meaning have to be learnt. There is a stylistic difference between enclitic and free particle verb combinations. The free combinations are usually considered as more colloquial, while the enclitic combinations reflect a higher degree of formality. Another property is the difference in meaning. The free particle verb combination is often concrete, while the enclitic combinations tend to be abstract (Norén 1996). Advice on how to produce ETR texts often em-

phasize that enclitic particle verbs should be avoided and substituted into a verb and an adverbial particle.

Written Swedish is characterized by its property to form concatenated **compounds**, which means for instance that an orthographical representation of one Swedish noun would demand two or more tokens in many other languages. This productive capacity implies that a complex concept can be elegantly expressed in a rather compact form, but with the drawback that the amount of words in a text, regardless of size, always contains a porportionally large amount of word tokens only appearing once (> 50%). For the target group of readers, low frequent compounds can imply significant difficulties in both decoding and understanding.

Sentence structure

In traditional readability measures, *sentence structure* has been proposed to manifest itself through sentence length (Dale and Tyler 1934; Gray and Leary 1935), number or ratio of simple sentences as compared with complex sentences in empirical evaluations (Vogel and Washburne 1928), or number of prepositional phrases (Dale and Tyler 1934; Gray and Leary 1935; Lorge 1944). Chall (1958) maintained that easy materials are characterized by short, simple sentences with few prepositional clauses. With the development of language technology tools such as part-of-speech taggers and robust syntactic parsers, more detailed linguistic features can be retrieved and exploited.

Analysis of syntactic variables is crucial in all text research with a scope beyond isolated words. The abstract properties of individual classes to which particular morphological forms belong, i.e. the morphosyntactic variables, are hereby annotated with part-of-speech tags in sets of different size. For Swedish, two main tag sets exist, the SUC (Ejerhed et al. 1992) and the PAROLE (Språkbanken, Göteborgs universitet 2000), tag sets, each containing 156 tags and mutually convertible by use of a mapping scheme. Probability counts of unique part-of-speech unigrams, indicating morphosyntactical properties, has proven useful in implementations of readability assessment tools (Pitler and Nenkova 2008), (Aluisio et al. 2010), and (Dell'Orletta, Montemagni and Venturi 2011).

The next step in moving from a string of words into its meaning is to assign sentence structure markers. Syntactic parsing is the process of finding the immediate constituents of a sentence and determining whether these can be grouped together. All recent readability studies,

including readability prediction (Heilman, Collins-Thompson and Eskenazi 2008; Chae and Nenkova 2009; Feng et al. 2010), readability assessment (Dell'Orletta, Montemagni and Venturi 2011), text simplification (Inui et al. 2003), language diagnostics (Roark, Mitchell and Hollingshead 2007) rest on some analysis of syntactic variables. In recent years, dependency-based syntactic parsing methods have become increasingly popular. The basic assumption behind these methods is that syntactic structure consists of lexical elements linked by binary asymmetrical relations called dependencies. Works by Gibson (1998) and Temperley (2007) show that the complexity of processing a sentence is related to the **length of the dependencies** within it. According to Gibson's Dependency Locality Theory (DLT), two factors are considered as predictors of complexity, namely the *storage cost* and the *integration cost*. The storage cost is the (human) memory load necessary to maintain the syntactic predictions of previous words, while the integration cost is the (human) effort of synctactically connecting a word to previous words with which it has dependent relations. Gibson shows that the DLT between different types of dependencies predicts comprehension of a number of syntactic relations. The probability of different types of syntactic dependencies have earlier been used as an indicator for automatic readability assessment, e.g. by Dell'Orletta, Montemagni and Venturi (2011).

Subordination and nominal modifiers are generally considered to indicate a higher degree of complexity in educational and linguistic research. In the work of Dell'Orletta, Montemagni and Venturi (2011) unconditional probabilities of different types of syntactic dependencies (e.g. subject, direct object, modifier, etc.) was implemented. **Prenominal modifiers** (or adjectival attributes) are adjectives or participles, inflected to express agreement. **Postnominal modifiers** include relative clauses, prepositional phrases, adverbials, and infinitive clauses.

Work on readability by Schwarm and Ostendorf (2005) incorporates grammatical surface variables such as parse tree depth and average number of verb phrases in support vector machines to produce a better method of assessing reading level. In an early paper, Yngve (1960) examined what we would call sentence comprehension difficulty. He describes the maximum number of symbols needed to be stored during the construction of a given sentence as the depth of that sentence. Also Miller and Chomsky (1963) propose a metric of syntactic complexity based on the ratio between non-terminal and terminal nodes of the syntactic tree of a sentence. In treebank terms this could be illustrated as **the height of the whole parse tree**, i.e. the number of nodes from

the root to the most distant leaf. The parse tree depth is the largest distance to the root, and the difference is 1, i.e. the height is always one larger than the distance to the root. Figure 2.2 presents the parse tree of sentence example 1.

Example 1 *Alla människor ska vara lika mycket värda.*
'All people should be valued equally.'

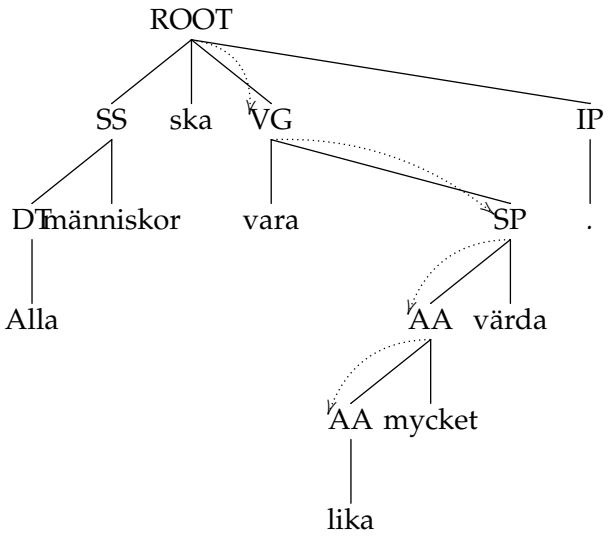


Figure 2.2: Illustration of parse tree height measurement

Heilman et al. (2007) studied the effect of combining grammatical and vocabulary features in a language modeling approach, aiming to improve readability measures.

Idea density

Morriss and Holversen (1938) claimed that difficulty in understanding what is read depends on the author's choice of words and the complexity of relation between the ideas. The general concept was that materials which are hard to understand contain more ideas to a given number of words. Chall (1958) formulated the belief that semantic complexity could be captured by surface features. These could for instance be the relative number of prepositional phrases (Dale and Tyler 1934; Gray and Leary 1935; Lorge 1944), as they usually act as qualifiers of, or additions to, simple statements. This, in turn also leads to the effect of

extended sentence lengths and, ultimately, to increased sentence difficulty. A significant indicator of information load could be the number of different content words (nouns, verbs, adjectives and adverbs), or by the proportion of content words to the total number of grammatical words used. A general measure is the ratio of content words (i.e. nouns, adjectives, and verbs) to grammatical words in a text, which is considered to be an indicator of information load. According to Sorvali (1984) no normalized scale exists for this way of calculating *lexical density* in Swedish text written for adults. There is, however, an assessment by Korkman (1995) proposing that a content word ratio between 40 and 45 should be indicative for Swedish.

Morriss and Holversen (1938) argued that instead of being inherently easy or difficult, many words can be regarded as conveying easy or hard ideas. Common practice in readability measurements resting on word lists would be to classify a word like *mått* 'measure' as easy, due to its length as well as its commonness. There is, however, a substantial difference in the use of this specific word in a phrase like *pH är ett mått på hur sur en produkt är* 'pH is a measure of the acidity of a product', which is easy, and *att vidta mått och steg* 'to take measures', which is difficult. According to Morriss and Holversen, the ease or difficulty of words should thus be decided upon by looking at the context in which they were used in the sentence. Furthermore they suggested that it would be challenging to find some means for appraising the differential meaning of words or ideas, yielding a more valid estimate of text difficulty. They classified words according to their properties of being content or non-content words, i.e. if they are conveying ideas or just serving a structural function. The non-content words, according to the authors are simple prepositions, simple conjunctions, pronouns, articles, all forms of the word *to be* and the auxiliary verb *to have*, simple exclamations, and proper names such as those which have no function except to name characters in fiction. Names of living people or historical persons are to be regarded as content words when they hold meaning in context. Content words were further subdivided into four groups, gradually increasing in difficulty and chronologically following the natural learning of words, as presented in table 2.6. Ultimately, a technique for analyzing books by this method was developed, although never published or put into any usable form. It has still made a contribution by giving importance to the meaning of words (Chall 1958).

Bringing Morriss and Holversen's theory a bit further, many studies of word reading have examined the effect of early learning on adult reading performance. The factor is known as age-of-acquisition (AoA)

and is basically a measure of how early in life a word is learned. The implication would be that the earlier a word is acquired in spoken language, the faster and easier the processing in adult reading. Although this hypothesis would support the principle of content word categorization in order of acquisition, it doesn't seem that the AoA effect is more than a natural loss in plasticity of many cognitive domains that limits the ability to acquire new information (Zevin and Seidenberg 2002). The preferred view in this thesis is that words are acquired and used according to naturally occurring hierarchies, and that the principles underlying the determination of which hierarchical level is basic are expected to be universal, as stated in section 2.1.

Class	**Description**	**Example**
I	Words representing fundamental or elemental experience in life of a people of a given culture. Common to all localities within the culture and learned almost unconsciously by the individual during the course of a normal childhood	*father* *water* *home* *hurt*
II	Words used by limited groups of the population and learned early in life	*corn* *tide*
III	Words signifying concrete ideas, the names of persons and places, things and processes Words representing concrete relationships and words descriptive of the working of machinery or experiments of scientists. All measures of quantity. These words are normally not learned before the end of the sixth grade	*Picasso* *Iraq* *oxidation*
IV	Words that signify abstractness, quality, states of mind, degree, shape, size, color. In addition, complex and sophisticated words	*platitudes* *concave* *culminate*

Table 2.6: Content word classification

Kintsch and Keenan (1973) found that the measures used for idea density counts hitherto had been arbitrary and theoretically unfounded, and hereby concluded that earlier attempts to show how it contribute to reading difficulty were generally unsuccessful. In line with several other researchers, they explored the hypothesis that sentences have a base structure, consisting of propositions which represent their semantic content. More specifically, they investigated subjects' reading rate and retension related to the **number of propositions** in the base struc-

ture of sentences, and found evidence for the hypothesis that propositions are a basic unit of memory for text. There was, however, differences in the retention time depending on the individual properties of the propositions, in that superordinate propositions seemed to be recalled better than subordinate. Based on these findings, and with the propositional text base construction formalized later on by Turner and Greene (1977), Brown et al. (2008) implemented a computer program that determines the propositional idea density (P-density) of English texts on the basis of part-of-speech tags. They suggest that propositions roughly correspond to verbs, adjectives, adverbs, prepositions, and subordinating conjunctions, and regard the approach suitable for readability style guides and applications.

The **noun/pronoun ratio** is a style indicator which can indicate text complexity. Graesser, McNamara and Kulikowich (2011) found that pronouns are important for cohesion and makes comprehension more difficult and are more prevalent in oral discourse than in written text.

Some parts-of-speech can be studied in terms of information load. Nouns and verbs do both hold a high degree of content, but they also have a specific mutual relationship in that they are complementary. A high degree of nouns implies a low ratio of verbs, and vice versa. A text with many verbs are considered to have a verbal style which is characteristic for spoken language. Nouns, on the other hand, are more frequent in informative and investigative texts. A tendency towards nominalization, i.e. when information is expressed with nouns instead of verbs, is typical marker for written text. The recognized way to get a rough idea about the information load in a specific text is to calculate the **nominal ratio** (NR), the simplest being the number of nouns divided by the number of verbs in the same text. A high NR indicates a more professional and stylistically developed level with high information density, thus more demanding and time-consuming to process (Melin and Lange 2000). A more elaborated NR is achieved by calculating the proportion of nouns, prepositions and participles in relation to verbs, pronouns and adverbs. The former are characterized as nominal classes and the latter as verbal classes. NR normal value is 1.0, and the levels for different genres are illustrated in table 2.7 (from Melin and Lange (2000)). An example is the noun *omhändertagande* 'custody', which is the participle form of the verb *omhänderta* 'take into custody'. Both word forms signal a formal style, and it also implies a judicial or penal safe-keeping as opposed to the verb construction with a free adverbial particle *ta hand om* 'take care of'.

Genre	NR
Leaflet	1.19
Textbook	1.18
Morning news	1.04
Tabloid	.99
Magazine	.85
High school prose	.72
Speech	.25

Table 2.7: NR levels per text genre

The lexical-semantic network Saldo (Borin and Forsberg 2009) provides information on associative relations between its entries. The entries are based on metaphorical kinships, hierarchically organized under an artificial most central entry, the PRIME. The distance between the a specific lexical entry and the PRIME can hence be considered as a measure of an entry's **semantic depth**, which will be further developed.

Human interest

It is widely recognized that interest has an important role in readers' text processing. A review performed by Hidi (2001) suggests that two distinct ways of investigating the role of interest in learning (and implicitly reading) exist: the first focusing on the impact of personal preferences, and the second being a text-based approach, i.e. how the interestingness of stimulus materials influences the individual's performance. The latter being a subtype of situational interest – the kind of interest evoked by something in the immediate environment. Researchers have also investigated topic interest, which can be regarded as a subtype of individual as well as situational interest. The role of interest in reading is a well-documented area, and the most significant issue concerns the effect of interest on readers' text processing and learning. Research has demonstrated that both individual and situational interest contribute to increased comprehension and learning. Another issue regards the question how text characteristics can make reading materials more interesting. Schank (1979) cited in Hidi (2001) referred to some concepts, such as 'death', 'danger', 'power', and 'violence' as emotional interests, distinguished from cognitive interests that result from events that play a role in complex cognitive structures or hold surprise. Research in automatic emotion analysis in texts has long been restrained

by the limited supply of relevant lexica. However, interesting work has recently been carried out by Mohammad and Turney (2011), who conducted experiments on people's opinions on 14,000 common English words. The questions regarded the words' associations with eight emotions (anger, fear, anticipation, trust, surprise, sadness, joy, and disgust) and two sentiments (negative and positive). The annotations were done manually through Amazon's Mechanical Turk service. The final result, the lexicon EmoLex, has entries for more than 10,000 word-sense pairs, and a list of associations of the pair with the above-mentioned basic emotions. In addition a subset of the lexicon was compiled into a list of 826 terms that refer directly to emotions. Proverbio et al. (2009) investigated neural processing of proper and common nouns, and found that person names were more emotional and sensory vivid than common noun semantic access in ERP patterns.

Factors denoting human interest in traditional readability studies were based either on the relative number of personal words, i.e. pronouns, words indicating a living person, and proper names, or "colorful words", the latter evidently less usable due to its vagueness. This concept seems to be more associated with reading comprehension than reading speed, and is included as a factor in the formulas by Gray and Leary (1935), Lewerentz (1929), and Flesch (1948).

In accordance with the group described above, few producers of readability indices have so far been concerned with elements indicating some influence on the human interest. Three formulas (Gray and Leary 1935; Lewerentz 1929; Flesch 1948), proposed one of the following measures as a factor:

- Relative number of personal pronouns
- Nouns with natural gender
- Proper names
- Colorful words
- Personal sentences, such as spoken sentences, questions, commands, requests, and other sentences directly addressed to the reader; exclamations; and grammatically incomplete sentences whose meaning has to be inferred from the context.

A text-based NLP approach able to identify general "interestingness" above the word level would demand context sensitive analyses like those developed within the field of information retrieval and text mining. Methods for analyses at a higher textual level rest on features captured by thematic relations, anaphora resolution and discourse mark-

ers, i.e. features that are not annotated in the present materials. It would also demand quite another text selection approach, where the original documents were analyzed from their beginning to the end. This study involves pseudodocuments, where sets of 30 sentences are selected randomly from the subcorpora. Of the items above, the ratio of proper names was chosen as a rough estimate of the degree of human interest. This under the presupposition that words that create more emotions might also arouse more human interest.

Personal interest

The personal interest relates to the reason for reading. Reading a newspaper means that you are interested in the latest updates and news, you read fiction because you want to be entertained, and informative text reading takes place because you want to know how to perform a task or get to grips with certain facts. In addition to the cognitive factors, features related to genre and topic are crucial for determining on a specific text for an individual reader. The target group of readers treated in this thesis are persons with mild intellectual disability. The research of Feng, Elhadad and Huenerfauth (2009) addresses two literacy impairments that distinguish people with intellectual disability from other low-literacy adults, namely limitations in working memory and in discourse representation. Relying only on the LIX scale, a readability level of texts directed towards children, i.e. below 30, seems to be the best fit. An adult person mostly seeks texts that fall within the scope of his/her personal interests, which for the major part are different from those of children. A consequence of this is that searches for texts limited to a certain LIX value often result in mismatches.

Feng, Elhadad and Huenerfauth (2009) analyzed the significance of discourse-level features related to probable cognitive factors underlying reading challenges for persons with MR. These features, novel in the sense that they are not considered in the traditional readability metrics, are related to the "entity density" of a text. The number of entities in a text that a reader has to keep in mind while reading each sentence and throughout an entire document, was one of the cognitively motivated features examined. These entities were regarded to consist of the union of the common nouns and the named entity noun phrases in a text. Additionally, parse-tree-related features suggested by Petersen and Ostendorf (2009) were implemented in the study, calculated from parsing the sentences in their corpus using the Charniak parser (Charniak 2000). It is an interesting study, and later research carried out by

Feng (Feng et al. 2010), (Feng 2010) have a similar view on readability assessment, in that aspects related to memory span and comprehension are regarded as having a profound impact.

2.4.4 Summary of features

The literature survey above can be summarized into a scheme presenting features acting at different textual levels. The major part of all published readability studies relate to English, and very few large-scale studies exist on Swedish. Although many findings regarding reading and its difficulties can be assumed to be universal, language-specific features should ideally also be taken into account. For each type of reading situation, i.e. isolated word reading (Word), sentence reading (Sent) and document reading (Doc), variables from the set of linguistic surface features are picked. They are listed in table 2.8, together with relevant deep structure feature variables for each situation. Additionally, some of the formulas and indices relevant for Swedish are included.

2.5 Matching texts to readers

As already mentioned, Feng presented studies targeted towards understanding the particular difficulties faced by readers with intellectual disabilities. All studies aimed at developing an automatic readability assessment tool that modeled the reading difficulty of a text for people with MR. For this purpose paired texts and texts graded by readability level were used to analyze and select features that distinguish the simplified texts most from the original ones.

In the first study (Feng 2009), language models based on either shallow features, parse-tree-related features suggested by Petersen and Ostendorf (2009), or global semantic (discourse level) properties of a text were implemented and compared. Preliminary results showed that the latter metric outperformed the previous ones with high margin on articles with lower grade levels. The following study (Feng, Elhadad and Huenerfauth 2009) proposed entity density features, based on named entity and lexical chain identification. The term "lexical chain" is here referred to as nouns in a document connected by relations like synonymy or hyponomy, and more precisely, chains that can indicate concepts that recur throughout a text. For adult persons with MR, the read-

ability of a text was supposed to be influenced by its entity density, since slow semantic encoding and working memory limitations are found to underpin their literacy challenges. The results showed that entity density features were useful in modeling the grade level of elementary school texts as well as correlating to small-scale MR reader comprehension tests. The findings were confirmed in a later study (Feng et al. 2010) where POS features, in particular nouns, seemed to have significant predictive power. This was supposed to explain the good performance of the entity density features, which were based primarily on nouns. More precisely, the selected POS features appeared to be more correlated to text complexity than syntactic features, shallow features and most discourse features, according to this study. It is, however, to be noted that this study was directed towards texts aimed at primary-school students, i.e. a different target group of readers. The same results were also presented in Feng (2010).

In addition to the challenge of matching texts to a specific target group of readers in terms of age and reading ability, the individual preferences must also be met. This issue was investigated by Miltsakaki and Troutt (2008) and Miltsakaki (2009) in a system designed to evaluate if text retrieved from the web was appropriate for the intended reader. The system performs keyword search, thematic classification and analysis of reading difficulty of internet texts and returns the results. Three readability algorithms were used for the latter task: Coleman-Liau, formula (6), LIX, formula (9), and a simplified LIX for English, called RIX (Anderson 1983). Thematic classification was performed by using a text corpus of prelabeled thematic categories and evaluating the suitability. A set of supercategories (Literature, Science, Sport) were further divided into 8 basic categories, and 41 subcategories. Three different classification algorithms were compared, out of which the one built on maximum entropy modeling presented a performance of 93 % in the supercategorization task.

The present study is concerned with readability assessment for specific target groups of readers. Our approach is to investigate factors influencing text complexity at different language levels, to implement selected language features into a vector model for text classification, and to evaluate the results with regard to text type and genre specifics. It is proposed that this approach will contribute to a method for determining upon texts suitable for persons with different kinds of reading impairments with respect to a) reading difficulties specific to some intended target group, b) personal needs in terms of text genre (fiction, news or information), and c) personal interest with regard to text topic.

We want to go beyond the approach by Feng and colleagues (Feng 2009; Feng, Elhadad and Huenerfauth 2009; Feng et al. 2010; Feng 2010) who investigated particular parts-of-speech that were supposed to indicate complexity in terms of entity density. It was suggested that the entity density would influence the working memory load and that a low level would implicitly benefit readability for persons with MR. We will extend these studies into a multilevel model for readability assessment which may be used to tailor reading materials according to different readers' needs and preferences. From a contrastive perspective we expect to find specific characteristics that distinguish easy-to-read texts from ordinary texts in a comparable corpus. The easy-to-read texts are primarily intended for persons with cognitive disabilities, which leads us to assume that the language model achieved reflects the complexity and implicitly readability of texts for a target audience included in this group.

Feature category	**Feature**	**Id**	**Word**	**Sent**	**Doc**
Surface	Word length in char	C_n	I	M	M
	Word length in syllables	Sl_n	I	M	M
	Sentence length in word tokens	W_n		I	M
	Sentences	S_n			I
	Word tokens > 6 char	LW_n		N	N
	Word tokens > 13 char	XLW_n		N	N
	Unique word tokens	UW_n		N	N
	Type-token ratio	TTR			R
	Lexical neighborhood density	LND	R		
	Lexical neighborhood frequency	LNF	N		
	OVIX	OVIX			**F**
	LIX	LIX			**F**
Vocabulary load	Unique lemmas	ULM_n		N	N
	Lemma variation index	LVIX			F
	SweVoc	SV	B	R	R
	Adverbial particles	Part	B	N	R
	Compound word	Comp	B	R	R
Sentence structure	Dependency distance	MDD		I	M
	Subordinate clauses	UA		N	R
	Prenominal modifiers	AT		N	R
	Postnominal modifiers	ET		N	R
	Parse tree height	PT		I	M
Idea density	Propositional density	Pr		N	R
	Nouns/pronouns	NoPr		R	R
	Nominal ratio	NR			R
	Semantic depth	Sa	I	M	M
Human interest	Personal nouns	PM		R	R

Table 2.8: Features and formulas involved in different reading situations, N = frequency, M = mean value, R = ratio, I = integer value, F = formula, B = boolean.

3 MATERIAL

This chapter will present the language resources used in the project, which embraces corpora, text collections and lexica. Language technology tools adopted at various stages of text processing are presented within the context of each resource.

3.1 Corpora

The majority of all studies on how humans process text by reading are made on humans in large- or small-scale experiments. By using corpora as the working material the focus is reversed and the problem attacked in the opposite way. Thus, the questions are put directly to the texts instead of the humans intended to read them. History has provided humanity with innumerable pages of written text, along with the knowledge and skill of how to produce and consume it. The great advantage of corpus linguistics is that if not all, then at least some of these pages can be consulted in order to get a more or less appropriate answer to specific questions on how language is composed. The object of study in this thesis is comparable corpus compiled from the subcorpora of three other resources: LäSBarT, SUC and Göteborgs-Posten. LäSBarT is a corpus of texts produced for easy consumption, i.e. they are meant to be read by children or young adults, or people with reading difficulties, principally cognitive disabilities. SUC and Göteborgs-Posten texts are used in order to distinguish between those properties that are to be considered to pertain specifically to the "easy-to-read" type of texts (in the following referred to as ETR texts) and those that are general for every-day Swedish (ordinary texts).

3.1.1 The LäSBarT corpus

Ideally, text corpus compilation is a task following two successive stages; design and implementation. In the design phase you are supposed to define the object to be created – the corpus – and to specify the texts that you aim to include. After the description phase you initiate the construction, trying to stay as close to the original design as possible. Design criteria for corpus compilation have been set up by for instance Atkins, Clear and Ostler (1992), Sinclair (1991) and Biber (1993). In reality, a sequential procedure is likely doomed to fail. In fact, after identification of the main kinds of texts that can be found in computerized form, the drawn-out and time-consuming undertaking of assembling material and negotiating with presumptive text suppliers often force you to intertwine the tasks in a circular process where you accumulate the content to a degree considered to be largely in line with the original design.

Broadly speaking there are two main types of corpora; balanced and specialized, although both types are designed for a particular purpose. The first corpus type is balanced with regard to text genres and domains that typically represent the language under consideration, while the latter includes a particular type of texts.

LäSBarT is a corpus of 1.3 million tokens, which with today's standards must be considered a small corpus. Based on the two different types of methodology involved, there is one major reason to distinguish between small and large corpora. A small corpus needs a lot of manual work and human supervision in the implementation stage, because the number of occurrences of most phenomena is not large enough to benefit from computer processing beyond selection and ordering. A large corpus might be liable to further stages of automation before organizing methods, hence human intervention must be delayed as long as possible (Sinclair 2000) in order not to invalidate the automatic processing.

To initiate the construction of the corpus, an initial subset of texts was selected. This subset, approximately 200,000 tokens, was compiled from three different genres; news text, fiction and information. It was originally collected for provision of a vocabulary of simple Swedish (Forslund 2004). Progressive collection work resulted in an enriched corpus, *LäSBarT*, an acronym for *Lättläst Svenska och Barnbokstext* 'Easy-to-read Swedish and Children's fiction Texts', which was purposefully constructed to serve as working material for this study. According to the standards, it is a specialized corpus.

The LäSBarT (henceforward referred to as LB) corpus was thus compiled from 2003 to 2008 with the objective to mirror simple language use. The subcorpora are supposed to reflect language use in different domains and genres. It includes different genres of fiction, official documents containing information from the government, parliament, county council, municipality and daily news. The common denominator for all the texts is that they are all intended to be read by persons that do not fully master everyday Swedish language. The texts pertaining to the daily news and fiction genres were produced by professional authors, specialized in ETR writing, while the authorship of information texts is generally unknown.

Published ETR texts, by definition shorter than most written material, are still rather scarce which obviously restrained the collection work. At the lexical level, studies have shown that a corpus of 1–3 million words only allows reliable estimates for high-frequency words. For words with a frequency smaller than 10 per million, a corpus of at least 16 million words is required (Brysbaert and New 2009), while Biber (1993) is of the opinion that "reliable information" on frequently occurring linguistic items such as nouns can be got from 120k-word sample, while an infrequently occurring construction such as conditional clause would need 2.4 million words. Although the present resource of 1.3 million tokens must be considered a very small corpus, it was considered appropriate for the present purpose. The limited amount of text was compensated for by making text representativeness be decisive during compilation. Representativeness is an assessment of to which degree a sample includes the entire range of variation of the population (Biber 1993). One reason to trust the representativeness of the corpus materials is that the supply of ETR texts is limited and subsequently, the variation range is quite narrow. Contrary to many other writing tasks, the production of ETR text is elicited by a specific need from society and we cannot expect a large variety of genres.

Three genres of easy-to-read texts were identified for obtaining a representative sample, namely fiction, news and information. The fiction genre was subdivided into texts targeted towards children and adults. A detailed description of the LB corpus composition is given in appendix A.

Fiction texts

The total extent of books in the fiction genre is difficult to evaluate, since the term "easy-to-read" does not exist in either sale or lending statistics. After applying to different publishing houses directed towards persons seeking simplified texts we were kindly permitted to use an appropriate amount of children's books as well as fiction written for the adult readership. In addition, a subcorpus of children's ordinary literature was included for comparative reasons. Depending on the origin of the texts, four different types are normally distinguished: *Original texts* might be written for a certain target audience, while *rewritten texts* are original texts adapted for a specific target audience. *Translated texts* are written in a source language and translated into a target language. Finally, *translated and rewritten texts* are originally written in a source language, translated into a target language and finally adapted for a certain audience. The influence of translation has been studied by for instance Gellerstam (1991), who argues that translation between two natural languages inevitably affects the target text in terms of *translationese*. Although some of the fiction books are translated into Swedish, the aspect of translationese can be ignored in this study since adaptation into a new literary style is by no way driven by ambitions to maintain all the characteristics of the source text.

Mean sample length of the entries in different LB fiction subcorpora, i.e. children's ordinary literature (table A.1), children's ETR fiction (table A.2), and adults' ETR books (table A.3), was 9,652 words[2], 15,609 words[3], and 8,269 words[4] respectively. Corpus examples of children's ETR fiction, children's ordinary fiction, and adults' ETR fiction are presented in B.1, B.2, and B.3, respectively.

Newspaper texts

The news genre is principally covered by the weekly ETR newspaper *8 SIDOR* '8 pages', supplemented with a website which is updated on a daily basis. The publication of 8 SIDOR was elicited by an initiative in the mid 1980's when the Swedish parliament proposed the government to allocate funds for a newspaper for intellectually disabled people (KU 1986/87:4). Other sources have existed over time, for instance *Invan-*

[2] σ=3,768
[3] σ=15,410
[4] σ=3,012

drartidningen 'The immigrants' newspaper', a newspaper targeted at readers of Swedish as a foreign language. Another source is *Klarspråk*, literally 'Plain language', radio-transmitted news in an easy manner. Sample length measure in words for each daily news' headline varies between 50 and 7,136[5]. Details regarding the news subcorpus are displayed in table A.4, and an example of ETR news text is given in B.5.

Information texts

On practical grounds, texts included in the information genre are all retrieved from public web sites, since the vast majority of all public information today is available via this medium. When the collection work started in 2003, only a limited amount of public information was accessible in an ETR version, but the amount has apparently increased during the last few years; most probably an effect of the Swedish commitment to the EU eInclusion agenda from 2007 (CEC 2007). The EU member states have agreed upon a wide range of measures for harnessing the potential of ICT to promote inclusion, deliver better public services and improve quality of life of its citizens. The goal was to step up actions to reduce gaps in digital literacy and e-skills by 2008, and make all public websites accessible by 2010. In a Swedish survey in 2006 only 12 % of the responding public authorities, county councils and communities presented easy-to-read information on their web sites (Falk and Johansson 2006). According to recent figures presented by the Swedish Centre for Easy-to-read, 95 out of 290 communities (33 %) had accessible information available at their web pages (Centrum för Lättläst. Lättläst-tjänsten 2012).

The public information texts are non-homogeneous in many respects. There are for instance large variations in sample length, measured in words[6]. Specifics of the text samples from municipalities are listed in table A.5, government and parliament in table A.6, and county councils in table A.7. A corpus example of ETR information text is given in B.8.

Corpus processing

After assembling and preprocessing, the corpus was tokenized, lemmatized, part-of-speech-tagged and syntactically parsed. Each of these

[5] $\overline{X}$=628, σ=348

[6] $\overline{X}$=2,408, σ=2182, N=83

steps was performed by means of specific tools, described below. The proportions of the entire LB corpus is displayed in table 3.1, and complete details regarding its composition are presented in Appendix A.

The *word* in its orthographical form is commonly denoted as a *token*. Before any analysis can be made on tokens, they have to be isolated from the original stream of characters. In running text, a number of structurally recognizable tokens contain ambiguous punctuation, which must be resolved before any further computational processing can be performed. The isolation of orthographical units into words and sentences is made in the preprocessing step of computational treatment known as tokenization. Although this seemingly uninteresting task can be made rather easily by use of either hand-made or publicly available tools, several problems may arise when working on large corpora. The first, and most obvious observation is that tokenization is language-specific. Another challenge is to resolve any ambiguity connected to punctuation, particularly since the period is an extremely ambiguous punctuation mark (Grefenstette and Tapanainen 1994).

After tokenization, the corpus was tagged with parts-of-speech with the TnT-tagger (Brants 2000a), and annotated with lemma and lexeme information from the Swedish Morphological Database (SMDB) (Berg and Cederholm 2001). The TnT-tagger is a statistical part-of-speech tagger that is trainable on different languages and virtually any tag set. The Swedish version of TnT-tagger is trained on SUC. Reported accuracy of the tagger for the mixed English Susanne Corpus of 150,000 tokens (Sampson 1993) was lowest with 94.5 %, considered as due to the small size of the corpus, and the large tag set of more than 160 multi-token tags. The accuracy of 96.7 % for the 1.2 million tokens newspaper corpus Penn Treebank (Marcus, Santorini and Marcinkiewicz 1993) is reported as state-of-the-art for English texts, and the same accuracy of 96.7 % holds for the German NEGRA newspaper corpus (Brants, Skut and Uszkoreit 1999) of 350,000 tokens, which was deemed as excellent (Brants 2000b, a). The first run of the part-of-speech tagger on the LB corpus yielded an accuracy of 91.8 % (105,537 errors), while the subsequent automatic lemmatizing prior to any correction of the erroneous automatic tagging presented 96.9 % correct lemmas assigned to each token. After the automatic processing, the material was checked and corrected manually.

The SMDB is derived from a version of the 13th edition of the Swedish Academy's Word List of the Swedish Language (SAOL 2006). The entries, amounting to approximately 120,000, are distributed into different categories depending on inflectional paradigm. In this way, all in-

Source	Category	Type	Genre	No. tokens	% tokens	No. words	% words
Bonnier Carlsen	Children's literature	Ord	Fiction	421,452	32.6	364,463	31.9
Hegas	Children's literature	ETR	Fiction	143,789	11.1	122,473	10.7
Lättlästförlaget	Adults' literature	ETR	Fiction	115,770	8.9	100,521	8.8
8 SIDOR	Daily news	ETR	News	357,444	27.6	322,660	28.2
Invandrartidningen	Immigrants' daily news	ETR	News	6,585	0.5	5,701	0.6
Klartext	Broadcast news	ETR	News	26,773	2.1	24,612	2.2
Kommuninformation	Municipality	ETR	Info	20,405	1.6	18,478	1.6
Statlig information	Parliament and government	ETR	Info	86,312	6.7	78,276	6.9
Länsstyrelse och landsting	County council	ETR	Info	53,958	4.1	48,828	4.3
Myndigheter	Public authorities	ETR	Info	47,518	3.7	43,468	3.8
Diverse	Church, organisations, cultural institutions	ETR	Info	14,044	1.1	12,737	1.1
				1,294,050		1,142,217	

Table 3.1: Corpus size and ratios of LäSBarT corpus

flectional forms are generated into a full form database. Considering the amount of tags in the tag set, SMDB is not as elaborate as SUC. The SMDB tag set contains 133 tags, while SUC has 153 tags; a difference that according to Johansson Kokkinakis (2002) does not negatively affect the performance or the output of the tagger.

Each token in the LB corpus was supplied with a part-of-speech and included in the frequency counts unless otherwise stated. In accordance with the standards set up for SUC 2.0, punctuation is defined as a token. A word is defined as a token that is not a punctuation token.

Many text properties are implicit to the human reader, but before further computer processing these features must be rendered explicit in a standardized way, at least when the goal is to make the texts and the data accessible and useful for a variety of disciplines. Different ways have been devised for such encoding, or mark-up; the majority with the drawback that they were too narrow and not extendible beyond the original format and purpose. Some encoding schemes focused on the theoretical view of the text, while others concentrated on the typographic appearance. In 1990 the Text Encoding Initiative (TEI), which was a major international project within the humanities and language industries, published the first draft of the TEI Guidelines. These are recommendations both on what features to encode and how to encode them. The TEI elements used for mark-up of LB, and a brief description of each, are listed in table C.1 in Appendix C.

Syntactic parsing of the corpus was performed by means of the MaltParser (Nivre et al. 2007). It is a language-independent system for data-driven dependency parsing, which means that it can be ported to new languages, provided that the necessary linguistic resources are available. A pretrained parsing model for Swedish, the *swemalt*, is trained on the professional prose section of Talbanken05 (Nivre et al. 2006), which is a Swedish treebank of roughly 300,000 words. The dependency relations used in this study are from Teleman (1974).

3.1.2 SUC 2.0

SUC 2.0 (Källgren 1998) is a balanced corpus of 1 million words in written Swedish, originating from the 1990's. It is designed according to the Brown corpus (Francis and Kučera 1979) and LOB corpus (Johansson, Leech and Goodluck 1978) principles, which means that it consists of 500 samples of text with a length of about 2,000 words each. The state-of-the-art markup language at the time of compilation was SGML, and

this annotation schema is kept also in the actual, revised version. All entries are annotated with parts-of-speech, morphological analysis and lemma, or rather base form. The corpus is also provided with a wide range of structural tags and functionally interpreted tags, according to the TEI standards (Sperberg-McQueen and Burnard 1994; TEI Consortium 2007). All SUC categories, together with the letter combinations that are used for naming them and the approximate number of tokens in each of them, are listed in table 3.2. Three of the SUC categories were chosen as reference materials. Category A, composed by press reports, was expected to match the ETR news texts in LB. Category K consists of adults' ordinary fiction texts, and was selected for comparison to the adults' ETR fiction subcorpus in LB. SUC category H (miscellaneous) contains information and regulations from local and national authorities, in addition to a minor part company information. In the following, SUC category A will be referred to as ordinary news (ON), SUC H as ordinary information (OI), and SUC K as adults' ordinary fiction (AOF). Corpus examples are given in B.6, B.9, and B.4, respectively.

Id	**Category**	**Number of tokens**
I.	Informative prose	746,000
A	Press: Reports	88,000
B	Press: Editorial	34,000
C	Press: Reviews	54,000
E	Skills, Trades and Hobbies	116,000
F	Popular Lore	96,000
G	Belles Lettres, Biography, Memoirs	52,000
H	Miscellaneous	140,000
J	Learned and Scientific Writing	166,000
II.	Imaginative prose	254,000
KK	General fiction	164,000
KL	Mysteries and Science Fiction	38,000
KN	Light reading	40,000
KR	Humour	12,000
	Total number of tokens	1,000,000

Table 3.2: Distribution of texts in SUC

3.1.3 Göteborgs-Posten

An additional portion of ordinary daily newspaper text was prepared for use as reference materials. It originates from the 2007 edition of Göteborgs-Posten (GP), chosen in order to at least to some extent match the news coverage of LB news text *8 SIDOR* from the same year. The GP excerpt consists of 118,703 words, distributed into 7,920 sentences. A text example is given in B.7.

3.1.4 A monolingual comparable corpus

With a somewhat lax view of the concept *comparable corpus*, the selected text materials from three different sources, i.e. LB, SUC and GP, will further on be referred to and used as one text collection constructed for the purpose of comparison of specific linguistic features across language type and genres. According to McEnery, Xiao and Tono (2006) a comparable corpus contains components from different languages that are collected using the same sampling frame and similar balance and representativeness, while Hunston (2002) also considers intralingual corpora collected according to the above specifics as comparable.

The approximate number of tokens in each text type and category are listed in table 3.3. Shorthand notation for the subcorpora will be used according to table 3.4.

Genre	**Ordinary**	**ETR**	**Total**
Children's fiction	421,452	143,789	565,241
Adults' fiction	254,000	115,770	369,770
Newspaper texts	206,703	390,802	597,505
Information texts	140,000	222,237	362,237
Total number of tokens	1,022,155	872,598	1,894,753

Table 3.3: Text distribution in the comparable corpus

3.2 Lexica

3.2.1 The NST Swedish Lexicon

In the early 2000's much effort was spent at the company Nordisk Språkteknologi Holding AS (NST) in Norway to collect and build acoustic

Notation	Subcorpus	Corpus origin
CEF	Children's ETR fiction	LB
COF	Children's ordinary fiction	LB
AEF	Adults' ETR fiction	LB
AOF	Adults' ordinary fiction	SUC
EN	ETR news	LB
ON	Ordinary news	SUC + GP
EI	ETR information	LB
OI	Ordinary information	SUC

Table 3.4: Subcorpora shorthand labels and origin

and lexical language resources for the three Scandinavian languages. The resource-building was a necessary groundwork for further production of HLT applications for Norwegian, Swedish and Danish. The starting point for production of the lexical databases was frequency-based unlemmatized wordlists, drawn from text corpora. The words in each lexicon were part-of-speech tagged, phonetically transcribed according to the standards for the current language, supplied with morphological information, and compounds were annotated with regard to the concatenation of and possible infixes between the various compound elements. Abbreviations and acronyms were identified, marked-up and possibly expanded.

After the closing-down of NST, the resources have been transferred to the custody of the Norwegian Language Bank. The material has been examined and quality checked with the conclusion that it is valuable material, substantially in compliance with international standards for similar language resources (Andersen 2005).

In this study, the Swedish pronunciation lexicon was used for the task of syllabification. It consists of 927,167 entries, all supplied with at least one phonetic transcription. About 63 % of the lexicon is covered by nouns, while verbs amount to 16 %, adjectives to 14 %, proper nouns to 5 %, adverbs to 0.3 % and other grammatical categories to 0.2 %. The vocabulary is general and no specific domain is represented. The format is based on Parole/SIMPLE for the morphosyntactic annotation and the SAMPA for phonetic transcription.

3.2.2 Saldo

Saldo (Borin and Forsberg 2009) is a modern Swedish semantic and morphological lexicon. The organization differs in a fundamental way from the widely used lexical-semantic database Princeton WordNet (Miller 1995; Fellbaum 1998a, b) even though both are based on psycholinguistic principles. While Princeton WordNet and its descendant Swedish WordNet (Viberg et al. 2002), are organized in encoded concepts in terms of sets of synonyms, called synsets, the associative relations between the entries in Saldo are based on metaphorical kinships that are specified as strictly hierarchical structures. Every entry in Saldo has a descriptor, either primary or secondary. At the top of the hierarchy is an artificial most central entry, the PRIM, which is used as the descriptor of 50 semantically unrelated entries. In this way, all entries become totally integrated into a single rooted tree without cycles.

3.2.3 Swedish Base Lemma Vocabulary Pool

The Swedish base lemma pool vocabulary (SBVP) Forsbom (2006) contains 8,215 entries ranked according to relative frequency weighted with dispersion, i.e. how evenly spread-out they are across the SUC subcorpora. The major drawback of this word list is that it provides information only at the base form level, i.e. only the part-of-speech of a word is used for disambiguation of two homographs. As an example, the word *vara* has five different lemma forms in SMDB, out of which three are nouns and two are verbs. The noun *vara* is either the lemma **1 vara** 'item', the **2 vara** as part of the multi-word-unit *ta vara på* 'take care of', and **3 vara** 'existence'. The verbal forms **4 vara** 'to be' is the most common, while **5 vara** has two separate lexeme forms 'to last' and 'to fester'. As a consequence, all the three nominal lemma forms are aggregated in the frequency counts. Obviously, the same holds for the verbal forms, despite their obvious differences in meaning.

3.2.4 SweVoc

Ogden (1930) described already in the 1930's how a 850 base word vocabulary could be used to describe in principle all the 25,000 words in a pocket dictionary. According to the author, this was possible through paraphrasing and exchange of complex words into more simple ones. A more elaborated proposal of a linguistic base vocabulary has been

given for Italian by De Mauro (1980), who suggested an amount of words which, depending on context, can be regarded as fundamental and necessary for everyday communication. The author divides the words in three different groups, depending on their use:

- Everyday words that reliability tests have shown to be comprehensible for primary students
- An intersection of these with high frequency use
- A set of words that do not appear very often in spoken or written language, but are tightly connected to everyday situations and objects, for instance *toothbrush*

For Swedish, no extensive list of base vocabulary words existed at the initial phase of this study, apart from the Swedish Base Vocabulary Pool described earlier, and access to a reliable list was regarded as an essential part of the thesis. This is why SweVoc (Heimann Mühlenbock and Johansson Kokkinakis 2012) was compiled. It is a comprehensive resource, based on material from four different sources. The backbone is the monolingual SBVP, which was enlarged with information from a translated version of the earlier mentioned work by De Mauro (1980). It can be argued that translation of a foreign word list is an old-fashioned and time-consuming way of gaining information that could easily be extracted from a large enough balanced corpus. A justification of this approach is to be found in the distinction between *base vocabulary* and *core vocabulary*. A language teaching situation might for instance involve a more extensive base vocabulary, while assistive technology applications mostly rely on a restricted core vocabulary, expandable with complementary vocabulary items from specific domains. From this follows that a core vocabulary should ideally contain words denoting universal concepts, central to the language as a whole, completed with language-specific function words. Another reason for the translation approach is that word lists from different languages can provide information about meaning (Dagan, Itai and Schwall 1991), otherwise hidden in a single base form of a word. Furthermore, a large amount of words denoting everyday objects do very seldom appear in writing.

The International Classification of Functioning, Disability and Health, known more commonly as ICF, is a classification of health and health-related domains, provided by WHO (World Health Organization 2001). The Swedish version is published by The Swedish National Board of

Health and Welfare (Socialstyrelsen 2003). The domains in ICF are classified from body to individual and societal perspectives. Since an individual's functioning and disability occurs in a context, the ICF also includes a list of environmental factors. Beyond its purpose to serve as a tool for medical and rehabilitation assessment, it is also produced as a means to describe situations regarding overall human functional states and limitations and it also serves as a frame for structuring knowledge and information within this domain. The ICF has among other things also been used for creation of medical dictionaries (Nyström et al. 2006). In the present work, it was used as a source for augmenting the SweVoc list with words belonging to the everyday vocabulary. It integrates words denoting body structure and functions, activities, self-care, getting along and interacting with other people, domestic life activities and participation in community activities.

The Kelly modern vocabulary list (Johansson Kokkinakis and Volodina 2011) of approximately 300 words translated between Italian and Swedish was aslo employed. It ensured that frequent words used in more modern settings than the list of De Mauro (1980) were included.

The result after combining these four sources is a word list with $\approx 8,000$ entries at the lemma level, pertaining to one of six different categories, as listed in table 3.5.

The amount of words and phrases that should be included in a base vocabulary obviously depends on the structure of the natural language. We can for example assume that a language like Italian uses more verbs than Swedish because the Romance languages' morphology adds the content that in present Swedish is expressed by verbs and adverbial particles, cfr. it *accompagnare* sv *följa med*.

One example of difference between SweVoc and SBVP is the word **torka**, which can be either a verb with one lemma and two lexeme forms **1 torka** 'becoming dry', **2 torka** 'cause to be dry', or a monosemous noun **3 torka** 'dry weather'. In SweVoc, these entries are represented as: **torka V C** 'becoming dry', **torka V H** 'cause to be dry' and **torka NCU S** 'dry weather'.

3.3 Language resources and information accessibility

Language technology tools and methods for improved accessibility of information are becoming a topic in a number of public initiatives. Several stakeholders have expressed the need to establish an openly ac-

Category	Description	Example and POS	No. of entries
C	Core vocabulary	och 'and' CC	2,201
D	Words referring to everyday objects and actions	lärare 'teacher' NCU	1,019
H	Highly frequent words	samtal 'conversation' NCN	1,518
K	Present in the Kelly word list	debatt 'debate' NCU	288
S	Supplementary words (from SBVP)	sammanhang 'context' NCN	3,442
M	Words related to the medical domain	trumhinna 'eardrum' NCU	139

Table 3.5: SweVoc categories and number of entries

cessible national language bank for storage of language databases and associated analysis tools, and the Swedish Language Council has been commissioned to design such a proposal.

Turning to the resources utilized in this study, they are all produced as open source materials, but with certain restrictions. The following resources are distributed by Språkbanken, University of Gothenburg <http://spraakbanken.gu.se>:

- SUC 2.0 is available for scientific research only, and under individual lincense
- LäSBarT and Göteborgs-Posten are accessible for on-line search in the corpus search interface *Korp* and sentence-transposed versions are free of use under CC BY 3.0
- Saldo is free of use under CC BY 3.0
- SweVoc will be free of use under CC BY 3.0

4 METHOD

This chapter is dedicated to the design of the study and the descriptive statistical methods used. The multilevel language feature model adopted is described and the set of linguistic features selected for further analysis is introduced. The text classification procedure and implemented algorithms are presented, followed by an account of the evaluation procedure. Finally, a new language model is proposed.

4.1 Design of the study

Chall's (1958) multilevel ordering was maintained when creating a framework for a new readability model. After a literature survey of features considered to have an impact upon text complexity, and implicitly on the degree of readability, appropriate instances were selected for further development into a general feature model. Descriptive statistical analyses at each representation level were made of ETR and ordinary texts in order to find discriminative, or "typical" properties. The discriminative properties were expected to be found within, as well as across genres. Some of the features resting on deep linguistic features were partially described in Mühlenbock and Johansson Kokkina-kis (2009). Mean ($\bar{X}$), standard deviation (σ), variation (Var), and variance range of each of the features listed in 4.1 was computed. Significance testing of differences in the materials was made using two-tailed t-test and the significance level was set to $p < 0.001$. The statistical computing was made with the R programming language (Ihaka96 1996).

In connection to the surface feature analysis, each of the traditional readability formulas was implemented and evaluated.

The text classification task was performed in order to build statistical language models that can be re-used for training of additional materials, but also to prove the fairness of the corpus-based approach. After the descriptive analysis, a second selection round was made where only

features acting at the document level were retained. Some of the features were expected to be genre specific, which was further explored by binary cross-comparison of all document sets. The feature candidates were fed as vector variables into three different text classification algorithms; the Naïve Bayes, a support vector machine implementation, and classification via regression. If the text classification task succeeded, i.e. if the classifier proved to show significantly better accuracy, precision and recall results than the traditional LIX formula, it could be taken as evidence for the assumption that the feature set was correctly chosen and that it was likely to influence readability. It was further to be explored whether the three mentioned algorithms showed similar results, or if one of them could be interpreted as superior to the others in some respect.

After descriptive statistical analyses, significance testings and principal component analyses of the impact of specific features on the classification results, a final, reduced feature model was achieved. It consists of a subset of the features presented in table 4.1.

4.2 Text classification

Three algorithms for text classification were selected and implemented in the Waikato Environment for Knowledge Analysis, or Weka, software (Hall et al. 2009). Weka is a suite of machine learning and data mining tools and algorithms implemented in Java. Only the suite of GUI applications was used, although another part exists, containing a library of Java classes for including Weka directly into Java applications. In order to find the best algorithm for each task, the three different algorithms were tested in each run of the experiment. The Naïve Bayes and support vector machines (SVMs) are two models often mentioned in data mining literature, where one difference is connected to the training time aspect. While the SVMs need roughly cn^2 times to train, where n is the number of training examples and c is an algorithm-dependent constant, the Naïve Bayes algorithm is linear. This can obviously be of importance in tasks involving very large data sets. Useful statistical methods for comparing classifiers are recommended in (Tan, Steinbach and Kumar 2006), but time constraints for the study limited evaluation to accuracy, precision, recall and F-score calculations.

In this study, based on other properties than purely lexical properties, we are interested to know whether the features suggested to be discriminative by prior statistical measurements, also correspond to

Feature category	Feature	Id
Surface	Mean word length in characters	MWLC
	Mean word length in syllables	MWLS
	Mean sentence length in word tokens	MSL
	Number of word tokens > 13 characters	XLW
	Type-token ratio	TTR
	Lexical neighborhood density	LND
	Lexical neighborhood frequency	LNF
	OVIX	OVIX
	LIX	LIX
Vocabulary	Lemma variation index	LVIX
load	% words in SweVoc	SV
	% SweVoc cat (C)	SVC
	% SweVoc cat (D)	SVD
	% SweVoc cat (H)	SVH
	% SweVoc cat (K)	SVK
	% SweVoc cat (S)	SVS
Sentence	Mean dependency distance	MDD
structure	Number of subordinate clauses	UA
	Number of prenominal modifiers	AT
	Number of postnominal modifiers	ET
	Mean parse tree height	PT
Idea density	Propositional density	Pr
	Ratio nouns/pronouns	NoPr
	Nominal ratio	NR
	Mean semantic depth	Sa
Human interest	Mean amount of personal nouns	PM

Table 4.1: The feature set investigated

the features selected by the classifier in a primary component analysis (PCA). A PCA is commonly used to reduce a large set of variables into a smaller, more informative set of uncorrelated components.

All document sets were thus cross-compared, which implies that 28 separate classification tasks were performed. The document pairs were assigned to one of the following four categories:

1. Documents belonging to the same genre and of the same type

2. Documents belonging to the same genre, but of different types

3. Documents belonging to different genres, but of the same type

4. Documents belonging to different genres, and of different types

4.2.1 Naïve Bayes

One established classifier is the naïve Bayes classification algorithm. It is based on Bayes' rule, which assumes the attributes $X_1 \cdots X_n$ to be conditionally independent of one another, given the class variable Y. To give an example, we assume that a document is considered to be less complex if it has a low mean sentence length (MSL), a low mean word length in syllables (MWLS), and a low mean syntactic dependency distance (MDD). Even if these variables depend on each other or upon the existence of other variables, a naïve Bayes classifier considers all of these properties to independently contribute to the probability that the document belongs to the ETR class.

4.2.2 SMO

One implementation of a SVM uses the Sequential Minimal Optimization (SMO) algorithm. It is directed towards the problem of optimizing a large quadratic function of several variables by breaking it down into a series of smallest possible problems that are analytically solvable (Platt 1999).

4.2.3 Classification via Regression

Classification via regression (CVR) is a method for modelling the conditional class probability function of each class. It uses the algorithm *model trees*, which is a combination of regression and tree induction. Regression refers to the process of estimating a numeric target value and it is used to solve a classification problem with a learner that can only produce estimates for a numeric target variable. During training, one function is learned for each class; the attribute values are used as input and with possible output values 1 and 0, indicating whether the current training instance belongs to this class or not.

4.2.4 Feature vectors

Twenty-two features supposed to be relevant for document classification were picked and fed into a 22-dimensional vector. Each feature set will be described below, along with distinguishing shorthand labels for each feature. The LIX value is adopted as baseline for all classification performance accounts.

4.3 Document classification

The document classification task required texts comparable in size. This was achieved by splitting the subcorpora into chunks of 30 sentences each, here referred to as "documents". The total number of documents within each text type and genre is presented in table 4.2. For each task, an appropriate number of documents were randomly picked from the larger set in order to achieve a match in size between each test set. Twenty-eight experiments were performed in order to compare the performance of three algorithms adopting two different language models. The base model is simply made up by LIX values, i.e. surface properties, while the second, SVIT model, consists of multilevel feature values related to surface properties, *vocabulary load*, *sentence structure*, *idea density* and degree of *human interest*. In order to eliminate the risk of bias, a reduced SVIT model was also tested. The reduced SVIT consisted of the feature set described above, but with the removal of all features signaling surface properties. Pairwise classification of texts across ages of the intended audience, genres and text types were made. The document pairs were grouped into four different categories with respect to text type and genre relation. Test set sizes and categorizations are presented in table 4.3. Finally, all eight document test sets in table 4.2 were used for multiclass classification.

4.4 Classification evaluation

Evaluation of the different classification tasks were performed by means of 7-fold cross validation. An n-fold cross validation splits the data set into a set of n equally large sets and runs the experiment n times, in which one fold of data is used as test set and the rest $n-1$ folds are used as training set. Another approach is to use a partition of the data set as training set and to hold out the rest as test set. For sufficiently large data

Subcorpus	**Notation**	**Total no of documents**
Children's ETR fiction	CEF	562
Children's Ord fiction	COF	1,416
Adults' ETR fiction	AEF	424
Adults' Ord fiction	AOF	700
ETR news	EN	1,192
Ord news (SUC)	ON	240
Ord news (GP)	ON	264
ETR information	EI	581
Ord information	OI	357

Table 4.2: Number of documents in each subcorpus

sets, the cross validation is usually preferred as it renders smoothed results and eliminates statistical anomalies that might arise in individual test runs.

Accuracy, precision, recall and F-score, based on the counts of test records correctly and incorrectly predicted by each model was calculated. Precision is calculated as the percentage of documents that were correctly classified, while recall is the percentage of relevant documents that were actually correctly classified. The F-score combines precision and recall into a harmonic mean according to the formula

$$Fscore = 2 * \frac{precision * recall}{precision + recall} \tag{12}$$

4.5 Principal component analysis

Usually, a principal component analysis (PCA) is performed in order to assess whether a few components account for a large proportion of the variation in a dataset. The basic idea is to describe the variation in terms of a set of new, uncorrelated variables, each of which defined to be a particular linear combination of the original raw data variables. In regression analysis the relationship between two variables can be seen as a line in a graph, so that for any point on the line, it is possible to extract the value for both of the correlated variables through the coordinates. A PCA can be regarded as a combination of multiple regression analyses, where a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space with 1 axis per variable.

Class type	Document sets	Notation	No of doc
Same genre and same type	Children's ETR fiction/Adults' ETR fiction	CEF/AEF	420/420
	Children's Ord fiction/Adults' Ord fiction	COF/AOF	700/700
Same genre and different types	Children's ETR fiction/Children's Ord fiction	CEF/COF	560/560
	Children's ETR fiction/Adult's Ord fiction	CEF/AOF	560/560
	Adults' ETR fiction/Children's Ord fiction	AEF/COF	420/420
	Adults' ETR fiction/Adults' Ord fiction	AEF/AOF	420/420
	ETR news/Ord news	EN/ON	500/500
	ETR info/Ord info	EI/OI	350/350
Different genres and same type	Children's ETR fiction/ETR news	CEF/EN	560/560
	Children's ETR fiction/ETR info	CEF/EI	560/560
	Adults' ETR fiction/ETR news	AEF/EN	420/420
	Adults' ETR fiction/ETR info	AEF/EI	560/560
	Children's Ord fiction/Ord news	COF/ON	240/240
	Children's Ord fiction/Ord info	COF/OI	350/350
	Adults' Ord fiction/Ord news	AOF/ON	240/240
	Adults' Ord fiction/Ord info	AOF/OI	350/350
	ETR news/ETR info	EN/EI	580/580
	Ord news/Ord info	ON/OI	240/240
Different genres and different types	Children's ETR fiction/Ord news	CEF/ON	240/240
	Children's ETR fiction/Ord info	CEF/OI	350/350
	Adults' ETR fiction/Ord news	AEF/ON	240/240
	Adults' ETR fiction/Ord info	AEF/OI	350/350
	Children's Ord fiction/ETR news	COF/EN	1190/1190
	Children's Ord fiction/ETR info	COF/EI	580/580
	Adults' Ord fiction/ETR news	AOF/EN	700/700
	Adults' Ord fiction/ETR info	AOF/EI	350/350
	Ord news/ETR info	ON/EI	240/240
	Ord info/ETR news	OI/EN	350/350

Table 4.3: Document sets and sizes for classification

By rotating the data space, a lower-dimensional picture can be viewed from its most informative viewpoint.

Weka's implementation of principal components does a Varimax rotation, which is a computerized algorithm for the objective, or analytic, transformation (orthogonal rotation) of factor axes. Its goal is to minimize the complexity of the components by making the large loadings larger and the small loadings smaller within each component. Each consecutive factor maximizes the variability that is not captured by the preceding factor, which means that consecutive factors are uncorrelated and orthogonal to each other.

4.6 SVIT - The proposed readability model

The first selection of features and formulas involved in different reading situations resulted in a set of 27 variables, described in table 4.1. Some features were found to be redundant, and were excluded in the following processing. Descriptive statistical analyses confirmed that a language feature model consisting of 22 variables was appropriate for further exploration. This model will be referred to as *SVIT*, an acronym for (S)entence structure, (V)ocabulary load, (I)dea density, and human and personal interes(T), but also of the corresponding Swedish words (S)yntax, (V)okabulär, (I)ntressegrad and idé(T)äthet (syntax, vocabulary, degree of interest and idea density).

5 DESCRIPTIVE ANALYSIS

This chapter will provide an overview of the descriptive statistics of the comparable corpus. Partial descriptions and comparisons will also be made of SUC and LB general characteristics.

An account will first be given of the texts by means of traditional readability formulas, operating on text surface and phonological levels. As mentioned earlier, the most widely used formula for Swedish is LIX, but there will also be a comparison of the outcome of LIX to prevalent readability measures for English. For LIX, the readability metrics are based on word length in characters and sentence length in words. The English metrics Flesch RE, Flesch-Kincaid, ARI, Coleman-Liau, Gunning Fog, and SMOG formulas and indices operate on the same units, or on word length as a function of syllables. The starting point will thus be a characterization of textual surface features. In order to give a full account of characteristics in ordinary text, the full set of SUC 2.0 subcorpora will also be described in terms of surface features. The test sets described in table 4.2 will undergo further analysis in that all feature values will be compared and tested for significance across text types and genres, according to the classification scheme in table 4.3. Significance testing was performed by means of Welch two-tailed two-sample T-test, and the level for significance was set to $p < 0.001$.

Some authentic examples from the corpora will serve as illustrations in the following account of features. A few short sentence examples are given in running text, while more exhaustive examples retrieved from the subcorpora are placed in appendix B.

Example 2 *Även/även om/om vi/jag har/ha rett/reda ut/ut det/den är/vara det/den som/som om/om något/någon liksom/liksom ligger/ligga kvar/kvar och/och gnager/gnaga.*

'Even if we have sorted it out, it feels like something is still gnawing.'

Example 2 is selected from the children's ETR fiction (CEF), and corresponds to a LIX value of 17, which is clearly within the range for that specific text genre and type.

Example 3 *Många/mången vill/vilja ha/ha texter/text som/som är/vara lätta/lätt att/att läsa/läsa med/med det/den allra/allra viktigaste/viktig och/och utan/utan svåra/svår ord/ord.*

'Many [people] want texts that are easy to read with the most important [things] and without difficult words.'

Example 3 has been retrieved from the ETR information (EI) part. It has a LIX value of 23, which is slightly above the lower limit for young adults' literature.

5.1 Surface text analysis

5.1.1 Word length in characters

Word length in characters was calculated as the mean length of all word tokens in the subsets of the comparable corpus. Mean word length in characters (MWLC), standard deviation (σ) and variance (Var) are given in tables 5.1 at the sentence level, and 5.2 at the document level.

The MWLC was significantly different ($p < 0.001$) for the two principal corpora, $MWLC = 4.64$ for LB and $MWLC = 5.32$ for SUC, when calculated as the mean of the mean values for each sentence. See figure 5.1 for plots of the raw frequencies. The texts showed consistent results, regardless of whether they were calculated on the sentence or document level.

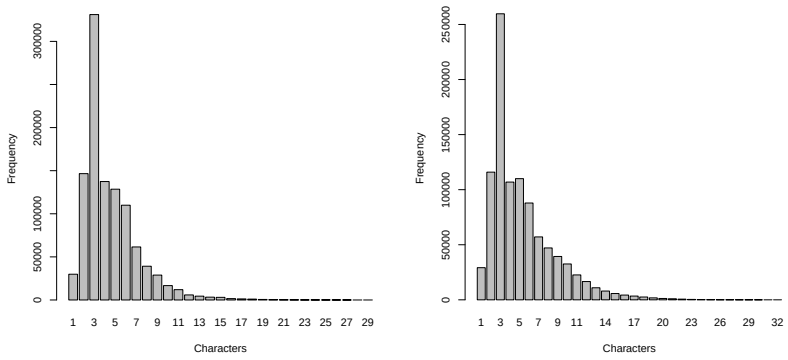
Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	1.0 – 17.0	4.2	0.82	0.68
Children's Ord fiction	COF	1.0 – 20.0	4.3	0.95	0.90
Adults' ETR fiction	AEF	2.0 – 13.0	4.4	0.84	0.70
Adults' Ord fiction	AOF	1.0 – 22.0	4.6	1.07	1.14
ETR news	EN	1.9 – 14.5	5.0	0.97	0.94
Ord news	ON	1.0 – 22.0	5.3	1.23	1.51
ETR information	EI	2.0 – 25.0	5.1	1.60	2.58
Ord information	OI	1.0 – 30.0	5.9	2.26	5.12

Table 5.1: Mean word length in characters (MWLC) in sentences

Subcorpus		**Label Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	3.6 – 4.6	4.1	0.18	0.03
Children's Ord fiction	COF	3.5 – 4.9	4.3	0.21	0.04
Adults' ETR fiction	AEF	3.9 – 5.2	4.4	0.20	0.04
Adults' Ord fiction	AOF	3.7 – 5.8	4.6	0.31	0.10
ETR news	EN	4.1 – 5.9	5.0	0.23	0.05
Ord news	ON	4.2 – 6.1	5.2	0.37	0.14
ETR information	EI	3.9 – 6.2	4.9	0.40	0.16
Ord information	OI	4.6 – 7.4	5.8	0.53	0.28

Table 5.2: Mean word length in characters (MWLC) in documents

Paired significance tests show a p-value < 0.001 across all subcorpora, i.e. there is a statistical difference in word length between all text types and genres.



(a) LB *(b)* SUC

Figure 5.1: Word length frequency distributions in LB and SUC

5.1.2 Word length in syllables

If we replicate Björnsson's calculations on LB and SUC, we find that 93 % (LB) and 96 % (SUC) of the polysyllabic words are indeed long. The corresponding figures for mono- and bisyllabic short words are 95 % (LB) and 94 % (SUC), respectively. See tables 5.3 and 5.4.

Word types	No. of characters					
	1–3	4–6	7–9	10–12	> 12	Sum
Monosyllabic	48	9	–	–	–	57
2-syllabic	–	25	4	–	–	29
3-syllabic	–	1	7	1	–	9
Polysyllabic	–	–	1	2	2	5
Sum %	48	35	12	3	2	100

Table 5.3: Percentage of words in LB with regard to syllables and characters

Word types	No. of characters					
	1–3	4–6	7–9	10–12	> 12	Sum
Monosyllabic	42	8	–	–	–	50
2-syllabic	–	22	5	–	–	27
3-syllabic	–	1	9	2	–	12
Polysyllabic	–	–	2	5	4	11
Sum %	42	31	14	4	5	100

Table 5.4: Percentage of words in SUC with regard to syllables and characters

A closer look at the text sets strengthens the theory of a high correlation between word length as a measurement of syllables and characters. The results show identical tendencies, i.e. the children's ETR texts has the lowest and the ordinary news texts the highest mean number of syllables per word, based on the sentence as well as the document level, see tables 5.5 and 5.6. The intermediate counts are also ranged accordingly. Furthermore, character and syllable counts in the entire LB and SUC corpora were plotted together in figure 5.3 and followed an almost dead straight line, which leads to the conclusion that syllable counting seems to be redundant in the presence of character counts.

At the sentence level, significance tests showed a p-value < 0.001 for all pairs, except for the pair ETR news/ETR information. It shall however be noted that fuzzy syllabification was performed for words outside the NST pronunciation lexicon, and that some of the ordinary texts contained a large proportion of previously unknown words.

As an attempt to somewhat illuminate the question whether Swedish orthographic words, in the same manner as English words tend to correlate in length with frequency and lexical neighborhood density, a scatterplot was made of all words in LB and SUC, respectively. The results are shown in figure 5.6.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	1.0 – 6.0	1.53	0.32	0.10
Children's Ord fiction	COF	1.0 – 7.0	1.56	0.34	0.12
Adults' ETR fiction	AEF	1.0 – 8.0	1.63	0.35	0.12
Adults' Ord fiction	AOF	1.0 – 8.0	1.67	0.38	0.14
ETR news	EN	1.0 – 10.0	1.90	0.42	0.17
Ord news	ON	1.0 – 9.0	1.94	0.48	0.23
ETR information	EI	1.0 – 13.0	1.90	0.59	0.35
Ord information	OI	1.0 – 11.0	2.19	0.73	0.54

Table 5.5: Mean word length in syllables (MWLS) in sentences

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	1.31 – 1.74	1.51	0.06	0.00
Children's Ord fiction	COF	1.30 – 1.76	1.54	0.07	0.01
Adults' ETR fiction	AEF	1.39 – 1.87	1.60	0.08	0.01
Adults' Ord fiction	AOF	1.37 – 2.12	1.65	0.10	0.01
ETR news	EN	1.60 – 2.16	1.85	0.09	0.01
Ord news	ON	1.53 – 2.30	1.91	0.14	0.02
ETR information	EI	1.45 – 2.53	1.81	0.16	0.03
Ord information	OI	1.70 – 2.64	2.14	0.20	0.04

Table 5.6: Mean word length in syllables (MWLS) in documents

5.1.3 Sentence length

Mean sentence length in words (MSL), standard deviation (σ), and variance (Var) for the materials are given in tables 5.7 and 5.8. Sentence lengths vary largely between all subcorpora, and a t-test shows a significant difference for all pairs with a $p - value < 0.001$.

Not surprisingly, children's ETR fiction turns out to have the shortest MSL with 7.3, calculated both at the sentence and document levels. Sentence length seems to be discriminative for the ETR texts, although the MSL values at sentence level in the ETR information and adults' ordinary fiction subcorpora are fairly close (11.5 and 11.9, respectively).

5.1.4 Comparison of readability formulas for Swedish and English

The Swedish LIX formula operates on word length in characters, as opposed to readability formulas for English which mainly operate on

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	1 – 30	7.25	3.60	12.99
Children's Ord fiction	COF	1 – 78	8.57	5.20	26.99
Adults' ETR fiction	AEF	1 – 48	7.88	3.72	13.85
Adults' Ord fiction	AOF	1 – 306	11.86	9.66	93.37
ETR news	EN	1 – 50	9.75	4.12	16.98
Ord news	ON	1 – 93	13.97	8.11	65.79
ETR information	EI	1 – 69	11.49	6.06	36.77
Ord information	OI	1 – 81	13.48	9.48	89.92

Table 5.7: Mean sentence length in words (MSL) in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	4.4 – 10.4	7.27	1.03	1.06
Children's Ord fiction	COF	5.3 – 13.7	8.61	1.54	2.38
Adults' ETR fiction	AEF	5.1 – 13.4	7.89	1.31	1.73
Adults' Ord fiction	AOF	3.7 – 33.4	12.28	4.07	16.64
ETR news	EN	6.6 – 14.9	9.86	1.08	1.17
Ord news	ON	6.8 – 24.0	14.00	2.69	7.25
ETR information	EI	4.5 – 17.6	11.56	1.91	3.67
Ord information	OI	2.4 – 25.5	13.63	3.41	11.66

Table 5.8: Mean sentence length in words (MSL) in documents

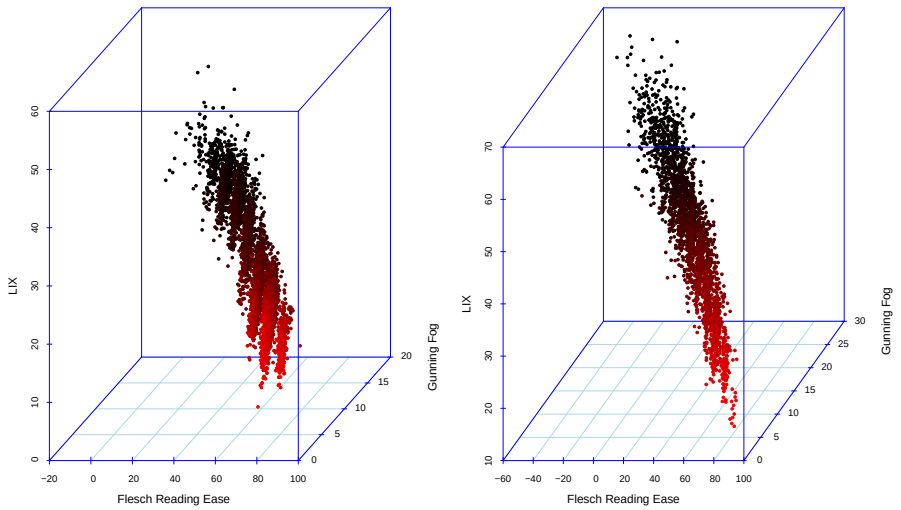
syllable counts. The reason is probably to find in the ease of character counting, as compared to the calculation of syllables. From the readability perspective, Björnsson (1968) regarded them as equal measures, based on the correlation between character and syllable counts in word length calculations, and the earlier mentioned analysis made on LB and SUC supports this hypothesis. A conclusion of this is that readability formulas operating on surface features could be interchangeable, at least between Swedish and English.

An inquiry was made on the LB and SUC subcorpora in order to test this idea. The seven readability formulas mentioned in 2.4.2 were applied on the total amount of texts in each subcorpus, split into chunks of 30 sentences each. This means that seven different readability indices were assigned to each of the five different genres in LB and twelve different genres in SUC. As opposed to the other readability indices, the lowest value of the Flesch RE is to be considered as indicating the highest reading difficulty. To make the values comparable, the Flesch RE

values were inverted and multiplied by 10. The values in table 5.10 were plotted in figures 5.4 and 5.5.

The first observation is that the inverted FRE, ARI, Coleman-Liau, Gunning Fog, and SMOG show almost identical behaviors, which is foreseen as they operate on similar variables and are expected to indicate a certain reading level. Another observation is that the LIX values follow the line of these RI fairly well for the SUC corpus, but behave slightly different for the LB corpus. Finally, we can see that the inverted Flesch RE values are reaching extreme levels for some of the SUC subcorpora. The SUC B (editorial press), SUC H (miscellaneous) and SUC J (learned and scientific writing) subcorpora are having the highest values in all RI measurements, and are also the materials showing the highest MWLS values in the SUC corpus. The fact that the FRE values are multiplied by 10 is most certainly contributing to the deviation from the other values in the plot.

In order to see how the three different types of readability formulas correlate for all sentences in the both corpora, scatterplots were made of the outcomes from FRE, ARI and Gunning Fog, displayed in figure 5.2. The plots show very neatly how the values correlate, but also a smaller variation in surface text complexity within the LB corpus.



(a) LB *(b)* SUC

Figure 5.2: Readability scores for LB and SUC

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	9.8 – 27.7	17.2	3.25	10.60
Children's Ord fiction	COF	9.0 – 35.5	21.6	3.81	14.49
Adults' ETR fiction	AEF	13.2 – 32.6	22.4	3.78	14.26
Adults' Ord fiction	AOF	12.4 – 52.5	30.0	7.01	49.10
ETR news	EN	17.2 – 46.7	34.8	3.34	11.13
Ord news	ON	22.7 – 58.2	40.9	6.12	37.43
ETR information	EI	17.9 – 49.4	33.3	5.14	26.39
Ord information	OI	29.2 – 64.1	46.8	6.70	44.72

Table 5.9: LIX reference values in documents

corpus	No. doc	FRE	FKF	ARI	Cole-man	FOG	SMOG	LIX
SUC corpus	2472	31.7	12.3	10.1	12.6	14.8	13.2	40.2
Press: Reports	241	32.0	12.0	9.6	12.6	14.5	13.0	40.1
Press: Editorial	79	20.5	14.2	12.2	14.7	16.7	14.5	45.2
Press: Reviews	132	31.1	12.5	10.7	13.2	15.0	13.5	42.8
Skills, trades, and hobbies	297	32.8	12.0	9.7	12.4	14.4	12.9	39.1
Popular lore	217	28.0	13.1	10.9	13.3	15.8	14.0	42.7
Belles lettres, biography, memoirs	119	27.5	13.1	11.0	13.3	15.7	13.9	42.5
Miscellaneous	361	12.2	14.9	12.8	16.0	17.8	14.6	46.8
Learned and scientific writing	319	8.3	16.5	14.7	16.5	19.5	16.3	51.1
General fiction	434	54.3	8.9	6.5	8.6	10.8	10.7	30.4
Mysteries and science fiction	135	58.2	7.5	4.6	7.6	9.5	9.7	27.0
Light reading	96	53.1	9.3	7.1	8.9	11.2	11.0	32.0
Humour	35	53.6	8.6	6.1	8.7	10.5	10.5	30.2

Table 5.10: Readability formulas applied on SUC

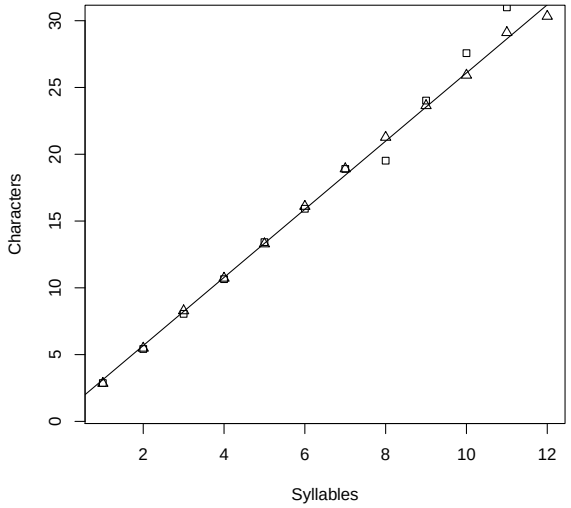


Figure 5.3: Mean length distribution of orthographical words in SUC = △ and LB = □ as a function of syllable and character counts

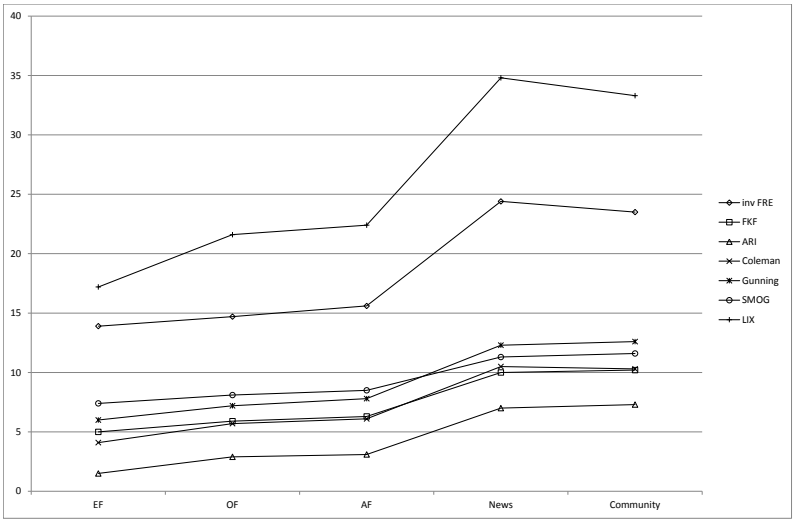


Figure 5.4: Comparison of different readability formulas applied on LB

corpus	**No. doc**	**FRE**	**FKF**	**ARI**	**Cole-man**	**FOG**	**SMOG**	**LIX**
LB corpus	4173	56.9	7.6	4.5	7.5	9.3	9.4	26.4
Children's ETR fiction	561	72.1	5.0	1.5	4.1	6.0	7.4	17.2
Children's Ord fiction	1415	68.1	5.9	2.9	5.7	7.2	8.1	21.6
Adults' ETR fiction	423	63.9	6.3	3.1	6.1	7.8	8.5	22.4
ETR info	580	42.5	10.2	7.3	10.3	12.5	11.6	33.3
ETR news	1191	41.0	10.0	7.0	10.5	12.3	11.3	34.8

Table 5.11: Readability indices applied on LB

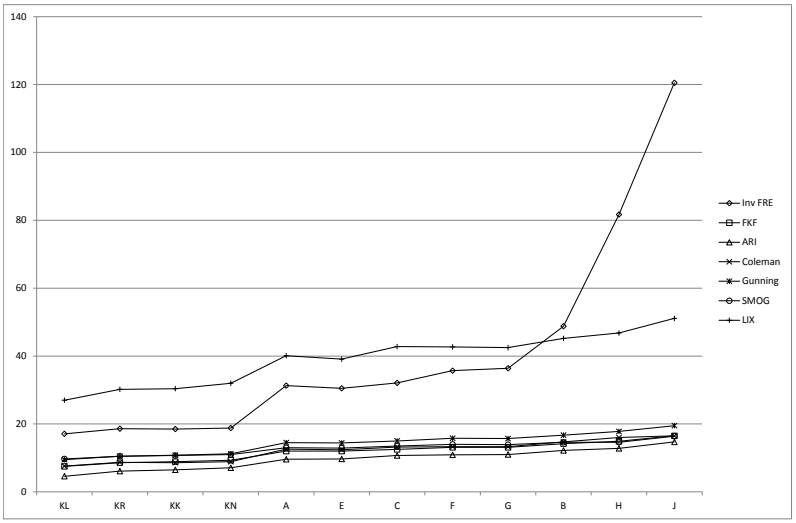


Figure 5.5: Comparison of different readability formulas applied on SUC

5.1.5 Extra long words

Words longer than 13 characters are regarded as extra long. For LB sentences, this means an average of 2 % of all words (range between 0 and 2), and 3 % for SUC (range between 1 and 6). This property is strongly related to genre – information, scientific writing and newspaper texts

tend to have the highest ratio of extra long words. The ratios of extra long words in documents are given in table 5.12, together with range and σ. Levels of variance were disregarded.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ
Children's ETR fiction	CEF	0.00 – 0.03	0.001	0.003
Children's Ord fiction	COF	0.00 – 0.03	0.002	0.004
Adults' ETR fiction	AEF	0.00 – 0.02	0.002	0.004
Adults' Ord fiction	AOF	0.00 – 0.04	0.008	0.008
ETR news	EN	0.00 – 0.07	0.016	0.011
Ord news	ON	0.00 – 0.07	0.027	0.014
ETR information	EI	0.00 – 0.10	0.027	0.020
Ord information	OI	0.00 – 0.16	0.054	0.028

Table 5.12: Mean ratios of extra long words in documents

5.1.6 Lexical neighborhood density and frequency

The LND and LNF counts for each word in example 2 are shown in table 5.13.

Word	**LND**	**LNF**	**Lemma**	**Pos**
även	0	0	även	AB
om	3	0	om	SN
vi	4	0	jag	PN
har	17	1	ha	VB
rett	15	11	reda	VB
ut	5	0	ut	PL
det	12	0	den	PN
är	6	0	vara	VB
det	12	0	den	PN
som	12	0	som	SN
om	3	0	om	SN
något	1	1	någon	PN
ligger	2	0	ligga	VB
kvar	4	0	kvar	PL
och	1	0	och	KN
gnager	0	0	gnaga	VB

Table 5.13: LND and LNF counts in example 2

Only three words have lexical neighbors in the reference language, i.e. the LB total word list, that are higher in frequency. The word *rett* has 15 lexical neighbors, and a lexical neighborhood frequency of 11. These words are, listed in order of frequency: *rätt*, *sett*, *gett*, *rött*, *bett*, *rent*, *fett*, *rest*, *lett*, *hett*, *ritt*. The remaining words, i.e. lexical neighbors with the same or lower frequency are: *rott*, *rått*, *vett*, *ratt*. In addition to the orthographical similarity, the two words *rett* and *rätt* are also homophones in standard Swedish, which increases the complexity. The word *har* has 17 neighbors, but only one word is higher in frequency, namely *han*. The word *något* is exceeded in frequency only by its lemma member *någon*.

Example 3 shows a different pattern, as can be seen in table 5.14.

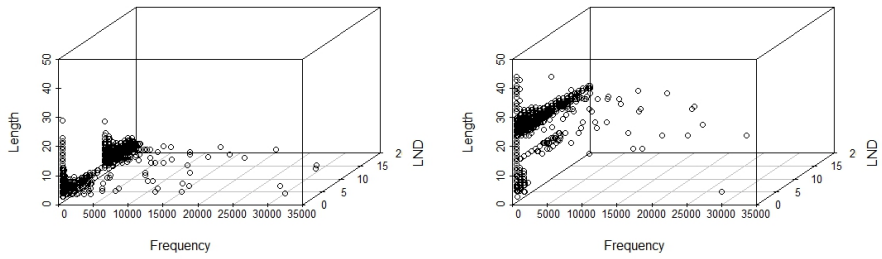
Word	**LND**	**LNF**	**Lemma**	**Pos**
många	2	0	många	JJ
vill	5	1	vilja	VB
ha	10	1	ha	VB
texter	3	1	text	NN
som	12	0	som	HP
är	6	0	vara	VB
lätta	7	2	lätt	JJ
att	2	0	att	IE
läsa	9	0	läsa	VB
med	9	0	med	PP
det	12	0	den	DT
allra	0	0	allra	AB
viktigaste	0	0	viktig	JJ
och	1	0	och	KN
utan	3	0	utan	PP
svåra	3	1	svår	JJ
ord	4	0	ord	NN

Table 5.14: LND and LNF counts in example 3

The words *vill*, *ha*, *texter*, *lätta*, and *svåra* have neighbors that are higher in frequency. The word *vill* is exceeded by the preposition *till*, the verb *ha* by another verb, namely *sa*, the noun *texter* by *texten*, which belongs to the same lemma, the adjective *lätta* by the verb *sätta*, and the adjective *svåra* by the verb *svara*.

Lexical neighborhood density and frequency was calculated as the mean of all words in each subcorpus, in relation to a reference vocabu-

lary consisting of the total set of words in the LB or SUC corpora. This means that all words in the ETR texts and the texts included in the children's ordinary fiction part of LB were compared against the total LB corpus, and all remaining ordinary texts were compared against SUC. There was a significant difference in mean lexical neighborhood density (LND) for all subcorpora studied.



(a) LB *(b)* SUC

Figure 5.6: Correlation plot between frequency, word length and LND in LB and SUC

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 22	6.61	5.51	30.32
Children's Ord fiction	COF	0 – 22	6.52	5.52	30.52
Adults' ETR fiction	AEF	0 – 22	6.25	5.58	31.18
Adults' Ord fiction	AOF	0 – 27	7.81	7.30	53.22
ETR news	EN	0 – 22	5.28	5.48	30.00
Ord news	ON	0 – 27	4.71	5.76	33.17
ETR information	EI	0 – 19	5.91	5.51	30.40
Ord information	OI	0 – 26	5.61	6.87	47.25

Table 5.15: Mean lexical neighborhood density (LND) for words

There was no significant internal difference in lexical neighborhood frequency (LNF) between the children's ETR fiction and children's ordinary fiction[7], neither between adults' ETR fiction and the ordinary news[8].

[7] $p = 0.571$
[8] $p = 0.145$

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 15	0.58	1.39	1.92
Children's Ord fiction	COF	0 – 19	0.58	1.43	2.05
Adults' ETR fiction	AEF	0 – 13	0.51	1.22	1.49
Adults' Ord fiction	AOF	0 – 18	0.72	1.75	3.06
ETR news	EN	0 – 17	0.36	1.08	1.17
Ord news	ON	0 – 20	0.51	1.46	2.13
ETR information	EI	0 – 16	0.47	1.28	1.63
Ord information	OI	0 – 18	0.53	1.56	2.45

Table 5.16: Mean lexical neighborhood frequency (LNF) for words

5.1.7 Type/token ratio

The value of raw frequency counts showing the ratio of types/tokens (TTR) can normally be questioned, since it lacks important information at the lemma/lexeme level. For our purposes, it can still be used as a means to illustrate the variation of graphical words at the text surface and within similar sizes of text chunks. As can be seen in table 5.17, the values are fairly much identical, except for the ETR information texts, which had a lower mean TTR. Corpus example B.8 has a TTR of 0.41, which indicates a low lexical variation.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0.46 – 0.73	0.62	0.04	0.002
Children's Ord fiction	COF	0.41 – 0.76	0.61	0.05	0.002
Adults' ETR fiction	AEF	0.48 – 0.71	0.60	0.04	0.002
Adults' Ord fiction	AOF	0.44 – 0.75	0.61	0.05	0.003
ETR news	EN	0.46 – 0.72	0.60	0.04	0.002
Ord news	ON	0.47 – 0.75	0.62	0.05	0.002
ETR information	EI	0.25 – 0.71	0.49	0.07	0.004
Ord information	OI	0.31 – 1.00	0.58	0.09	0.008

Table 5.17: Mean type/token ratios (TTR) in documents

5.1.8 OVIX

There was a significant difference in OVIX for all text sets. See table 5.18.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	41 – 94	63.7	8.33	69.37
Children's Ord fiction	COF	40 – 101	66.3	9.88	97.54
Adults' ETR fiction	AEF	41 – 86	61.5	7.06	49.83
Adults' Ord fiction	AOF	41 – 113	73.5	11.04	121.95
ETR news	EN	47 – 101	66.5	7.36	54.24
Ord news	ON	51 – 120	81.2	11.46	131.25
ETR information	EI	28 – 83	51.1	7.84	61.54
Ord information	OI	33 – 191	70.4	15.71	246.94

Table 5.18: Mean OVIX values in documents

5.2 Deeper linguistic analysis

5.2.1 Vocabulary

5.2.1.1 Lexical variation

The lemma variation indices for the different text sets are shown in table 5.19. Statistically, there was no significant difference in this measure for the pairs children's ETR fiction/ETR news and children's ETR fiction/adults' ETR fiction. According to the standards set by Hultman and Westman (1977), values below 60 would denote a low lexical variation, which seems to hold for all the subcorpora in LB.

At this point, it is time to compare the outcome of three different methods for determining the variation of lexical representations in the materials. At the surface, it is analyzed as the ratio of word types to word tokens, calculated as raw frequencies of orthographic representations. Adjustment of the figures with regard to hapax words and differences in sample size was made by using the OVIX formula. In order to capitalize on the information supplied by the lemmatizer, we have also calculated the lemma variation index, which resulted in a plot displaying a smaller set of units that are expected to mirror lexical variation at the semantic level. For readability analysis purposes, the first method relates to the decoding of graphical words, i.e. how many words are encountered during the reading of a specific text passage, and how many of these are repeated? The second method is merely an adjustment of the previous one, while the third method seems to be a practical way to measure the variation of meaning in the text in question. Plots of the

outcome of all the three calculations on the ETR news subcorpus are shown in figure 5.7.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	34 – 76	52.7	6.34	40.21
Children's Ord fiction	COF	34 – 80	55.5	7.72	59.66
Adults' ETR fiction	AEF	35 – 74	51.4	5.83	34.02
Adults' Ord fiction	AOF	35 – 96	62.3	9.17	84.08
ETR news	EN	37 – 75	53.2	5.42	29.37
Ord news	ON	47 – 101	67.8	9.39	88.11
ETR information	EI	27 – 70	43.1	5.94	35.25
Ord information	OI	31 – 162	59.1	12.49	156.00

Table 5.19: Lemma variation index (LVIX) in documents

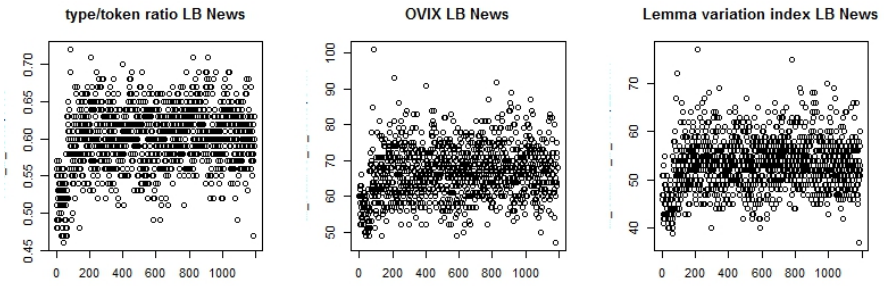
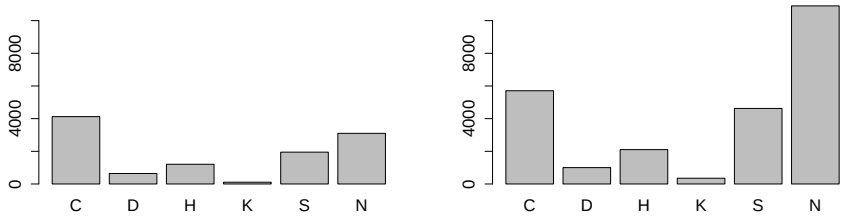


Figure 5.7: Lexical variation measured as type/token ratio, OVIX and lemma variation index

5.2.1.2 Vocabulary rate

The vocabulary rate, i.e. the internal composition of the vocabulary of LB and SUC, was measured in comparison to the reference word list SweVoc. Figure 5.8 illustrates the overall distribution of lemmas referred to specific SweVoc categories in LB and SUC. The label "N" denotes out-of-vocabulary lemmas.

Returning to the initial examples, 2 has a SweVoc general rate of 94 % of tokens, the category (C) words rate is 76 %, and (S) 18 %. The only out-of-vocabulary word is *gnager* 'gnaws'. Example 3 has a SweVoc general rate of 100 %, a category (C) rate of 94 %, and 6 % (H) rate



(a) LB *(b)* SUC

Figure 5.8: Vocabulary rate related to SweVoc in LB and SUC

(*texter*, 'texts'). A more illustrative example of SweVoc structure and content can be retrieved from the text sample B.1 in appendix B, where a chunk of 30 sentences from a children's ETR fiction book is presented, together with feature values. It consists of 107 different lemmas, listed in table 5.21, out of which 96 are present in the SweVoc total vocabulary. As measured in tokens, it equals 91 % of the total amount of words. Of the eleven out-of-vocabulary lemmas, four are proper nouns, three common nouns, three adjectives, and two verbs, according to table 5.20.

Lemma	PoS	English
tjejband	NN	girl's band
tjafs	NN	fuss
snack	NN	chat
skum	JJ	weird
cool	JJ	cool
ball	JJ	super
repa	VB	rehearse
diskutera	VB	discuss

Table 5.20: Out-of-vocabulary lemmas in children's ETR fiction (CEF) example B.1

The ratios of words belonging to SweVoc categories were measured in the text sets. The results of measuring the percentage of total SweVoc vocabulary lemmas at the sentence and document levels are shown in table 5.22 and 5.23; core vocabulary words (C) in tables 5.24 and 5.25; high frequency words (H) in tables 5.26 and 5.27; words of daily use (D) in tables 5.28 and 5.29; words from the Kelly wordlist (K) in tables

Category	Lemma	Freq	Category	Lemma	Freq
C	vara VB	12	C	den PN	10
C	jag PN	8	C	ha VB	7
C	och KN	6	C	han PN	6
C	i PP	5	C	tid NN	4
C	som HP	4	C	komma VB	4
C	inte AB	4	H	band NN	4
C	spela VB	3	C	lite AB	3
C	hon PN	3	C	en DT	3
C	att IE	3	C	som RH	2
H	rock NN	2	C	mycken AB	2
C	min PS	2	C	men KN	2
C	in PL	2	H	ibland AB	2
C	bruka VB	2	C	bli VB	2
C	bara AB	2	C	att SN	2
C	all PI	2	C	värld NN	1
C	viss AB	1	C	väg NN	1
C	varför AB	1	C	ur PL	1
C	undra VB	1	K	tydligen AB	1
D	trumma NN	1	C	tröttna VB	1
C	tala VB	1	C	svår AB	1
C	så AB	4	C	som SN	1
C	sjunga VB	1	C	åsikt NN	1
C	säga VB	2	C	sedan AB	1
C	risk NN	1	C	på PP	5
C	onödig JJ	1	C	om PP	1
C	om PL	1	C	olik JJ	1
C	ofta AB	1	C	också AB	1
C	nu AB	1	C	när RH	3
C	mycken JJ	1	C	måste VB	1
C	mest AB	1	C	mer JJ	1
C	med PP	1	C	med PL	1
C	massa NN	1	C	man PI	1
C	lust NN	1	C	liv NN	1
C	kunna VB	1	H	kör NN	1
S	kompis NN	1	C	känna VB	1
C	kanske AB	1	S	invända VB	1
C	hålla VB	1	C	hitta VB	1
C	hinna VB	1	C	gå VB	1
C	göra VB	1	D	gitarr NN	1
C	för PP	1	S	förresten AB	1
C	före PP	1	C	för KN	1
S	fotboll NN	1	S	fast SN	1
C	fall NN	1	C	där AB	1
C	dröm NN	1	C	bäst JJ	1
C	börja VB	2	C	bestämd PC	1
C	bas NN	1	C	annat JJ	1
C	all DT	1	C	aldrig AB	1
C	vart RH	1			

Table 5.21: Lemmas in children's ETR fiction (CEF) example B.1.

5.30 and 5.31; and finally all the supplementary words (S) in tables 5.32 and 5.33. Significance testings of each category at the document level will be summarized below. The corresponding figures at the sentence level did not seem to add anything to the total results, and will hence be left aside.

There was no difference in total SweVoc lemma percentage between children's ETR fiction and adults' ETR fiction (CEF/AEF)[9].

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 100	92.5	15.76	248.39
Children's Ord fiction	COF	0 – 100	90.6	15.70	246.53
Adults' ETR fiction	AEF	0 – 100	93.4	12.17	148.19
Adults' Ord fiction	AOF	0 – 100	86.3	16.70	278.75
ETR news	EN	0 – 100	91.5	11.26	126.70
Ord news	ON	0 – 100	82.6	18.71	350.24
ETR information	EI	0 – 100	90.6	15.25	232.42
Ord information	OI	0 – 100	75.2	29.31	858.87

Table 5.22: Mean percentage of SweVoc lemmas in sentences

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	82.8 – 99.0	90.8	2.95	8.72
Children's Ord fiction	COF	69.7 – 97.8	86.9	4.29	18.43
Adults' ETR fiction	AEF	72.4 – 98.5	90.9	3.89	15.14
Adults' Ord fiction	AOF	60.6 – 96.6	83.9	5.25	27.52
ETR news	EN	74.1 – 97.1	88.6	3.51	12.31
Ord news	ON	55.6 – 94.3	81.7	4.76	22.65
ETR information	EI	55.3 – 98.5	89.5	5.37	28.79
Ord information	OI	60.6 – 96.6	83.9	5.25	27.52

Table 5.23: Mean percentage of SweVoc lemmas in documents

Core vocabulary

There was a significant difference in core vocabulary percentage between all texts, except amongst ETR fiction for children and for adults

[9] $p = 0.454$

(CEF/AEF)[10].

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 100	82.4	19.74	389.68
Children's Ord fiction	COF	0 – 100	80.4	19.33	373.49
Adults' ETR fiction	AEF	0 – 100	83.1	16.92	286.22
Adults' Ord fiction	AOF	0 – 100	75.8	19.43	377.56
ETR news	EN	0 – 100	78.8	15.27	233.21
Ord news	ON	0 – 100	68.8	19.53	381.79
ETR information	EI	0 – 100	79.2	17.89	320.20
Ord information	OI	0 – 100	59.2	26.49	701.62

Table 5.24: Mean percentage of SweVoc core vocabulary lemmas in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	66.4 – 90.2	78.5	4.01	16.11
Children's Ord fiction	COF	56.1 – 89.7	74.1	5.48	30.08
Adults' ETR fiction	AEF	61.7 – 93.3	79.4	5.17	26.77
Adults' Ord fiction	AOF	45.5 – 89.0	69.4	6.76	45.65
ETR news	EN	61.0 – 89.0	75.4	4.34	18.87
Ord news	ON	45.2 – 80.5	64.1	5.56	30.95
ETR information	EI	41.9 – 90.9	77.3	6.56	43.03
Ord information	OI	0.0 – 76.8	58.1	9.01	81.17

Table 5.25: Mean percentage of SweVoc core vocabulary lemmas in documents

Highly frequent words

There was no significant difference in the fiction texts regarding the text pairs children's vs. adults' ETR fiction (CEF/AEF)[11], children's ETR vs. ordinary fiction (CEF/COF)[12], or children's ordinary fiction vs. adults' ETR fiction (COF/AEF)[13]. Statistical significance testing of adults' ordinary fiction vs. ETR information (AOF/EI) was $p = 0.080$, i.e. no significant difference was found.

[10] $p = 0.003$
[11] $p = 0.859$
[12] $p = 0.157$
[13] $p = 0.291$

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 100	2.1	6.86	47.12
Children's Ord fiction	COF	0 – 100	1.9	6.28	39.40
Adults' ETR fiction	AEF	0 – 100	2.2	6.87	47.27
Adults' Ord fiction	AOF	0 – 100	2.4	6.14	37.65
ETR news	EN	0 – 100	3.9	7.38	54.43
Ord news	ON	0 – 100	3.8	7.18	51.51
ETR information	EI	0 – 100	3.3	7.26	52.75
Ord information	OI	0 – 100	4.0	7.65	58.45

Table 5.26: Mean percentage of SweVoc highly frequent lemmas in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0.0 – 8.6	2.5	1.51	2.28
Children's Ord fiction	COF	0.0 – 7.5	2.6	1.41	1.99
Adults' ETR fiction	AEF	0.0 – 10.0	2.5	1.52	2.30
Adults' Ord fiction	AOF	0.0 – 8.1	3.4	1.43	2.06
ETR news	EN	0.0 – 11.1	4.2	1.75	3.05
Ord news	ON	0.5 – 11.0	4.7	1.78	3.18
ETR information	EI	0.0 – 9.7	3.6	1.75	3.07
Ord information	OI	0.0 – 10.8	5.2	1.89	3.59

Table 5.27: Mean percentage of SweVoc highly frequent lemmas in documents

Words in daily use

Turning to SweVoc words assumed to be frequently used in a person's daily life, there was no significant difference between the fiction texts in children's ETR fiction vs. adults' ETR fiction (CEF/AEF)[14], children's ordinary fiction vs. adults' ordinary fiction (COF/AOF)[15], children's ETR fiction vs. children's ordinary fiction (CEF/COF)[16], or children's ETR fiction vs. adults' ordinary fiction (CEF/AOF)[17]. Comparison of news texts and information in the following sets did not show any significant difference either: ETR news vs. ETR information (EN/EI)[18],

[14] $p = 0.004$
[15] $p = 0.874$
[16] $p = 0.829$
[17] $p = 0.751$
[18] $p = 0.224$

ETR news vs. ordinary information (EN/OI)[19], ordinary news vs, ETR information (ON/EI)[20], ordinary news vs. ordinary information (ON/OI)[21], and ETR information vs. ordinary information (EI/OI)[22].

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 100	0.78	4.06	16.55
Children's Ord fiction	COF	0 – 100	0.96	4.32	18.64
Adults' ETR fiction	AEF	0 – 100	0.87	3.94	15.55
Adults' Ord fiction	AOF	0 – 100	1.02	4.29	18.41
ETR news	EN	0 – 100	1.80	5.19	26.89
Ord news	ON	0 – 100	1.46	4.99	24.86
ETR information	EI	0 – 100	1.30	4.64	21.49
Ord information	OI	1 – 100	1.61	5.63	31.73

Table 5.28: Mean percentage of SweVoc lemmas in daily use in sentences

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 6.67	1.92	1.22	1.50
Children's Ord fiction	COF	0 – 8.11	1.93	1.14	1.30
Adults' ETR fiction	AEF	0 – 7.96	1.68	1.29	1.66
Adults' Ord fiction	AOF	0 – 5.95	1.94	1.14	1.29
ETR news	EN	0 – 6.72	2.40	1.31	1.71
Ord news	ON	0 – 6.35	2.19	1.07	1.14
ETR information	EI	0 – 7.10	2.32	1.29	1.67
Ord information	OI	0 – 11.54	2.36	1.49	2.23

Table 5.29: Mean percentage of SweVoc lemmas in daily use in documents

Words in Kelly list

As indicated in table 3.5, the Kelly word list is very restricted, and the coverage in different subcorpora is shown in tables 5.30 and 5.31. A significant difference in the proportion of lemmas belonging to SweVoc category K was only seen in document sets not involving ETR materials. Mean values vary between 0.82 % and 2.15 % in the subcorpora.

[19] $p = 0.648$
[20] $p = 0.081$
[21] $p = 0.072$
[22] $p = 0.674$

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 100	0.68	3.94	15.53
Children's Ord fiction	COF	0 – 100	0.74	4.60	21.20
Adults' ETR fiction	AEF	0 – 100	0.62	3.54	12.55
Adults' Ord fiction	AOF	0 – 100	0.80	4.23	17.92
ETR news	EN	0 – 100	0.55	3.02	9.16
Ord news	ON	0 – 100	1.04	3.79	14.33
ETR information	EI	0 – 100	0.47	2.76	7.60
Ord information	OI	0 – 100	1.49	5.19	26.89

Table 5.30: Mean percentage of SweVoc lemmas from Kelly word list in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 4.3	1.07	0.92	0.84
Children's Ord fiction	COF	0 – 6.4	1.04	0.92	0.85
Adults' ETR fiction	AEF	0 – 4.4	0.94	0.87	0.75
Adults' Ord fiction	AOF	0 – 4.7	1.12	0.83	0.69
ETR news	EN	0 – 4.0	0.84	0.78	0.61
Ord news	ON	0 – 4.8	1.60	0.94	0.88
ETR information	EI	0 – 5.8	0.82	0.82	0.66
Ord information	OI	0 – 10.5	2.15	1.30	1.69

Table 5.31: Mean percentage of SweVoc lemmas from Kelly word list in documents

Supplementary words

The percentage of SweVoc supplementary words in the documents varied from 5.9 % (EI) to 11.5 % (OI). This category of words consisted of all items from the Swedish base lemma vocabulary pool (SBVP) that were not included in SweVoc categories C, D, H, or K. SBVP is derived from SUC, which gives the texts from this source an unfair advantage. Excluding the pair children's ETR fiction vs. adults' ETR fiction (CEF/AEF)[23], all sets had a significant difference in SweVoc category S coverage.

[23] $p = 0.010$

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 100	6.55	11.95	142.87
LB Children's Ord fiction	COF	0 – 100	6.64	10.98	120.61
Adults' ETR fiction	AEF	0 – 100	6.64	11.08	122.86
Adults' Ord fiction	AOF	0 – 100	6.35	9.97	99.42
ETR news	EN	0 – 100	6.48	9.11	82.94
Ord news	ON	0 – 100	7.50	9.71	94.28
ETR information	EI	0 – 100	6.37	9.84	96.80
Ord information	OI	0 – 100	8.80	11.27	127.06

Table 5.32: Mean percentage of SweVoc supplementary lemmas in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	1.0 – 14.0	6.8	2.26	5.09
Children's Ord fiction	COF	1.2 – 16.2	7.3	2.29	5.22
Adults' ETR fiction	AEF	0.0 – 12.8	6.4	2.35	5.54
Adults' Ord fiction	AOF	1.5 – 14.6	8.1	2.28	5.20
ETR news	EN	0.7 – 13.2	5.9	1.99	3.94
Ord news	ON	4.0 – 16.4	9.1	2.42	5.87
ETR information	EI	0.0 – 16.1	5.6	2.21	4.89
Ord information	OI	0.0 – 21.9	11.5	3.20	10.23

Table 5.33: Mean percentage of SweVoc supplementary lemmas in documents

5.2.2 Sentence structure

The Swedish MaltParser was used for the syntactic analysis. It is trained on the dependency version of Swedish Treebank (Nivre et al. 2006) with SUC-annotated parts-of-speech. Consider an input sentence example in 4.

Example 4 *Hela gatan kommer att vakna!*
'The entire street will wake up!'

The output from the MaltParser is in the CoNLL data format, illustrated in table 5.34. Each line consists of tab-separated fields with information about each token's position in the sentence, word form, lemma form, part-of-speech tag and morphological features, a pointer to the head word in the dependency tree and type of dependency relation to the head. Parse tree output of example 4 is shown in figure 5.9.

1	Hela	hel	JJ	POS|UTR|SIN|DEF|NOM	2	DT
2	gatan	gata	NN	UTR|SIN|DEF|NOM	3	SS
3	kommer	komma	VB	PRS|AKT	0	ROOT
4	att	att	IE	_	3	VG
5	vakna	vakna	VB	INF|AKT	4	IF
6	!	!	MAD	_	3	IU

Table 5.34: Output after parsing

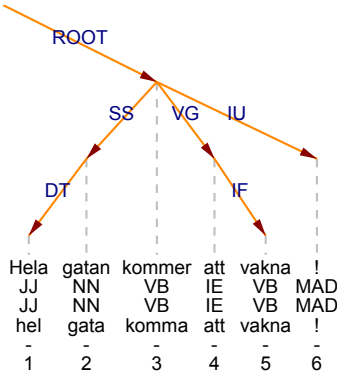


Figure 5.9: Parse tree output of example 4

Graphic parse trees of example sentences 2 and 3 are presented in figures 5.10 and 5.11, respectively.

5.2.2.1 *Mean dependency distance*

Under the assumption suggested earlier that the syntactic structure of a sentence consists of dependencies between individual words, the dependency distance can be calculated as the linear distance between a governor and its dependent. Our method to calculate the mean dependency distance (MDD) of a sentence is as follows:

Theorem 5 *Formally, let $W_1 \ldots W_i \ldots W_n$ be a word string. If word Wa is a governor, and word Wb its dependent, the dependency distance (DD) between them can be defined as the difference $a - b$.*

The DD is a negative number when $a < b$, i.e. the governor precedes the dependent and the parse tree is "right-branching". When $a > b$ it is a positive number; the governor follows the dependent and the parse

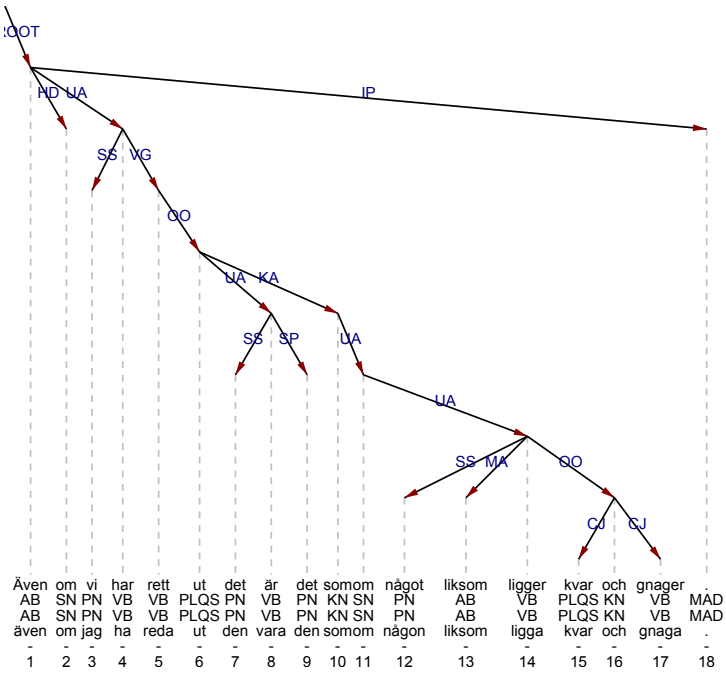


Figure 5.10: Parse tree output of example sentence 2

tree is "left-branching". Positive and negative numbers are used when the directions of a DD is relevant, otherwise the absolute dependency distance is decisive. Examples 6 and 7, illustrated in figures 5.12 and 5.13 illustrate a right vs. a left branching sentence. Figure 5.14 illustrates a mixture of branchings, indicating longer dependency lengths.

Example 6 *[...] att försöka undvika att bli närsynt*
'[...] try to avoid being nearsighted'

Example 7 *Hur mycket pengar föräldrarna får [...]*
'How much money the parents get [...]'

The mean dependency distance (MDD) of a sentence S is formally described as follows:

$$MDD(S) = \frac{1}{n-1} \sum_{i=1}^{n} |DD_i| \tag{13}$$

Liu (2008) has found a complexity metric based on dependency distance to bear on several examples using the above formula. In center-embedded clauses for instance, subject-relative sentences were found

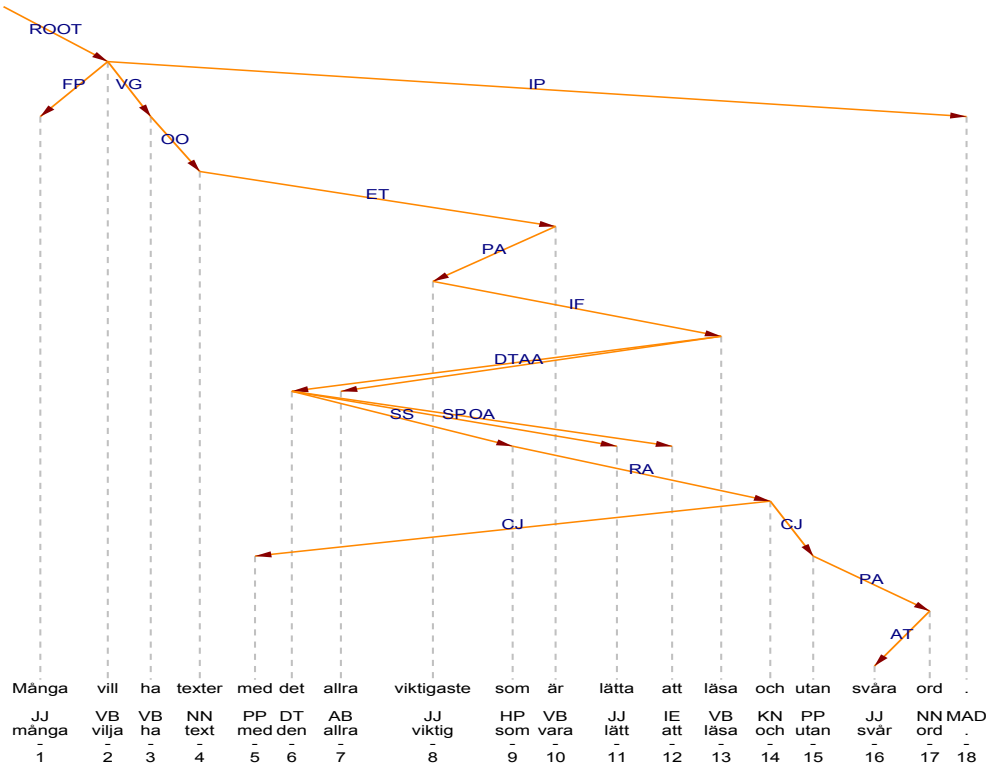


Figure 5.11: Parse tree of example sentence 3

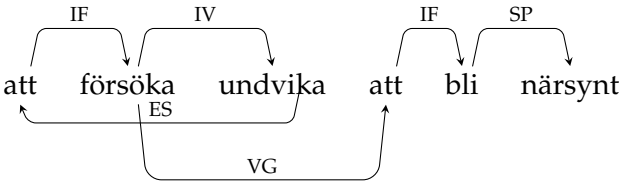


Figure 5.12: Right-branching dependency

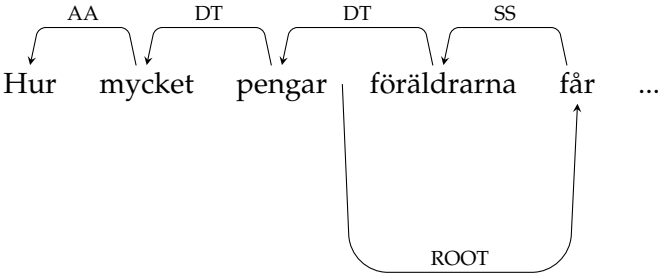


Figure 5.13: Left-branching dependency

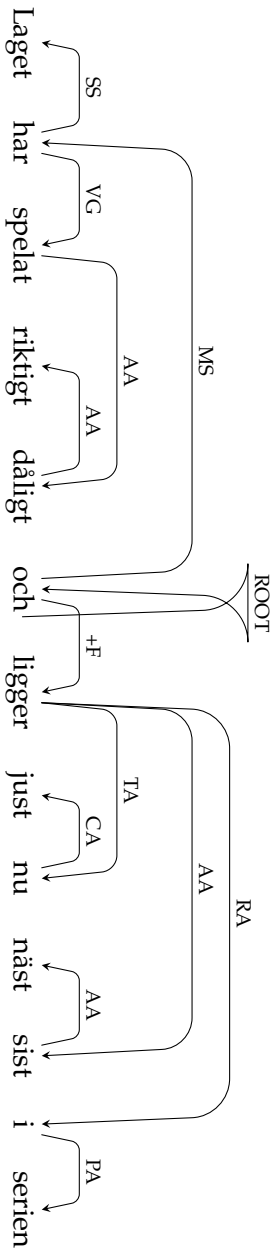


Figure 5.14: Mixed dependency branching

to be easier to process than object-relative sentences. This is illustrated in examples 8 and 9. Example 8, parsed in figure 5.15, has an MDD = 12/7 = 1.71. Meanwhile, example 9, parsed in figure 5.16, has MDD = 15/7 = 2.14. 1.71 < 2.14, which suggests that 8 is easier than 9.

Example 8 *En kund som dödade föraren tog hans pengar*
'A client that killed the driver took his money'

Example 9 *En kund som föraren dödade tog hans pengar*
'A client that the driver killed took his money'

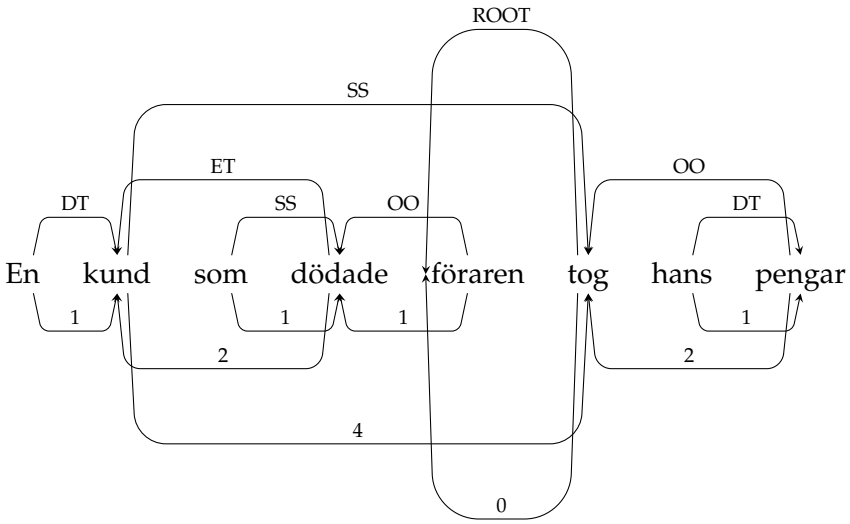


Figure 5.15: Subject-relative center embedding sentence

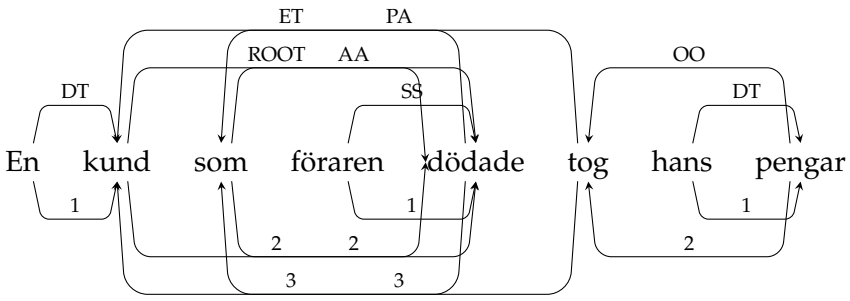


Figure 5.16: Object-relative center embedding sentece

Example sentence 2 has an MDD of 1.43 and 3 has an MDD of 1.75.

Mean dependency distance in parse trees showed significant differences for all pairs, except at the sentence level for children's ETR fiction vs. ETR news and adults' ETR fiction vs. ETR news. At the document level, children's ETR fiction, adults' ETR fiction and ETR news showed similar internal values. See tables 5.35 and 5.36.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 5.7	2.10	0.42	0.18
Children's Ord fiction	COF	0 – 9.0	2.19	0.46	0.21
Adults' ETR fiction	AEF	0 – 4.4	2.09	0.40	0.16
Adults' Ord fiction	AOF	1 – 15.5	2.44	0.82	0.66
ETR news	EN	0 – 18.7	2.11	0.41	0.17
Ord news	ON	0 – 13.2	2.38	0.69	0.47
ETR information	EI	1 – 10.5	2.19	0.48	0.23
Ord information	OI	1 – 23.0	2.46	0.89	0.79

Table 5.35: Mean dependency distance (MDD) in sentences

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	1.8 – 2.3	2.02	0.09	0.01
Children's Ord fiction	COF	1.8 – 2.4	2.10	0.10	0.01
Adults' ETR fiction	AEF	1.8 – 2.3	2.01	0.09	0.01
Adults' Ord fiction	AOF	1.9 – 3.8	2.44	0.24	0.06
ETR news	EN	1.8 – 2.5	2.01	0.08	0.01
Ord news	ON	1.9 – 3.4	2.39	0.18	0.03
ETR information	EI	1.8 – 2.6	2.06	0.12	0.01
Ord information	OI	1.6 – 3.4	2.47	0.22	0.05

Table 5.36: Mean dependency distance (MDD) in documents

Disregarding punctuation marks, 50 grammatical relations were identified in the material in total. Three of these – subordinate clauses and pre- and postnominal modifiers – were further explored in the SVIT model.

A rough estimate of the subordinate rate counts the frequencies of subordinate clauses (grammatical relation UA) in the text. Sentence example 10 from the ETR fiction for adults (AEF) in B.3 will serve as the illustration of this syntactical construction, rendered in upright font:

Example 10 *Karin skulle bli full av sorg* om hon kunde läsa Simons tankar
'Karin would be filled by sorrow if she could read Simon's thoughts'

A sentence with three prenominals and one postnominal is retrieved from the ETR information document (EI) in B.8, and presented in example 11. The prenominal modifiers are rendered in upright font, and the postnominal in bold.

Example 11 *Det kan vara genom* bra *sjukvård, att människor äter* nyttig *mat, att vaccinera mot* farliga *sjukdomar, att se till att sjukdomar* ***som smittar*** *inte sprider sig.*

'It could be by good *care, that people eat* healthy *food, to vaccinate against* dangerous *diseases, to ensure that infectious diseases do not spread.'*

Sentence example 2 contains 4 subordinate clauses, while example 3 has one prenominal modifier and one postnominal modifier, see dependency parse trees in figure 5.10 and 5.11.

5.2.2.2 Subordinate clauses

There was no significant difference seen in the frequency of subordinate clauses (UA) between neither sentences nor documents in ordinary news vs. ETR news, ETR news vs. children's ordinary fiction, or adults' ETR fiction vs. children's ETR fiction. Values are shown in tables 5.37 and 5.38.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	0 – 4	0.20	0.47	0.22
Children's Ord fiction	COF	0 – 10	0.25	0.55	0.30
Adults' ETR fiction	AEF	0 – 3	0.19	0.45	0.20
Adults' Ord fiction	AOF	0 – 10	0.31	0.64	0.41
ETR news	EN	0 – 4	0.25	0.51	0.26
Ord news	ON	0 – 4	0.29	0.58	0.33
ETR information	EI	0 – 6	0.35	0.64	0.41
Ord information	OI	0 – 5	0.28	0.59	0.35

Table 5.37: Mean frequencies of subordinate clauses (UA) in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 19	5.94	3.25	10.56
Children's Ord fiction	COF	0 – 25	7.49	3.86	14.89
Adults' ETR fiction	AEF	0 – 20	5.74	3.35	11.20
Adults' Ord fiction	AOF	0 – 36	9.35	5.39	29.09
ETR news	EN	0 – 23	7.52	3.80	14.47
Ord news	ON	0 – 27	8.69	4.66	21.69
ETR information	EI	0 – 29	10.61	5.34	28.54
Ord information	OI	0 – 32	8.52	5.33	28.36

Table 5.38: Mean frequencies of subordinate clauses (UA) in documents

5.2.2.3 Modifiers

The mean absolute frequency of prenominal modifiers (AT) showed significant overall differences at the sentence level as well as the document level. See tables 5.39 and 5.40.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 4	0.11	0.35	0.12
Children's Ord fiction	COF	0 – 6	0.16	0.44	0.20
Adults' ETR fiction	AEF	0 – 5	0.14	0.39	0.15
Adults' Ord fiction	AOF	0 – 13	0.41	0.80	0.64
ETR news	EN	0 – 5	0.30	0.53	0.29
Ord news	ON	0 – 13	0.58	0.86	0.74
ETR information	EI	0 – 5	0.27	0.54	0.29
Ord information	OI	0 – 10	0.71	0.97	0.94

Table 5.39: Mean frequencies of prenominal modifiers (AT) in sentences

The mean absolute frequency of postnominal modifiers (ET) also exposed significant overall differences at both the sentence level and the document level. See tables 5.41 and 5.42.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 12	3.44	2.38	5.68
Children's Ord fiction	COF	0 – 17	4.88	3.32	11.00
Adults' ETR fiction	AEF	0 – 17	4.07	2.78	7.75
Adults' Ord fiction	AOF	0 – 63	12.23	8.67	75.24
ETR news	EN	0 – 23	8.84	3.67	13.50
Ord news	ON	2 – 58	17.62	8.57	73.44
ETR information	EI	0 – 36	8.01	4.51	20.37
Ord information	OI	2 – 67	21.71	10.77	116.05

Table 5.40: Mean frequencies of prenominal modifiers (AT) in documents

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 4	0.20	0.46	0.21
Children's Ord fiction	COF	0 – 7	0.26	0.56	0.31
Adults' ETR fiction	AEF	0 – 4	0.22	0.49	0.24
Adults' Ord fiction	AOF	0 – 27	0.45	0.84	0.71
ETR news	EN	0 – 20	0.53	0.73	0.54
Ord news	ON	0 – 12	0.91	1.14	1.30
ETR information	EI	0 – 10	0.61	0.82	0.67
Ord information	OI	0 – 12	1.05	1.33	1.78

Table 5.41: Mean frequencies of post-nominal modifiers (ET) in sentences

5.2.2.4 *Parse tree height*

The parse tree height (PT) is the mean number of nodes from the root to the most distant leaf. Consider example 2, which has a parse tree height of 9, and example 3, which has a height of 10.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 16	5.93	3.03	9.18
Children's Ord fiction	COF	0 – 20	7.74	3.74	14.01
Adults' ETR fiction	AEF	1 – 23	6.69	3.45	11.91
Adults' Ord fiction	AOF	0 – 51	13.46	8.11	65.91
ETR news	EN	4 – 41	15.78	4.87	23.71
Ord news	ON	3 – 63	27.34	10.21	104.30
ETR information	EI	6 – 52	18.16	6.57	43.21
Ord information	OI	6 – 76	31.78	12.75	162.75

Table 5.42: Mean frequencies of post-nominal modifiers (ET) in documents

The PT heights were significantly different at the sentence (table 5.43) as well as document (5.44) level for all pairs except for the news texts of different types. Histograms of average PT heights in LB (figure 5.17) and SUC (figure 5.18) illustrate the differences between the corpora.

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	2 – 15	4.23	1.74	3.02
Children's Ord fiction	COF	2 – 27	4.61	2.16	4.67
Adults' ETR fiction	AEF	2 – 15	4.45	1.74	3.03
Adults' Ord fiction	AOF	0 – 21	5.07	2.32	5.36
ETR news	EN	2 – 20	5.33	1.98	3.91
Ord news	ON	2 – 23	5.92	2.53	6.41
ETR information	EI	2 – 34	5.93	2.60	6.75
Ord information	OI	1 – 21	5.60	2.84	8.12

Table 5.43: Mean parse tree height (PT) in sentences

Subcorpus	**Label**	**Range**	$\overline{X}$	σ	**Var**
Children's ETR fiction	CEF	3.13 – 5.53	4.24	0.46	0.21
Children's Ord fiction	COF	3.37 – 6.67	4.61	0.59	0.35
Adults' ETR fiction	AEF	3.33 – 6.77	4.45	0.53	0.28
Adults' Ord fiction	AOF	2.80 – 8.50	5.09	0.89	0.78
ETR news	EN	4.07 – 7.33	5.33	0.53	0.28
Ord news	ON	3.37 – 8.27	5.93	0.86	0.74
ETR information	EI	2.87 – 8.60	5.93	0.81	0.66
Ord information	OI	2.23 – 9.43	5.65	1.03	1.06

Table 5.44: Mean parse tree height (PT) in documents

5.2.3 Idea density

5.2.3.1 Propositional percentage

The propositional density was estimated by calculations of content words with respect to verbs, adjectives, adverbs, prepositions and subordinating conjunctions, following the suggestions of Brown et al. (2008). Ex-

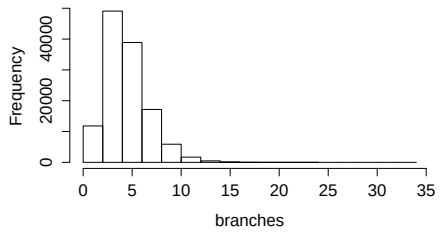


Figure 5.17: Histogram of PT heights in LB

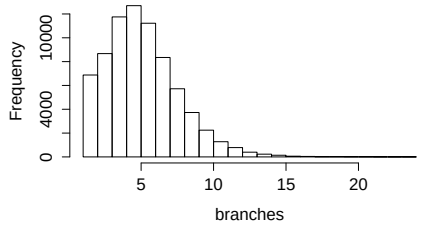


Figure 5.18: Histogram of PT heights in SUC

ample 2 has a P-density of 0.60, while 3 has a slightly higher value of 0.65.

There was a significant difference in the percentage of these parts-of-speech regarding the pairs ETR news vs. ordinary news, children's ETR fiction vs. children's ordinary fiction, and ETR news vs. children's ETR fiction. At the document level there was a significant difference for all pairs, except for children's ordinary fiction vs. ETR information, ETR news vs. adults' ETR fiction, and adults' ETR fiction vs. children's ETR fiction. See the summaries in tables 5.45 and 5.46.

5.2.3.2 *Noun/pronoun ratio*

The noun/pronoun ratio (NoPr) (Graesser, McNamara and Kulikowich 2011) showed a significant difference at the sentence level for all texts compared, except for ETR fiction texts for adults vs. children's ordinary fiction. At the document level there was no significant difference

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 86	43.5	14.06	197.67
Children's Ord fiction	COF	0 – 86	44.2	13.71	187.84
Adults' ETR fiction	AEF	0 – 86	43.8	12.25	150.01
Adults' Ord fiction	AOF	0 – 83	41.6	13.27	175.97
ETR news	EN	0 – 83	43.9	11.17	124.62
Ord news	ON	0 – 80	42.1	13.02	169.52
ETR information	EI	0 – 83	43.4	13.57	184.16
Ord information	OI	0 – 75	35.7	17.59	309.42

Table 5.45: Mean propositional percentage (Pr) in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	36.7 – 54.6	45.3	2.96	8.75
Children's Ord fiction	COF	34.7 – 55.4	46.2	3.14	9.87
Adults' ETR fiction	AEF	34.6 – 55.2	44.8	3.09	9.55
Adults' Ord fiction	AOF	33.1 – 54.1	43.6	2.80	7.82
ETR news	EN	35.6 – 52.8	44.7	2.87	8.22
Ord news	ON	24.4 – 70.2	50.2	8.75	76.53
ETR information	EI	19.3 – 58.3	45.9	4.48	20.03
Ord information	OI	16.0 – 51.5	41.0	4.82	23.19

Table 5.46: Mean propositional percentage (Pr) in documents

regarding the ordinary vs. ETR news texts. Complete results are shown in tables 5.47 and 5.48.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 7	0.53	0.80	0.65
Children's Ord fiction	COF	0 – 17	0.64	0.97	0.94
Adults' ETR fiction	AEF	0 – 10	0.65	0.89	0.79
Adults' Ord fiction	AOF	0 – 18	0.97	1.35	1.82
ETR news	EN	0 – 10	0.86	1.27	1.60
Ord news	ON	0 – 19	1.30	1.89	3.56
ETR information	EI	0 – 20	1.22	1.60	2.56
Ord information	OI	0 – 23	1.20	2.29	5.24

Table 5.47: Mean noun/pronoun ratio (NoPr) in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0.10 – 2.09	0.83	0.36	0.13
Children's Ord fiction	COF	0.15 – 3.53	0.96	0.43	0.18
Adults' ETR fiction	AEF	0.21 – 3.65	1.06	0.47	0.22
Adults' Ord fiction	AOF	0.26 – 5.35	1.30	0.69	0.48
ETR news	EN	0.93 – 15.4	4.03	1.64	2.69
Ord news	ON	0.72 – 103.00	4.63	5.58	31.12
ETR information	EI	0.00 – 25.00	3.18	2.26	5.13
Ord information	OI	0.50 – 160.00	13.63	16.81	282.54

Table 5.48: Mean noun/pronoun ratio (NoPr) in documents

5.2.3.3 *Nominal ratio*

There was a significant difference in nominal ratio (NR) (Melin and Lange 2000) for all texts, except ETR news vs. ETR information and the ETR fiction texts for adults vs. children's ordinary fiction. See table 5.49.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0.09 – 0.86	0.43	0.13	0.02
Children's Ord fiction	COF	0.12 – 1.18	0.48	0.16	0.27
Adults' ETR fiction	AEF	0.16 – 1.52	0.51	0.17	0.03
Adults' Ord fiction	AOF	0.19 – 1.89	0.65	0.25	0.06
ETR news	EN	0.50 – 1.95	1.08	0.23	0.05
Ord news	ON	0.40 – 4.45	1.27	0.44	0.20
ETR information	EI	0.32 – 2.44	0.93	0.31	0.10
Ord information	OI	0.63 – 41.00	2.36	2.88	8.01

Table 5.49: Mean nominal ratio (NR) in documents

5.2.3.4 *Semantic depth*

The semantic depth was calculated as the number of entries from the first descriptor of a lexeme to the artificial PRIM entry in the semantic and morphological lexicon Saldo (Borin and Forsberg 2009). For polysemous words, the mean value of all possible words was calculated. This was supposed to in some way reflect how humans process words in natural reading situations. An example of such an entry is the noun

läge ('situation', 'position'). Table 5.50 shows two different paths. They both have depth 3, so the mean semantic depth is set to 3.

Lexeme	**Depth**	**Path in Saldo**		
läge	3	← plats	← var	← PRIM
läge	3	← situation	← vara	← PRIM

Table 5.50: Saldo depths of the noun *läge*

Further illustrations of the semantic depth in Saldo are found below. Table 5.51 presents the semantic depth for each word in example sentence 2. A lexically more complex illustration from example 3 is shown in table 5.52.

The semantic depth in Saldo as a frequency of lexical type frequencies in LB and SUC is presented in figure 5.19. At the sentence and document level, the mean Saldo depth in each text set is displayed in tables 5.53 and 5.54, respectively. There is a significant difference between all sets except for the pair involving ordinary vs. ETR news.

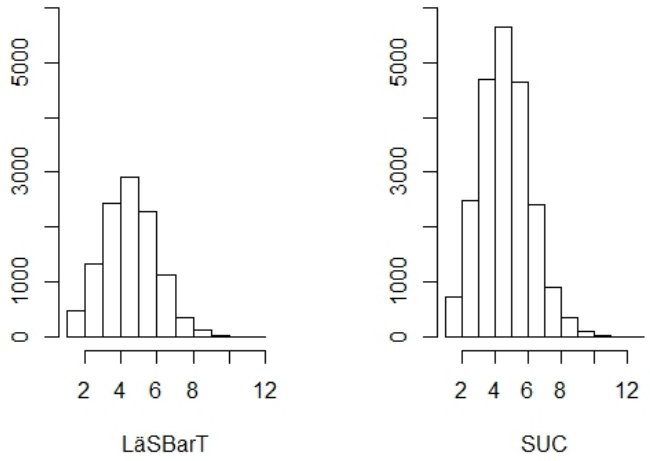


Figure 5.19: Saldo depth in LB and SUC as a function of lexical frequencies

Lemma	**No of homo-nyms**	**Semantic depth**	**Path in Saldo**
även	1	5	← också ← och ← med ← ha ← PRIM
om	5	1	← PRIM
vi	3	2	← vem ← PRIM
ha	2	1	← PRIM
reda_ut	1	4	← reda ← ordning ← bra ← PRIM
den	2	2	← vem ← PRIM
vara	4	1	← PRIM
som	3	7	← lik ← samma ← en ← tal ← mången ← mycken ← PRIM
om	5	1	← PRIM
någon	1	2	← vem ← PRIM
liksom	3	8	← som ← lik ← samma ← en ← tal ← mången ← mycken ← PRIM
ligga	3	4	← läge ← plats ← var ← PRIM
kvar	1	3	← här ← var ← PRIM
och	1	3	← med ← ha ← PRIM
gnaga	1	6	← bita ← ihop ← samman ← med ← ha ← PRIM

Table 5.51: Semantic depth in Saldo for sentence example 2

Lemma	**No of homo-nyms**	**Semantic depth**	**Path in Saldo**
mången	1	2	← mycken ← PRIM
vilja	2	1	← PRIM
ha	2	1	← PRIM
			↙ veta ← PRIM
text	1	4	← läsa ← se ← ljus ← PRIM
som	3	2	← vad ← PRIM
vara	4	1	← PRIM
lätt	3	2	← bra ← PRIM
att	2	3	← den ← vad ← PRIM
			↙ veta ← PRIM
läsa	2	3	← se ← ljus ← PRIM
med	5	2	← ha ← PRIM
viktig	2	3	← stor ← mycken ← PRIM
och	1	3	← med ← ha ← PRIM
			↙ motsats ← motsatt ← mot ← med ← ha ← PRIM
utan	2	4	← inte ← nej ← ja ← PRIM
			↙ motsats ← motsatt ← mot ← med ← ha ← PRIM
svår	2	3	← lätt ← bra ← PRIM
ord	1	4	← språk ← tala ← säga ← PRIM

Table 5.52: Semantic depth in Saldo for sentence example 3

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	1 – 10.0	3.15	0.83	0.69
Children's Ord fiction	COF	1 – 10.0	3.11	0.78	0.61
Adults' ETR fiction	AEF	1 – 7.8	3.24	0.74	0.55
Adults' Ord fiction	AOF	1 – 10.0	3.20	0.83	0.68
ETR news	EN	1 – 10.5	3.57	0.68	0.47
Ord news	ON	1 – 10.0	3.45	0.77	0.59
ETR information	EI	1 – 11.0	3.51	0.73	0.54
Ord information	OI	1 – 10.0	3.35	1.28	1.64

Table 5.53: Mean Saldo depth (Sa) in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	3.17 – 3.96	3.59	0.14	0.02
Children's Ord fiction	COF	3.04 – 4.07	3.60	0.14	0.02
Adults' ETR fiction	AEF	3.23 – 4.13	3.62	0.15	0.02
Adults' Ord fiction	AOF	3.22 – 4.26	3.73	0.16	0.03
ETR news	EN	3.56 – 4.37	3.91	0.13	0.02
Ord news	ON	3.51 – 4.41	3.90	0.14	0.02
ETR information	EI	3.41 – 4.40	3.76	0.16	0.03
Ord information	OI	3.59 – 5.03	4.02	0.19	0.03

Table 5.54: Mean Saldo depth (Sa) in documents

5.2.4 Human interest

5.2.4.1 Personal noun percentage

For all texts there was a significant difference in the occurrence of personal nouns (PM) at the sentence level. This also holds at the document level except for ordinary vs. ETR news, which showed identical values. For details, see tables 5.55 and 5.56.

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 100	5.1	8.9	0.80
Children's Ord fiction	COF	0 – 100	4.1	7.8	0.61
Adults' ETR fiction	AEF	0 – 75	5.7	9.3	0.86
Adults' Ord fiction	AOF	0 – 75	3.2	7.4	0.55
ETR news	EN	0 – 90	8.1	10.7	1.14
Ord news	ON	0 – 100	7.7	12.9	1.67
ETR information	EI	0 – 100	2.8	7.8	0.61
Ord information	OI	0 – 100	3.6	10.5	1.09

Table 5.55: Mean personal noun (PM) percentage in sentences

Subcorpus	Label	Range	$\overline{X}$	σ	Var
Children's ETR fiction	CEF	0 – 14	4.7	2.2	0.05
Children's Ord fiction	COF	0 – 13	3.7	2.1	0.04
Adults' ETR fiction	AEF	0 – 14	5.4	3.0	0.87
Adults' Ord fiction	AOF	0 – 12	2.9	2.1	0.05
ETR news	EN	1 – 21	7.9	3.2	0.10
Ord news	ON	0 – 30	6.9	4.2	0.17
ETR information	EI	0 – 18	2.4	2.7	0.07
Ord information	OI	0 – 33	2.9	3.5	0.12

Table 5.56: Mean personal noun (PM) percentage in documents

6 DOCUMENT CLASSIFICATION

The performance of three functionally different algorithms for text classification were evaluated in 28 experiments, where members of eight document test sets were pairwisely compared. The algorithms adopted were Naïve Bayes algorithm (NB), the sequential minimal optimization (SMO), and Classification via Regression (CVR). In each experiment, vector values from equal amounts of documents from two different sets were fed into the Weka software. Evaluation was performed by means of 7-fold cross validation, and accuracy, precision, recall and F-score were calculated for each experiment. Two models were used, the base model, consisting of the LIX values, and the SVIT model, constituted by 22 textual features. Finally, a reduced SVIT model was evaluated, where all features signaling surface properties were ignored. More precisely, features indicating word length (MWLC and MWLS), sentence length (MSL), ratio of extra-long words (XLW) and OVIX values were excluded from the language model.

The classification sets were grouped into four categories, depending on text genre and type, in accordance with table 4.3. In addition to 28 binary classification tasks, a final experiment was made on all of the previously used test sets in order to evaluate the performance of multiclass classification. Results from pair-wise classification of ordinary and ETR texts from the same genres will be presented in more detail, while the tasks involving texts of different genres will be addressed comprehensively. Detailed results from each experiment are presented in tables D.1–D.12 in Appendix D. Accuracy of the best-performing language model and classification algorithm for each task is indicated by the highlight.

6.1 Same genre and type

6.1.1 Fiction across ages

This test suite is populated of four different document sets - fiction intended for children and for adults in both ETR and ordinary format. As can be seen in table D.1, the SVIT language model performed best on ordinary texts, i.e. it seemed easier to distinguish between ordinary fiction texts intended for adults and those intended for children than between the corresponding ETR texts. The difference in performance between the SVIT model and the base model was also larger for the former test set than the latter.

6.2 Same genre and different types

Eight different document sets were here compared pairwise with the purpose to distinguish between texts of the same genre in an ordinary and ETR fashion.

6.2.1 Fiction

Even this test suite contains fiction texts directed towards different ages, but the task was to classify documents as being either ordinary or ETR. Four classification tasks were performed where the age of the targeted audience was also considered. The results, shown in table D.2, indicate that ordinary texts intended for adults were easier to distinguish from ETR texts for children and ETR texts for adults, than was the case for children's ordinary fiction texts. Furthermore, it is clear that the SVIT model, as compared to the base model, enhanced the classification results considerably for the task of separating adults' ETR fiction from childrens ordinary fiction.

6.2.2 News

The next experiment was carried out with a test set containing documents from the ETR news subcorpus (EN) and from an equal amount of documents from ordinary news (ON). Overall results, shown in table D.3, are impressively good in that the SVIT model performed 23.6 percentage points better than the base model.

6.2.3 Information

In this experiment, the test suite was equally distributed between ETR information texts (EI) and their ordinary counterpart (OI). Also this classification task was performed with best results for the SVIT model, as can be seen in table D.4. It performed 11.2 percentage points (pp) better than the base model.

6.3 Different genres and same type

A separate category was constructed for the evaluation of texts differing in genre but converging in type. We will start by looking at the ETR fiction texts as compared to corresponding texts from the news and information genres. As can be seen in table D.5 the classification results with the SVIT model were just marginally better than base model classification. The only noteworthy improvement is in the task of distinguishing between adults' ETR fiction (AEF) and ETR information documents (EI). The difference in performance between SVIT and the base model was 9.7 pp.

Ordinary fiction text documents were also compared to corresponding texts from the news and information genres. In line with the results presented above, the best results for the SVIT model were found in the experiment involving adults' fiction and information documents, but here also in adults' fiction and news.

The next experiment in this category included documents from ETR news (EN) and from the ETR information subcorpus (EI). The SMO classifier yielded 91.7 % accuracy for the SVIT model, i.e. 32.3 pp better than the base model. This seems to be one of the rare instances where the NB algorithm with the base model marginally superceded the other two algorithms. See table D.7 for details.

Finally, documents from ordinary news (ON) and ordinary information (OI) were used for the last experiment in this category. As in the findings presented in D.7, the SVIT model produced excellent results, exceeding the accuracy obtained with the base model with 34.7 pp. Details are given in D.8.

6.4 Different genres and different types

To complete the tests, a final classification category was constructed out of texts that differ in both aspects, i.e. it contains data that are cross-

compared over genre and type. The following is an attempt to present the results in a reasonably structured manner.

ETR fiction texts targeted towards children as well as adults were compared to ordinary news and information texts. Very little is gained with SVIT as compared to the base model - only the task of separating adults' ETR fiction from ordinary news seems to somewhat exploit the additional power of SVIT. As can be seen in table D.9, the difference is 4.2 pp.

The next experiment suite in this mixed category concerns documents from children's and adults' ordinary fiction as compared to ETR news and information texts. A considerable advantage of the SVIT model can be seen in three of the experiments, the largest achieved for the task of separating adults' ordinary fiction from ETR news. In this case a 55.4 pp improvement was obtained. See further table D.10.

Finally, two experiments with information text documents and news texts were made. The first relates to ordinary news and ETR information. Table D.11 displays the results, indicating that the SVIT model performed 26.6 pp better than the base model. A switch of text types was also made, and ordinary information documents were classified together with ETR news. In this case, a difference of 6.9 pp was presented as can be seen in table D.12.

6.5 Document classification with all test sets

In the experiment involving all the test sets, the presentation is limited to the results from the SMO algorithm, as it proved to be the overall best-performing algorithm. The base model reached an accuracy level of 40.5 %, while the SVIT model produced 78.8 %, and a reduced SVIT model 72.4 % accuracy. Precision, recall and F-score for each of the classes are given in table 6.1. The lowest precision score was obtained by SVIT on children's ETR fiction (CEF), at 61.2 %. The difference in precision score between the base model and the SVIT model ranges between 4.4 for children's ETR fiction (CEF) and 63.7 pp for adults' ordinary fiction texts (AOF) and ETR information (EI).

Confusion matrices in tables 6.2 and 6.3 provide figures for further analysis. Looking at the results from base model classification, we can see that the ETR information texts were frequently mixed up with news texts, both in ordinary and ETR fashion, but also with ordinary and ETR fiction texts for adults.

Test set	Model	Performance		
		Prec	Recall	F-score
CEF	Base	56.8	80.0	66.4
	SVIT	61.2	76.3	67.9
	Red SVIT	45.9	58.8	51.6
COF	Base	35.6	25.8	30.0
	SVIT	63.7	65.0	64.3
	Red SVIT	62.4	58.8	60.5
AEF	Base	30.8	36.7	33.5
	SVIT	70.4	65.4	67.8
	Red SVIT	53.1	46.7	49.7
AOF	Base	31.1	27.1	29.0
	SVIT	94.8	75.8	84.3
	Red SVIT	91.0	76.3	83.0
EN	Base	36.7	65.8	47.2
	SVIT	86.0	92.1	88.9
	Red SVIT	79.5	87.1	83.1
ON	Base	30.6	20.4	24.5
	SVIT	85.6	85.4	85.4
	Red SVIT	83.8	88.3	86.0
EI	Base	27.7	7.5	11.8
	SVIT	91.4	84.6	87.9
	Red SVIT	84.7	80.8	82.7
OI	Base	56.2	60.4	58.2
	SVIT	85.1	85.8	85.5
	Red SVIT	86.1	82.5	84.3

Table 6.1: Detailed accuracy by test set

	Classified as							
Class	**CEF**	**COF**	**AEF**	**AOF**	**EN**	**ON**	**EI**	**OI**
CEF	192	24	23	1	0	0	0	0
COF	69	62	93	16	0	0	0	0
AEF	63	56	88	29	1	0	3	0
AOF	12	23	50	65	54	15	10	11
EN	0	0	1	27	158	31	17	6
ON	0	1	8	20	66	49	15	81
EI	1	6	22	48	105	25	18	15
OI	1	2	1	3	46	40	2	145

Table 6.2: Confusion matrix of document classification with the base model

	Classified as							
Class	**CEF**	**COF**	**AEF**	**AOF**	**EN**	**ON**	**EI**	**OI**
CEF	183	24	32	0	1	0	0	0
COF	54	156	19	10	0	0	1	0
AEF	50	25	157	0	7	0	1	0
AOF	10	39	5	182	2	2	0	0
EN	0	0	1	0	221	1	16	1
ON	1	0	1	0	2	205	0	31
EI	1	1	8	0	23	0	203	4
OI	0	0	0	0	1	32	1	206

Table 6.3: Confusion matrix of document classification with the SVIT model

	Classified as							
Class	**CEF**	**COF**	**AEF**	**AOF**	**EN**	**ON**	**EI**	**OI**
CEF	141	28	65	1	5	0	0	0
COF	62	141	20	14	0	0	3	0
AEF	90	19	112	1	14	0	4	0
AOF	13	37	2	183	1	1	1	2
EN	0	0	4	0	209	1	26	0
ON	0	0	1	0	3	212	0	24
EI	1	1	7	2	29	0	194	6
OI	0	0	0	0	2	39	1	198

Table 6.4: Confusion matrix of document classification with the reduced SVIT model

6.6 Summary of classification results

The SMO algorithm with the SVIT model was found to be superior in almost all the experiments. The best average accuracy for the base model in pairwise classification was 84.2 %, while the SVIT model presented a 96.4 % best average accuracy. Generally, the difference in accuracy between the base model and SVIT was found to be smallest for pairs involving ETR fiction texts for children, and largest for pairs involving ETR information. Analyses of the confusion matrices confirm the figures from pairwise classification, in that there are more instances of confusion with both models between fiction texts, ETR as well as ordinary texts, although much less with the SVIT model.

7 CONCLUDING RESULTS

This chapter will provide combined results from the descriptive statistical analysis and the document classification task. Presentation of the results is structured in the following manner: Each classification category, grouped with regard to correspondences in genre and style, will first be commented on individually. After that, a discussion of the impact of different features will follow. Finally, a comprehensive summary of the results will be given.

7.1 Overview of the combined results

An overview of the combined results from the statistical descriptive analyses of documents and principal component analyses is presented in tables 7.1, 7.2, and 7.3. Significance tests were made on the entire subcorpora, while the PCAs were performed as a subtask of the classification with the smaller test sets given in 4.3. The feature values in the two document sets could be either positively correlated, denoted by (+) or negatively (-), which only means that the variable in question correlates positively or negatively to the axis. Generally, the PCAs listed 15 features out of 22 to be behind $\approx$ 95 % of the variations in data. Mostly, a primary and secondary loading revealed 10-11 features to account for 40–60 % of the variation in the data sets. In some cases, a tertiary loading was also included. The outcome of significance testing of feature value differences by means of Welch two-sample T-test at a significance level of $p < 0.001$ is marked as either significantly different (1) or no difference (0). A general observation is that SweVoc categories (D), (H), (K), and (S) never had any impact on the total variation in the data, although they differed significantly between the classes. These features will hence be excluded in the comprehensive remarks.

Feat cat	Feature	Same genre and type		Same genre and different types					
		CEF vs AEF	**COF vs AOF**	**CEF vs COF**	**CEF vs AOF**	**AEF vs COF**	**AEF vs AOF**	**EN vs ON**	**EI vs OI**
Surface	MWLC	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1	1 p+
	MWLS	1 p+	1 p+	1 p+	1 p+	1 p+	1	0	1 p+
	MSL	1 s+	1 s-	1 s-	1	1 s-	1	1 s-	1 s-
	XLW	1	1	1	1	0	1	1 p+	1
	OVIX	1	1	1 p+	1	1 t+	1	1	1
Voc load	LVIX	0	1	1 p+	1	1 p+	1 p+	1 p+	1
	SV	0 s+	1 s-	1 s-	1 s-	1 s-	1 s-	1	1 p-
	SVC	0 s+	1 s-	1	1 s-	1 t-	1 s-	1 p-	1 p-
	SVD	0	0	0	0	1	1	1	0
	SVH	0	1	0	1	0	1	1	1
	SVK	0	0	0	0	0	1	1	1
	SVS	0	1	1	1	1	1	1	1
Sent str	MDD	0	1	1	1	1	1	1 t+	1
	UA	0 s+	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-
	AT	1	1 p+	1	1 p+	1	1 p+	1 p+	1
	ET	1	1	1	1 p+	1 s-	1 p+	1 p+	1 s-
	PT	1 s+	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-
Idea dens	Pr	0	1	1 s-	1 s-	1	1	1 s-	1 s-
	NoPr	1 p+	1 p+	1 p+	1	1 p+	1	0 t-	1
	NR	1 p+	1 p+	1 p+	1 p+	0 p+	1 p+	1 s+	1
	Sa	1 p+	1	0	1	1	0	1 t-	1 p+
Human int	PM	1	1	1	1	1 t-	1	1	0
SVIT performance		83.9	92.9	84.6	97.5	84.2	96.7	99.6	97.1
Diff. SVIT-Base pp		6.5	15.8	11.7	7.7	28.8	23.0	23.6	11.2

Table 7.1: Results from statistical significance tests and principal component analysis of features in categories 1 and 2. 1 = sign diff ($p < 0.001$), 0 = no sign diff. p = primary, s = secondary, t = tertiary loading.

Feat cat	Feature	Different genres and same type									
		CEF vs EN	**CEF vs EI**	**AEF vs EN**	**AEF vs EI**	**COF vs ON**	**COF vs OI**	**AOF vs ON**	**AOF vs OI**	**EN vs EI**	**ON vs OI**
Surface	MWLC	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1 s+	1 p+
	MWLS	1 p+	1	1 p+	1 s-	1 p+	1 p+	1 p+	1 p+	1 s+	1 p+
	MSL	1	1 p+	1 s-	1 p+	1 s-	1 s-	0 s-	1 s-	1 s+	0 s-
	XLW	1	1	1	1 p+	1	1 p+	1	1 p+	1 s+	1 p+
	OVIX	1	1	1	1	1	1	1	0	1	1
Voc load	LVIX	0 s+	1 s-	1	1	1	1	1	1	1 p+	1
	SV	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1 p-	1 s-
	SVC	1 s-	1 s-	1 s-	1 s-	1	1	1	1	1 p-	1 s-
	SVD	1	1	1	1	1	1	1	1	0	0
	SVH	1	1	1	1	1	1	1	1	1	1
	SVK	1	1	0	0	1	1	1	1	0	1
	SVS	1	1	1	1	1	1	1	1	0	1
Sent str	MDD	0	1	0	1	1	1	0	0	1	1
	UA	1 s-	1 s-	1 s-	1	0 s-	1 s-	1 s-	0 s-	1	0
	AT	1	1	1	1	1	1	1	1	1	1
	ET	1	1 p+	1	1 p+	1 p+	1 p+	1 p+	1 p+	1 s+	1 p+
	PT	1	1 p+	1 s-	1 p+	1 s-	1 s-	1 s-	1 s-	1 p-	1 s-
Idea dens	Pr	1 s-	0 s-	0	1 s-	1 s-	1 s-	1	1 s-	1	1 s-
	NoPr	1 p+	1	1 p+	1	1	1	1	1	1	1
	NR	1 p+	1 p+	1 p+	1	1 p+	1	1 p+	1	0	1
	Sa	1 p+	1	1 p+	1 s+	1 p+	1 p+	1 p+	1 p+	1	1 p+
Hum int	PM	1	1	1	1	1	1	1	0	1	1
SVIT performance		99.7	99.5	98.6	98.0	98.8	99.4	89.8	99.1	91.7	94.6
Diff. SVIT-Base pp		0.2	2.5	2.2	9.7	2.3	0.8	9.6	14.0	32.3	34.7

Table 7.2: Results from statistical significance tests and principal component analysis of features in category 3. 1 = sign diff ($p < 0.001$), 0 = no sign diff. p = primary, s = secondary, t = tertiary loading.

Feat cat	Feature	Different genres and different types									
		CEF vs ON	**CEF vs OI**	**AEF vs ON**	**AEF vs OI**	**COF vs EN**	**COF vs EI**	**AOF vs EN**	**AOF vs EI**	**ON vs EI**	**OI vs EN**
Surface	MWLC	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1 p+	1 s+	1 p+	1 p+
	MWLS	1 p+	1 p+	1 p+	1 p+	1 p+	1	1 p+	1 s+	1 t-	1 p+
	MSL	1 p+	1	1 s-	1 s-	1 s-	1 p+	1 s-	1	1 s-	1 s-
	XLW	1	1 p+	1 p+	1 p+	1	1 p+	1	1 p+	0 s-	1 p+
	OVIX	1	1	1	1	0	1	1	1 p+	1 p+	1
Voc load	LVIX	1	1	1	1	1	1	1	1 p+	1 p+	1
	SV	1 s-	1 s-	1 s-	1 s-	1 s-	1 s-	1	1 p+	1	1 s-
	SVC	1	1	1	1	1 s-	1 s-	1	1 s-	1 p+	1 p+
	SVD	1	1	1	1	1	1	1	1	0	1
	SVH	1	1	1	1	1	1	1	0	1	1
	SVK	1	1	1	1	1	1	1	1	1	1
	SVS	1	1	1	1	1	1	1	1	1	1
Sent str	MDD	1	1	1	1	1	1	1	1 t-	1 t+	1
	UA	1 s-	1 s-	1 s-	1 s-	0 s-	1 s-	1 s-	1 t-	1 s-	0 s-
	AT	1	1 p+	1	1 p+	1	1	1 s-	1 t-	1	1 p+
	ET	1 p+	1 p+	1 p+	1 p+	1	1 p+	1 s-	1 t-	1 s-	1
	PT	1 s-	1 s-	1 s-	1 s-	1 s-	1 p+	1 s-	1 p+	1 s-	1 s-
Idea dens	Pr	1 s-	1 s-	1 s-	1 s-	1	0 s-	1	1	1	1 s-
	NoPr	1	1	1	1	1 p+	1	1 p+	1 s+	1 t-	1
	NR	1 p+	1 s-	1 p+	1	1 p+	1	1 p+	1	1 p+	1
	Sa	1	1	1	1	1 p+	1 s+	1 p+	1 s+	1 t-	1
Human int	PM	1	1	1	1	1	1	1	1	1	1
SVIT performance		99.4	99.4	99.4	99.6	99.7	99.1	100.0	99.4	98.1	99.0
Diff. SVIT-Base pp		0.4	0.7	4.2	0.7	1.3	8.8	55.4	36.0	26.6	6.9

Table 7.3: Results from statistical significance tests and principal component analysis of features in category 4. 1 = sign diff ($p < 0.001$), 0 = no sign diff. p = primary, s = secondary, t = tertiary loading.

7.2 Category 1. Same text genre and same text type

The first category to be described contains texts from the fiction genre of both ETR type and ordinary types, aimed at adults or children. At the *surface level*, mean word length in characters and syllables, and mean sentence length showed overall significant differences. This was also the case for extra-long words and OVIX values. At the *vocabulary level*, the lemma variation index was similar between the ETR texts. The percentage of words present in the total SweVoc word list and core vocabulary showed a significant difference in the ordinary documents. *Sentence structure* showed significant difference for three features, i.e. parse tree height, and pre- and post-nominal modifiers. The frequency of subordinate clauses and mean dependency distance were similar in the ETR text set at the sentence level as well as the document level. The *idea density* varied in difference with respect to propositional percentage, but was significantly different with respect to noun/pronoun ratio, nominal ratio, and semantic depth in Saldo. At the level of *human interest*, the ratio of personal nouns differed significantly.

Text classification with the three selected algorithms showed a significantly better result for the complex feature model both for the ETR texts and ordinary texts. The SMO classifier performed best, and gave 6.5 % better accuracy with the SVIT model compared to LIX for the former test set, and 15.8 for the latter. A principal component analysis revealed that the combined *idea density* features noun/pronoun ratio, the nominal ratio and the semantic depth in Saldo had the highest impact on the success of the classifier. Together with sentence length and word length features they explained 23 % of the variation in the data set at a primary loading on class for the ETR text pair, while variation in the ordinary text pair also depended on the frequency of prenominal modifiers. Total distributions of noun/pronoun ratio, nominal ratio and Saldo depth in text set are displayed in figures 7.1, 7.2, and 7.3, respectively. Distribution of prenominal modifier frequencies in ordinary texts is shown in figure 7.4.

7.3 Category 2. Same text genre and different text types

7.3.1 Fiction

Children's fiction of two different types were compared; ETR texts, and ordinary texts. All *surface level* features differed significantly. At the *vo-*

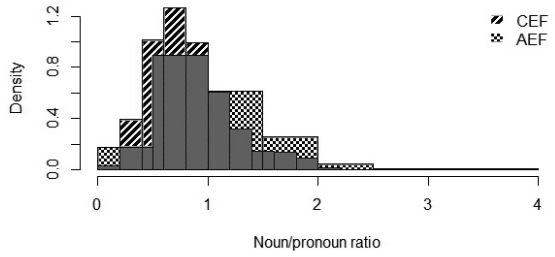


Figure 7.1: Noun/pronoun ratio distribution in children's and adults' ETR fiction

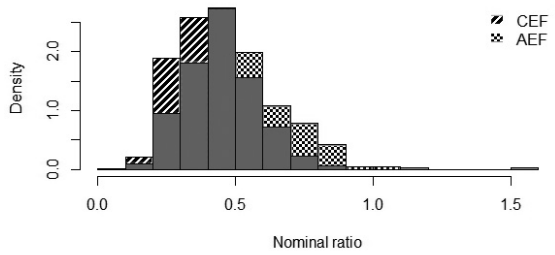


Figure 7.2: Nominal ratio distribution in children's and adults' ETR fiction

cabulary level, testing showed that the ratios of SweVoc core vocabulary words differed and also the lemma variation index. The semantic depths in Saldo, indicating *idea density* were similar in the two document sets, but the propositional and noun/pronoun densities differed, as well as the personal noun ratio at the level of *human interest*.

The SMO algorithm presented 11.7 pp better result for the SVIT model compared to base line (72.9 % and 84.6 %, respectively). The principal components contributing to the results were OVIX, figure 7.5 and lemma variation index, figure 7.6, i.e. *vocabulary diversity* indicators, to a lesser degree *vocabulary difficulty* (presence in SweVoc) and the noun/pronoun, figure 7.7 and nominal ratios, figure 7.8, marking the *idea density* level. Together with word length and sentence length these features answered for 43 % of the variation.

Children's ETR fiction was also compared to adult's ordinary fiction. Most features seemed to differ significantly, while PCA showed that

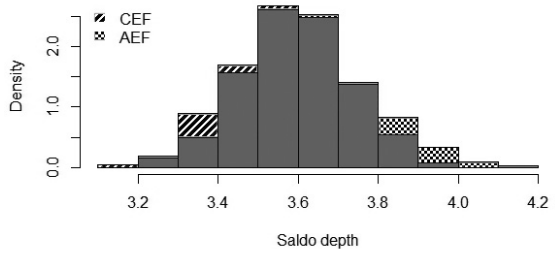


Figure 7.3: Saldo depth distribution in children's and adults' ETR fiction

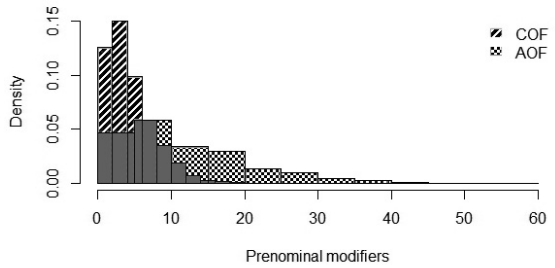


Figure 7.4: Prenominal modifier distribution in children's and adults' ordinary fiction

indicators of *sentence structure* complexity, i.e. number of subordinate features and pre- and post-nominal modifiers, along with parse tree height, prevailed. The feature set contributing to a variation of 52 % in the data set also included nominal ratio, propositional density, and SweVoc coverage, which means that even the *idea density* and *vocabulary load* were decisive, word length and sentence length excluded.

Children's ordinary texts were also compared to adults' ETR fiction. These texts were found to share properties similar to the former set. It showed statistical differences at the *idea density* level, proposition density, the noun/pronoun ratio, and the Saldo depth.

This classification task rendered interesting results. As in most of the experiments, the SMO classifier produced the highest accuracy, but from a very low base line. Classification performed only by means of LIX values gave a 55.4 % accuracy, i.e. around randomness, while the SVIT model produced a 84.2 % overall accuracy. After principal com-

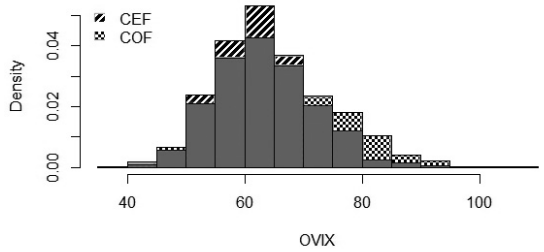


Figure 7.5: OVIX distribution in children's ETR and ordinary fiction

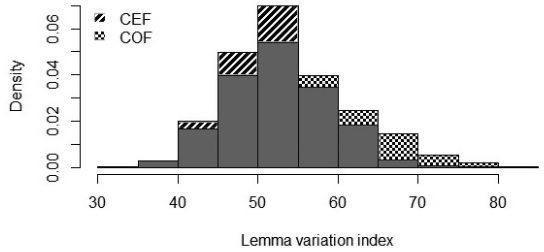


Figure 7.6: Lemma variation index distribution in children's ETR and ordinary fiction

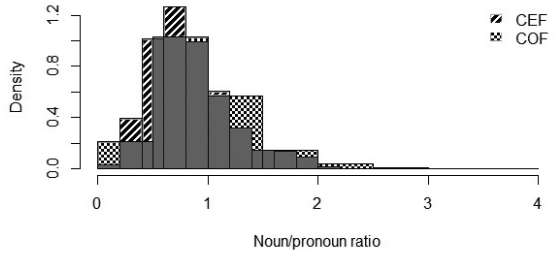


Figure 7.7: Noun/pronoun ratio distribution in children's ETR and ordinary fiction

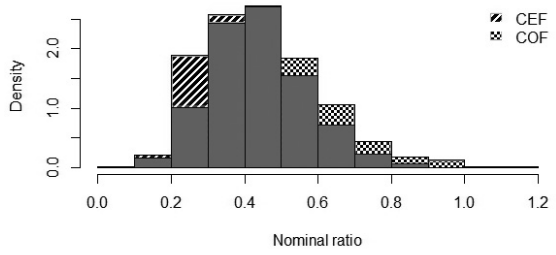


Figure 7.8: Nominal ratio distribution in children's ETR and ordinary fiction

ponent analysis, it was evident that the SweVoc core vocabulary ratio, figure 7.9, the nominal ratio, figure 7.10, and the noun/pronoun ratio, figure 7.11, together with word length had the strongest impact on the results of the classifier. This feature combination contributed to ≈ 51% to the variation in the data set.

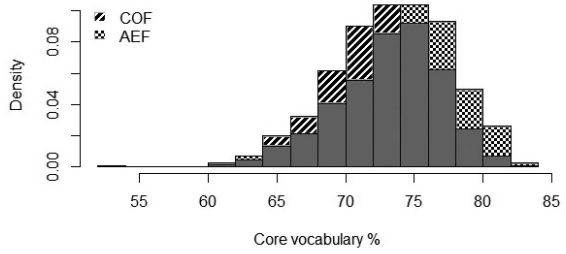


Figure 7.9: SweVoc category C distribution in children's ordinary and adults' ETR fiction

A comparison of ETR and ordinary fiction text for adults completes the group. The features contributing to the accuracy percentage of 96.7 % are to a large extent the same as in the previous test set. However, the lemma variation index also seems to have an impact.

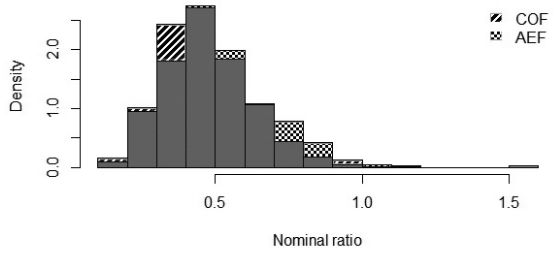


Figure 7.10: Nominal ratio distribution in children's ordinary and adults' ETR fiction

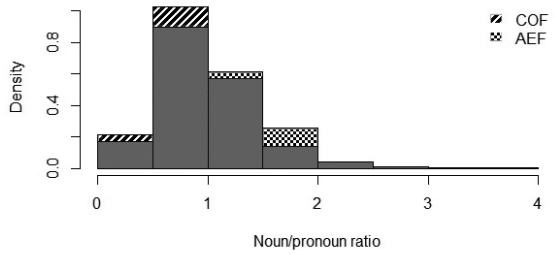


Figure 7.11: Noun/pronoun ratio distribution in children's ordinary and adults' ETR fiction

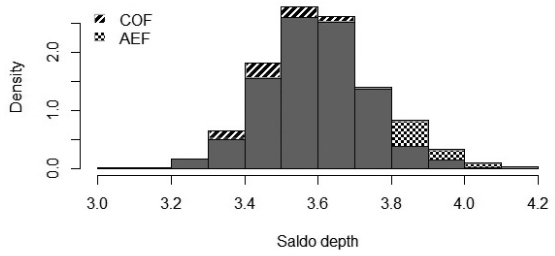


Figure 7.12: Saldo depth distribution in children's ordinary and adults' ETR fiction

7.3.2 News

At the *surface level*, the news ETR and ordinary texts differed significantly in word length as well as sentence length. The ratio of extra-long words and OVIX values differed significantly. Turning to the *vocabulary load*, lemma variation index and SweVoc total word list differed, as did all SweVoc categories. The *sentence structure* showed similar tendencies regarding mean dependency distance and number of pre- and post-modifiers showed a significant difference between the ordinary and the ETR materials, as did the number of subordinate clauses and the parse tree height. At the *idea density* level, the ratios of noun/ pronouns were similar, but the propositional density, the nominal ratio and the semantic depth in Saldo differed.

Text classification showed a significantly better result for the SVIT model with 99.6 % accuracy for SVIT, compared to the base model with 76.0 %. Principal component analysis showed that the *vocabulary load* features, mirroring vocabulary diversity (lemma variation index, figure 7.13) and difficulty (presence in SweVoc core vocabulary, figure 7.14) had major impact. Features related to *sentence structure*, *idea density* mirrored by nominal ratio (figure 7.15), together with the *surface level* features in terms of word length and sentence length, contributed to 56 % of the variation in the data sets at tertiary loading on class.

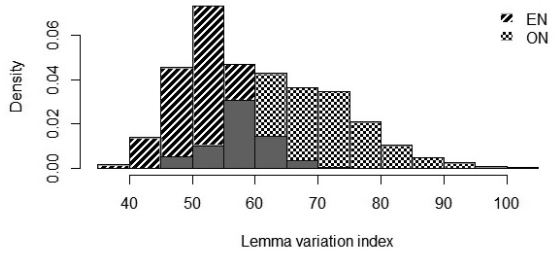


Figure 7.13: Lemma variation index distribution in ETR and ordinary news

7.3.3 Information

Turning to the information genre, signficant differences were seen at the *surface level* between ETR and ordinary tests. Also the *vocabulary*

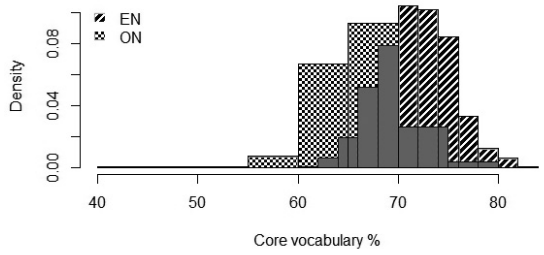


Figure 7.14: SweVoc core vocabulary distribution in ETR and ordinary news

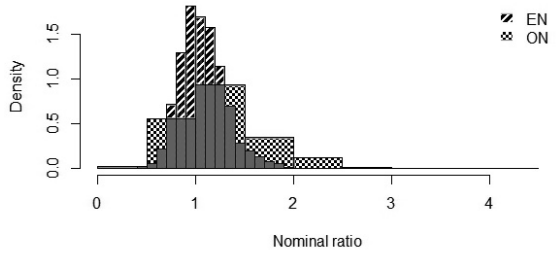


Figure 7.15: Nominal ratio distribution in ETR and ordinary news

load, *sentence structure*, and *idea density* features presented significant differences for almost all features. However, at the level of *human interest*, there was no difference found in personal noun ratios between the two text types.

The text classification experiment showed results to the advantage of the SVIT model. Overall accuracy was 97.1 %, compared to the base model accuracy of 85.9 %. The features contributing to the better results were primarily related to *vocabulary load* (SweVoc core vocabulary) and idea density (Saldo depth). Features acting at the level of *sentence structure* contributed, but only at secondary loading.

7.4 Category 3. Different text genres and same text types

7.4.1 News and information

The first set evaluated was the ETR news texts compared to the ETR information texts. Descriptive statistics gave that the ratio of entries in the total SweVoc word list as well as the subcategories categories C, D and H showed a difference between the text genres, indicating that the *vocabulary load* in terms of vocabulary difficulty seemed to differ. The *idea density*, measured as the propositional density and the nominal ratio, were also alike. The *surface features* word length, sentence length, ratio of extra-long words and OVIX were all significantly different. All features expressing *sentence structure* differed, as did all *idea density* indicators examined, except for the nominal ratio. Finally, the *human interest* marker PM, i.e. ratio of personal nouns, differed significantly.

Results from the classification experiment gave also in this case strong evidence for the advantage of the SMO algorithm and the SVIT feature model. Base line figures showed an overall accuracy of 59.4 %, while the SVIT model improved the correct classification rate to 91.7 %. The *vocabulary load* features contributed largely to this favorable result in terms of diversity (lemma variation index, figure 7.16) and difficulty (SweVoc, figure 7.17), and the *sentence structure* measured as parse tree height, figure 7.18. These features explained ≈ 42% of the variation in the entire data set, together with the *surface level* feature expressed as sentence length.

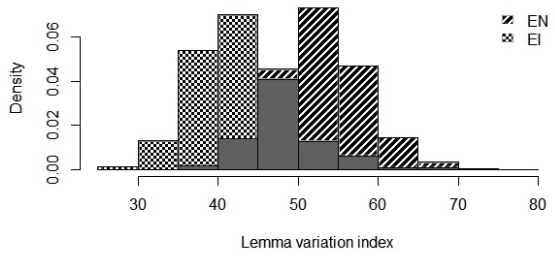


Figure 7.16: Lemma variation index distribution in ETR news and information

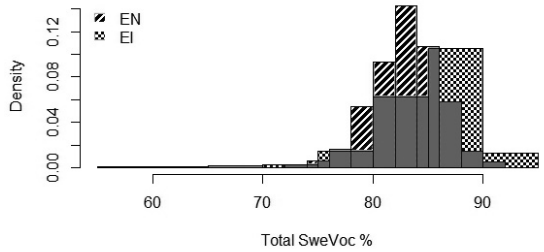


Figure 7.17: SweVoc ratio distribution in ETR news and information

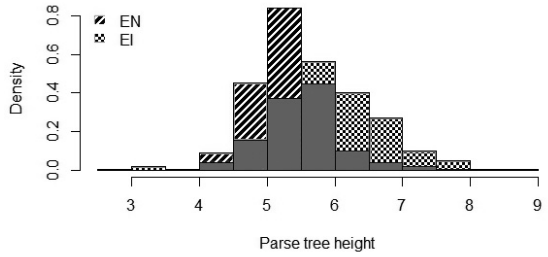


Figure 7.18: Parse tree height distribution in ETR news and information

7.4.2 News and fiction

The second set in this category was the ETR news texts and the children's ETR fiction texts. The statistical analysis showed significant differences for all features with the exception of lemma variation index, which suggested a similar *vocabulary load* with regard to this variable, and mean dependency distance, indicating a *sentence structure* agreement.

Classification evaluation with SMO showed 99.5 % accuracy at base line, and 99.7 % with the SVIT feature set, hence no difference. Analysis of principal components indicated that word length, noun/pronoun ratio, figure 7.19, together with the nominal ratio, figure 7.20, and the Saldo depth, figure 7.21, had an influence of $\approx 54\%$ on the overall variation. This indicates a clear overweight for features signaling *idea density*.

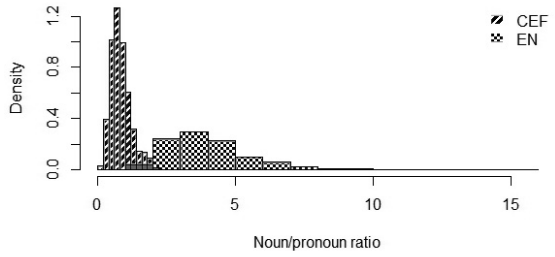


Figure 7.19: Noun/pronoun ratio distribution in children's ETR fiction and ETR news

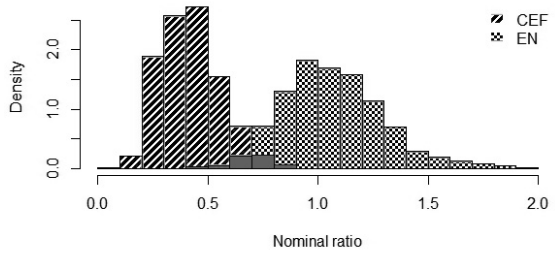


Figure 7.20: Nominal ratio distribution in children's ETR fiction and ETR news

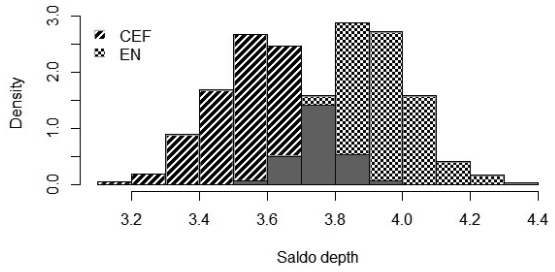


Figure 7.21: Saldo depth distribution in children's ETR fiction and ETR news

The third set shows largely the same figures as the previous one. The ETR news texts were compared to the ETR fiction targeted towards adults. Regarding the statistics, there was a significant difference for all features. Even in this case, mean dependency distance was similar for both sets (2.01), but the *idea density* feature propositional density was significantly different.

Turning to the classification results, we find that the base line accuracy with SMO was 96.4 %, compared to 98.6 % accuracy with the SVIT features. The principal components accountable for the positive results are identical to the previous test set, namely the noun/pronoun, figure 7.22, and nominal ratios, figure 7.23, and the Saldo depth, see figure 7.24, indicating the degree of *idea density*. These features, in combination with word length, explained 52 % of the variation in data.

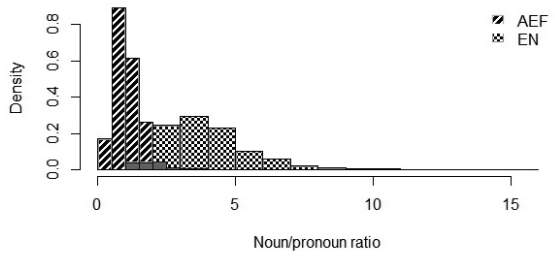


Figure 7.22: Noun/pronoun ratio distribution in adults' ETR fiction and ETR news

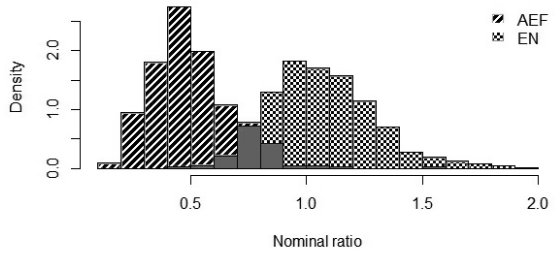


Figure 7.23: Nominal ratio distribution in adults' ETR fiction and ETR news

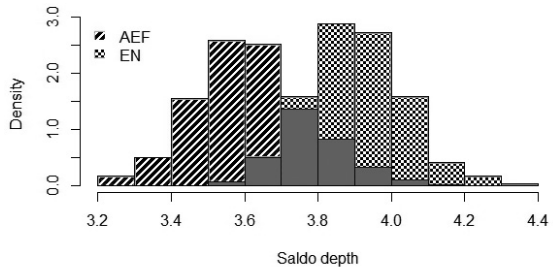


Figure 7.24: Saldo depth distribution in adults' ETR fiction and ETR news

7.4.3 Information and fiction

For the information texts and children's ETR fiction, there was a significant difference between all identified features. One exception was the propositional percentage, which was 45.9 % in the information texts and 45.3 % in the fiction texts, indicating some similarity for the *idea density* factors.

The classification results showed 97.0 % accuracy for the base model and 99.5 % for the SVIT model. Principal components in this case were word length, sentence length, and extra-long words at the *surface level*, parse tree height and frequency of post-nominal modifiers, contributing to the *sentence structure* complexity. Figure 7.25 illustrates density curves for postnominals. In all, 53 % of the variation in the data set was dependent upon these features.

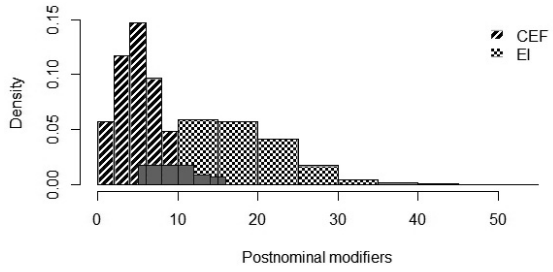


Figure 7.25: Postnominal modifier frequency in children's ETR fiction and ETR information

Next experiment was performed on the information texts and adults' ETR fiction. Descriptive statistics showed that the features differed in all aspects.

Text classification with SMO rendered 88.3 % accuracy for base line, compared to 98.0 % for the SVIT model. Principal components were found to conform to the previous experiment, see figures 7.26 and 7.27, but also the *idea density*, measured in terms of nominal ratio, contributed together with the *surface level* features mean sentence length and extra-long words. Totally 48 % of the variation was explained by this feature combination.

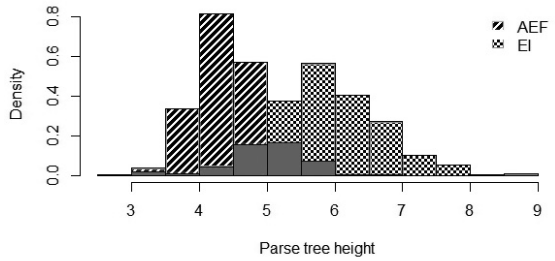


Figure 7.26: Parse tree height density in adults' ETR fiction and ETR information

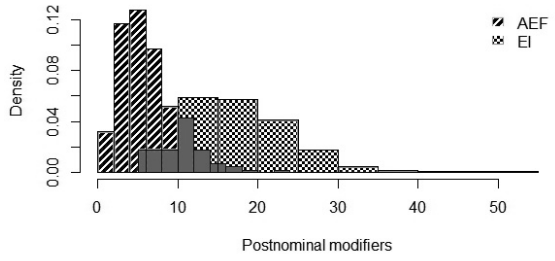


Figure 7.27: Postnominal modifier frequency density in adults' ETR fiction and ETR information

Ordinary news texts compared to ordinary fiction texts targeted towards adults and children showed very similar results, although the difference between the accuracy of SMO with Base and SVIT mod-

els differed. It seems that the task of distinguishing between texts for adults and children was the easiest, since the accuracy was 96.5 % already at base line and 98.8 % with SVIT. The corresponding accuracies for adults' fiction texts of different types were 80.2 % and 89.8 %, respectively. Principal component analyses were strikingly alike, in that *sentence structure* and *idea density* features seemed to have the largest impact.

7.5 Category 4. Different text genres and different text types

The final group to be reported contains text documents differing in genre as well as text type. Some of the results are discussed below.

7.5.1 Children's ordinary fiction and ETR information

The two members of this classification set differed in all respects, except for the propositional density that showed similar values.

Considering the classification results, we find that base line accuracy was 90.3 % with SMO, and that the SVIT model produced a result of 99.1 %. The principal component feature set is highly similar to the one shown for ETR information and ETR children's fiction, i.e. word length, sentence length, extra long words, parse tree height and frequency of post-nominal modifiers. Inclusion of *idea density* features, i.e. propositional density and Saldo depth, contributed to 48 % of the variation.

7.5.2 ETR fiction and ordinary news

These two data sets, i.e. ordinary news text and ETR fiction for two separate age groups, were similar in all respects. Classification accuracy for the pair involving children's fiction was 99.0 % for base line, and 99.4 % for the SVIT model. Corresponding figures for adults' fiction as compared to news texts were 95.2 % and 99.4 %. Nearly all features differed significantly, and PCAs showed predominance for features indicating *sentence structure* and *idea density*. *Vocabulary difficulty*, measured as SweVoc coverage, together with subordinate clause frequency, post-nominal modifiers, parse tree height, propositional density, nominal ratio, and word and sentence length accounted for 63 % and 61 % of the total variation.

7.5.3 Adults' ordinary fiction and ETR information

This test set exhibited the largest variation between classification with the base and SVIT models. The first task rendered 63.4 % accuracy and the second 99.4 %, hence an improvement of 36 pp. Significant differences between the texts were found for nearly all feature values. Principal component analysis revealed that features belonging to the levels of *vocabulary load* (lemma variation index and SweVoc coverage), *sentence structure* (all relevant features), and *idea density* (noun/pronoun ratio and Saldo depth), in addition to word length, contributed to 57 % of the variation.

7.5.4 Children's ordinary fiction and ETR news

The document set with children's ordinary fiction and ETR news texts differed in all respects, except for a few features. The number of subordinate clauses, both at the sentence level and the document level, kept close to each other. Mean frequency of subordinate clauses in children's ordinary fiction and ETR news sentences was 0.25 %, and in documents 7.49 % and 7.52 %, respectively. Also OVIX seems to be alike for the two text sets.

Classification with the SMO algorithm showed very good results at base line with 96.4 % accuracy, although the SVIT model enhanced the results to the excellent performance of 99.7 %. The principal components were found to be noun/pronoun ratio, nominal ratio and Saldo semantic depth, all belonging to the *idea density* category. Together with *surface structure* features, calculated as word length, they explained totally 53 % of the variation. See figures 7.28, 7.29 and 7.30.

7.6 General impact of different features

7.6.1 Surface level

As was seen in tables 7.1, 7.2 and 7.3, mean word length in characters (MWLC) differed statistically between all the text pairs according to the classification scheme. The PCAs provided further flesh on the bones, since it was included almost invariably in the set of primary features loaded on classes. It is quite clear that mean word length has a strong correlation to texts types, within as well as across genres.

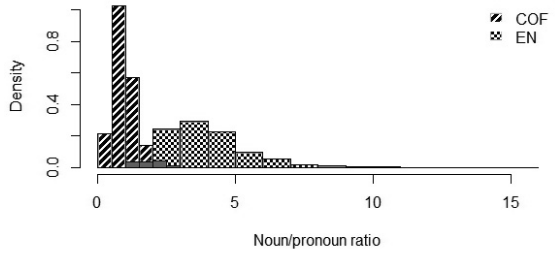


Figure 7.28: Noun/pronoun ratio distribution in children's ordinary fiction and ETR news

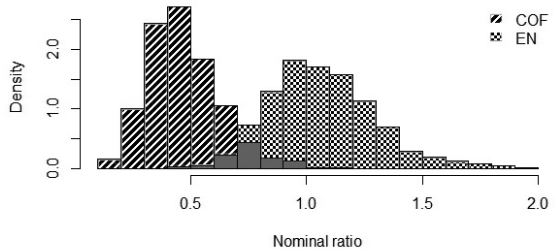


Figure 7.29: Nominal ratio distribution in children's ordinary fiction and ETR news

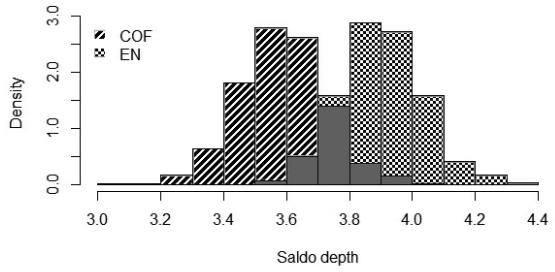


Figure 7.30: Saldo depth distribution in children's ordinary fiction and ETR news

Word length calculated as number of syllables (MWLS) was also found to differ significantly, but as demonstrated in section 5.1.2, syllable counts can be regarded as redundant in the presence of character counts.

Mean sentence length (MSL) showed significant differences for nearly all data sets, and it was included among the principal components at primary or secondary loadings in the majority of the classification sets.

The ratios of extra-long words (XLW) were also found to differ significantly for all data sets except for the pairs consisting of ETR fiction texts targeted for adults and ordinary fiction texts for children, and ETR information and ordinary news. From the classification point of view, it was found to have impact, again only for pairs involving ETR information and ordinary news texts.

Finally, the OVIX values differed significantly for all the text pairs examined, except ETR news and children's ordinary fiction where it had similar values.

7.6.2 Vocabulary load

The lemma variation index (LVIX) was found to be statistically different for all text pairs except for the children's ETR fiction compared to texts for fiction for adults, and ETR news texts, where it tended to be alike. PCA showed primary loadings on ETR news texts compared to ordinary news, and also compared to information texts. There was also a primary loading on children's fiction split between ETR and ordinary texts. This might indicate that the lemma variation index, signaling the degree of vocabulary diversity, is a strong candidate for inclusion into a combined readability measure.

From the vocabulary difficulty perspective, the ratio of words existing in the total SweVoc word list (SV) were significantly different within 10 of the 12 test pairs examined. The ones to break the regularity were children's ordinary vs. ETR fiction, and ETR news vs. fiction targeted towards adults. Turning to the loadings of PCA's on classes, it was selected at primary loading on the ETR news and ETR information texts. From the perspective of feature selection, a subset of SweVoc entries, consisting of words belonging to the core vocabulary (C) generally proved to be more useful than the total SweVoc list.

7.6.3 Sentence structure

Five features denoting sentence structure complexity were tested. The first was mean dependency distance (MDD), which turned out to be significantly different for all the classification sets from different text types. The couples consisting of texts from same text type varied. ETR fiction for children and for adults had similar MDD values, which was the case also for ETR news texts compared to the corresponding fiction texts.

The number of subordinate clauses (UA) showed dissimilarity for all classification pairs in category 3, while significance testings of the text pairs in the remaining categories gave varied results.

Pre- and post-nominal modifiers (AT and ET) differed significantly for all the texts, although the PCA turned out to include the prenominal modifiers only on classes of ETR texts vs. ordinary news texts. Post-nominal modifiers were also among the preferred features for this pair, as was the case for information texts as compared to fiction.

Parse tree height (PT) differed significantly at all instances. Principal component analysis favored this feature at primary loadings on all the classification sets involving information texts.

7.6.4 Idea density

The propositional density (Pr) was found to differ in all of the sets included in category 2, i.e. texts belonging to the same genre but of different types. The results of significance testings of the instances in categories 3 and 4 varied between significant difference and no difference.

Noun/pronoun ratio (NoPr) was significantly different for all texts except ETR news vs. ordinary news. This feature showed to be selected at primary loadings on all fiction texts compared crosswise, and also on ETR news texts as compared to the three different fiction text sets.

Turning to the nominal ratio (NR), it was found to differ for all texts except for children's ETR and ordinary fiction, and for ETR news and information. Eight classification sets proved to show nominal ratios that merited an inclusion among the primary selected features, namely all the sets involving ETR news compared to fiction and to ordinary news texts, all fiction texts cross-compared, and ETR information compared to ETR adults' fiction.

Finally, semantic depth in Saldo (Sa) differed significantly except in the set containing children's ETR texts and children's ordinary texts. In

PCA primary loadings, Saldo depth was chosen for all sets including ETR texts vs. fiction texts. It was also found to be decisive in classification of ETR texts directed towards adults as compared to the children's ETR and ordinary fiction.

7.6.5 Human interest

Human interest, measured as the ratio of personal nouns (PM), proved to be significantly different for all sets, with the exclusion of the smaller set of ETR news text compared to ordinary news text. In PCA, it was not distinguished at primary or secondary loadings at any instance.

7.7 Dominant features in the ETR subcorpora

After a survey of results from descriptive statistical analysis as well as principal component analysis, it is time to investigate the findings from the original point of view, i.e. how well do different features supposed to influence readability correspond to the actual findings? The following sections will provide a summary of the results as to the properties of the four ETR subcorpora in LB. The discussion will mainly touch upon findings that are supposed to substantially improve the performance of a document readability assessment tool as compared to LIX metrics.

7.7.1 Children's ETR fiction

ETR fiction texts directed towards children (CEF) were distinguishable from the corresponding texts directed towards adults (AEF) due to differences related to vocabulary load, sentence structure, and idea density. Mean word length as well as sentence length were found to be shorter in children's texts. The number of subordinate clauses were fewer, and the parse trees were also generally lower. The ratio between nouns and pronouns and the nominal ratio indicated a style slightly more directed towards oral discourse, and mean semantic depth was shallower in the children's texts.

Compared to children's ordinary fiction (COF), i.e. across text types, the findings were about the same, although also lemma variation index contributed to an additional impact on the vocabulary load. Hence, the vocabulary exhibits a higher degree of diversity for the ordinary fiction. The propositional density was generally higher in the ordinary texts,

but the semantic depth in Saldo seemed to correlate between the two texts types.

7.7.2 Adults' ETR fiction

Apart from the superficial features, vocabulary load, sentence structure, and idea density differed substantially between adults' ETR fiction (AEF) and children's ordinary fiction (COF). The vocabulary was generally easier, the number of subordinate clauses and postnominal modifiers were fewer, and the parse trees lower for the ETR texts. Regarding idea density, the noun/pronoun ratio and the nominal ratio pointed at a less colloquial style in adults' ETR fiction as compared to children's texts. The former texts were also found to contain more words that exceeded the semantic depth of words in children's texts.

7.7.3 ETR information

This text genre (EI) exhibited major differences in relation to all the other texts inspected. In comparison to ETR news texts (EN), the superficial features indicated that words were generally shorter, sentences longer, but the ratio of extra-long words were higher in the information texts. The vocabulary load was generally lower with regard to both vocabulary diversity and difficulty, but the sentence structure was more complex as judged by the number of post-nominal modifiers and the parse tree heights.

Compared to fiction texts, word length and sentence length indicated substantially higher values for the information texts, as did the frequency of extra-long words. The vocabulary load was found to correspond to the findings above, i.e. a lower vocabulary diversity and difficulty. The sentence structure was remarkably more complex, also in line with the above-mentioned findings. However, in contrast to the comparison between information and news texts, semantic depth seemed to provide an important clue regarding complexity, in that the depth in Saldo was much higher than in the fiction texts.

7.7.4 ETR news

In comparison to ordinary news texts (ON), the ETR news subcorpus (EN) exhibited differences at all levels. Word length, sentence length,

and OVIX displayed much lower values for the ETR materials, indicating differences distinguishable already at the surface level. At the level of vocabulary load, a reduced diversity and difficulty was seen, judging by lemma variation index and SweVoc ratios. Sentence structure was less complex in terms of subordinate clause frequencies, pre- and post-nominal modifiers, and parse tree height. The latter findings were particularly explicit in comparison to the GP news text collection. The features indicating differences at the idea density level were however not unambiguous. The nominal ratio was considerably lower in comparison to the ordinary news texts, while mean semantic depth actually was higher in the ETR news.

Lastly, we will take a look into the matter of what differentiates the ETR news texts from fiction texts. Superficial features in terms of word length, sentence length, and ratio of extra-long words showed higher values for the news texts. Vocabulary diversity differed in that a higher lemma variation index was seen for the news texts in comparison to ETR fiction texts, but it was lower as compared to ordinary fiction. Vocabulary difficulty seemed to be higher in the news texts in that the coverage of words in SweVoc core vocabulary was higher for the fiction texts. At the sentence structure level, the number of subordinate clauses and the parse tree heights showed higher values for the news than the fiction texts. At the idea density level, the propositional density actually showed a slightly lower overall value for the news materials, but the noun/pronoun ratio, the nominal ratio, and the semantic depth showed unambiguously higher values.

7.8 Diagnosticity of specific features

7.8.1 Surface level

Mean word length and sentence length are undoubtedly indicators of text complexity, but only if the analysis is superficial in scope. It can, however, be used as a supplement to deep linguistic features when measuring readability. Frequency calculations of extra-long words showed to be useful when separating specific genres. The OVIX values generally seemed to be of minor utility in the presence of lemma variation index values.

7.8.2 Vocabulary load

Lemma variation index seemed to be useful, especially for distinction of texts differing in type. The SweVoc general word list turned out to be a valuable asset for identifying core vocabulary items.

7.8.3 Sentence structure

The syntactic mean dependency distance was significantly different between ETR texts and ordinary texts. Frequency measurements of subordinate clauses were good indicators of sentence complexity, as was the measures of postnominal modifiers. Parse tree heights also seemed to correlate well with complexity.

7.8.4 Idea density

The noun/pronoun ratio showed uneven results for the texts differing in type but within the same genre. The nominal ratios and the semantic depths seemed generally to be good indicators of idea density across the materials.

7.8.5 Human interest

The only feature investigated was the ratios of personal nouns. It was found to differ in a significant way between almost all test sets, but did not contribute to the classification accuracy.

7.9 Feature selection

We also wanted to identify a representative set of features from which to construct a robust classification model. This issue is central in most machine learning tasks, especially for datasets with large amounts of variables, as redundant features duplicate much of the information contained in other attributes, and irrelevant features can reduce the accuracy. Tables 7.4 and 7.5 show the PCA loadings of principal and secondary components on all subcorpora, while tables 7.6 and 7.7 display the loadings in the absence of surface features. It appears that superficial features in terms of mean word length in characters (MWLC),

mean word length in syllables (MWLS), mean sentence length (MSL), or frequency of extra-long words (XLW) are present in all the feature set loadings. Since the LIX value is a product of word length and sentence lengths, we were interested in to which degree these features influence the classification results in the presence of more elaborated feature sets. All the classification tasks were subsequently repeated with feature vector attributes limited to deep linguistic features. The results are shown in 7.8.

Feat cat	**Feature**	**CEF**	**AEF**	**EN**	**EI**
Surface	MWLC	p+	p+	s+	
	MWLS			s+	p+
	MSL	s-	s+	p- s+	s+
	XLW			s+	s+
	OVIX				
Voc load	LVIX	p+	p+	p+	
	SV	s-	s+	p-	p-
	SVC	s-	p-	p-	p-
	SVD				
	SVH				
	SVK				
	SVS				
Sent struct	MDD				
	UA	s-			p-
	AT		s+		s+
	ET		s+		s+
	PT	s-	s+	p- s+	s+
Idea dens	Pr				p-
	NoPr	p+	p+		
	NR	p+	p+		
	Sa	p+			
Human interest	PM				

Table 7.4: Results from principal component analysis of features in ETR subcorpora.p = primary loading on PCA, s = secondary loading. A plus (+) or minus (-) sign designates a loading in either positive or negative direction.

Feat cat	Feature	COF	AOF	ON	OI
Surface	MWLC	p+	p+	p+	s+
	MWLS	p+	p+	p+	s+
	MSL	s+	s-	s-	p-
	XLW				s+
	OVIX				
Voc load	LVIX	p+			
	SV	s+	s-		p-
	SVC		s-	p-	p-
	SVD				
	SVH				
	SVK				
	SVS				
Sent struct	MDD				
	UA	s+	s-	s-	
	AT		p+		s+
	ET			s-	s+
	PT	s+	s-	s-	p-
Idea dens	Pr	s+		s-	p-
	NoPr	p+	p+		
	NR	p+	p+	p+	
	Sa		p+	p+	
Human interest	PM				

Table 7.5: Results from principal component analysis of features in ordinary subcorpora.p = primary loading on PCA, s = secondary loading. A plus (+) or minus (-) sign designates a loading in either positive or negative direction.

7.10 Word reading

Linguistic features that have been discussed upon, but not implemented, concern single-word reading. Although sentence and document reading can be regarded as a function of the reading of words constituting the text, different processes interact at different levels of comprehension. Word length and frequency are evidently of importance in isolated word reading. Other variables that seem to affect the identification of words are lexical neighborhood size and frequency. These features were examined in relation to sentence readability in section 5.1.6,

Feat cat	Feature	CEF	AEF	EN	EI
Voc load	LVIX	p+			
	SV	s-	p- s+	p-	p-
	SVC	s-	p- s+	p-	p-
	SVD				
	SVH				s+
	SVK				
	SVS				
Sent struct	MDD				s-
	UA	s-		p-	p-
	AT	p+	s+	s+	s+
	ET		s+	s+	s+
	PT	s-	s+	p- s+	p- s+
Idea dens	Pr	s-	p-	p-	p-
	NoPr	p+	p+	s+	s+
	NR	p+	p+	s+	
	Sa	p+			
Human interest	PM			s+	

Table 7.6: Results from principal component analysis of deep linguistic features in ETR subcorpora.p = primary loading on PCA, s = secondary loading. A plus (+) or minus (-) sign designates a loading in either positive or negative direction.

where the lexical neighborhood densities and frequencies were found to differ significantly between some subcorpora, but these are mere tendencies. In the morphological domain of word reading, there are several measures that have been scientifically documented, but as said by way of introduction, they are only mentioned at face value in section 2.4.3. The importance of language specific features such as adverbial particles and compounds were also discussed in the same section, but no further analyses on the matter have been done.

7.11 Sentence reading

A subset of the features implemented in SVIT was considered for assessment of sentence readability. These are listed in table 2.8. Since there does not exist any gold standard for Swedish sentences from the perspective of readability, no classification was made at the sentence level.

Feat cat	Feature	COF	AOF	ON	OI
Voc load	LVIX	p+	p+	p+	s+
	SV	s+	s-		p-
	SVC	p-	s-	p-	p-
	SVD				
	SVH				
	SVK				
	SVS				
Sent struct	MDD				
	UA	s+	s-	s-	p-
	AT		p+	s-	
	ET	s+	p+	s-	s+
	PT	s+	s-	s-	p-
Idea dens	Pr	s+	s-	s-	p-
	NoPr	p+	p+	p+	s+
	NR	p+	p+	p+	
	Sa	p+	p+	p+	s+
Human interest	PM				s-

Table 7.7: Results from principal component analysis of deep linguistic features in ordinary subcorpora.p = primary loading on PCA, s = secondary loading. A plus (+) or minus (-) sign designates a loading in either positive or negative direction.

Evaluation of the feasibility of the suggested features in a sentence reading situation should ideally be made by human readers, which is beyond the scope of this thesis.

7.12 The final SVIT model for text complexity assessment

It has been shown that generally, the full SVIT feature model performed best. It presented on average 95.6 % accuracy in 28 different experiments, where the same model with reduced feature set achieved an accuracy of on average 89.3 %. This is contrasted to the base model built on LIX values, which reached an accuracy of on average 82.3 %, but with large variations. It has also been demonstrated that the full SVIT model performed best when distinguishing between texts of different genres, but that it was only marginally better than LIX in the task of separating between texts with highly distinctive superficial features.

Classification task	Base model	SVIT model	Reduced feature set
CEF/AEF	77.4	83.9	67.4
COF/AOF	77.1	92.9	90.3
CEF/COF	72.9	84.6	77.0
CEF/AOF	89.8	97.5	96.5
AEF/COF	55.4	84.2	82.3
AEF/AOF	73.7	96.7	96.5
EN/ON	76.0	99.6	98.3
EI/OI	85.9	97.1	97.1
CEF/EN	99.5	99.7	99.3
CEF/EI	97.0	99.5	98.5
AEF/EN	96.4	98.6	97.6
AEF/EI	88.3	98.0	98.5
COF/ON	96.5	98.8	97.7
COF/OI	98.6	99.4	98.9
AOF/ON	80.2	89.8	90.2
AOF/OI	85.1	99.1	99.3
EN/EI	59.4	91.7	89.8
ON/OI	59.9	94.6	95.0
CEF/ON	99.0	99.4	99.2
CEF/OI	99.4	99.4	99.3
AEF/ON	95.2	99.4	99.4
AEF/OI	98.9	99.6	99.3
COF/EN	96.4	99.7	99.5
COF/EI	90.3	99.1	98.5
AOF/EN	44.6	100.0	99.9
AOF/EI	63.4	99.4	99.3
ON/EI	71.5	98.1	97.9
OI/EN	92.1	99.0	99.0
All test sets	40.5	78.8	72.4

Table 7.8: Comparison of the accuracy of different classification tasks with the SMO algorithm, base model, SVIT model and reduced feature set

The strongest difference between the full SVIT and LIX amounted to 36 pp, and was seen in the task of singling out ETR information from ordinary fiction targeted towards adults. Also the LIX figures for ETR news texts and information texts are close to each other, to the extent

that the classification task performed with the base model performed 59.4 %, compared to full SVIT which reached 91.7 %. We have shown that a feature set of 22 variables, selected from different language levels, proves to give an overall result of 95–100 % accuracy for 21 out of 28 classification tasks. The model produced an accuracy of 90–95 % in three tasks, and 80–90 % in the remaining four tasks.

Feature reduction is normally performed in order reduce the computational cost and produce a classifier with good generalizability. The time taken to build a SVIT model in Weka with the SMO algorithm, and an input of 1,000 vector instances was 0.1 second for the full model, and 0.06 seconds for the model with reduced feature set, consequently only a minimal time cost for both sets. For the NB algorithm, the corresponding time with both full and reduced feature set was 0.01 second, while finally the CVR algorithm required 0.26 seconds for the full SVIT model, and 0.20 seconds for the reduced model. Given the fact that the accuracy of reduced feature model never exceeded the outcome from tasks performed with the full feature model, and that the computational costs seem differ only marginally, we can conclude that the full feature SVIT model implemented in the SMO algorithm, seems to be the optimal choice for readability assessment of documents across genres and text types.

8 Discussion and conclusions

The general purpose of the study was to investigate which factors influence the readability of various texts, and to what extent these factors in aggregate can be regarded as worthy competitors to traditional readability measures. The ultimate goal was to identify specific linguistic properties that could be assumed to influence a specific reader's understanding of the text, or from a more general point of view, to correspond to the reading level of a target group of readers.

Statistical analyses and text classification experiments showed that a combination of features observed at different language levels could actually predict text complexity, and ultimately readability. It was found that binary text classification accuracy with the SVIT model varied between 83.9 % and 100.0 % correctness, and that the difference between SVIT and a base model varied between 0.2 and 55.4 percentage points.

Input data for the base model were LIX calculations, indicating mean word and sentence length in the test documents. The ETR texts examined were probably produced according to current guidelines involving LIX measurements during the writing phase, which implies that the authors might have made explicit attempts to limit sentence length in order to keep to certain standards. Texts such as the ETR news that all originate from the same source might thus be unfairly advantaged compared to less homogenic texts. In fact, binary classification involving ETR news showed the smallest difference in performance between the two models – 0.2 % and 2.2 % respectively, compared to children's and adults' ETR fiction. However, it is also the fact that the highest difference in accuracy between LIX and SVIT (55.4 %) was found in ETR news text compared to adults' ordinary fiction. This can probably be attributed to the presence of sentence structure features in the SVIT model, which were not captured by LIX counts. The SVIT values in classification results with smaller differences compared to LIX indicated a higher impact of idea density features. It is likely that the large

variations in sentence length data for documents in the adults' ordinary fiction test set[24] contributed to better statistical modelling. Anther hypothesis is that there is a stronger correlation between idea density features and word length than between sentence structure features and sentence length.

Mean LIX values in ETR subcorpora indicate a rising degree of complexity from 17.2 in children's fiction, 22.4 for adults' fiction, and to 34.8 and 33.3 for news and information texts respectively. Corresponding values in ordinary texts can be ordered from 21.6 in children's fiction, 30.0 for adults' fiction, 40.9 for news, and 46.8 for information texts. Generally, it appears that the four subcorpora of each text type, i.e. the ETR texts and the ordinary texts, can be graded along a rising complexity scale with respect to predominant surface level features, while the vocabulary load seems to be more related to genre. Another observation is that the ETR fiction texts targeted towards adults generally have a lower vocabulary load than children's fiction of both types as well as other genres of the ETR text type. Sentence structure feature values are not unambiguously pointing towards neither a general complexity trend nor genre relation. Dependency distance between sentence elements are alike in all ETR subcorpora, while they seem to be genre-specific in the ordinary subcorpora. The frequency of subordinate clauses is generally lower in the adults' ETR fiction than any of the other texts, ETR as well as ordinary. Pre- and postnominal modifier frequencies and parse tree heights are arranged according to a complexity scale from lower to higher for both text types. Idea density levels in terms of noun/pronoun ratio and nominal ratio are low in all fiction texts, higher in news texts and rise to extremely high levels in ordinary information texts, as expected. Propositional density does not differ much neither within text type nor across genre. General mean semantic depth follows the same rising tendency curve for both ETR and ordinary text types, but with a certain drop for information texts. This is most probably due to the fact that many new compound words that are frequent in information texts are missing in Saldo.

The ratio of proper nouns did not have any impact on the success of the classifier, although it was found to differ significantly between almost all test sets. The benefit of proper noun ratio measurement as an indicator of human interest is thus not verified.

Adaptation of texts to specific target groups was regarded as another major challenge of the study. Public initiatives promoting readability

[24] MSL $\overline{X}$=12.28,σ=4.07

are often influenced by an old-fashioned view of reading, ignoring the complexity of the reading process and its manifestations at the individual or group level. In most cases, the adult reader can access texts from different genres and types without any major difficulties. There is, however, a considerable amount of people that experience substantial problems when it comes to reading unfamiliar Swedish texts. The International Adult Literary Survey (IALS), carried out during 1994-1996 by means of face-to-face interviews and literacy tests, reported that 25 % of the Swedish adult population were not able to read unfamiliar texts unless they had a simple structure and an unambiguous message (The National Agency for Education 2000). This is a very heterogeneous group of people, which means that non-native readers are brought together with persons affected by acquired brain injuries or intellectual disability. Beginning readers are also included, given that not all Swedish immigrants are originally literate. The group also contains persons with developmental dyslexia, which is a specific reading disability with a genetic component. Depending on the nature of etiology, different remedial efforts may be called for, and the readability assessment of specific text materials is a central issue. If the weak component in the reading skill is properly identified, texts matching certain criteria at specific textual levels might be called for (Aaron, Joshi and Williams 1999).

Constraining readers' needs and wishes to the individual level, we think that the personal interest might be satisfied by assessment of the textual genre. Again, a person with limited reading skills might want to have reading materials adapted to his or her personal interest in terms of genre and topic. We have demonstrated that it is possible to identify features that differentiate genres both within and across text types and that the SVIT model for classification of an equal amount of documents from all test sets was able to assign correct text genre and type with an average accuracy of 79.2 %. The corresponding value with the base model was 38.6 %. The SVIT model was not the indisputable winner in all classification tasks, but it is demonstrated to reflect text complexity characteristics unrecognized by LIX.

Some words are also to be said about the corpus based approach adopted in the study. The corpus material is restricted to written texts, and a more complete view of language use would certainly have been achieved with the addition of spoken language data. The fact that two different corpora are used also rises the question whether the data from each of the sources are equally treated at the preprocessing stage. The importance of consistency with regard to taggers, parsers and feature

selection is stressed by among others Biber (1990), and is clearly vital for the reliability of statistical results. The SUC corpus was assembled and processed about twenty years earlier than the LäSBarT corpus. The tagsets used are mutually interchangeable, but the automatic lemmatization turned out to give different results in the two materials. A considerable amount of manual work was thus investigated into the matter of harmonizing the corpora. Generally, it can be admitted that no statistical part-of-speech tagger, lemmatizer or parser is 100 % correct, and that errors in the preprocessing phase give rise to incorrect data which may invalidate further statistical analysis. It is, however, our belief that the present material is satisfactorily preprocessed and that the manual checking is adequately performed.

A general conclusion of this study is that readers presumably can be provided with text materials adapted for his or her language limitations, given that the group metaphor is acknowledged and that texts are produced or checked with NLP support acting at different language levels. The optimal method for proving the fairness of the present approach is to validate the findings with data from human evaluation. In the future, we envisage further studies including data from human studies in an authentic reading environment.

REFERENCES

Aaron, P.G., Malatesha Joshi and Kathryn A. Williams 1999. Not all reading disabilities are alike. *Journal of Learning Disabilities* 32: 120–137.

Aluisio, Sandra, Lucia Specia, Caroline Gasperin and Carolina Scarton 2010. Readability assessment for text simplification. *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '10, 1–9. Stroudsburg, PA, USA: Association for Computational Linguistics.

Andersen, Gisle 2005. Gjennomgang og evaluering av språkressurser fra NSTs konkursbo. Technical Report, Avdeling for kultur, språk og informasjonsteknologi (Aksis) ved UNIFOB/Universitetet i Bergen.

Anderson, Jonathan 1983. Lix and rix: Variations of a little-known readability index. *Journal of Reading* 26 (6): 490–496.

Aro, Mikko 2004. Learning to read. The effect of orthography. Ph.D. diss., University of Jyväskylä, Helsinki.

Atkins, Sue, Jeremy Clear and N. Ostler 1992. Corpus design criteria. *Literary and Linguistic Computing* 7 (1): 1–16.

Baayen, R. Harald and Rochelle Lieber 1996. Word frequency distributions and lexical semantics. *Computers and the Humanities* 30: 281–291.

Baddeley, A. 1990. *Human memory: Theory and practice*. Hove: Erlbaum.

Berg, Sture and Yvonne Cederholm 2001. Att hålla på formerna. Om framväxten av Svensk morfologisk databas. Sture Allén et al. (ed.), *Gäller stam, suffix och ord. Festskrift till Martin Gellerstam*, 58–69. Göteborg.

Biber, D. 1990. Methodological issues regarding corpus-based analyses of linguistic variation. *Literary and Linguistic Computing* 5 (4): 257–269.

Biber, D. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8 (4): 243–258.

Biber, Douglas and Susan Conrad 2009. *Register, genre, and style*. Cambridge University Press.

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

Bliss, Charles K. 1949. *International Semantography: A non-alphabetical symbol writing readable in all languages. A Practical Tool for General International Communication, Especially in Science, Industry, Commerce, Traffic, etc. and for Semantical Education, Based on the Principles of Ideographic Writing and Chemical Symbolism*. Sydney: Institute of Semantography.

Borin, Lars and Markus Forsberg 2009. All in the family: A comparison of SALDO and WordNet. *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources – between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*. Odense.

Bornkessel, Ina D., Matthias Schlesewsky and Angela D. Friederici 2002. Grammar overrides frequency: evidence from the online processing of flexible word order. *Cognition* 85: B21–B30.

Bornkessel-Schlesewsky, Ina and Matthias Schlesewsky 2008. The role of prominence information in the real-time comprehension of transitive constructions: A cross-linguistic approach. *Language and Linguistics Compass*, vol. 2.

Brants, Thorsten 2000a. TnT: a statistical part-of-speech tagger. *ANLC '00 Proceedings of the sixth conference on Applied language processing*.

Brants, Thorsten 2000b. TnT – A Statistical Part-of-Speech Tagger. Technical Report, Saarland University, Computational Linguistics.

Brants, Thorsten, Wojciech Skut and Hans Uszkoreit 1999. Syntactic annotation of a German newspaper corpus. *Proceedings of the ATALA Treebank Workshop*, 69–76. France.

Brown, Cati, Tony Snodgrass, Susan J. Kemper, Ruth Herman and Michael A. Covington 2008. Automatic measurement of propositional idea density from part-of-speech tagging. *Behavior Research Methods* 40 (2): 540–545.

Brown, R. 1958. How shall a thing be named? *Psychological Review* 65: 14–21.

Brysbaert, Marc and Boris New 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41 (4): 977–990.

Cassirer, P. 1970. Deskriptiv stilistik. En begrepps- och metoddiskussion.

CEC 2007. European i2010 initiative on e-Inclusion "To be part of the Information society" Impact Assessment.

Centrum för Lättläst. Lättläst-tjänsten 2012. Kommunundersökningen 2012: Lättläst information på kommunernas webbplatser. Technical Report.

Chae, Jicun and Ani Nenkova 2009. Predicting the fluency of text with shallow structural features: Case studies of machine translation and human-written text. *Proceedings of the 12th Conference of the European Chapter of the ACL*, 139–147. Athens, Greece.

Chall, Jeanne S 1958. *Readability – an appraisal of research and application*. The Ohio State University.

Chapman, K. L. and C. B. Mervis 1989. Patterns of object-name extension in production. *Journal of Child Language* 16 (3): 561–571.

Charniak, Eugene 2000. A maximum-entropy-inspired parser. *Proceedings NAACL*, pp. 132–139.

Chotlos, John W. 1944. A statistical and comparative analysis of individual written language samples. http://search.proquest.com/docview/916872144?accountid=11162.

Cohen, David, Monique Plaza, Fernando Perez-Diaz, Odile Lanthier, Dominique Chauvin, Nicole Hambourg, Anna J. Wilson, Michel Basquin, Philippe Mazet and Jean Philippe Rivière 2006. Individual cognitive training of reading disability improves word identification and sentence comprehension in adults with mild mental retardation. *Research in Developmental Disabilities* 27: 501–516.

Cohen, David, Jean Philippe Rivière, Monique Plaza, Caroline Thompson, Dominique Chauvin, Nicole Hambourg, Odile Lanthier, Philippe Mazet and Martine Flament 2001. Word identification in adults with mild mental retardation: Does IQ influence reading achievement? *Brain and Cognition* 46: 69–73.

Colé, Pascale, Annie Magnan and Jonathan Grainger 1999. Syllable-sized units in visual word recognition: Evidence from skilled and beginning readers of French. *Applied Psycholinguistics* 20 (04): 507–532.

Coleman, M. and T.L. Liau 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology* 60: 283–284.

Coltheart, M., E. Davelaar, J.F. Jonasson and D. Besner 1977, *Access to the internal lexicon*. Hillsdale, NJ: Erlbaum.

Cutting, L.E., A.M. Clements, S. Courtney, S.L. Rimrodt, J.G.B. Schafer, J. Bisesi, J.J. Pekar and K.R. Pugh 2006. Differential components of sentence comprehension: Beyond single word reading and memory. *Neuroimage* 29: 429–438.

Dagan, Ido, Alon Itai and Ulrike Schwall 1991. Two languages are more informative than one. *Proceedings of the 29th annual meeting on Association for Computational Linguistics*, ACL '91, 130–137. Stroudsburg, PA, USA: Association for Computational Linguistics.

Dale, Edgar and Jeanne S Chall 1948. A formula for predicting readability. *Educational Research Bulletin* 27: 37–54.

Dale, Edgar and R.W. Tyler 1934. A study of the factors influencing the difficulty of reading materials for adults of limited reading ability. *Library Quarterly* 4: 384–412.

Dell'Orletta, Felice, Simonetta Montemagni and Giulia Venturi 2011. READ-IT: Assessing readability of Italian Texts with a View to Text Simplification. *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*. Association for Computational Linguistics, Edinburgh, Scotland, UK.

De Mauro, Tullio 1980. *Guida all'uso delle parole*. Roma: Editori Riuniti.

Dolch, E. W. 1936. A basic sight vocabulary. *The Elementary School Journal* 36: 456–460.

Dolch, E. W. 1949. The use of vocabulary lists in predicting readability and in developing reading materials. *Elementary English* XXVI: 142–149, 177.

Dollaghan, C. A. 1994. Children's phonological neighbourhoods: Half empty or half full? *Journal of Child Language* 21: 257–271.

Duabeitia, Jon Andoni and Eduardo Vidal-Abarca 2008. Children like dense neighborhoods: Orthographic neighborhood density effects in novel readers. *The Spanish Journal of Psychology* 11 (1): 26–35.

Ejerhed, Eva, Gunnel Källgren, Ola Wennstedt and Magnus Åström 1992. The linguistic annotation system of the Stockholm-Umeå corpus project. Technical Report, Department of General Linguistics, University of Umeå.

Elert, Claes-Christian 1997. *Allmän och svensk fonetik*. 7. Norstedts Förlag.

Falk, Lena and Stefan Johansson 2006. Hur fungerar lättlästa texter på webben? Undersökning av lättlästa offentliga texter på offentliga webbplatser. Technical Report, Funka.nu.

Falk, Lena Erika 2003. *Skriv så att folk förstår*. Centrum för Lättläst.

Falkenjack, Johan and Katarina Heimann Mühlenbock 2012. Using the probability of readability to order Swedish texts. *Proceedings of Swedish Language Technology Conference, SLTC*. Lund.

Farr, James N., James J. Jenkins and Donald G. Paterson 1951. Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology* XXXV: 333–37.

Fellbaum, Christiane 1998a. A semantic network of english – the mother of all wordnets. *Computers and the Humanities* 32: 209–220.

Fellbaum, Christiane (ed.) 1998b. *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press.

Feng, Lijun 2009. Automatic readability assessment for people with intellectual disabilities. *SIGACCESS Access. Comput.*, no. 93 (January): 84–91.

Feng, Lijun 2010. Automatic readability assessment. Ph.D. diss., City University of New York.

Feng, Lijun, Noémie Elhadad and Matt Huenerfauth 2009. Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, 229–237. Stroudsburg, PA, USA: Association for Computational Linguistics.

Feng, Lijun, Martin Jansche, Matt Huenerfauth and Noémie Elhadad 2010. A comparison of features for automatic readability assessment. *Proceedings of the 23rd international conference on computational linguistics: Posters*, COLING '10, 276–284. Stroudsburg, PA, USA: Association for Computational Linguistics.

Fish, Stanley 1970. Literature in the Reader: Affective Stylistics. *New Literary History* 2 (1): 123–162.

Flesch, Rudolf 1948. A new readability yardstick. *Journal of Applied Psychology* 32 (3): 221–233.

Flesch, Rudolf Franz 1943. Marks of readable style. Technical Report, New York.

Forbes, Fritz W. and William C. Cottle 1953. A new method for determining readability of standardized tests. *Journal of Applied Psychology* XXXVII (June): 185–90.

Forsbom, Eva 2006. A Swedish Base Vocabulary Pool. *Proceedings of Swedish Language Technology Conference, SLTC*. Dept. of Linguistics and Philology, Uppsala University.

Forslund, Ann-Charlotte 2004. LÄST, OLÄST och symbolspråk på svenska. Master's Thesis, University of Gothenburg.

Francis, W. Nelson and Henry Kučera 1979. Manual of information to accompany a standard corpus of present-day edited American English for use with digital computers. Technical Report, Department of Linguistics, Brown University.

Friederici, Angela D., Christian J. Fiebach, Matthias Schlesewsky, Ina D. Bornkessel and D. Yves von Cramon 2006. Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex* 16: 1709–1717.

Gellerstam, Martin 1991. Modern Swedish text corpora. Jan Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium*, Volume 82, 149–169. Stockholm.

Gibson, Edward 1998. Linguistic complexity: locality of syntactic dependencies. *Cognition* 68 (1): 1–76.

Goodman, Kenneth S. and Yetta M. Goodman 2009. Helping readers make sense of print. Susan E. Israel and Gerald G. Duffy (eds), *Handbook of Research on Reading Comprehension. Research that Supports a Whole Language Pedagogy*. Routledge.

Göransson, Kerstin 1985. *Att skriva lätt. Utvärdering av nyhetstidningen 8 sidor*. Stockholm: ALA Hls. Ped. Inst.

Goswami, Usha 2008. The development of reading across languages. *Annals of the New York Academy of Science* 1145: 1–12.

Gough, Philip B. and W.E. Tunmer 1986. Decoding, reading and reading disability. *Remedial and Special Education* 7: 6–10.

Graesser, Arthur C., Danielle S. McNamara and Jonna M. Kulikowich June/July 2011. Coh-metrix. *Educational Researcher* 40 (5): 223–234.

Grainger, Jonathan, J. O'regan, Arthur Jacobs and Juan Segui 1989. On the role of competing word units in visual word recognition: The neighborhood frequency effect. *Attention, Perception, & Psychophysics* 45: 189–195.

Grainger, Jonathan and Juan Segui 1990. Neighborhood frequency effects in visual word recognition: A comparison of lexical decision and masked identification latencies. *Attention, Perception, & Psychophysics* 47: 191–198.

Gray, J.S. and B. Leary 1935. *What makes a book readable.* Chicago: Chicago University Press.

Grefenstette, Gregory and Pasi Tapanainen 1994. What is a word, what is a sentence? Problems of Tokenization. *3rd Conference on Computational Lexicography and Text Research*. Budapest.

Grigorenko, Elena L. and Adam J. Naples (eds) 2007. *Single-Word Reading. Behavioral and Biological Perspectives*. New Directions in Communication Disorders Research: Integrative Approaches. Lawrence Erlbaum Associates.

Gunning, Robert 1952. *The technique of clear writing*. New York, NY: McGraw-Hill International Book Co.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Bernhard Reutemann, Peter Reutemann and Ian H. Witten 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, vol. 11.

Halliday, M.A.K 1985. *An introduction to functional grammar*. London: Edward Arnold.

Heilman, Michael J., Kevyn Collins-Thompson, Jamie Callan and Maxine Eskenazi 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proceedings of NAACL HLT*, 460–467. Association for Computational Linguistics, Rochester, NY.

Heilman, Michael J., Kevyn Collins-Thompson and Maxine Eskenazi 2008. An analysis of statistical models and features for reading difficulty prediction. *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, EANL '08, 71–79. Stroudsburg, PA, USA: Association for Computational Linguistics.

Heimann Mühlenbock, Katarina and Sofie Johansson Kokkinakis 2012. SweVoc – A Swedish vocabulary resource for CALL. *Proceedings of the SLTC 2012 workshop on NLP for CALL*, 28–34. Lund: Linköping University Electronic Press.

Hidi, Suzanne 2001. Interest, Reading, and Learning: Theoretical and Practical Considerations. *Educational Psychology Review* 13, no. 3.

Hirsh, David and Paul Nation 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language* 8 (2): 689–696.

Hodges, J.R., N. Graham and K. Patterson 1995. Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory* 3: 463–495.

Honoré, A. 1979. Some simple measures of richness of vocabulary. *Association of Literary and Linguistic Computing Bulletin*, pp. 172–179.

Hoover, Wesley A. and Philip B. Gough 1990. The simple view of reading. *Reading and Writing: An Interdisciplinary Journal* 2: 127–160.

Hultman, T. G. and M. Westman 1977. *Gymnasistsvenska*. Lund: Liber Läromedel.

Hunston, Susan 2002. *Corpora in applied linguistics*. The Cambridge applied linguistics series. Cambridge: Cambridge University Press.

Ihaka96 1996. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314.

Inui, Kentaro, Atsushi Fujita, Tetsuro Takahashi, Ryu Iida and Tomoya Iwakura 2003. Text simplification for reading assistance: a project note. *Proceedings of the second international workshop on paraphrasing - volume 16*, PARAPHRASE '03, 9–16. Stroudsburg, PA, USA: Association for Computational Linguistics.

Joachims, Thorsten 1998. Text categorization with support vector machines: Learning with many relevant features. *Lecture Notes in Computer Science (ECML -98* 1398: 137–42.

Johansson, Stig, G. Leech and H. Goodluck 1978. Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers. Technical Report, Unversity of Oslo.

Johansson Kokkinakis, Sofie 2002. En studie över påverkande faktorer i ordklasstaggning. Baserad på taggning av svensk text med EPOS. Ph.D. diss., Göteborgs universitet.

Johansson Kokkinakis, Sofie and Elena Volodina 2011. Corpus-based approaches for the creation of a frequency based vocabulary list in the EU project Kelly. Issues on reliability, validity and coverage. *eLex Conference*. Slovenia.

Jorm, Anthony F. and David L. Share 1983. An invited article: Phonological recoding and reading acquisition. *Applied Psycholinguistics* 4 (02): 103–147.

Juel, C 1988. Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology* 80: 437–447.

Juhasz, Barbara J. and Keith Rayner 2003. Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading. *Journal of Experimental Psychology. Learning, Memory & Cognition* 29 (6): 1312–1318.

Just, Marcel Adam and Patricia A. Carpenter 1980. A theory of reading: from eye fixations to comprehension. *Psychological Review* 87 (4): 329–354.

Källgren, Gunnel 1998. Documentation of the Stockholm-Umeå Corpus.

Kincaid, J. P., R. P. Fishburne, R. L. Rogers and B.S. Chissom 1975. Derivation of new readability formulas: (Automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel. Technical Report, Millington, Tenn.

Kintsch, Walter and Janice Keenan 1973. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology* 5: 257–274.

Klare, George R. 1963. *The measurement of readability*. Iowa State University Press.

Korkman, Christina 1995. *Tvåspråkighet och skriftlig framställning. en undersökning av tvåspråkiga elevers uppsatser i den finlandssvenska grundskolan*. Helsingfors: Skrifter utgivna av Svenska litteratursällskapet i Finland.

Lewerentz, A.S 1929. Measurement of difficulty of reading materials. *Los Angeles Educational Research Bulletin* 8: 11–16.

Liu, Haitao 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science* 9 (2): 159–191.

Lively, B.A. and S.L. Pressey 1923. A method of measuring vocabulary burden of textbooks. *Educational Administration and Supervision* 9: 389–398.

Lorge, Irving 1944. Predicting readability. *Teachers College Record* XLV: 404–419.

Lundberg, Ingvar and Monica Reichenberg 2008. *Vad är lättläst?* Specialpedagogiska skolmyndigheten.

Marcus, Mitchell, Beatrice Santorini and Mary Ann Marcinkiewicz 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19 (2): 313–330.

McCarthy, Philip M. and Scott Jarvis 2007. vocd: A theoretical and empirical evaluation. *Language Testing* 24 (4): 459–488.

McEnery, Tony, Richard Xiao and Yukio Tono 2006. *Corpus-based language studies: an avanced resource book*. London: Routledge Applied Linguistics.

McKee, Gerard, David Malvern and Brian Richards 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing* 15 (3): 323–337.

McLaughlin, G. Harry 1969. SMOG grading – a New Readability Formula. *Journal of Reading* 12 (8): 639–646.

Melin, Lars and S. Lange 2000. *Att analysera text. Stilanalys med exempel*. Lund: Studentlitteratur.

Miller, G. A. and N. Chomsky 1963. Chapter 13: Finitary Models of Language Users of . R. Luce, R. Bush and E. Galanter (eds), *Handbook of Mathematical Psychology*. New York: Wiley.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11): 39–41.

Miltsakaki, Eleni 2009. Matching Readers' Preferences and Reading Skills with Appropriate Web Texts. *Proceedings of the EACL 2009 Demonstration Session*, 49–52. Athens, Greece.

Miltsakaki, Eleni and Audrey Troutt 2008. Real-time Web Text Classification and Analysis of Reading Difficulty. *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications*, 89–97. Columbus, Ohio.

Mohammad, Saif and Peter Turney 2011. Crowd-sourcing a word-emotion association lexicon. *Computational Intelligence*, vol. 59.

Morriss, E.C. and D. Holversen 1938. Idea analysis technique.

Mühlenbock, Katarina and Sofie Johansson Kokkina-kis 2009. LIX 68 revisited – An extended readability measure. *Proceedings of Corpus Linguistics 2009*. Liverpool.

Murphy, G.L. and M.E. Lassaline 1997. Hierarchical structure in concepts and the basic level of categorization. K. Lamberts and D. Shanks (eds), *Knowledge, concepts and categories*, 93–132. Hove, East Sussex, U.K.: Psychology Press.

Nagy, William E. and Judith A Scott 2000. Chapter 18 of . P. David Pearson (ed.), *Vocabulary processes*, 269–284. Lawrence Erlbaum Associates.

New, Boris, L. Ferrand, C Pallier and Marc Brysbaert 2006. Reexamining word length effects in visual word recognition: New evidence from the English Lexicon Project. *Psychonomic Bulletin and Review* 13: 45–52. 10.3758/BF03193811.

Nivre, Joakim, , Jens Nilsson and Johan Hall 2006. Talbanken05: A Swedish Treebank with Phrase Structure and Dependency annotation. *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC2006)*. Genoa, Italy.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Eryt Gül sen, Sandra Kübler, Svetoslav Marinov and Erwin Marsi 2007. Malt-

Parser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13 (2): 95–135.

Norén, Kerstin 1996. Svenska partikelverbs semantik. Ph.D. diss., Göteborgs universitet.

Nyström, Mikael, Magnus Merkel, Lars Ahrenberg, Pierre Zweigenbaum, Håkan Petersson and Hans Åhlfeldt 2006. Creating a medical English-Swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making* 6, no. 35.

Ogden, C. K. 1930. *Basic English: A general introduction with rules and grammar.* London: Paul Treber & Co. Ltd.

Patty, W.W. and W.I. Painter 1931. Improving our method of selecting high-school textbooks. *Journal of Educational Research* XXIV (June): 23–32.

Perea, Manuel and Eva Rosa 2000. The effects of orthographic neighborhood in reading and laboratory word identification tasks: A review. *Psicológica* 21: 327–340.

Petersen, Sarah E. and Mari Ostendorf 2009. A machine learning approach to reading level assessment. *Computer Speech & Language* 23 (1): 89–106.

Pitler, Emily and Ani Nenkova 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality. *2008 Conference on Empirical Methods in Natural Language Processing*, 186–195. Honolulu.

Platt, John C. 1999. Using analytic QP and sparseness to speed training of support vector machines. M.S. Kearns, S.A. Solla and D.A. Cohn (eds), *Advances in neural information systems*, Volume 11. MIT Press.

Platzack, Christer 1974. *Språket och läsbarheten: en studie i samspelet mellan läsare och text*. Skrifter utgivna av Svensklärarföreningen. Lund: Gleerup.

Proverbio, Alice Mado, Serena Mariani, Alberto Zani and Roberta Adorni 2009. How are 'Barack Obama' and 'President Elect' differentially stored in the brain? An ERP investigation on the processing of proper and common noun pairs. *PLoS ONE*, vol. 4 (September).

Rayner, Keith 1998. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin* 124 (3): 372–422.

Reichenberg, Monica 2000. Röst och kausalitet i lärobokstexter. en studie av elevers förståelse av olika textversioner. Ph.D. diss., Department of Education, Göteborg University, Sweden.

Reichle, Erik D. and Charles A. Perfetti 2003. Morphology in word identification: A word-experience model that accounts for morpheme frequency effects. *Scientific Studies of Reading* 7 (3): 219–237.

Richards, Todd L. 2001. Functional Magnetic Resonance Imaging and Spectroscopic Imaging of the Brain: Application of fMRI and fMRS to Reading Disabilities and Education. *Learning Disability Quarterly* 24 (3): pp. 189–203.

Roark, Brian, Margaret Mitchell and Kristy Hollingshead 2007. Syntactic complexity measures for detecting mild cognitive impairment. *BioNLP 2007: Biological, translational and clinical language processing*, 1–8. Association for Computational Linguistics, Prague.

Rogers, T.T and J.L. McClelland 2004. *Semantic cognition – a parallel distributed processing approach*. Cambridge, Mass.: The MIT Press.

Rondal, Jean A. and Susan Edwards 1997. *Language in mental retardation*. London: Whurr Publishers Ltd.

Rosch, Eleanor 1978. Principles of categorization. E. Rosch and B.B. Lloyd (eds), *Cognition and categorization*. New York: Erlbaum.

Sampson, G. 1993. The need for grammatical stocktaking. *Literary and Linguistic Computing* 8 (4): 267–273.

Sandberg, Karin, Sara Spånning-Westerlund and Karin Wejderot 2005. Prova din text – en rapport om textanpassning och utprovning för olika målgrupper. Technical Report, Centrum för lättläst i samarbete med Institutionen för Innovation, Design och Produktutveckling vid Mälardalens högskola i Eskilstuna.

SAOL 2006. Svenska Akademiens Ordlista över svenska språket. Norstedts.

Schank, Roger 1979. Interestingness: Controlling inferences. *Artificial Intelligence* 12: 379–390.

Schwarm, Sarah E. and Mari Ostendorf 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, 523–530. Stroudsburg, PA, USA: Association for Computational Linguistics.

Senter, R.J. and E. A. Smith 1967. Automated readability index. Technical Report, Cincinnati Univ. Ohio.

Siddharthan, Advaith 2006. Syntactic simplification and text cohesion. *Research on Language & Computation* 4 (1): 77–109.

Siegel, Linda S. 1989. IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities* 22: 469–486.

Sinclair, John M. 1991. The automatic analysis of corpora. Jan Svartvik (ed.), *Directions in Corpus Linguistics. Proceedings of Nobel Symposium* , Volume 82, 379–400. Stockholm.

Sinclair, John M. 2000. Preface. Mohsen Ghadessy, Robert L. Roseberry and Alex Henry (eds), *The use of small corpora in the teaching of language*. Amsterdam: John Benjamins.

Sjöholm, Johan 2012. Probability as readability. Master's thesis, Department of Computer and Information Science, Linköping University.

Socialstyrelsen 2003. Klassifikation av funktionstillstånd, funktionshinder och hälsa.

Sorvali, Irma 1984. *Tentamenssvenska*. Helsingfors: Gaudeamus.

Spache, George March, 1953. A new readability formula for primary-grade reading materials. *Elementary School Journal* LIII: 410–413.

Sperberg-McQueen, C.M. and Lou Burnard 1994. *Guidelines for text encoding and interchange (TEI P3)*.

Språkbanken, Göteborgs universitet 2000. The Swedish PAROLE Lexicon. A language engineering resource with access to morphological and syntactic information in Swedish. "`http://spraakdata.gu.se/parole/lexikon/swedish.parole.lexikon.html`".

Stone, Clarence 1938. Measures of simplicity and beginning texts in reading. *Journal of Educational Research* XXXI (February): 447–50.

Svensén, Bo 2004. *Handbok i lexikografi. Ordböcker och ordboksarbete i teori och praktik*. 2nd Edition. Stockholm: Norstedts Akademiska Förlag.

Taft, Marcus 1979. Recognition of affixed words and the word frequency effect. *Memory and Cognition*.

Tan, Pan-Ning, Michael Steinbach and Vipin Kumar 2006. *Introduction to data mining*. Pearson International Edition (ed.). Addison Wesley.

TEI Consortium 2007. TEI P5: Guidelines for Electronic Text Encoding and Interchange.

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

Temperley, David 2007. Minimization of dependency length in written english. *Cognition* 105 (2): 300–333.

The National Agency for Education 2000. The Foundation for Lifelong Learning. A comparative international study of adult skills in

understanding and using printed and written information. Report No. 188. Technical Report.

Thorndike, Edward L. 1917. Reading as reasoning: A study of mistakes in paragraph reading. *Journal of Educational Psychology* 8 (6): 323–332.

Thorndike, Edward L. 1921. *The teacher's word book*. New York: Teacher's College, Columbia University.

Thorndike, Edward L. and I. Lorge 1944. *The teacher's word book of 30,000 words*. New York: Columbia University Press.

Tunmer, William E. and Wesley A. Hoover 1992. Philip B. Gough, Linnea C. Ehri and Rebecca Treiman (eds), *Cognitive and linguistic factors in learning to read*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Turner, A. and E. Greene 1977. *The construction and use of a propositional text base*. Tech. Rep. No. 63. University of Colorado, Institute for the Study of Intellectual Behavior.

Viberg, Åke, K Lindmark, A Lindvall and I Mellenius 2002. The Swedish WordNet project. *Proceedings of Euralex 2002*, 407–412. Copenhagen University.

Vogel, M. and C. Washburne 1928. An objective method of determining grade placement of children's reading material. *Elementary School Journal* 28: 373–381.

World Health Organization 2001. International Classification of functioning, disability and health (ICF).

World Health Organization 2008. International statistical classification of diseases and related health problems. Technical Report, World Health Organization, New York, NY.

World Health Organization 2011. Mental Retardation from knowledge to action. "`http://www.searo.who.int/en/Section1174/Section1199/Section1567/Section1825.htm`". "[Online; accessed 25-Aug-2011]".

Wren, Sebastian 2001. The cognitive foundations of learning to read. Technical Report.

Yap, Melvin J. and David A. Balota 2009. Visual word recognition of multisyllabic words. *Journal of Memory and Language* 60: 502–529.

Yates, Mark, John Friend and Danielle M. Ploetz 2008. The effect of phonological neighborhood density on eye movements during reading. *Cognition* 107: 685–692.

Yildirim, Kasim, Mustafa Yildiz and Seyit Ates 2011. Is vocabulary a strong variable predicting reading comprehension and does the prediction degree of vocabulary vary according to text types. *KURAM VE UYGULAMADA EGITIM BILIMLERI* 11 (3): 1541–1547 (SUM).

Yngve, V. 1960. A model and a hypothesis for language structure. *Proceedings of the American Philosophical Society*, Volume 104, 444–466.

Yoakam, Gerald A. 1939. A technique for determining the difficulty of reading materials. Unpublished study.

Yu, Bei 2008. An evaluation of text classification methods for literary study. *Literary and Linguistic Computing* 23 (3): 327–343.

Zaidel, Eran, Andrew Hill and Scott Weems 2008. EEG correlates of hemispheric word recognition. Zvia Breznitz (ed.), *Brain Research in Language*, Volume 1, 225–245. College Station, USA: Springer.

Zevin, Jason D. and Mark. S. Seidenberg 2002. Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language* 47: 1–29.

Zipf, George K. 1932. *Selective studies and the principle of relative frequency in language*. Cambridge, Mass.

Zipf, George K. 1949. *Human behavior and the principle of least-effort*. Cambridge, Mass.: Addison-Wesley, 1965.

Appendices

A Composition of the LäsBarT corpus

Author/ title	Publishing year	Ages	No. tokens	No. words
Gull Åkerblom:				
Inez värsta vecka	2006	6-9	11,847	10,305
Inez långa väg till mormor	2006	6-9	10,324	8,860
Trolleri, Inez	2006	6-9	6,744	5,782
Inez och spökhuset	2005	6-9	9,371	8,073
Ulf Stark:				
Min vän Percy, Buffalo Bill och jag	2004	9-12	49,051	42,016
Märklin och Turbin	2005	9-12	15,494	13,198
Kan du vissla Johanna	2003	9-12	4,660	3,955
När jag besökte himlen	2003	6-9	2,431	2,057
Fullt med flugor i klassen	2003	6-9	1,677	1,434
Helena Bross:				
Hemliga kompisar	2006	6-9	2,097	1,783
En rejäl ryggsäck	2006	6-9	1,683	1,398
Spöket i skolan	2008	6-9	2,208	1,867
Felix	2004	6-9	1,506	1,208
Jan Lööf:				
Pelle träffar en lokförare	2004	6-9	547	452
Annika Thor:				
Pirr i magen och klump i halsen	2003	6-9	18,970	16,432
Rött hjärta blå fjäril	2002	6-9	20,404	17,539
Petter Lidbeck:				
När Johan vaknar en morgon är han stark	2005	6-9	1,456	1,226
En dag i prinsessan Victorias liv	2004	6-9	3,525	2,890
Vår Vinni	2002	6-9	14,851	12,626
Ulf Nilsson:				
Den döde talar	2006	9-12	9,063	7,643
En halv tusenlapp	2005	9-12	8,014	6,787
Pia Hagmar:				
Drömponnyn	2003	9-12	37,247	32,904
Vänner	2003	9-12	37,147	32,516
Drömmen om en häst	2001	9-12	39,411	34,711
Elsie Johansson:				
Mormorsmysteriet	2004	9-12	50,876	43,308
Veronica Wägner:				
Tillträde förbjudet	2005	9-12	29,656	26,069
Kråköns hemlighet	2004	9-12	31,192	27,424
Total			421,452	364,463

Table A.1: Childrens fiction – ordinary literature from Bonnier Carlsen Publishing Company

Author/title	Publishing year	Ages	No. tokens	No. words
Thomas Dömstedt:				
Bellas drömmar	2007	9-12	12,732	10,700
Glenn Ringtved:				
Drömlaget 1 – mot nya mål	2007	9-12	11,045	9,497
Drömlaget 2 – offside	2007	9-12	10,591	9,182
Drömlaget 3 – tuffa tag	2007	9-12	10,473	8,995
Torsten Bengtsson:				
Galne Hugo och de två flygarna	2007	11-14	14,921	12,850
Kirsten Sonne Harlid:				
I full galopp	2004	9-12	13,544	11,382
Ponny & Co	2007	9-12	13,306	11,322
Uppdrag Maja	2007	9-12	12,242	10,378
Ingrid Mühlow:				
Lisa på landet	2007	9-12	6,028	5,151
Åsa Storck:				
Milos flykt	2008	9-12	5,395	4,690
Vikarien	2008	14-	2,942	2,505
Bente Bratlund:				
Skolresan	2008	9-12	3,098	2,615
Eva Christina Johansson:				
Tornet föll	2007	10-14	11,409	9,725
Gunnar Åberg:				
Varning! Livsfara!	2007	11-	8,625	7,390
Jan-Olof Ekholm:				
MVG, grabben	2006	10-	7,438	6,091
Total			143,789	122,473

Table A.2: Childrens fiction – easy-to-read literature from Hegas Publishing Company

Author/title	**Publ year**	**No. tokens**	**No. words**
Maj Sjöwall & Per Wahlöö:			
Mannen på balkongen	2002	8,061	6,993
Roseanna	2000	8,034	6,991
Håkan Nesser:			
Det fruktansvärda	2001	8,316	7,209
Ove Magnusson:			
Vittnet	2001	3,987	3,462
Matteo Bandello:			
Romeo och Julia	2001	5,676	5,104
Marianne Fredriksson:			
Simon och ekarna	2003	12,996	11,341
Katarina Runeson:			
Dubbelspel i Barcelona	2003	5,817	5,089
Charles Dickens:			
Spöket i skåpet och andra hemska historier	2004	6,394	5,527
PC Jersild:			
Barnens ö	2000	13,481	11,466
Selma Lagerlöf:			
Kejsarn av Portugallien	2001	11,391	9,894
Tösen från Stormyrtorpet	1999	2,975	2,553
Per Anders Fogelström:			
Mina drömmars stad	1998	9,838	8,504
I en förvandlad stad	1999	9,391	8,208
Reidar Jönsson:			
Mitt liv som hund	2002	9,413	8,180
Total		115,770	100,521

Table A.3: Adults' fiction – easy-to-read literature from Lättlästförlaget

Source	**Category Category**	**Publ. year**	**No. tokens**	**No. words**
8 sidor	Written general	Dec. 2003 – Jan. 2004	20,745	18,696
8 sidor	Written general	2006	178,637	160,984
8 sidor	Written general	2007	158,062	142,980
Invandrar-tidningen	Written for immigrants	1997	6,585	5,701
Klartext	Written to be read	2004	26,773	24,612
Total			390,802	352,973

Table A.4: The news genre. Sources, publishing year and sample size

Source	**Access date**	**No. tokens**	**No. words**
Skellefteå	2003-12-08	1,701	1,561
Kungsör	2003-12-14	697	631
Falköping	2003-12-13	1,025	885
Stockholm	2003-12-08	9,846	8,944
Lidköping	2003-12-13	3,389	3,071
Örebro	2003-12-13	541	482
Åre	2003-12-08	322	293
Töreboda	2003-12-13	58	54
Mariestad	2003-12-13	558	503
Karlstad	2007-03-03	2,268	2,054
Total		20,405	18,478

Table A.5: The information genre. Municipality information per access date and sample size

Source sw	Source en	Access date	No. tokens	No. words
Livsmedelsv	The National Food adm.	2003-12-14	11,208	9,938
Tullverket	Swedish Customs	2003-12-14	3,024	2,742
Sveriges Riksd.	Swedish Parliament	2003-11-12	9,135	8,357
FN	UN	2006-01-20	3,965	3,629
Centrum för Lättläst	Centre for Easy-to-read	2006-08-25	3,371	3,051
Socialdep	Ministry of Health and Social Affairs	2004	2,294	2,114
Socialdep	Ministry of Health and Social Affairs	2006-01-19	15,105	13,765
Justitiedep	Ministry of Justice	2006-01-19	12,332	11,271
Justitiedep	Ministry of Justice	2006-08-08	4,026	3,661
Regeringskansl	Government offices of Sweden	2006-01-19	10,074	9,092
Regeringskansl	Government offices of Sweden	2006-08-04	5,809	5,187
Jordbruksdep	Ministry of Agriculture	2004	3,070	2,849
EU-upplysn	EC information	2007-04-09	2,899	2,620
Total			86,312	78,276

Table A.6: The information genre. Government and parliament information per access date and sample size

Source sw	Source en	Access date	No. tokens	No. words
Västra Götalandsreg	Region Västra Götaland	2007-02-10	48,059	43,522
Örebro Landsting	Örebro County Council	2003-12-08	2,018	1,782
Kalmar Landsting	Kalmar County Council	2003-12-08	2,512	2,264
Länsstyr Gotland	County Council Gotland	2007-04-09	1,369	1,260
Total			53,958	48,828

Table A.7: The information genre. County council information per access date and sample size

Source sw	**Source en**	**Access date**	**No. tok**	**No. words**
Sisus	Ped Resources Liberal Adult Educ Net	2003-12-14	793	722
BO	Children's Ombudsm	2003-12-14	1,229	1,139
BO	Children's Ombudsm	2007-03-12	3,402	3,120
HO	Disability Ombudsm	2003-12-14	1,381	1,234
Socialstyr	Nat Board of Health and Welfare	2003-12-14	1,029	938
CSN	Financial aid for studies and housing	2006-06-07	6,093	5,586
F-kassan	Social insurance auth	2006-08-08	8,131	7,453
Polisen	Swedish Police	2006-06-25	1,865	1,729
Rikspolisstyr	Nat Police Board	2006	394	369
Banverket	Swedish Rail Admin	2003-07-12	251	223
Allm Rekl.nämnden	Swedish Nat Board for Consumer Compl	2007-03-12	1,017	917
Boverket	Nat Board of Housing, Building and Planning	2007-03-12	6,412	5,910
Domstolsverket	Nat Board of Swedish Courts	2007-03-12	682	609
Krisberedskaps-myndigheten	Swedish Civil Contingencies Agency	2007-03-12	1,064	967
Läkemedelsv	Medical Prod Agency	2007-03-12	941	872
SKI	Swedish Radiation Safety Authority	2007-03-12	1,385	1,238
Vägverket	The Swedish Road Administration	2007-03-12	2,977	2,710
Stockholmsförs	Stockholm trials	2007-04-09	900	828
PTS	Swedish Post and Telecom Agency	2007-04-09	1,884	1,723
Konkurrensv	Swedish Competition Authority	2007-04-09	228	204
JO	Parliam Ombudsm	2007-04-09	410	377
Fiskeriv	Swedish Board of Fisheries	2007-04-09	983	891
DI	Swedish Data inspection board	2007-04-09	1,926	1,779
Åklagarmyndigh	Swedish Prosecution authority	2007-04-09	2,141	1,930
Total			47,518	43,468

Table A.8: The information genre. Public authorities' information per access date and sample size

Source sw	**Source en**	**Access date**	**No. tokens**	**No. words**
HSO	Swedish Disability Federation	2006-06-22	4,565	4,196
Svenska kyrkan	Church of Sweden	2006-03-12	4,728	4,227
Biblioteken i Göteborg	Public libraries in Göteborg	2006-08-04	351	322
Riksteatern	Sweden's nat theatre	2007-03-03	935	845
Storstockholms lokaltrafik	Stockholm transport	2007-04-09	3,465	3,147
Total			14,044	12,737

Table A.9: The information genre. Miscellaneous information per access date and sample size

B CORPUS EXAMPLES

B.1 Children's ETR fiction (CEF) text

Source: Bellas drömmar by Thomas Dömstedt. Adapted for children 9–12 years. Published by Hegas.
Distribution: LäSBarT corpus, children's easy fiction part.

Det är Tove på gitarr, Amanda på bas och Isak som sjunger. Och så är det jag, Bella. Som spelar trummor. Det är vi som är Bellas Band. Som är världens bästa. I alla fall i mina drömmar. Repar gör vi så ofta vi hinner. När vi har lust. Inte på bestämda tider. Fast det där har vi lite olika åsikter om. Det är mer rock att spela när man känner för det, brukar Tove säga. Risken är att det blir mycket snack och lite rock, brukar jag invända. För det blir mycket diskuterande och onödigt tjafs. Svårt att hitta tider när alla kan. Det är mest Tove som inte har tid. Hon har så mycket annat. Spelar fotboll. Är med i en kör. Och så har hon en massa coola kompisar som hon visst bara måste vara med. Och allt går tydligen före bandet. Isak har förresten också börjat balla ur. Det är lite skumt. Ibland bara kommer han inte. Ibland säger han att han inte har tid. Men han talar aldrig om varför. Bellas Band började som ett tjejband. Sedan kom Isak. Kom in i bandet, kom in i mitt liv. Men jag undrar vart han är på väg nu. Kanske håller han på att tröttna.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	208	6.93	3.90	1.45	0	54

Table B.1: Superficial features, chunk CEF404

LVIX	SweVoc	SV C	SV D	SV H	SV K	SV S
45	89.7	77.6	1.9	3.7	0.9	5.6

Table B.2: Features indicating vocabulary load, chunk CEF404

MDD	UA	AT	ET	PT
2.06	5	4	3	4.2

Table B.3: Features indicating sentence structure, chunk CEF404

Pr	NoPr	NR	Sem. depth	PM	LIX
47.5	0.9	0.39	3.13	0.05	14.6

Table B.4: Features indicating idea density, human interest, and LIX value, chunk CEF404

B.2 Children's ordinary fiction (COF) text

Source: Kan du vissla, Johanna by Ulf Stark. Adapted for children 9–12 years. Bonnier Carlsen förlag.
Distribution: LäSBarT corpus, children's ordinary fiction part.

En eftermiddag när Berra och jag gungar på vår hemmagjorda gungbräda säger jag att jag ska till min morfar. Jag ska dit och äta tårta. För att han fyller år den kvällen.
– Då får jag fem kronor, säger jag.
– Får du pengar när han fyller år? säger Berra förundrat.
– Ja, säger jag. Det får jag varje gång som jag träffar honom.
– Oj då. Han är visst snäll han? säger Berra.
– Ja, nickar jag. Och han ska få en stor cigarr av mej.
Då kikar Berra längtansfullt upp mot molnen.
– Jag skulle också vilja ha en morfar, mumlar han. Vad gör såna egentligen?
– Tja, dom bjuder en på kaffe, säger jag. Och så äter dom grisfötter.
– Nu skojar du, va? säger Berra.
– Nä, det är säkert, säger jag. Grisfötter i gelé. Och ibland så tar dom en till en sjö och metar fisk.
– Varför har inte jag en morfar? undrar Berra.
– Det vet jag inte, svarar jag. Men jag vet i alla fall var du kan få tag på en.
– Var då? säger Berra.
– Det får du se i morgon, säger jag. För nu måste jag in och sätta på mej en vit skjorta och kamma håret.
När jag kliver av gungbrädan så åker Berra ner och slår i sin haka.
Nästa dag tar jag med mig Berra. Då har han tvättat sig. Han har ett rent plåster på hakan och i handen håller han en ringblomma som han hittat i Gustavssons trädgård.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	242	8.07	3.87	1.40	0	51

Table B.5: Superficial features, chunk COF156

LVIX	SweVoc	SV C	SV D	SV H	SV K	SV S
45	87.7	77.7	0.1	4.5	1.8	3.6

Table B.6: Features indicating vocabulary load, chunk COF156

MDD	UA	AT	ET	PT
2.15	5	5	7	4.1

Table B.7: Features indicating sentence structure, chunk COF156

Pr	NoPr	NR	Sem. depth	PM	LIX
40.1	0.6	0.38	3.02	0.04	15.1

Table B.8: Features indicating idea density, human interest, and LIX value, chunk COF156

B.3 Adults' ETR fiction (AEF) text

Source: Simon och ekarna by Marianne Fredriksson. LL-förlaget.
Distribution: LäSBarT corpus, adults' ETR fiction part.

De trivdes bra ihop, de sjöng och snickrade. Isak trivdes också med Karin. Hon var alltid snäll och såg till att han hade det bra. Karin gav honom allt det som han inte hade fått av sin mor. Simon blev svartsjuk. Simon och Erik hade aldrig så roligt tillsammans. Och när Karin daltade med Isak blev Simon arg. Han tyckte att föräldrarna bara såg hans vän. Simon rusade till sina ekar. När han var liten brukade träden trösta honom, men nu stod de tysta.

- Jag ska ta livet av Isak! skrek Simon.

Sedan skämdes han. Isak var hans bästa vän. Det var synd om honom.

- Jag ska ta livet av mig! skrek Simon.

Då skulle han få Karin och Erik att gråta. Men Simon ville leva. Sedan skämdes Simon igen. Han var ett lyckligt barn. Hur kunde han tänka så? Karin skulle bli full av sorg om hon kunde läsa Simons tankar. Då blev Simon arg igen. Han hatade sin mamma! Nej, det kunde han inte, mamma var god. Då tänkte Simon på Eriks kusin, Inga. Ibland tvingade föräldrarna med honom till hennes torp, men han gillade inte stället. Dessutom var Inga fet, och hon vågade aldrig se Simon i ögonen.

- Hon är en subba! skrek han till ekarna.

Simon förstod inte varför han hatade Inga. Inga hade inte gjort honom något illa.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	223	7.43	4.20	1.55	0	53

Table B.9: Superficial features, chunk AEF123

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
44	88.4	73.2	0.9	4.5	0.9	8.9

Table B.10: Features indicating vocabulary load, chunk AEF123

MDD	UA	AT	ET	PT
2.06	4	2	4	4.1

Table B.11: Features indicating sentence structure, chunk AEF123

Pr	NoPr	NR	Sem. depth	PM	LIX
46.6	0.6	0.29	3.20	0.11	15.1

Table B.12: Features indicating idea density, human interest, and LIX value, chunk AEF123

B.4 Adults' ordinary fiction (AOF) text

Source: Rosalie by Rita Tornborg. Albert Bonniers förlag.
Distribution: SUC KK.

När svullnaden lagt sig, när hon åter kunde se och röra på läpparna, blundade hon och teg. Genia och hennes väninnor hjälptes åt att vaka över mor dygnet runt, de var rädda att hon annars skulle ta sitt liv. De hade fel, men bara till hälften.

Varje gång hon gick på toaletten fylldes skålen med hennes blod. Inga män tilläts komma in. Mannen som hittade henne, Zawadzki Piotr, presenterade han sig militäriskt, stannade självmant i farstun, när han kom på sina dagliga visiter för att förhöra sig om mors tillstånd, och då talade han genom sin basker. Snart visade det sig att hon hade gonorré och var gravid. Den stränga gynekologen föreslog att hon själv skulle skrapa henne ren från sådden hemma, men mor bara skakade på huvudet, lyssnade inte till några argument, gav inga skäl, förklarade sig inte. En gång såg Genia hur hon förundrad strök över sin putande mage. Herr Zawadzki kom och sade genom baskern, han hade fått för vana att kommunicera med Genia på det viset, att han kunde ha misstagit sig, att de kanske bara varit tre. Han menade det som tröst. Hon skulle till en väninna med vårt familjealbum på kvällen. Väninnan var scenograf och sökte inspiration till en Tjechovpjäs. De hade hållit på med mor hela natten. När de fått nog slängde de ut henne och körde därifrån. Genia kom aldrig på tanken att vända sig till milisen. Albumet kom aldrig till rätta. De hade tagit med sig minnen av våra minnen och lämnade en oäkting efter sig. Mig, med fyra eller kanske bara tre pappor. Tanken på barnet med fyra eller kanske bara tre fäder fick Genia att för första och sista gången i sitt liv dunka huvudet i väggen. Herr Zawadzki har berättat det för mig. Genom baskern. Jag har träffat honom flera gånger. Det är en sak att ta en kvinna med våld, klagade Genias väninnor, men varför misshandla henne också? Ingen av dem hade blivit våldtagen. Ingen av dem förstod att det var nödvändigt. Barnet växte ohejdbart. Jag rörde mig ogärna, växte och växte men sparkade inte. Genias förtvivlan växte i kapp med mig, tills hon en salig dag såg mor leende, hur hon leende med båda händerna lyfte sin stora mage. Lyfte bördan som fyra eller kanske bara tre män lämnat efter sig, och hur hon böjde sitt huvud för att lyssna till barnet. Det gick ju inte, men sen den dagen började Genia studera alla tillgängliga almanackor och göra upp listor på passande flicknamn. Hon var övertygad om att det skulle bli en flicka.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	427	14.2	4.50	1.66	0	62

Table B.13: Superficial features, chunk AOF81

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
52	83.5	74.5	1.15	2.3	0.0	5.5

Table B.14: Features indicating vocabulary load, chunk AOF81

MDD	**UA**	**AT**	**ET**	**PT**
2.3	9	8	18	5.4

Table B.15: Features indicating sentence structure, chunk AOF81

Pr	**NoPr**	**NR**	**Sem. depth**	**PM**	**LIX**
42.0	1.16	0.70	3.71	0.03	29.9

Table B.16: Features indicating idea density, human interest, and LIX value, chunk AOF81

B.5 ETR news (EN) text

Source: 8 SIDOR from Oct 27, 2003.
Distribution: LäSBarT corpus, ETR news part.

Det har sprängts flera bomber i Irak den senaste tiden.
Idag dödades sex människor av en bomb.
Det hände i staden Falludja i som ligger i närheten av huvudstaden Bagdad.
Bomben sprängdes av en person i en bil.
Även han som sprängde bomben dog.
Idag har det också blivit känt att den som tillfälligt var ledare för Bagdad sköts till döds i Söndags.
På ett annat ställe i Irak sprängdes också en bomb idag.
Tre människor skadades i den attacken.

Varje höst är det björnjakt i Sverige.
Då är det tillåtet att döda björnar.
Däremot är det förbjudet att skjuta björnhonor som har ungar.
Ändå är det flera jägare som gör det.
5 björnhonor har blivit dödade i höst trots att de hade ungar.
De jägare som dödar en björnhona som har ungar bryter mot lagen.
Och kan få fängelse som straff.
Men det är få jägare som får sitta i fängelse för att de skjutit en björnhona med ungar.
För ofta är det svårt för poliserna att bevisa att jägaren visste om att honan hade ungar när han sköt.
Och då får jägaren behålla björnen.
Nu vill myndigheten Naturvårdsverket ändra på lagen.
De vill att poliserna ska kunna ta den skjutna björnen från jägaren.
Då skulle färre björnhonor med ungar skjutas, tror Naturvårdsverket.

Flera tusen människor har flytt bort från elden.
Det är mest i skogen det brinner.
Men många hus har också brunnit ner.
Och minst 13 människor har dött.
Bränderna är i närheten av den stora stan Los Angeles.
Mer än 7000 brandmän försöker stoppa bränderna.

För ett tag sedan krigade USA och några andra länder mot landet Irak i Mellanöstern.

USA sa att kriget tog slut i maj.
Men sedan dess har det ändå varit väldigt oroligt i Irak.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	293	9.77	4.47	1.65	2	58

Table B.17: Superficial features, chunk ENS45

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
44	86.6	68.1	0.0	5.9	0.0	12.6

Table B.18: Features indicating vocabulary load, chunk EN45

MDD	**UA**	**AT**	**ET**	**PT**
2.03	8	4	12	5.2

Table B.19: Features indicating sentence structure, chunk EN45

Pr	**NoPr**	**NR**	**Sem. depth**	**PM**	**LIX**
46.4	3.53	0.95	3.54	0.04	28.9

Table B.20: Features indicating idea density, human interest, and LIX value, chunk EN404

B.6 Ordinary news (ON) text from SUC

Source: Svenska Dagbladet from June 5, 1990.
Distribution: SUC 2.0 part A - Press reports.

4-0 är klart missvisande och hade Tyresö haft 1-1 då knappt tio minuter återstod hade säkert resultatet stått sig tiden ut. Nu flyttade södergänget upp sin ene mittback i slutet i hopp om kvittering. Men den åtgärden straffade sig snabbt. Vasalund fick ytor för blixtsnabba kontringar och det utnyttjades på ett imponerande sätt.
- Det här var en av våra bästa matcher och resultatet speglar inte händelserna på plan, suckade Tyresös tränare Tommy Davidsson. Det verkar löjligt med tanke på resultatet, men vi var inte långt ifrån en poäng.
Kollegan Bosse Petersson:
– Det här var en mycket stark seger. Vi utnyttjade motståndarens svaghet på kontringar. Tre av de fyra hemmamålen kom på blixtsnabba attacker – det fjärde var en donation av gästernas målvakt som totalt missbedömde en frispark.
Vasalunds spel dominerades av två ex-allsvenskar:
Thomas Bergman regisserade på ett underbart sätt mittfältsspelet och låg bl a bakom tre av målen. Mästerliga passningar i rätt ögonblick präglade Bergmans spel. Peter Gerhardsson visade vägen med ett proffsigt 1-0-mål och svarade sedan för en målgivande passning. 2,5 år i division I har inte satt några spår – han är fortfarande Stockholms kvickaste och mest explosive spelare. "Bobban" kommer.
I nästa omgång gör Slobodan Krcmarevic efterlängtad årsdebut för Vasalund. Blir det månne revansch på Grimsta för 0-3 i fjol eller etablerar sig också BP som ett topplag? Två ödesmatcher väntar Tyresö.
Först Väsby på Bollmoravallen och sedan Holmsund i Umeå. Fyra poäng är en nödvändighet om luften inte pyser ur ballongen redan i mitten av juni.
Bosse Petersson:
– ... och nu blir det fest på Grimsta.
Vasalund i serietopp, men laget för dagen i fotbollens norretta är ett annat Stockholmsgäng. Efter 2-0 på Eskilstuna har BP tagit 13 poäng av 15 möjliga på de fem senaste matcherna. Fyra segrar och en oavgjord (Väsby) med mersmak. Så visst är BP i praktform just nu. Det bästa av allt är seriemakarnas sätt att pussla ihop programmet. I nästa omgång (söndag) möts nämligen BP och Vasalund på Grimsta. Räkna med en publikfest av format i detta toppderby.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	335	11.17	5.10	1.87	4	85

Table B.21: Superficial features, chunk ON161

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
74	76.7	66.0	0.3	1.9	1.3	7.1

Table B.22: Features indicating vocabulary load, chunk ON161

MDD	**UA**	**AT**	**ET**	**PT**
2.51	3	14	24	5.0

Table B.23: Features indicating sentence structure, chunk ON161

Pr	**NoPr**	**NR**	**Sem. depth**	**PM**	**LIX**
39.94	4.94	1.41	3.58	0.10	40.4

Table B.24: Features indicating idea density, human interest, and LIX value, chunk ON161

B.7 Ordinary news (ON) text from GP

Source: Göteborgs-Posten from January 4, 2007.
Distribution: Korp, Språkbanken, University of Gothenburg.

Och Ingrid är en sådan person men även den här gruppen är viktig för mig. Utan den skulle mina tankar var mer tragiska. Ingrid nickar instämmande.
– Visst är det så, då skulle livet inte vara lika roligt. Att ha en nära förtrogen som man kan dela sorger och glädjeämnen med är en viktig del i det goda åldrandet.
– En vän kan hjälpa en att återknyta till verkligheten när något uppjagande har hänt, säger Bo G Eriksson, forskare och sociolog vid Göteborgs universitet.
Han är med i H70-projektet i Göteborg som har studerat äldres hälsa sedan början av 70-talet. I sina studier har Bo G Eriksson sett hur viktigt det är för hälsan att ha vänner, inte minst för att reducera vardagsoro. Han berättar om en gammal dam som hade hemtjänst. Vid ett tillfälle påpekade ett vårdbiträde att damen hade en så fin lägenhet. "Här vill jag bo " sa vårdbiträdet.
När nästa vårdbiträde kom dagen därpå var damen uppriven.
"Ann sa att jag måste flytta, för här vill hon bo".
– Om kvinnan haft en vän att prata med hade vännen kunnat förklara för henne vad vårdbiträdet menade med det hon sa. Då hade inte kvinnan behövt ligga sömnlös den natten. Vi behöver andra människor för att återknyta till verkligheten. Forskning visar att det är viktigt för minnet att återberätta upplevelser. Vännen blir en spegel, en person som tar till sig den andras tankar och reflektioner.
– Och genom att dela sin upplevelse med någon kan man i framtiden minnas mer då man tillsammans återberättar det man varit med om. Det blir ett slags hjärngympa.
Internationella studier pekar på att kvinnor i högre utsträckning är män värdesätter vänskap på äldre dar, medan män tycks uppleva äktenskapet som viktigast. Sofie Ghazanfareeon Karlsson är forskare vid institutionen för socialt arbete, Umeå universitet. Hon har skrivit avhandlingen "Tillsammans men var för sig. Om särboenderelationer mellan äldre kvinnor och män i Sverige". De kvinnor hon studerat beskriver hur viktigt det är att ingå i mer flexibla relationsformer när man blivit gammal.
– De föredrar särborelationer därför att äktenskap eller samboskap med en person av motsatt kön innehåller många förpliktelser som är förknip-

pad med ofrihet och krav. Exempelvis vad gäller service och sysslor som på något sätt ingår i ett traditionellt äktenskap. Det kan vara en orsak till att särborelationer och vänskapsrelationer är viktigare för kvinnor. Det är relationer som de i större utsträckning kan forma som de vill. I det goda åldrandet ingår även bra matvanor, känslan av att känna sig behövd och fysisk aktivitet.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	417	13.9	4.96	1.82	6	65

Table B.25: Superficial features, chunk GP165

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
51	81.6	70.7	0.9	3.5	0.5	6.0

Table B.26: Features indicating vocabulary load, chunk GP165

MDD	**UA**	**AT**	**ET**	**PT**
2.29	8	16	24	6.4

Table B.27: Features indicating sentence structure, chunk GP165

Pr	**NoPr**	**NR**	**Sem. depth**	**PM**	**LIX**
44.72	2.49	0.91	3.22	0.04	38.4

Table B.28: Features indicating idea density, human interest, and LIX value, chunk GP165

B.8 ETR information (EI) text

Source: Skellefteå handikapprogram (Skellefteå disability program). Accessed December 8, 2003.
Distribution: LäSBarT corpus, ETR information part.

Detta är en lättläst sammanfattning av handikapprogrammet.
Det viktigaste finns med.
Hela programmet kan beställas från handikappavdelningen vid socialkontoret.
Handikappolitiken bestämmer hur människor med funktionshinder ska leva i kommunen.
Det är viktigt att vi är överens om vad kommunen ska göra för personer med funktionshinder.
Handikapprogrammet ska visa det.
Programmet ska ge oss mer kunskap.
Det ska innehålla planer för vad alla ska göra och mål för hur det ska bli.
Funktionshinder kan bero på skador i kroppen, i själen eller i hjärnan.
Det kan också bero på sjukdomar.
Men det är människor omkring och hur allting är ordnat som bestämmer hur mycket handikappad man är.
Om hus, affärer och bussar är handikappvänliga blir man inte så handikappad.
Det är viktigt att på olika sätt hindra så att folk inte får funktionshinder.
Det kan vara genom bra sjukvård, att människor äter nyttig mat, att vaccinera mot farliga sjukdomar, att se till att sjukdomar som smittar inte sprider sig.
Arbetsplatser ska inte vara farliga så att människor skadas i olyckor.
Det är också viktigt att göra allt i kommunen så att människor med funktionshinder kan leva som andra.
Alla människor ska vara lika mycket värda.
De har rätt att bli behandlade med respekt.
Det ska inte ha någon betydelse om man har ett funktionshinder eller inte.
Människor med funktionshinder har rätt att leva lika bra liv som andra.
Alla måste få bättre kunskap om vad personer med funktionshinder har rätt till, vad de behöver och vilka möjligheter de har.
Det behövs mer information och utbildning om funktionshinder.

Kommunen ska informera och utbilda människor om funktionshinder.
Då förstår de bättre hur människor med funktionshinder ska kunna leva som andra.
Det är viktigt att man upptäcker sjukdomar och skador tidigt så att man kan behandla dem på en gång.
Människor med funktionshinder ska få behandling med jämna mellanrum.
De ska få de mediciner de behöver.
Då kommer de att klara lika mycket i framtiden eller kanske bli bättre.
Kommunen ska se till att all personal i vården och omsorgen har tillräcklig utbildning, så att de kan upptäcka vilken vård människor med funktionshinder behöver.
De ska kunna ge rätt vård eller föreslå rätt service.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	360	12.0	5.09	1.81	20	45

Table B.29: Superficial features, chunk EI16

LVIX	**SweVoc**	**SV C**	**SV D**	**SV H**	**SV K**	**SV S**
40	93.0	80.1	2.1	2.8	0.0	8.0

Table B.30: Features indicating vocabulary load, chunk EI16

MDD	**UA**	**AT**	**ET**	**PT**
2.01	17	9	18	6.4

Table B.31: Features indicating sentence structure, chunk EI16

Pr	**NoPr**	**NR**	**Sem. depth**	**PM**	**LIX**
52.3	2.26	0.74	3.27	0.0	38.9

Table B.32: Features indicating idea density, human interest, and LIX value, chunk EI16

B.9 Ordinary information (OI) text

Source: Uminova stad. Published by Umeå kommun, 1991.
Distribution: SUC 2.0 part H - Miscellaneous.

Och den moderna teknologin gör att avstånden krymper. I Uminova stad kommer det att finnas en centralpunkt, en "telehamn", där företag och forskare har tillgång till all tänkbar kommunikationsteknisk utrustning.
Vem får bo? Det blir 500 nya bostäder i området? Vem ska bo i dem? Är det bara de som jobbar i området som har en chans att komma hit?
– Nejdå, säger Ingemar Engman, alla bostäder fördelas genom kommunens bostadsförmedling. Men det hindrar naturligtvis inte att de som har sin verksamhet här söker en bostad inom stadsdelen.
Och det ska väl också sägas: Det här blir ett fint område att arbeta och bo i. Mycket grönytor, litet biltrafik. Bilarna parkeras under husen. Naturen i området tas tillvara. Små näckrosdammar här och där, frisk luft, stora fria ytor. Gedigen och god arkitektur. Byggnaderna smälter in i miljön. Inga höga hus. Med ett undantag: Det stora entréhuset på sjutton våningar som bildar en djärv båge och ett utropstecken just vid infarten till Uminova stad.
Entréhuset som bland annat innehåller bostäder, byggs först, troligen redan till hösten. Dessförinnan ska den vanliga proceduren ske: Planen ska ställas ut, antagligen i maj-juni, och sen hoppas man att fullmäktige ger klartecken i augusti. Men redan laddar projektgruppen för presentation av de sammanlagt 150000 kvadratmetrarna lokaler. Där kommer cirka 3000 personer att arbeta när allt är klart.
Alla är fyllda av optimism. Kommunen vet att Skanska och AP-fonden gjorde en ingående analys av olika svenska städer, innan de beslöt sig för Umeå som samarbetspartner. Här finns hög kompetens när det gäller utveckling och forskning. Och framtidstro. Umeå är som bekant en av de få städer i Sverige som växer varje år. Sist men inte minst, enigheten bland kommunens tyngre partier är total – Uminova stad är en av de stora satsningar man vill göra under 90-talet.

S_n	$\sum W_n$	**MSLW**	**MWLC**	**MWLS**	XLW_n	**OVIX**
30	299	10.0	5.47	1.97	15	72

Table B.33: Superficial features, chunk OI186

LVIX	SweVoc	SV C	SV D	SV H	SV K	SV S
63	77.8	64.0	0.3	5.2	0.6	7.7

Table B.34: Features indicating vocabulary load, chunk OI186

MDD	UA	AT	ET	PT
2.69	6	18	19	5.2

Table B.35: Features indicating sentence structure, chunk OI186

Pr	NoPr	NR	Sem. depth	PM	LIX
42.8	5.94	1.33	3.88	0.0	37.7

Table B.36: Features indicating idea density, human interest, and LIX value, chunk OI186

C TEI ELEMENTS FOR CORPUS TAGGING

Tag	Description
abbr	contains an abbreviation of any sort
address	contains a postal or other address
age	contains information about the age of the expected audience
author	contains the name of the author(s) of the work
byline	contains the primary statement of responsibility of a work
chapter	contains the number of a text chapter
closer	groups together dateline, byline, salutation, and similar phrases appearing as a final group at the end
date	contains a date in any format
ellipsis	marks the position where text is omitted
emph	marks words or phrases which are stressed or emphasized for linguistic or rhetorical effects
foreign	indentifies a word or phrase as belonging to some language other than that of the surrounding text
head	contains the heading of a book chapter or a news headline
item	contains one component of a list
l	contains a single, possibly incomplete, line of verse
label	contains the label associated with an item in a list
lg	contains a group of verse lines functioning as a formal unit e.g. a stanza, refrain, verse paragraph, etc.
list	contains any sequence of items organized as a list
email	contains an e-mail address identifying a location to which e-mail messages can be delivered
name	contains a proper noun or noun phrase. Attributes can indicate its type, give a normalized form, or associate it with a specific individual or thing by means of a unique identifiers
opener	groups together dateline, byline, salutation, and similar phrases appearing as a preliminary group at the start
q	contains a quotation or apparent quotation
ref	defines a reference to another location
source	describes the original source for the information contained
subtype	contains the category of a fiction text
text	contains a corpus sample
title	contains the title of a work, whether article, book, journal, or series. Optional type can take the value "subtitle"
type	describes the genre of the text contained
volume	contains publishing information of the text

Table C.1: TEI elements for tagging of the LäsBarT corpus

D Detailed classification results

Children's ETR fiction vs. adults' ETR fiction								
Algo-rithm	**Model**	**Acc**	**CEF**			**AEF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	77.1	75.2	81.0	78.0	79.4	73.3	76.2
	SVIT	75.8	73.6	80.5	76.9	78.5	71.2	74.7
SMO	Base	77.4	79.2	74.3	76.7	75.8	80.5	78.1
	SVIT	83.9	82.6	86.0	84.2	85.4	81.9	83.6
CVR	Base	77.1	75.4	80.5	77.9	79.1	73.8	76.4
	SVIT	82.5	82.0	83.3	82.6	83.1	81.7	82.4
Children's ordinary fiction vs. adults' ordinary fiction								
Algo-rithm	**Model**	**Acc**	**COF**			**AOF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	76.4	84.2	64.9	73.3	71.5	87.9	78.8
	SVIT	81.6	89.3	71.3	79.5	76.4	91.4	83.2
SMO	Base	77.1	81.9	69.4	75.1	73.5	84.7	78.7
	SVIT	92.9	95.3	90.1	92.6	90.7	95.6	93.0
CVR	Base	77.1	81.0	70.7	75.5	74.0	83.4	78.4
	SVIT	90.4	91.7	88.8	90.3	89.2	92.0	90.6

Table D.1: Classification results of fiction documents across ages

Children's ETR fiction vs. children's ordinary fiction								
Algorithm	**Model**	**Acc**	**CEF**			**COF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	72.9	70.5	78.6	74.3	75.8	67.1	71.2
	SVIT	73.1	70.0	80.9	75.1	77.4	65.4	70.9
SMO	Base	72.9	71.4	75.9	73.6	74.3	69.6	71.9
	SVIT	84.6	82.5	87.7	85.0	86.9	81.4	84.1
CVR	Base	72.5	71.4	75.0	73.2	73.7	70.0	71.8
	SVIT	81.7	80.4	83.6	82.0	82.9	79.8	81.3
Children's ETR fiction vs. adults' ordinary fiction								
Algorithm	**Model**	**Acc**	**CEF**			**AOF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	89.7	86.3	94.5	90.2	93.9	85.0	89.2
	SVIT	90.5	85.7	97.3	91.1	96.9	83.8	89.2
SMO	Base	89.8	86.1	95.0	90.3	94.4	84.6	89.3
	SVIT	97.5	95.7	99.5	97.5	99.4	95.5	97.4
CVR	Base	90.1	88.1	92.7	90.3	92.3	97.5	89.8
	SVIT	96.1	94.2	98.2	96.4	98.1	93.9	96.0
Adults' ETR fiction vs. children's ordinary fiction								
Algorithm	**Model**	**Acc**	**AEF**			**COF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	55.0	54.4	62.4	58.1	55.9	47.6	51.4
	SVIT	82.4	82.4	82.4	82.4	82.4	82.4	82.4
SMO	Base	55.4	55.8	51.9	53.8	55.0	58.8	56.8
	SVIT	84.2	84.2	84.3	84.2	84.2	84.0	84.1
CVR	Base	55.0	55.2	53.3	54.2	54.8	56.7	55.7
	SVIT	79.8	79.9	79.5	79.7	79.6	80.0	79.8
Adults' ETR fiction vs. adults' ordinary fiction								
Algorithm	**Model**	**Acc**	**AEF**			**AOF**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	73.8	69.3	85.5	76.5	81.1	62.1	70.4
	SVIT	88.5	83.9	95.2	89.2	94.5	81.7	87.6
SMO	Base	73.7	69.8	83.6	76.1	79.5	63.8	70.8
	SVIT	96.7	95.0	98.6	96.7	98.5	94.8	96.6
CVR	Base	74.9	72.9	79.3	75.9	77.3	70.5	73.7
	SVIT	95.4	94.2	96.7	95.4	96.6	94.0	95.3

Table D.2: Classification results of fiction documents across text types

ETR news vs. ordinary news								
Algo-rithm	**Model**	**Acc**	**EN**			**ON**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	76.3	72.0	86.2	78.4	82.8	66.4	73.7
	SVIT	97.9	97.6	98.2	97.9	98.2	97.6	97.9
SMO	Base	76.0	71.7	86.0	78.2	82.5	66.0	73.3
	SVIT	99.6	98.0	99.2	98.6	99.2	98.0	98.6
CVR	Base	76.1	70.0	91.2	79.2	87.4	61.0	71.8
	SVIT	97.3	96.6	98.0	97.3	98.0	96.6	97.3

Table D.3: Classification results of news documents across text types

ETR information vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**EI**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	86.0	84.1	88.9	86.4	88.2	83.1	85.6
	SVIT	94.9	95.6	94.0	94.8	94.1	95.7	94.9
SMO	Base	85.9	83.5	89.4	86.3	88.6	82.3	85.3
	SVIT	97.1	97.1	97.1	97.1	97.1	97.1	97.1
CVR	Base	84.7	81.4	90.0	85.5	88.8	79.4	83.9
	SVIT	96.6	97.1	96.0	96.6	96.0	97.1	96.6

Table D.4: Classification results of ETR community information (EI) vs. ordinary community information (OI)

Childrens ETR fiction vs. ETR news								
Algo-rithm	**Model**	**Acc**	**CEF**			**EN**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	99.5	99.3	99.6	99.5	99.6	99.3	99.5
	SVIT	99.3	99.6	98.9	99.3	99.9	99.6	99.3
SMO	Base	99.5	99.3	99.6	99.5	99.6	99.3	99.5
	SVIT	99.7	99.8	99.6	99.7	99.6	99.8	99.7
CVR	Base	99.4	98.9	99.8	99.4	99.8	98.9	99.4
	SVIT	99.3	99.1	99.5	99.3	99.5	99.1	99.3
Children's ETR fiction vs. ETR information								
Algo-rithm	**Model**	**Acc**	**CEF**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	97.1	96.8	97.5	97.2	97.5	96.8	97.1
	SVIT	99.1	99.1	99.1	99.1	99.1	99.1	99.1
SMO	Base	97.0	95.8	98.2	97.0	98.2	95.7	96.9
	SVIT	99.5	98.9	100.0	99.5	100.0	98.9	99.5
CVR	Base	97.1	96.6	97.7	97.2	97.7	96.6	97.1
	SVIT	98.9	98.8	99.1	98.9	99.1	98.8	98.9
Adults' ETR fiction vs. ETR news								
Algo-rithm	**Model**	**Acc**	**AEF**			**EN**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	96.7	97.6	95.7	96.6	95.8	97.6	96.7
	SVIT	97.1	98.5	95.7	97.1	95.8	98.6	97.2
SMO	Base	96.4	97.6	95.2	96.4	95.3	97.6	96.5
	SVIT	98.6	98.6	98.6	98.6	98.6	98.6	98.6
CVR	Base	95.7	96.6	94.8	95.7	94.9	96.7	95.8
	SVIT	98.1	98.1	98.1	98.1	98.1	98.1	98.1
Adults' ETR fiction vs. ETR information								
Algo-rithm	**Model**	**Acc**	**AEF**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	88.2	86.7	90.2	88.4	89.8	86.2	88.0
	SVIT	96.4	95.3	97.6	96.5	97.6	95.2	96.4
SMO	Base	88.3	86.1	91.4	88.7	90.9	85.2	88.0
	SVIT	98.0	97.6	98.3	98.0	98.3	97.6	98.0
CVR	Base	88.3	86.9	90.2	88.6	89.9	86.4	88.1
	SVIT	97.5	96.7	98.3	97.5	98.3	96.7	97.5

Table D.5

Children's ordinary fiction vs. ordinary news								
Algo-rithm	**Model**	**Acc**	**COF**			**ON**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	95.6	94.0	97.5	95.7	97.4	93.8	95.5
	SVIT	97.9	97.1	98.8	97.9	98.7	97.1	97.9
SMO	Base	96.5	93.4	100.0	96.6	100.0	92.9	96.3
	SVIT	98.8	97.6	100.0	98.8	100.0	97.5	98.7
CVR	Base	96.9	94.5	99.6	97.0	99.6	94.2	96.8
	SVIT	98.5	98.3	98.8	98.5	98.7	98.3	98.5
Children's ordinary fiction vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**COF**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	98.4	97.7	99.1	98.4	99.1	97.7	98.4
	SVIT	99.7	99.4	100.0	99.7	100.0	99.4	99.7
SMO	Base	98.6	97.8	99.4	98.6	99.4	97.7	98.6
	SVIT	99.4	98.9	100.0	99.4	100.0	98.9	99.4
CVR	Base	98.6	97.8	99.4	98.6	99.4	97.7	98.6
	SVIT	99.3	98.9	99.7	99.3	99.7	98.9	99.3
Adults' ordinary fiction vs. ordinary news								
Algo-rithm	**Model**	**Acc**	**AOF**			**ON**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	80.0	80.0	80.0	80.0	80.0	80.0	80.0
	SVIT	86.9	83.9	91.3	87.4	90.4	82.5	86.8
SMO	Base	80.2	79.8	80.8	80.3	80.6	79.6	80.1
	SVIT	89.8	86.3	94.6	90.3	94.0	85.0	89.3
CVR	Base	80.0	79.5	80.8	80.2	80.5	79.2	79.8
	SVIT	88.1	87.7	88.8	88.2	88.6	87.5	88.1
Adults' ordinary fiction vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**AOF**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	85.4	86.3	84.3	85.3	84.6	86.6	85.6
	SVIT	96.3	95.3	97.4	96.3	97.4	95.1	96.2
SMO	Base	85.1	86.8	82.9	84.8	83.6	87.4	85.5
	SVIT	99.1	98.9	99.4	99.1	99.4	98.9	99.1
CVR	Base	84.3	91.1	76.0	82.9	79.4	92.6	85.5
	SVIT	98.4	98.8	98.0	98.4	98.8	98.9	98.4

Table D.6

ETR news vs. ETR information								
Algo-rithm	**Model**	**Acc**	**EN**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	62.3	59.1	80.2	68.0	69.2	44.5	54.1
	SVIT	88.2	85.6	91.9	88.6	91.2	84.5	87.7
SMO	Base	59.4	57.1	75.3	65.0	63.8	43.4	51.7
	SVIT	91.7	91.7	91.7	91.7	91.7	91.7	91.7
CVR	Base	61.9	58.3	83.3	68.6	70.8	40.5	51.5
	SVIT	90.3	89.6	91.0	90.3	90.9	89.5	90.2

Table D.7: Classification results of ETR news texts (EN) vs. ETR community information texts (EI)

Ordinary news vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**ON**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	61.5	59.4	73.4	65.7	65.0	49.6	56.3
	SVIT	85.2	79.7	94.6	86.5	93.3	75.8	83.7
SMO	Base	59.9	59.0	65.6	62.1	61.0	54.2	57.4
	SVIT	94.6	94.6	94.6	94.6	94.6	94.6	94.6
CVR	Base	63.0	58.4	90.5	71.0	78.7	35.4	48.9
	SVIT	96.3	98.7	93.8	96.2	94.0	98.8	96.3

Table D.8: Classification results of ordinary news (ON) vs. ordinary information (OI)

Children's ETR fiction vs. ordinary news								
Algo-rithm	**Model**	**Acc**	**CEF**			**ON**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	99.2	99.2	99.2	99.2	99.2	99.2	99.2
	SVIT	99.4	99.2	99.6	99.4	99.6	99.8	99.4
SMO	Base	99.0	98.4	99.6	99.0	99.6	98.3	99.0
	SVIT	99.4	99.2	99.6	99.4	99.6	99.2	99.4
CVR	Base	98.3	98.3	98.3	98.3	98.3	98.3	98.3
	SVIT	99.2	99.2	99.2	99.2	99.2	99.2	99.2
Children's ETR fiction vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**CEF**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	99.4	98.9	100.0	99.4	100.0	98.9	99.4
	SVIT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
SMO	Base	99.0	98.0	100.0	99.0	100.0	98.0	99.0
	SVIT	99.7	99.4	100.0	99.7	100.0	99.4	99.7
CVR	Base	99.7	99.4	100.0	99.7	100.0	99.4	99.7
	SVIT	99.9	100.0	99.7	99.9	99.7	100.0	99.9
Adults' ETR fiction vs. ordinary news								
Algo-rithm	**Model**	**Acc**	**AEF**			**ON**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	95.0	93.8	97.1	95.1	97.0	92.9	94.9
	SVIT	98.5	97.9	99.2	98.6	99.2	97.9	98.5
SMO	Base	95.2	92.9	97.9	95.3	97.8	92.5	95.1
	SVIT	99.4	99.2	99.6	99.4	98.6	98.2	99.4
CVR	Base	94.1	93.1	95.4	94.2	95.3	92.9	94.1
	SVIT	98.1	98.7	97.5	98.1	97.5	98.8	98.1
Adults' ETR fiction vs. ordinary information								
Algo-rithm	**Model**	**Acc**	**AEF**			**OI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	99.1	98.3	100.0	99.2	100.0	98.3	99.1
	SVIT	99.4	99.3	98.2	98.7	99.1	99.7	99.4
SMO	Base	98.9	97.8	100.0	98.9	100.0	97.7	98.8
	SVIT	99.6	99.2	100.0	99.6	100.0	99.1	99.6
CVR	Base	98.9	98.0	99.7	98.9	99.7	98.0	98.8
	SVIT	98.9	98.9	98.9	98.9	98.9	98.9	98.9

Table D.9

Children's ordinary fiction vs. ETR news								
Algo-rithm	**Model**	**Acc**	**COF**			**EN**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	96.4	97.2	95.6	96.4	95.7	97.7	96.5
	SVIT	98.7	99.3	98.2	98.7	98.2	99.3	98.7
SMO	Base	96.4	97.2	95.5	96.4	95.6	97.2	96.4
	SVIT	99.7	99.7	99.8	99.7	99.8	99.7	99.7
CVR	Base	96.5	96.1	97.0	96.5	96.9	96.1	96.5
	SVIT	99.2	99.1	99.2	99.2	99.2	99.1	99.2
Children's ordinary fiction vs. ETR information								
Algo-rithm	**Model**	**Acc**	**COF**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	90.3	88.7	92.2	90.4	91.9	88.3	90.1
	SVIT	97.6	95.8	99.5	97.6	99.5	95.7	97.5
SMO	Base	90.3	89.0	91.9	90.4	91.6	88.6	90.1
	SVIT	99.1	98.5	99.7	99.1	99.7	98.4	99.0
CVR	Base	90.3	89.0	91.9	90.4	91.6	88.6	90.1
	SVIT	98.1	97.9	98.3	98.1	98.3	97.9	98.1
Adults' ordinary fiction vs. ETR news								
Algo-rithm	**Model**	**Acc**	**AOF**			**EN**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	70.4	77.8	56.8	65.7	66.1	83.9	73.9
	SVIT	99.7	99.4	100.0	99.7	99.4	100.0	99.7
SMO	Base	44.6	45.7	58.5	51.3	42.6	30.7	35.7
	SVIT	100.0	100.0	100.0	100.0	100.0	100.0	100.0
CVR	Base	70.4	77.8	57.1	65.8	66.2	83.7	73.9
	SVIT	99.4	99.4	99.4	99.4	99.4	99.4	99.4
Adults' ordinary fiction vs. ETR information								
Algo-rithm	**Model**	**Acc**	**AOF**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	63.7	69.0	49.8	57.9	60.7	77.6	68.1
	SVIT	99.5	99.0	100.0	99.5	100.0	99.0	99.5
SMO	Base	63.4	65.2	57.8	61.2	62.1	69.1	65.4
	SVIT	99.4	99.5	99.3	99.4	99.3	99.5	99.4
CVR	Base	64.6	66.8	57.9	62.0	62.9	71.2	66.8
	SVIT	97.7	97.3	98.1	97.7	98.1	97.8	97.7

Table D.10

Ordinary news vs. ETR information								
Algo-rithm	**Model**	**Acc**	**ON**			**EI**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	71.7	72.8	69.2	70.9	70.6	74.8	72.4
	SVIT	95.8	94.0	97.9	95.9	97.8	93.8	95.7
SMO	Base	71.5	72.7	68.8	70.7	70.4	74.2	72.2
	SVIT	98.1	97.9	98.3	98.1	98.3	97.9	98.1
CVR	Base	71.7	73.0	68.8	70.8	70.5	74.6	72.5
	SVIT	98.1	97.9	98.3	98.1	98.3	97.3	98.1

Table D.11: Classification results of ordinary news texts (ON) vs. ETR community information texts (EI)

Ordinary information vs. ETR news								
Algo-rithm	**Model**	**Acc**	**OI**			**EN**		
			Prec	**Recall**	**F-score**	**Prec**	**Recall**	**F-score**
NB	Base	92.3	96.3	88.0	91.9	88.9	96.6	92.6
	SVIT	97.9	98.0	97.7	97.9	97.7	98.0	97.9
SMO	Base	92.1	97.4	86.6	91.7	87.9	97.7	92.6
	SVIT	99.0	99.4	98.6	99.0	98.6	99.4	99.0
CVR	Base	91.6	95.9	86.9	91.2	88.0	96.3	92.0
	SVIT	98.3	98.8	97.7	98.3	97.7	98.9	98.3

Table D.12: Classification results of ordinary information (OI) vs. ETR news (EN)