An Architecture for Rating and Controlling Text Readability

Ingo Glöckner, Sven Hartrumpf, Hermann Helbig, Johannes Leveling, Rainer Osswald
Intelligent Information and Communication Systems
Department of Mathematics and Computer Science
FernUniversität in Hagen, Germany
http://pi7.fernuni-hagen.de/

Abstract

In the so-called information society with its strong tendency towards individualization, it becomes more and more important to have all sorts of textual information available in a simple and easy to understand language. We present an approach that allows to automatically rate the readability of German texts and also provides suggestions how to make a given text more readable. Our system, called DeLite, employs a powerful NLP component that supports the syntactic and semantic analysis of German texts.

1 Introduction

How to make documents readable and understandable has been a topic almost since writing exists. Especially in the Anglo-American language community there is a long tradition of style guides. Two areas where text readability and comprehensibility traditionally have drawn special interest are education and technical documentation.

Within an information society, where everyone should be able to access all information he or she requires, the need of guidelines and control mechanism for simple and understandable language becomes important to a broad community. Administrational and governmental information, for example, is relevant to every member of the society, including people that have reading difficulties because of a low education level or because the language in question is not their mother tongue. Health related topics provide another information domain important to a large and heterogenous target group. The reading and understanding of medical instructions, for instance, is vital to many people.

Accessibility in general and readability in particular is also an important issue for web page design. The Web Content Accessibility Guidelines (WCAG) proposed by the Web Accessibility Initiative of the W3C require the text content of a web page to be "readable and understandable"; see also Jenge et al. (2006).

Part of this work has been financially supported by the EU project *Benchmarking Tools for the Web* (BenToWeb, FP6-004275).

2 Measuring Readability and Comprehensibility

The systematic research on how to measure the readability of written material started in the 1920's in the United States with the advent of so-called *readability formulas*. According to George Klare (1963, p. 34), "a readability formula is a method of estimating the probable success a reader will have in reading and understanding a piece of writing." These formulas are often based on variables such as the mean word length and the mean sentence length. So, the influential Flesch Reading Ease score (Flesch, 1948) is a linear formula with two variables, namely average sentence length and average number of syllables per word. Average sentence length gives a rough approximation of syntactic complexity whereas the average number of syllables per word serves as a measure for vocabulary familiarity (in accordance with Zipf's law). Over the years, a considerable number of readability formulas has been proposed (Klare, 1963; DuBay, 2004).

The procedure for developing readability formulas typically shows the following three-step pattern (Klare, 1963): Select elements in a text that are related to readability; correlate element occurrences with text readability (measured by established comprehension tests); combine the variables into a regression equation. Gray and Leary (1935) in their landmark study of readability came up with more than forty indicators including linguistically motivated factors such as the number of prepositional phrases and the number of pronouns, both of which were part of the five factor formula resulting from regression analysis. Later studies, however, re-

vealed that eliminating the linguistically motivated factors only slightly decreases the correlation with the test ratings. This observation justified the restriction to simple factors such as average sentence and word length.

There are various proposals for readability formulas suitable to German. For instance, Flesch's Reading Ease formula has been adapted to German by adjusting the constants of the formula appropriately. A systematic approach in the line of Gray and Leary has been pursued by Dickes and Steiwer (1977), who derived three versions of their formula, one by starting the regression with all variables, one with variables suitable to computation (in 1977), and one with variables easily determinable by hand. The third formula only employs mean sentence length, mean word length, and the type-token ratio, whereas the first one includes variables such as the percentage of verbs with animate subject and the percentage of nouns denoting concrete objects. Again, the correlation of the simplified formula with the test score decreases only marginally compared to the original formula.

Nowadays, readability formulas are widely accepted as screening devices. A different question frequently addressed is whether readability can be improved by rewriting texts in a way that aims at changing the values of formula variables to get a better score. Reducing word and sentence length *per se* does not necessarily improve readability although it improves the readability score determined by most of the formulas. This observation has been a point of critique since the very introduction of readability formulas. Simple word and sentence variables are at best *indications* but not *causal factors* of semantic and syntactic difficulty. Rewriting texts for better readability must thus aim at changing the "causal" properties of words and sentences, which include such factors as familiarity, concreteness, and derivational complexity (as with nominalized verbs) at the level of words, and syntactic embedding depth, passive voice, and clause length at the sentence level. In other words, one has to focus on the linguistic features that affect reading and comprehension according to psycholinguistic evidence and models.

Readability and comprehensibility depend on the *content organization* of the text as well. Content features are not directly correlated to surface properties like sentence length. Instead, the cognitive aspects of reading and comprehension have to be taken into account, which have been studied intensively e.g. by Walter Kintsch and his colleagues (van Dijk and Kintsch, 1983; Kintsch, 1998). They developed a cognitive model of comprehension based on the idea that the sentences in the text are turned into *propositions*, i.e. predicate-argument structures, which represent the *meaning* of the text. The propositions form a propositional net which, roughly speaking, encodes the *coherence* relation between propositions defined by argument overlap and embedding. Although a more detailed description of the model is beyond the scope of the present paper, we note that the propositional level of a text and its coherence structure is an important factor for its comprehensibility.

To sum up, a system for rating a given text with respect to readability and comprehensibility should ideally be based on a detailed model about the cognitive and psycholinguistic capacities of a typical member of the intended audience. Such a model would incorporate all findings of psycholinguistic research at all levels of linguistic analysis ranging from word identification and lexical access to discourse and story processing. Surely no such "grand unified" model of language processing is available at present nor will presumably be in the near future; moreover it is to be expected that psycholinguistic research still has not said the last word about human language processing.

Nevertheless, we regard it as possible with state-of-the-art NLP techniques to automatically rate readability and comprehensibility based on much more complex variables of syntactic and semantic difficulty than involved in classical readability formulas.

3 Readability Criteria and Indicators

The above discussion of readability measures has shown that in order to overcome the limitations inherent in the traditional readability formula approach one has take into account linguistic features at all levels of linguistic analysis: morphology, lexicology, syntax, semantics, and discourse. For if not only screening of texts is at issue but guided rewriting then the obstacles to readability have to be isolated. In addition, discourse related features of texts such as coherence can hardly be detected by calculating statistics over mere surface characteristics.

Our approach towards measuring readability is thus to automatically extract as many as possible of the linguistic features that are known or suspected to

be of psycholinguistic relevance to readability and comprehensibility. For methodological reasons, we distinguish between *readability criteria* and *readability indicators*, where the purpose of latter is to operationalize the evaluation of the former. That is, readability criteria are properties of linguistic units such as sufficiently low syntactic complexity or unambiguity of anaphoric reference whereas readability indicators are specific numerical features such as the length of a noun phrase in words or the number of candidate antecedents of a pronoun. A subset of the readability criteria and indicators of the DeLite system (see Section 4) is sketched in the following.

Morphological complexity, especially when combined with low frequency, can be an obstacle to readability (Clahsen et al., 2003). In particular, this is the case for *compounds* and *derived words*. Readability indicators for this criterion are the number of simplicia within a compound and the number of deverbal and deadjectival nouns in a sentence.

Readability criteria related to the lexical level include *vocabulary variety*, *lexical ambiguities*, *idioms*, and *vocabulary abstractness*. Possible indicators are here the type-token ratio, the mean number of different readings per word, and the number of words denoting abstract entities.

The negative influence of *syntactic ambiguity* and *complexity* on readability has been thoroughly investigated in psycholinguistic research (Gibson, 1998). Another factor known to affect readability is *constituent ordering* (Hawkins, 1994). Associated readability indicators include clause embedding depth, number of attachment candidates, and the distance in words between verb and subject.

Readability criteria at the semantic level, as we understand it here, are mainly concerned with the propositional organization within single sentences, where propositions are seen as predicate-argument schemas (Kintsch, 1998). Possible indicators are the number of propositions per sentence and the number of propositions the conceptual entity denoted by a given nominal phrase is involved in.

Readability at the discourse level is essentially the same as discourse comprehensibility. Textual *coherence* is important here, which means, for instance, that coreferences can be resolved easily, including bridging references, and that the propositions expressed by the text are connected by discourse relations in an appropriate way (Halliday and Hassan, 1976; van Dijk and Kintsch, 1983). The indicators we use at this level are at present focused on coreference phenomena. They include the number of candidate antecedents of a pronoun, the distance in words and sentences between coreferring expressions, and the number of newly introduced discourse entities per sentence.

4 The DeLite System

Existing text analysis systems may be able to determine some of the more surface related readability indicators mentioned in the previous section. The DeLite system presented in this section, in contrast, aims at integrating all indicators from the morphological to the discourse level. DeLite is designed both as an *evaluation tool* that rates German texts with respect to readability and as an *authoring tool* that identifies the location and the nature of individual readability problems within the text.

Any system meeting these requirements calls for a powerful and robust NLP component which is not only capable of syntactic parsing but also of detailed semantic analysis. The DeLite system employs a syntactico-semantic parser (Hartrumpf, 2003) that has already been successfully employed for information retrieval and question answering tasks (Hartrumpf, 2005). The parser makes use of a large semantic lexicon for German (Hartrumpf et al., 2003) and is able to semantically analyze German natural language sentences and texts. The resulting semantic representations are based on the so-called MultiNet knowledge representation formalism (Helbig, 2006). The parser also contains a module for coreference resolution at the text level (Hartrumpf, 2001).

In order to analyze a given text passage with respect to readability, DeLite first applies the syntactico-semantic parser to the passage. The parser returns a data structure that includes, among other things, the following information: compound and derivation structures of words and lexical alternatives (word level); syntactic dependency trees and semantic structures in logical form (sentence level); the set of possible coreference pairs and a coreference partition (text level).

For further processing, the DeLite system maintains these data in a hierarchical annotation structure of linguistic units that allows to insert additional information at the word, phrase, sentence, and text level. The annotation structure is then successively traversed to calculate the values of the basic and derived readability indicators, which are stored in turn at the respective linguistic unit. For instance, the

clause embedding depth is a basic indicator associated with sentence units whereas the derived indicator of mean clause embedding depth is associated with the text unit.

Based on the indicator values, DeLite evaluates the readability of the input text. If an indicator value exceeds a certain threshold then the respective linguistic unit is marked as critical with respect to readability. In addition, DeLite calculates various readability scores including a single score for the whole input text: DeLite first determines scores for each readability criterion by normalizing the corresponding derived indicator values and then takes them to calculate readability scores for the five levels of linguistic description, which in turn are combined into an overall readability score for the input text. At each step, the combination of scores is calculated as a weighted sum. At its present state of development, DeLite uses a preliminary, heuristic setting for the thresholds and weights. It is part of current and future work to develop methods for adjusting the parameters automatically on the basis of pre-classified text passages (see Section 5). To this end, we started to build a test bed for DeLite based on a corpus of municipality texts partly obtained directly from local authorities, partly retrieved semi-automatically from the web.

The implemented prototype of DeLite determines about 40 readability indicators at all linguistic levels. In addition to giving heuristic readability scores for whole texts along the lines described in the previous section, the prototype returns a detailed readability report in XML format. Moreover, the prototype is equipped with a graphical user interface that allows one to highlight text passages which are problematic with respect to specific readability criteria interactively selected by the user.

5 Conclusion and Future Work

We have presented an architecture called DeLite for automatically rating and controlling the readability of German texts. Our approach overcomes the limitations of traditional readability formulas in that it employs powerful NLP techniques to extract causal factors of readability instead of approximating them by surface features such as sentence length. Consequently, DeLite is able to support the rewriting of texts towards better readability.

The next issue to be dealt with in the development of DeLite is the adjustment of the system parameters in a more systematic way. We plan to use texts with existing readability judgments for adjusting the parameters by means of machine learning techniques. In addition, we will investigate the suitability of other sorts of corpora for this purpose. For instance, non-fiction texts written for children may be assumed from the outset to be more easily readable than texts written for adults.

References

Harald Clahsen, Ingrid Sonnenstuhl, and James P. Blevins. 2003. Derivational morphology in the German mental lexicon: A dual-mechanism account. In *Morphological Structure in Language Processing*, pages 125–155. Mouton de Gruyter.

Paul Dickes and Laure Steiwer. 1977. Ausarbeitung von Lesbarkeitsformeln für die deutsche Sprache. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 9:20–28.

William H. DuBay. 2004. Principles of readability. Online available at http://www.impact-information.com/impactinfo/readability02.pdf.

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

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

William S. Gray and Bernice Leary. 1935. *What Makes a Book Readable*. University of Chicago Press.

Michael A. K. Halliday and Ruqaiya Hassan. 1976. *Coherence in English*. Longman.

Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. 2003. The semantically based computer lexicon HaGenLex. *TAL*, 44(2):81–105.

Sven Hartrumpf. 2001. Coreference resolution with syntactico-semantic rules and corpus statistics. In *Proc. of CoNLL-2001*, pages 137–144.

Sven Hartrumpf. 2003. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag.

Sven Hartrumpf. 2005. Question answering using sentence parsing and semantic network matching. In *Proc. of CLEF 2004*, LNCS 3491, pages 512–521. Springer.

John A. Hawkins. 1994. *A Performance Theory of Order and Constituency*. Cambridge University Press.

Hermann Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer.

Constantin Jenge, Sven Hartrumpf, Hermann Helbig, and Rainer Osswald. 2006. Automatic control of simple language in web pages. In *Proc. of ICCHP 2006*, LNCS 4061, pages 207–214, Springer.

Walter Kintsch. 1998. *Comprehension*. Cambridge University Press.

George R. Klare. 1963. *The Measurement of Readability*. Iowa State University Press.

Teun A. van Dijk and Walter Kintsch. 1983. *Strategies of Discourse Comprehension*. Academic Press.