A Readability Checker with Supervised Learning using Deep Syntactic and Semantic Indicators

Tim vor der Brück, Sven Hartrumpf, Hermann Helbig

Intelligent Information and Communication Systems (IICS)
FernUniversität in Hagen
58084 Hagen, Germany
firstname.lastname@fernuni-hagen.de

Abstract

Checking for readability or simplicity of texts is important for many institutional and individual users. Formulas for approximately measuring text readability have a long tradition. Usually, they exploit surface-oriented indicators like sentence length, word length, word frequency, etc. However, in many cases, this information is not adequate to realistically approximate the cognitive difficulties a person can have to understand a text. Therefore we use deep syntactic and semantic indicators in addition. The syntactic information is represented by a dependency tree, the semantic information by a semantic network. Both representations are automatically generated by a deep syntactico-semantic analysis. A global readability score is determined by applying a nearest neighbor algorithm on 3,000 ratings of 300 test persons. The evaluation showed, that the deep syntactic and semantic indicators lead to quite comparable results to most surface-based indicators. Finally, a graphical user interface has been developed which highlights difficult-to-read text passages, depending on the individual indicator values, and displays a global readability score.

1. Introduction

Readability checkers are used to highlight text passages that are difficult to read. They can help authors to write texts in an easy-to-read style. Furthermore they often display a global readability score which is derived by a readability formula. Such a formula describes the readability of a text numerically. There exists a large amount of readability formulas (Klare, 1963). Most of them use only surface-oriented indicators like word frequency, word length, or sentence length. Such indicators have only indirect and limited access to judging real understandability. Therefore, we use deep syntactic and semantic indicators[1] in addition to surface-oriented indicators. The semantic indicators operate mostly on a semantic network (SN); in contrast, the syntactic indicators mainly work on a dependency tree containing linguistic categories and surface text parts. The SNs and the dependency trees are derived by a deep syntactico-semantic analysis based on word-class functions.

Furthermore, we collected a whole range of readability criteria from almost all linguistic levels: morphology, lexicon, syntax, semantics, and discourse[2] (Hartrumpf et al., 2006). To make these criteria operable, each criterion is underpinned by one or more readability indicators that have been investigated in the (psycho-)linguistic literature and can be automatically determined by NLP tools (see (Jenge et al., 2005) for details). Two typical readability indicators for the syntactic readability criterion of *syntactic ambiguity* are the *center embedding depth of subclauses* and the *number of argument ambiguities* (concerning their syntactic role[3]).

2. Related Work

There are various methods to derive a numerical representation of text readability. One of the most popular readability formulas was created in 1948: the so-called Flesch Reading Ease (Flesch, 1948). The formula employs the average sentence length and the average number of syllables for judging readability. The sentence length is intended to roughly approximate sentence complexity, while the number of syllables approximates word frequency since usually long words are less used. Later on, this formula was adjusted to German (Amstad, 1978). Despite of its age, the Flesch formula is still widely used.

Also, the revised Dale-Chall readability index (Chall and Dale, 1995) mainly depends on surface-type indicators. Actually, it is based on sentence length and the occurrences of words in a given list of words which are assumed to be difficult to read.

Recently, several more sophisticated approaches which use advanced NLP technology were developed. They determine for instance the embedding depth of clauses, the usage of active/passive voice or text cohesion (McCarthy et al., 2006; Heilman et al., 2007; Segler, 2007). The method of (Chandrasekar and Srinivas, 1996) goes a step beyond pure analysis and also creates suggestions for possible improvements.

Usually, those approaches are based on surface or syntactic structures but not on a truly semantic representation which represents the cognitive difficulties for text understanding more adequately. Moreover, readability checkers normally focus on English texts which means that grammatical phenomena typical for German like separable prefixes are not handled (see Sect. 5.2.). Moreover, only few of those approaches (e.g., (Rascu, 2006)) integrate their readability checkers into a graphical user interface, which is vital for practical usage.

Readability formulas usually combine several so-called readability indicators like sentence or word length by a pa-

[1] In this paper, an indicator is called *deep* if it depends on a deep syntactico-semantic analysis.

[2] In this paper, discourse criteria are subsumed under the heading semantic because they form only a small group and rely directly on semantic information.

[3] Such ambiguities can occur in German because of its relatively free constituent order.

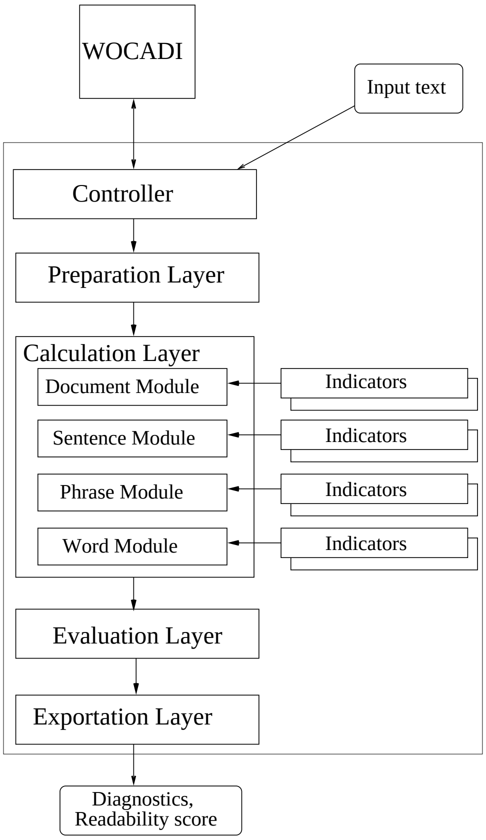


Figure 1: System architecture of the readability checker DeLite.

rameterized sum. Non-linear readability formulas are currently quite rare. Examples of the latter type are the nearest neighbor approach of (Heilman et al., 2007) and the employment of support vector machines by (Larsson, 2006). Larsson used them to separate the vectors of indicator values for given texts into the three different readability classes *easy*, *medium*, and *difficult*. A drawback of this method is that the classification into only three levels is rather rough.

3. System Architecture

A text is processed in several steps (see Figure 1) by our readability checker DeLite (an association of *Lite* as in light/easy reading and *De* as in Deutsch/German; there is also a prototype EnLite for English). First, the Controller passes the text to a deep syntactico-semantic analysis (WOCADI[4] parser, (Hartrumpf, 2003)), which is based on a word-class functional analysis and is supported by a large semantically oriented lexicon (Hartrumpf et al., 2003). The parser output for each sentence is a morpho-lexical analysis, one or more (in case of ambiguities) syntactic dependency trees, one or more SNs, and intrasentential and intersentential coreferences determined by a hybrid rule-statistical coreference resolution module. An example of the resulting SNs, which follow the MultiNet formalism (multilayered extended semantic network, (Helbig, 2006)), is shown in Figure 2. On the basis of this analysis, the text is divided into sentences, phrases, and words in the Preparation Layer.

The individual indicator values are determined by the Calculation Layer. DeLite currently uses 48 morphological, lexical, syntactic, and semantic indicators; in the following sections, we concentrate on some deep syntactic and semantic ones. Each indicator is attached to a certain processing module depending on the type of required information: words, phrases, sentences, or the entire document. Each module iterates over all objects of its associated type that exist in the text and triggers the calculation of the associated indicators. Examples for indicators operating on the word level are the indicators *number of characters* or *number of word readings*. Semantic and syntactic indicators usually operate on the sentence level. As the result of this calculation step an association from text segments to indicator values is established.

In the Evaluation Layer, the values of each indicator are averaged to the so-called *aggregated* indicator value. Note that there exists for each indicator only one aggregated indicator value per text. The readability score is then calculated (see Sect. 4.) by the k-nearest neighbor algorithm of the machine learning toolkit RapidMiner (Mierswa et al., 2006) . In spite of surface-type indicators a deep indicator can usually only be determined for a given sentence (usually, deep indicators operate on sentences) if certain prerequisites are met (e.g., full or chunk parse available). If this is not the case, the associated sentence is omitted for determining the aggregated indicator value. If an indicator could not be calculated for any sentence of the text at all, its value is set to some fixed constant.

Finally, all this information is marked up in XML and in a user-friendly HTML format and is returned to the calling process by the Exportation Layer.

4. Deriving a Readability Score Using the k-Nearest Neighbor Algorithm

A nearest neighbor algorithm is a supervised learning method. Thus, before this method can be applied to new data, a training phase is required. In this phase, a vector of aggregated indicator values is determined by RapidMiner (see previous section) for each text of our readability study. The vector components are normalized and multiplied by weights representing the importance of the individual indicators where the weights are automatically learned by an evolutionary algorithm. All vectors are stored together with the average user ratings for the associated texts.

To derive a readability score for a previously unseen text, the vector of weighted and normalized aggregated indicator values is determined for this text first (see above). Afterwards, the k vectors of the training data with the lowest distance to the former vector are extracted. The readability score is then given as a weighted sum of the user ratings associated with those k vectors (the k nearest neighbors).

5. Syntactic Indicators

5.1. Clause Center Embedding Depth

A sentence is difficult to read if the syntactic structure is very complex (Groeben, 1982). One reason for a high com-

[4] WOCADI is the abbreviation of **Wo**rd-**C**l**a**ss based **Di**sambiguating.

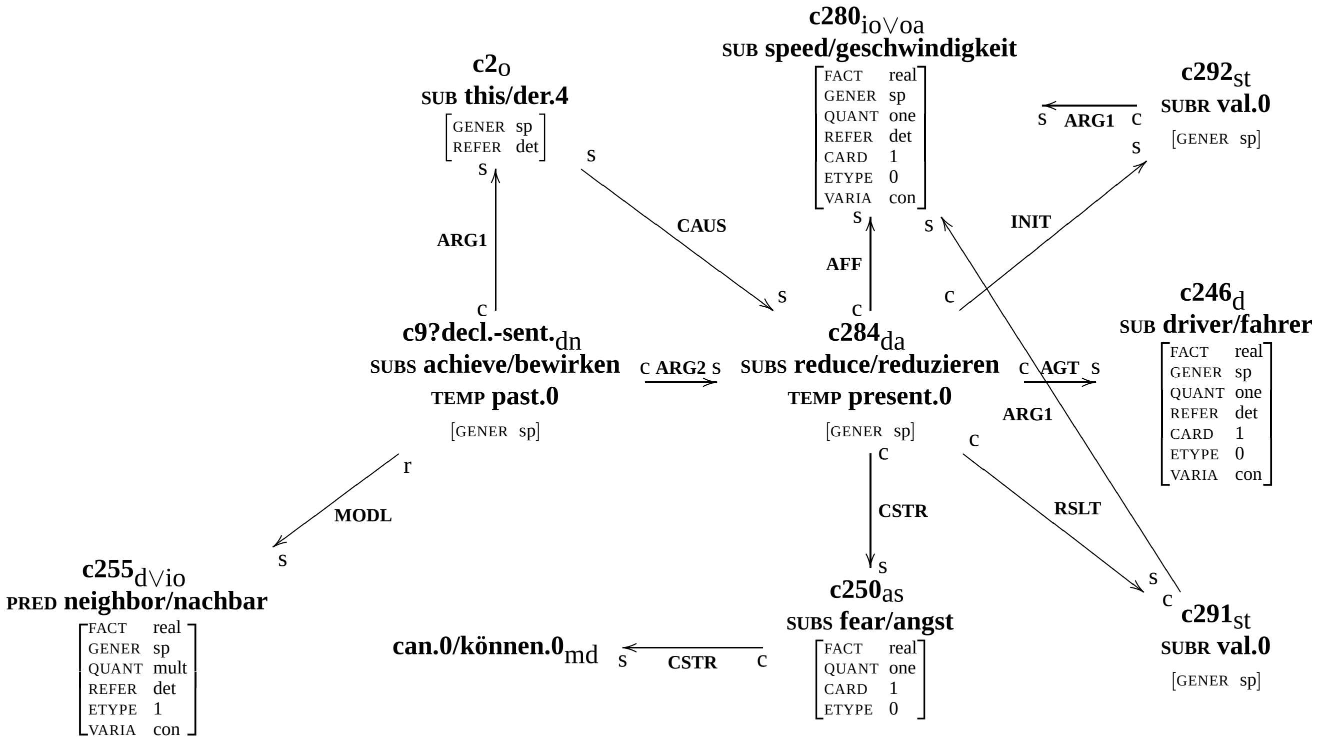


Figure 2: SN for the corpus sentence *Das könnte bewirken, dass der Fahrer aus Angst vor den Nachbarn die Geschwindigkeit reduziert.* (*This could achieve that the driver reduces the speed for fear of the neighbors.*)

plexity can be that the sentence contains deeply embedded subordinate clauses. The difficulty can be increased if the subordinate clause is embedded into the middle of a sentence since the reader has to memorize the superior clause until its continuation after the termination of the subordinate clause, for example: *Er verließ das Haus, in dem die Frau, die er liebte, wohnte, sofort.* (literally: *He left the house where the woman he loved lived immediately.*) Thus, we employ the center embedding depth of a main verb as a readability indicator and calculate its value in the following way. First, we determine the path from the root of the dependency tree to each main verb. Then, we count the occurrences of the dependency relations for relative or other subordinated clauses on this path. However, we only take them into account if the embedded clause is not located on the border of the superior clause which we can verify by comparing the start/end character indices of both clauses.

5.2. Distance between Verb and Separable Prefix

In German, so-called separable prefix verbs are split into two words in clauses with main clause word order. Example: *einladen* (*invite*) ⇒ *Er lädt*… *ein.* (*He invites*….). If the verb is far away from the verb prefix, it can be difficult to associate both parts.

5.3. Number of Words per Nominal Phrase

According to Miller (Miller, 1962), long NPs degrade readability. Hence, a part of the information given in the long NP should better be placed in a subordinate clause or a new sentence. Therefore we count the average number of words contained in an NP where a larger number results in a worse readability score. Note that we only consider maximal NPs (i.e., NPs not contained in a larger NP). Otherwise a large indicator value for the long NP could be compensated by small indicator values for the contained NPs which should be avoided.

6. Semantic Indicators

6.1. SN Quality

The fact that a sentence could not be completely parsed is caused mainly by syntactic or semantic defects since the parser builds the syntactic structure as a dependency tree and the semantic representation as an SN in parallel. Therefore, the indicator *SN quality* is a mixed one: semantic and syntactic. Consider for instance the two sentences *Das Werk kam vor allem bei jungen Theatergängern an. Schulbusse reisten an, um es sich anzusehen.*[5] (*The work was very well accepted by young visitors of the theater. School buses arrived to watch it.*) The second sentence, which is syntactically correct, is semantically incorrect and therefore difficult to read. The semantic lexicon, which is employed by the parser, requires that the first argument (which plays the semantic role of the agent) *ansehen.1.1*[6] (*to watch*) is of type *human*. Thus, this sentence is rejected by the parser as incorrect. In other cases the sentence might be accepted but considered as semantically improbable. This information, which is provided by the parser, is used by the readability checker DeLite and turned out to be very valuable for estimating text readability.

Three parse result types are differentiated: complete parse (around 60% of the sentences; note that this means

[5] from the newspaper *Schleswig-Holstein am Sonntag, 2007*

[6] Note that the readings of a lexeme are distinguished by numerical suffixes.

complete syntactic structure *and* semantic representation at the same time), chunk parse (25%), failure (15%).[7] Those three cases are mapped to different numerical values of the indicator *SN quality*. Additionally, if a full parse or a chunk parse is available, the parser provides a numerical value specifying the likelihood that the sentence is semantically correct which is determined by several heuristics. This information is incorporated into the quality score of this indicator too. Naturally, this indicator depends strongly on the applied parser. A different parser might lead to quite different results.

6.2. Number of Propositions per Sentence

DeLite also looks at the number of propositions in a sentence. More specifically, all SN nodes are counted which have the ontological sort *si*(tuation) (Helbig, 2006, p. 412) or one of its subsorts. In a lot of cases, readability can be judged more accurately by the number of propositions than by sentence length or similar surface-oriented indicators. Consider for instance a sentence containing a long list of NPs: *Mr. Miller, Dr. Peters, Mr. Schmitt, Prof. Kurt, … were present.* Although this sentence is quite long it is not difficult to understand (Langer et al., 1981). In contrast, short sentences can be dense and contain many propositions, e.g., concisely expressed by adjective or participle clauses.

6.3. Number of Connections between SN Nodes/Discourse Entities

The average number of nodes which are connected to an SN node is determined. A large number of such nodes often indicates a lot of semantic dependencies. For this indicator, the arcs leading to and leaving from an SN node are counted. Note that the evaluation showed that better results (stronger correlation and higher weight) have been achieved if only SN nodes are regarded which are assigned the ontological sort *object* (Helbig, 2006, p. 409–411). In this case, these SN nodes roughly represent the discourse entities of a sentence.

6.4. Length of Causal and Concessive Chains

Argumentation is needed to make many texts readable. But if an author puts too many ideas in too few words, the passage becomes hard to read. For example, the following sentence from a newspaper corpus has been automatically identified as pathologic because it contains three causal relations (CAUS and CSTR in Figure 2) chained together: *Das könnte bewirken, dass der Fahrer aus Angst vor den Nachbarn die Geschwindigkeit reduziert.* (*This could achieve that the driver reduces the speed for fear of the neighbors.*). Again, length measurements on the surface will not help to detect the readability problem, which exists for at least some user groups. Splitting such a sentence into several ones is a way out of too dense argumentation.

Indicator	Weight	Type
Number of words per sentence	0.679	Sur
Passive without semantic agent	0.601	Syn
Number of readings	0.520	Sem
Distance between verb and complement	0.518	Syn
SN quality	0.470	Syn/Sem
Number of connections between discourse entities	0.467	Sem
Inverse concept frequency	0.453	Sem
Clause center embedding depth	0.422	Syn
Number of sentence constituents	0.406	Syn
Maximum path length in the SN	0.395	Sem
Number of causal relations in a chain	0.390	Sem
Number of compound simplicia	0.378	Sur
…	…	…
Word form frequency	0.363	Sur
…	…	…
Number of connections between SN nodes	0.326	Sem

Table 1: Indicators with largest weights in our readability function (Syn=syntactic, Sem=semantic, and Sur=surface indicator type).

7. Evaluation

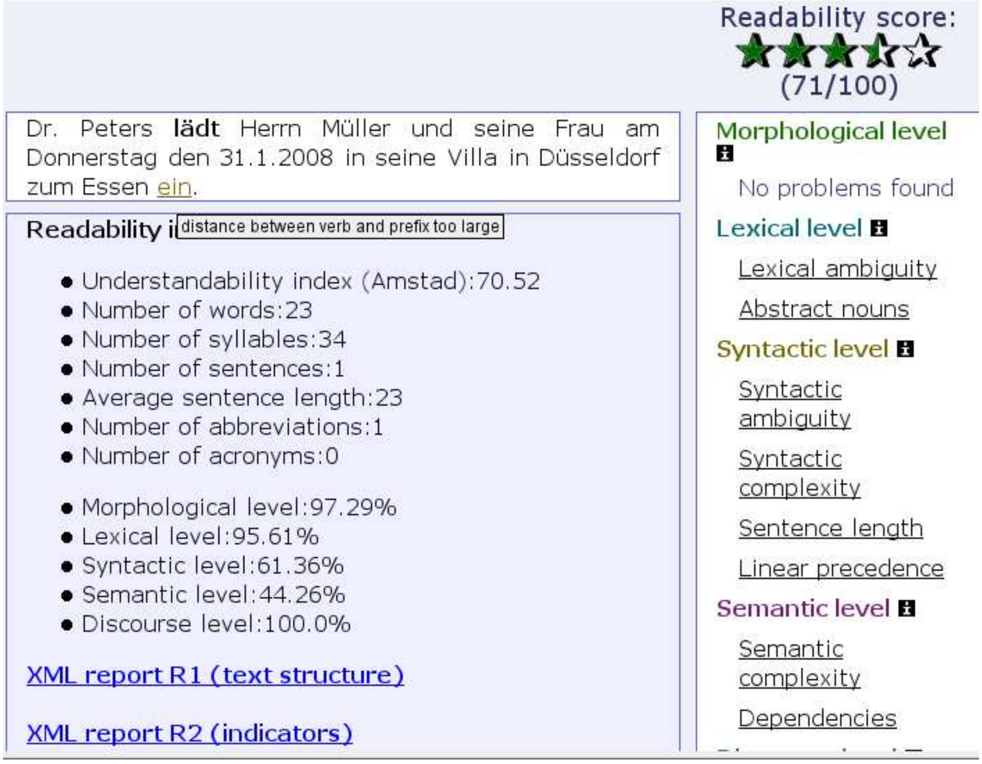
To judge the viability of our approach, we conducted an online readability study with 500 texts, more than 300 participants, and around 3,000 human ratings for individual texts where the participants rated the text readability on a 7 point Likert scale (Likert, 1932).

Almost 70 % of the participants were between 20 and 40 years old; the number of participants over 60 was very small (3 %). The participants were mainly well-educated. 58 % of them owned a university or college degree. There is none who had no school graduation at all.

Our text corpus originated from the municipal domain and differs significantly from newspaper corpora, which are widely used in computational linguistics. So the text corpus we used contains a lot of ordinances with legal terms and abbreviations, e.g., § *65 Abs. 1 Satz 1 Nr. 2 i.V.m.* § *64 Abs. 1 Satz 2 LWG NRW* (*section 65.1.1 (2) in connection with section 64.1.2 LWG NRW*). This corpus has been chosen because local administrations in Germany have committed themselves to make their web sites accessible; one central aspect of accessibility is simple language.

Figure 4 shows the mean average error (MAE) and the root mean square error (RMSE) of DeLite's global readability score in contrast to the average user rating determined by a 10 fold cross-validation over all 500 test documents. The ordinate contains MAE and RMSE, the abscissa, on a logarithmic scale, the number of neighbors used. The lowest errors (MAE: 0.126, RMSE: 0.153) were obtained when using the 40 nearest neighbors. The nearest neighbor algorithm determined the weights of each indicator using an evolutionary algorithm. The resulting indicator weights are given in Table 1.

[7] Note that the absence of a complete parse is problematic only for a part of the indicators, mainly deep syntactic and semantic ones. And even for some of these indicators, one can define fallback strategies to approximate indicator values by using partial results (chunks).



Readability score:

★★★⯪☆

(71/100)

Dr. Peters **lädt** Herrn Müller und seine Frau am Donnerstag den 31.1.2008 in seine Villa in Düsseldorf zum Essen ein.

Readability i distance between verb and prefix too large

- Understandability index (Amstad):70.52
- Number of words:23
- Number of syllables:34
- Number of sentences:1
- Average sentence length:23
- Number of abbreviations:1
- Number of acronyms:0

- Morphological level:97.29%
- Lexical level:95.61%
- Syntactic level:61.36%
- Semantic level:44.26%
- Discourse level:100.0%

XML report R1 (text structure)

XML report R2 (indicators)

Figure 3: DeLite screenshot showing a sentence which contains a large distance between verb (*lädt*) and separable verb prefix (*ein*). English translation for the example sentence: *Dr. Peters invites Mr. Müller and his wife for dinner on Thursday, Jan. 31, 2006 to his villa in Düsseldorf.*

Indicator	Correlation	Type
Number of words per sentence	0.430	Sur
SN quality	0.399	Syn/Sem
Inverse concept frequency	0.330	Sem
Word form frequency	0.262	Sur
Number of reference candidates for a pronoun	0.209	Sem
Number of propositions per sentence	0.180	Sem
Clause center embedding depth	0.157	Syn
Passive without semantic agent	0.155	Syn
Number of SN nodes	0.148	Sem
Pronoun without antecedent	0.140	Sem
Number of causal relations in a chain	0.139	Sem
Distance between pronoun and antecedent	0.138	Sem
Maximum path length in the SN	0.132	Sem
Number of connections between discourse entities	0.132	Sem

Table 2: Indicators most strongly correlated with user ratings (Syn=syntactic, Sem=semantic, and Sur=surface indicator type).

The correlations of the indicators in comparison with the user ratings are displayed in Table 2. Correlation and weights of deep syntactic and semantic indicators turned out to be quite comparable to surface-type indicators.

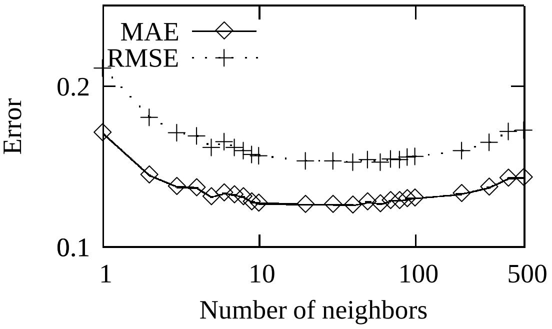


Figure 4: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) between the DeLite readability score and the average user ratings of a text depending on the number of neighbors.

Finally as a baseline, DeLite was compared to the readability index resulting from employing the nearest neighbor approach only on the indicators of the Flesch readability index, i.e. average sentence length and number of syllables per word. The correlation of DeLite with the user ratings is 0.501 which clearly outperforms the Flesch indicators (0.432).

8. User Interface

Besides a low-level server interface, DeLite provides a graphical user interface for comfortable usage. In Figure 3, a screenshot of this interface is shown.[8] The types of readability problems found in the text are displayed on the right side. If the user clicks on such a type, the associated difficult-to-read text segments are highlighted. Additional support for the user is provided if he/she wants to have more information about the readability problem. Moving the mouse pointer over the highlighted text segment, a flyover help text with a more detailed description is displayed. Moreover, if the user clicks on the highlighted segment, additional text segments are marked in bold face. These additional segments are needed to fully describe and explain specific readability problems.

The example in Figure 3 shows the readability analysis of a verb which is too far away from its separable prefix (see Sect. 5.2.). The prefix *ein-* is highlighted as problematic and additionally the main verb *lädt* is marked in bold face for better understanding.

9. Conclusion

An overview of some typical examples of deep syntactic and semantic readability indicators has been given. In our evaluation, it turned out that these indicators have comparable weights and correlations to most surface-type indicators in accurately judging readability.

In the future, the parser employed in DeLite will be continually improved. Currently, DeLite is only an authoring tool; we will investigate the addition of the ability to reformulate a sentence to be better to understand. Finally, the automatic distinction between real ambiguities that exist for humans and spurious ambiguities that exist only for machines (e.g., NLP methods like PP attachment and interpretation) must be sharpened.

Deep syntactic and semantic indicators turned out to be quite valuable for assessing readability and are expected to be a vital part of future readability checkers.

Acknowledgments

We wish to thank our colleagues Christian Eichhorn, Ingo Glöckner, Johannes Leveling, and Rainer Osswald for their support for this work. The research reported here was in part funded by the EU project Benchmarking Tools and Methods for the Web (BenToWeb, FP6-004275).

10. References

T. Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich, Zurich, Switzerland.

J. Chall and E. Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books, Brookline, Massachusetts.

R. Chandrasekar and B. Srinivas. 1996. Automatic induction of rules for text simplification. Technical Report IRCS Report 96-30, University of Pennsylvania, Philadelphia, Pennsylvania.

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

N. Groeben. 1982. *Leserpsychologie: Textverständnis – Textverständlichkeit*. Aschendorff, Münster, Germany.

S. Hartrumpf, H. Helbig, and R. Osswald. 2003. The semantically based computer lexicon HaGenLex – Structure and technological environment. *Traitement automatique des langues*, 44(2):81–105.

S. Hartrumpf, H. Helbig, J. Leveling, and R. Osswald. 2006. An architecture for controlling simple language in web pages. *eMinds: International Journal on Human-Computer Interaction*, 1(2):93–112.

S. Hartrumpf. 2003. *Hybrid Disambiguation in Natural Language Analysis*. Der Andere Verlag, Osnabrück, Germany.

M. J. Heilman, K. Collins-Thompson, J. Callan, and M. Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the Human Language Technology Conference*, Rochester, New York.

H. Helbig. 2006. *Knowledge Representation and the Semantics of Natural Language*. Springer, Berlin.

C. Jenge, S. Hartrumpf, H. Helbig, G. Nordbrock, and H. Gappa. 2005. Description of syntactic-semantic phenomena which can be automatically controlled by NLP techniques if set as criteria by certain guidelines. EU-Deliverable 6.1, FernUniversität in Hagen.

G. Klare. 1963. *The Measurement of Readability*. Iowa State University Press, Ames, Iowa.

I. Langer, F. Schulz von Thun, and R. Tausch. 1981. *Sich verständlich ausdrücken*. Reinhardt, München, Germany.

P. Larsson. 2006. Classification into readability levels. Master's thesis, Department of Linguistics and Philology, University Uppsala, Uppsala, Sweden.

R. Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.

P. McCarthy, E. Lightman, D. Dufty, and D. McNamara. 2006. Using Coh-Metrix to assess distributions of cohesion and difficulty: An investigation of the structure of high-school textbooks. In *Proc. of the Annual Meeting of the Cognitive Science Society*, Vancouver, Canada.

I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. 2006. Yale: Rapid prototyping for complex data mining tasks. In *Proc. of KDD*, Philadelphia, Pennsylvania.

G. Miller. 1962. Some psychological studies of grammar. *American Psychologist*, 17:748–762.

E. Rascu. 2006. A controlled language approach to text optimization in technical documentation. In *Proc. of KONVENS 2006*, pages 107–114, Konstanz, Germany.

T. M. Segler. 2007. *Investigating the Selection of Example Sentences for Unknown Target Words in ICALL Reading Texts for L2 German*. Ph.D. thesis, School of Informatics, University of Edinburgh.

[8] Note that the classification of indicators is slightly different in the screenshot than in this paper. This is caused by the fact that we want to evaluate surface-oriented indicators in comparison to linguistically informed indicators.