# A Semantically Oriented Readability Checker for German

## Tim vor der Brück, Sven Hartrumpf

Intelligent Information and Communication Systems (IICS)
FernUniversität in Hagen
58084 Hagen, Germany
{tim.vorderbrueck,sven.hartrumpf}@fernuni-hagen.de

## Abstract

One major reason that readability checkers are still far away from judging the understandability of texts consists in the fact that no semantic information is used. Syntactic, lexical, or morphological information can only give limited access for estimating the cognitive difficulties for a human being to comprehend a text. In this paper however, we present a readability checker which uses semantic information in addition. This information is represented as semantic networks and is derived by a deep syntactico-semantic analysis. We investigate in which situations a semantic readability indicator can lead to superior results in comparison with ordinary surface indicators like sentence length. Finally, we compute the correlations and absolute errors for our semantic indicators related to user ratings collected in an online evaluation.

## 1. Introduction

Basically, a readability[1] checker has two major application areas. First, it can be used to automatically identify easy-to-read texts in a text corpus. In this case it suffices to provide a global score which is usually calculated by a readability formula.

Second, a readability checker can be used to support authors to make their texts easy to read. In this case, more support is desirable than to compute only a global readability score. Instead, text passages which are difficult to read should be highlighted. A readability checker of this type can be found in (Rascu, 2006)[2]. The calculation of a global score can here be useful too in order to give an estimation of the understandability of a text.

In this paper, we will describe both application areas. Therefore we describe how semantic information can improve both the calculation of a global readability score and the identification of difficult text passages.

Readability checkers can compute a global score by applying a readability formula on several indicator values. Note that traditional readability formulas (Flesch, 1948; DuBay, 2004) use only surface type indicators like average word/sentence length or word frequency and do not exploit semantic or syntactic information. Especially without using any semantic information, access to actual understandability is only limited and indirect.

## 2. Semantic Networks

Semantic networks (SNs) of the MultiNet (Multilayered Extended Semantic Networks) formalism (Helbig, 2006) allow to homogeneously represent the semantics of single words, phrases, sentences, texts, or text collections. Such SNs are chosen as the semantic representation in our DeLite readability checker described in this paper.

An SN node represents a concept, while an SN arc expresses a relation between two concepts. In MultiNet, each node is semantically classified by a *sort* from a hierarchy comprising 45 sorts. Furthermore, a node has an inner structure (depending on its sort) containing *layer features* like CARD (cardinality) and REFER (referential determinacy). Figure 2. shows the graphical form of an SN for the sentence *Dr. Peters lädt Herrn Müller zum Essen ein, da heute sein Geburtstag ist.* ('*Dr. Peters invites Mr. Müller for dinner since it is his birthday today.*'). It was generated by the parser described in the following section.

## 3. Automatic Generation of Semantic Networks

The WOCADI parser (Hartrumpf, 2003), which is employed in DeLite, can construct SNs of the MultiNet formalism for German phrases, sentences, or texts. The text that is analyzed for readability is parsed sentence by sentence. During this process, SNs and syntactic dependency structures are built.

An important component of our deep syntactico-semantic analysis of natural language is HaGenLex, a semantically based computer lexicon (Hartrumpf et al., 2003). This lexicon not only lists verb valencies, but also their syntactic and semantic types. Consider for example the German verb *essen* ('*eat*'). Sentences like *Die Birne isst den Apfel.* ('*The pear eats the apple.*') are rejected because semantic selectional restrictions are violated. Besides this comprehensive lexicon with around 27,000 entries, we employ a flat lexicon, many name lexicons, and a sophisticated compound analysis to achieve the parser coverage required for applications like readability checkers.

Disambiguation is realized by specialized modules which work with symbolic rules and disambiguation statistics derived from annotated corpora. Currently, such modules exist for (intrasentential and intersentential) coreference resolution, the attachment of prepositional phrases, and the interpretation of prepositional phrases.

---

[1] In this paper, we use readability in the sense of understandability. We are aware that there exist other definitions where readability (or better: legibility) only relates to the form, but not to the contents of a text.

[2] Readability was not the only objective in this system. One further aspect was to ensure the fulfillment of certain formulation standards.

**c8$_{ta}$** SUB **birthday**
$$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \\ \text{VARIA} & \text{con} \end{bmatrix}$$

**c1$_{d}$** SUB **doctor**
$$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \\ \text{VARIA} & \text{con} \end{bmatrix}$$

**c2$_{na}$** SUB **last_name**
$$\begin{bmatrix} \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \end{bmatrix}$$

c ATTCH c — between c8 and c1 — c ATTR c — VAL → **peters.0$_{fe}$**

ARG1 — AGT

**c7$_{st}$** SUBS **be** [GENER sp] — S REAS S — **c3?decl-sent$_{da}$** SUBS **invite** [GENER sp] — c BENF s — **c4$_{d}$** SUB **mister**
$$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \\ \text{VARIA} & \text{con} \end{bmatrix}$$

c ATTR c — **c5$_{na}$** SUB **last_name**
$$\begin{bmatrix} \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \end{bmatrix}$$

TEMP — TEMP — TEMP — PURP — VAL

**today$_t$** — **present.0$_t$** — **c6$_{ad}$** SUBS **dinner**
$$\begin{bmatrix} \text{FACT} & \text{real} \\ \text{GENER} & \text{sp} \\ \text{QUANT} & \text{one} \\ \text{REFER} & \text{det} \\ \text{CARD} & 1 \\ \text{ETYPE} & 0 \\ \text{VARIA} & \text{con} \end{bmatrix}$$
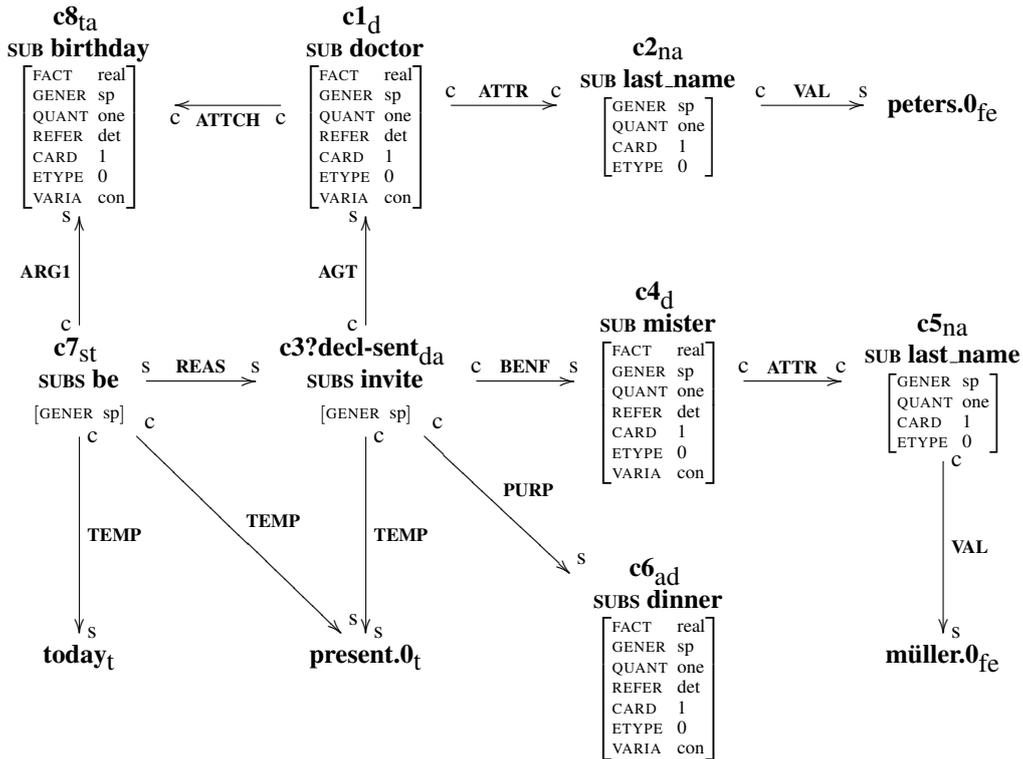
**müller.0$_{fe}$**

Figure 1: Example of an SN from the WOCADI parser (node names are translated from German). The sort of a node is written as a subscript of the node name. The feature structure below a node name shows layer features and their values. Each arc is labeled with a MultiNet relation. To simplify the graph topology, SUB and SUBS arcs are folded below the start nodes. Labels at the start and at the end of an arc indicate the so-called knowledge type, e.g. categorical (c) and situational (s); see for details.

## 4. Conception of Our Readability Checker

Readability can be measured by way of numerous *readability criteria*. Each criterion (like semantic complexity) can be realized or approximated by one or more operable (i.e. implementable) *readability indicators* (like number of propositions per sentence, longest path in the SN, etc.). Note that an indicator can only be applied on a specific type of text segments which we call the *segment type*[3] of this indicator, e.g. the indicator *Number of Propositions per Sentence* can only be applied on an entire sentence but not on single words. We differentiate between the segment types *word*, *phrase*, *sentence*, and *text*.

### 4.1. Calculating a Global Readability Score

A single indicator value has only limited account for describing the readability of a text. However we can calculate a global readability score by combining the individual readability indicators. To do this each indicator value has to be known at the text level. For some indicators this information is not known a priori (e.g. the binary indicator *Abstract Noun* operates on the word level) and has to be derived first. We do this by averaging the values of such indicators on all text segments those indicators are applicable to (having the correct segment type).

It should be kept in mind that the value ranges of readability indicators can be quite different. The sentence length, for example, usually varies from 5 to 30, while the ratio of nouns which have an abstract meaning is constrained between zero and one. Thus all indicator values have to be mapped to a common interval before they can be combined. We achieve this with a sigmoid function.

### 4.2. Highlighting Text Segments

We compute an indicator value for each text segment this indicator is applicable to. If that value exceeds a certain threshold the associated text segment is highlighted. For example, if the threshold for the indicator *Number of Concept Nodes in the SN* is 10, all sentences having 11 or more concept nodes will be highlighted.

We experienced that this approach did not suffice. Sometimes it is important for an exact understanding of the readability problem to highlight additional text segments. We call these text segments *supplementary highlight segments* in contrast to the *primary highlight segments* which directly refer to the found readability defect. Note that the segment type of a supplementary highlight segment does not have to match the segment type of the associated primary highlight segment.

In the following section, we describe some of the most important semantic readability indicators. For more motivation and references to the literature (e.g. from psycholinguistics) please see (Hartrumpf et al., 2006).

---

[3]In some rare cases the applicability is further restricted, e.g. the indicator *Number of Reference Candidates* is not applicable to all kinds of words but only to pronouns.

# 5. Semantically Oriented Readability Indicators

## 5.1. Abstract and Concrete Nouns

A high proportion of abstract nouns can deteriorate text readability (Groeben, 1982). A noun is considered as abstract if it does not directly refer to a visible object. The binary information whether a noun is abstract or not is available from our semantically oriented lexicon. The annotation is made on concepts and not on words since a word can have both abstract and concrete readings. For example, the German word *Platz* can mean *a place in a city* (like a plaza) which is a visible, concrete object. Alternatively, it can mean space like in the sentence: *Im Englisch-Kurs ist kein Platz frei.* (*'There is no space left in the English course.'*).

## 5.2. Multiple Negations

Multiple negations can make a sentence more difficult to understand (Groeben, 1982) and should be avoided if a positive formulation is possible. There exist many possibilities to convey negation in German (Drosdowski, 1995). Negation can be expressed by special words, e.g. *nicht* (*'not'*) and *niemals* (*'never'*), or prefixes, e.g. *unmöglich* (*'impossible'*) is the antonym of *möglich* (*'possible'*). While special words are quite easy to recognize this is not the case for negation prefixes. First, such a prefix is not trivial to recognize, e.g. the German word *unterirdisch* does not contain the negation prefix *un*, but the prefix *unter* (*'under'*), which has a completely different meaning. Second, in some cases a word contains actually a negation prefix, but it is not used as a negation, e.g. the adjective *unheimlich* (*'weird'*) is not an antonym of *heimlich* (*'secret'*). However if semantic information is available this problem can be handled quite easily. Consider we have some word $w$ which is the concatenation of the prefix *un* and a word $v$. We can infer that $w$ is a negated adjective if $w$ is an antonym of $v$ which means that the lexicon contains an ANTO (antonymy) relation connecting $v$ and $w$. Note that there exist several algorithms to extract semantic relations like ANTO by analyzing large text corpora. These methods would save the work to manually add ANTO relations to the lexicon; however, for cases like *unheimlich* and *heimlich* above, special treatment (or manual correction) is needed.

A special case of negations are *double negations*. A sentence contains a double negation if a similar (but not the same) semantics can be achieved by dropping two negations occurring in this sentence. This effect takes place if one negation is in the scope of another. Note that there are also sentences which contain triple or quadruple negations, e.g. the sentence *Ich glaube nicht, dass Peter nicht denkt, dass der Film nicht uninteressant ist.* (*'I do not believe that Peter does not think that the movie is not uninteresting.'*) contains a quadruple negation. In almost all cases, double negations are redundant and should be avoided. A double negation can relate to a sentence, to a phrase, or only to a word. Our readability checker can recognize several different kinds of double negations, e.g. a double negation occurs in a sentence if the sentence node

is associated to the facticity (layer feature FACT) *non-real* and is connected to the modality *non.0* by a MODL (modality) relation; see (Helbig, 2006) for details on the semantic representation.

## 5.3. Indicators Concerning Anaphors

Several readability problems can concern anaphors. Consider again the following sentence: *Dr. Peters lädt Herrn Müller zum Essen ein, da heute sein Geburtstag ist.* (*'Dr. Peters invites Mr. Müller for dinner since it is his birthday today.'*). The possessive determiner *sein* (*'his'*) can either relate to the antecedent candidate *Dr. Peters* or to the antecedent candidate *Mr. Müller*. (Figure 2. shows the SN of this sentence, where the first antecedent candidate has been chosen.) For a better understanding this sentence should be reformulated, e.g. by replacing the anaphor by either *Dr. Peters* or *Mr. Müller*. Thus we introduced a readability indicator counting the number of possible antecedents for each anaphor. If used in an authoring tool, we propose to mark the anaphor as primary and the antecedents as supplementary highlight segments if this indicator value exceeds the associated threshold (e.g. 1).

Furthermore an anaphoric reference can be difficult to resolve if the antecedent is too far away from the anaphor. The distance can be measured in words, sentences, or—more semantically and psycholinguistically motivated—by intervening entities (or discourse referents). Finally, we also use an indicator to check if there exists at least one antecedent for each anaphor.

## 5.4. Number of Propositions per Sentence

A further measure for sentence complexity is the number of SN nodes which bear the semantic sort *si* (situation, like *to discuss*) or *abs* (abstract situation; for nominalized verbs like *discussion*) or one of their subsorts. Such nodes correspond to the propositions in a given sentence. This indicator is correlated to the sentence length since a long sentence usually contains also several propositions. However this is not always the case. Consider for example the following long sentence: *Anwesend waren Dr. Schulz, Dr. Peters, Herr Werner, Frau Brand, Herr Mustermann, Herr Frank, Dr. Grainer, [. . . ].* (*'Dr. Schulz, Dr. Peters, Mr. Werner, Mrs. Brand, Mr. Mustermann, Mr. Frank, Dr. Grainer, [. . . ] were present.'*) which contains only a single proposition. Long item lists usually do not degrade readability (Langer et al., 1981). Therefore in such situations the readability can more appropriately be judged by the indicator *Number of Propositions per Sentence* than by sentence length.

Also the opposite effect can be found: a quite short sentence can contain many propositions (for example expressed by participle constructions). The indicator *Sentence Length* would not be violated, while the sentence is definitely hard to read, e.g. *The man* running *downhill and* meeting *the colleague* walking *to the office* fell *over a dog* chased *by a boy*. This sentence contains five propositions and is definitely hard to understand.

## 5.5. Longest Path in the SN

Information is often more difficult to understand if the constituents depend on each other and therefore a sequential interpretation is necessary. Consider for example the easy-to-read sentence *Ich besuche meine Schwiegermutter, meinen Onkel und meine Kusine.* (*'I visit my mother-in-law, my uncle, and my cousin.'*). Since the constituents in the coordination do not depend on each other they can be interpreted in parallel which makes the sentence easy to understand. However, this is not the case for the following sentence where the constituents have to be interpreted sequentially: *Ich besuche die Schwiegermutter des Onkels meiner Kusine.* (*'I visit the mother-in-law of the uncle of my cousin.'*) Similar effects can be observed in connection with negations where the special phenomena of double negations can emerge (see Section 5.2.). Sequentially interpreted sentences usually lead to longer paths in the SN. Thus we measure the length of the longest path that the SN contains.

## 5.6. Other Semantic Readability Indicators

We evaluated further semantic indicators. For instance, DeLite counts the concepts appearing in a sentence as well as the concepts that were newly introduced in a sentence. We also investigated an indicator determining the average number of arcs the discourse entities of the SN were connected to (*Connectivity of Discourse Entities*). For concessive and causal clauses, DeLite counts the causal and concessive relations in a chain.

## 6. Other Readability Indicators using Semantic Information

In addition to purely semantic indicators, we employed semantic information to improve indicators which were not originally semantic. Furthermore some indicators combine information from several linguistic levels (e.g. semantics and syntax). Two of the most important ones of such indicators are described below.

**Quality of the Semantic Network** The case that the SN for some sentence could not be constructed or is assigned a low quality score is often caused by the fact that the associated sentence is either syntactically or semantically complex or even incorrect. Thus we provide an indicator for this information. Note that this indicator is not purely semantic since the construction of the SN can fail if the syntactic structure of the sentence is invalid.

**Passive Construction** Usually sentence formulations in active voice are easier to understand than equivalent formulations in passive voice (Groeben, 1982). To convert a sentence into active voice the direct object and the subject have to change roles. We call the new subject the *semantic subject*. Passive constructions are very common in German. Thus we want to highlight a passive sentence (or reduce the readability score) only if it is obvious that an active formulation would be better.

There exist some exceptions to the rule that active formulations should be preferred. In some cases the semantic subject might not be known (or might be irrelevant), e.g. *Peter wurde rechtzeitig benachrichtigt.* (*'Peter was*

*informed on time.'*). In this case, the impersonal pronoun *man* (*'one'*) can be inserted to convert the sentence into active: *Man benachrichtigte Peter rechtzeitig.* However this formulation is usually not better than the original. Moreover, sometimes a passive formulation will be preferred if the semantic subject is neither a human being nor an animal. For example, the sentence *Peter wurde vom Blitz erschlagen.* (*'Peter was struck by a lightning.'*) need not be converted into *Der Blitz erschlug Peter.* (*'The lightning struck Peter.'*).

Since a complete linguistic treatment of all cases is not trivial we used a heuristic. We only penalized passive if the semantic subject is uttered and is connected to the sentence by the semantic relation AGT (agent). In this case the semantic subject usually performs some sort of action and an active formulation should always be possible. This heuristic conforms to (Helbig and Kempter, 1997) who propose that an active formulation should be preferred if the sentence is agent-oriented. In spite of the incomplete linguistic treatment we observed a high correlation of this indicator in comparison with the other semantically oriented indicators (see Section 7.).

## 7. Implementation and Evaluation

We evaluated our algorithm as implemented in DeLite on a text corpus of 500 texts from the local administration domain. More than 300 people participated in this user evaluation performed on the web. Each participant rated the readability and understandability of several texts on a 7-point Likert scale. Indicator weights were learned on this annotated corpus using a robust regression algorithm with linear optimization (Bertsimas and Tsitsiklis, 1997). We adopt the usual definition that the weights are constrained to numbers between zero and one and all weights together sum up to one.

It turned out that the purely semantic indicators as determined by our algorithm had only a total weight of 8% which was caused by the fact that many corpus texts were so complex that the SN construction failed. However, if we restrict the texts to the ones where most sentence SNs could be constructed, the overall weight of the semantic indicators raised up to 20%.

We further determined the correlation and absolute error for the semantically oriented indicators on our whole text corpus of 500 texts. The indicators with the best correlation to user ratings are shown in Table 1. Furthermore we also determined the average absolute error of each indicator. Besides the highly correlated indicators listed in

Table 1: Indicators, using semantic information, which are most strongly correlated to user ratings.

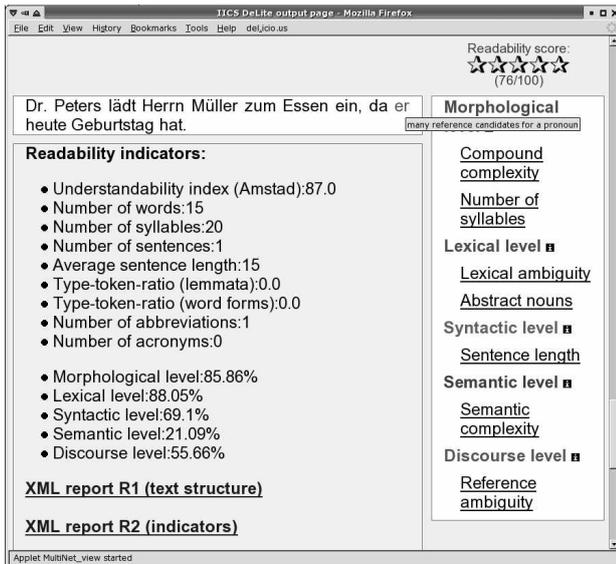| Indicator | Correlation |
|---|---|
| Quality of the SN | 0.360 |
| Passive with AGT | 0.209 |
| Pronoun Reference Distance | 0.203 |
| Number of Propositions per Sentence | 0.201 |
| (Double) Negations | 0.189 |
| Connectivity of Discourse Entities | 0.186 |

Figure 2: Screenshot of DeLite's GUI.

Table 1 we got small errors for the indicators *Number of Concept Nodes in the SN* (0.190), *Number of Introduced Concepts* (0.211) and *Longest Path in the SN* (0.289).

We implemented the readability checker DeLite (Hartrumpf et al., 2006) that calculates a global readability score and highlights (by color) text passages which are difficult to read according to at least one of our indicators (see Figure 2). If the user moves the mouse on such a text passage, the readability problem type will be described briefly. Supplementary highlight segments (if any) are printed in bold typeface if the user clicks on the colored text passage. In the upper right corner, a global readability score is provided which is calculated by a readability formula over all readability indicators. The user communication is realized over a web server which allows it to access our system by any JavaScript-enabled web browser.

## 8. Conclusion and Future Work

We proposed a new kind of readability indicators which are semantic and predominantly operate directly on semantic representations (SNs). We further investigated correlation and absolute errors of these indicators in comparison with user ratings. The evaluation showed that, although the SN could not be constructed for several sentences of our domain-specific corpus, semantic indicators can often yield scores that are more accurate than traditional, surface-oriented readability indicators. Therefore we expect that semantic readability indicators will play an important role for future readability checkers.

## Acknowledgments

## 9. References

Bertsimas, Dimitris and John Tsitsiklis, 1997. *Introduction to Linear Optimization*. Belmont, USA: Athena Scientific.

Drosdowski, Günter (ed.), 1995. *Duden - Grammatik der deutschen Gegenwartssprache*. Mannheim, Germany: Dudenverlag.

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

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

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

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

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

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

Helbig, Gerhard and Fritz Kempter, 1997. *Das Passiv. Zur Theorie und Praxis des Deutschunterrichts für Ausländer*. Berlin, Germany: Langenscheidt.

Helbig, Hermann, 2006. *Knowledge Representation and the Semantics of Natural Language*. Berlin, Germany: Springer.

Langer, Inghard, Friedemann Schulz von Thun, and Reinhard Tausch, 1981. *Sich verständlich ausdrücken*. München, Germany: Reinhardt.

Rascu, Ecaterina, 2006. A controlled language aproach to text optimization in technical documentation. In *Proceedings of KONVENS 2006*. Konstanz, Germany.