All Mixed Up? Finding the Optimal Feature Set for General Readability Prediction and Its Application to English and Dutch

Orphée De Clercq*
Ghent University

Véronique Hoste**
Ghent University

Readability research has a long and rich tradition, but there has been too little focus on general readability prediction without targeting a specific audience or text genre. Moreover, although NLP-inspired research has focused on adding more complex readability features, there is still no consensus on which features contribute most to the prediction. In this article, we investigate in close detail the feasibility of constructing a readability prediction system for English and Dutch generic text using supervised machine learning. Based on readability assessments by both experts and crowdsourcing, we implement different types of text characteristics ranging from easy-to-compute superficial text characteristics to features requiring deep linguistic processing, resulting in ten different feature groups. Both a regression and classification set-up are investigated reflecting the two possible readability prediction tasks: scoring individual texts or comparing two texts. We show that going beyond correlation calculations for readability optimization using a wrapper-based genetic algorithm optimization approach is a promising task that provides considerable insights in which feature combinations contribute to the overall readability prediction. Because we also have gold standard information available for those features requiring deep processing, we are able to investigate the true upper bound of our Dutch system. Interestingly, we will observe that the performance of our fully automatic readability prediction pipeline is on par with the pipeline using gold-standard deep syntactic and semantic information.

1. Introduction

In Western society, the literacy level of the general public is often assumed to be of such a level that adults understand all texts they are confronted with on an average day. Many studies, however, have revealed that this is not the case. In the United States, for

* LT3, Faculty of Arts and Philosophy, Groot-Brittanniëlaan 45, 9000 Ghent, Belgium.
E-mail: orphee.declercq@ugent.be.
** LT3, Faculty of Arts and Philosophy, Groot-Brittaniëlaan 45, 9000 Ghent, Belgium.
E-mail: veronique.hoste@ugent.be.

Submission received: 11 September 2014; revised version received: 2 November 2015; accepted for publication: 28 February 2016.

doi:10.1162/COLI_a_00255

example, the 2003 National Assessment Adult Literacy showed that only 13% of adults were maximally proficient in understanding texts they encounter in their daily life. The European Commission has also been involved in extensive investigations of literacy after research had revealed that almost one in five adults in the European society lack the literacy skills to successfully function in a modern society (Wolf 2005).

Every day we are confronted with all sorts of texts, some of which are easier to process than others. Moreover, it seems that the documents that are potentially the most important for adult readers are also the more difficult ones to process, such as mortgage files, legal texts, or patient information leaflets. According to a recent OECD study where the literacy of adults from 23 Western countries or regions was rated on a five-point scale, these specific texts genres all require a literacy level of at least four. The findings of this study for participants from the Dutch language area show that only 12.4% of adults in Flanders and 18.2% in the Netherlands reach the two highest levels of proficiency (OECD 2013).

Readability research and the automatic prediction of readability has a very long and rich tradition (see surveys by Klare 1976; DuBay 2004; Benjamin 2012; and Collins-Thompson 2014). Whereas superficial text characteristics leading to on-the-spot readability formulas were popular until the last decade of the previous century (Flesch 1948; Gunning 1952; Kincaid et al. 1975), recent advances in the field of computer science and natural language processing have triggered the inclusion of more intricate characteristics in present-day readability research (Si and Callan 2001; Collins-Thompson and Callan 2005; Schwarm and Ostendorf 2005; Heilman, Collins-Thompson, and Eskenazi 2008; Feng et al. 2010). The bulk of these studies, however, have focused on readability as perceived by specific groups of people, such as children (Schwarm and Ostendorf 2005), second language learners (François 2009), or people with intellectual disabilities (Feng et al. 2010), and on the readability of texts in specific domains, such as the medical one (Leroy and Endicott 2011). The investigation of the readability of a wide variety of texts without targeting a specific audience has not received much attention (Benjamin 2012).

Moreover, when it comes to current state-of-the art systems, it can be observed that even though more complex features trained on various levels of complexity have proven quite successful when implemented in a readability prediction system (Pitler and Nenkova 2008; Feng et al. 2010; Kate et al. 2010), there is still no consensus on which features are actually the best predictors of readability. As a consequence, when institutions, companies, or other research disciplines wish to use readability prediction techniques, they still rely on the more outdated superficial characteristics and formulas (see, for example, the recent work by van Boom [2014] on the readability of mortgage terms).

In this article, we investigate the creation of a fully automatic readability assessment system that can assess generic text material in two languages, English and Dutch. We use a supervised machine learning approach and investigate both a regression and classification set-up reflecting the two possible readability prediction tasks: scoring individual texts or comparing two texts. This requires general evaluation corpora of English and Dutch generic text comprising various text genres and levels of readability. As well as a suitable corpus, the investigation also requires a methodology to assess readability: In this respect, we were the first to explore crowdsourcing as an alternative to using expensive expert labels (De Clercq et al. 2014).

In our system various text characteristics have been implemented ranging from easy-to-compute superficial text features to features requiring deep linguistic processing. We investigate to what extent automatically derived features can be considered

optimal for predicting readability in both languages under consideration. We envisage finding the optimal mix of these readability predictors by exploiting a wrapper-based approach to feature selection using a genetic algorithm. We will show that going beyond correlation calculations for readability optimization using genetic algorithms is a promising task that provides considerable insights in which feature combinations contribute to the overall readability prediction.

Another aspect of this research is to investigate in closer detail the contribution of those features requiring deep linguistic processing. Though many advances have been made in NLP, the more difficult text-understanding tasks still achieve moderate performance rates. Think, for example, of coreference resolution where a combined F-measure of 60% is considered state-of-the-art.[1] Implementing such features in a full-fledged readability prediction system is thus risky as the automatically derived features might not truly represent the information at hand. Because we have gold standard deep syntactic and semantic information available for our Dutch readability data set, we were able to investigate in close detail its added value in predicting readability. Interestingly, we will observe that the performance of our fully automatic readability prediction pipeline is on par with the pipeline using gold-standard deep syntactic and semantic information.

The remainder of this article is organized as follows. After describing the related research with a specific focus on features that have been used in previous readability research (Section 2), we explain in Section 3 how the English and Dutch data were collected and assessed. Section 4 describes the methods used to perform the actual optimization experiments, the results of which are described and analyzed in Section 5. We end with a concluding general discussion in Section 6.

2. Related Work

What makes a particular text easy or difficult to read has been the central question in reading research over the past century. There seems to be a consensus that readability depends on complex language comprehension processes between a reader and a text (Davison and Kantor 1982; Feng et al. 2010). This implies that reading ease can be determined by looking at both intrinsic text properties as well as aspects of the reader. Since the first half of the 20th century, however, readability formulas have been developed to automatically predict the readability of an unseen text based only on superficial text characteristics such as the average word or sentence length. Over the years, many objections have been raised against these traditional formulas: their lack of absolute value (Bailin and Grafstein 2001), the fact that they are solely based on superficial text characteristics (Davison and Kantor 1982; DuBay 2004, 2007; Feng, Elhadad, and Huenerfauth 2009; Kraf and Pander Maat 2009), the underlying assumption of a regression between readability and the modeled text characteristics (Heilman, Collins-Thompson, and Eskenazi 2008), and so forth. Furthermore, there seems to be a remarkably strong correspondence between the readability formulas themselves, even across different languages (van Oosten, Tanghe, and Hoste 2010).

These objections have led to new quantitative approaches of doing readability prediction that adopt a machine learning perspective to the task. Advancements in these fields have introduced more intricate prediction methods such as naive Bayes classifiers (Collins-Thompson and Callan 2004), logistic regression (François 2009) and

1 See the results of the CoNLL-2011 Shared Task at `http://conll.cemantix.org/2011/`.

support vector machines (Schwarm and Ostendorf 2005; Feng et al. 2010; Tanaka-Ishii, Tezuka, and Terada 2010), and especially more complex features ranging from lexical features over syntactic to semantic and discourse features.

The **vocabulary** used in a text largely determines its readability (Alderson 1984; Pitler and Nenkova 2008). Until the millennium, lexical features were mainly studied by counting words, measuring lexical diversity using the type token ratio, or by calculating frequency statistics based on lists (Flesch 1948; Kincaid et al. 1975; Chall and Dale 1995). In later work, a generalization over this list look-up was made by training unigram language models on grade levels (Si and Callan 2001; Collins-Thompson and Callan 2005; Heilman et al. 2007). Subsequent work by Schwarm and Ostendorf (2005) compared higher-ordered *n*-gram models trained on part-of-speech sequences with those using information gain and found that the latter gave the best results. To this purpose they used two paired corpora (one complex and one simplified version) to train their language models. Using the same corpora, these findings were corroborated by Feng et al. (2010) when they investigated readability targeted to people with intellectual disabilities. These results were thus achieved when training and testing different language models that are built on various levels of complexity. Pitler and Nenkova (2008) were the first to train language models using background material complying with the genre the readability of which they were trying to assess (newspaper text). Kate et al. (2010) conducted similar experiments, but they used higher-ordered language models and normalized over document length. In subsequent work as well, language models have proven a successful technique for readability prediction (Feng et al. 2010; François 2011).

In addition, the structure or **syntax** of a text is seen as an important contributor to its overall readability. Because longer sentences have proven to be more difficult to process than short ones (Graesser et al. 2004), this traditional feature also persists in recent work (Feng et al. 2010; Nenkova et al. 2010; François 2011). Schwarm and Ostendorf (2005) were the first to introduce more complex syntactic features based on parse trees, such as the parse tree height, phrase length (NP, PP, VP), and the amount of subordinating conjunctions. Nenkova et al. (2010) were the first to study structural features in isolation and introduced some additional syntactic features that should be able to reflect sentence fluency. According to their findings particularly, features encoding the length of both sentences and phrases emerge as important readability predictors. POS-based features, which are less difficult to compute, have also been used and have proven to be effective, too (Heilman et al. 2007), especially features based on noun and preposition word class information (Feng et al. 2010) or features representing the amount of function words present in a text (Leroy et al. 2008). Overall, Schwarm and Ostendorf's parse tree features have been reproduced frequently and were found effective when combined with *n*-gram modeling (Heilman et al. 2007; Petersen and Ostendorf 2009; Nenkova et al. 2010) and discourse features (Barzilay and Lapata 2008).

This brings us to a final set of features, namely, those relating to **semantics**, which has been a popular focus in modern readability research (Pitler and Nenkova 2008; Feng et al. 2010; François 2011). Whereas the added value of the lexical and syntactic features has been corroborated repeatedly in the computational approaches to readability prediction that have surfaced in the last decade, it has proven much more difficult to unequivocally determine the added value of semantic features. Capturing semantics can be done from two different angles. The first angle relates to features that are used to describe semantic concepts. The complexity and density with which concepts are included in a text can be studied by looking at the actual words that are used to describe

these. Complexity was investigated in the framework of the Coh-Metrix by calculating the level of concreteness or lexical ambiguity of words against a database (Graesser et al. 2004). The validity of this approach for readability research, however, was not further investigated. Density was calculated by Feng et al. (2010) by performing entity recognition and has proven a useful feature in her work.

A second angle is to investigate how these concepts are structured within a text—for example, finding semantic representations of a text or elements of textual coherence. In this respect, reference can be made to both local and global coherence, which translates to looking at the coherence between adjacent sentences (local) and then extrapolating this knowledge to reveal something about the overall textual coherence (global). This type of semantic representation can also be referred to as discourse analysis. An intuitive and straightforward way to implement this is to simply count the number of connectives included in a text based on lists or to calculate the causal cohesion by focusing on connectives and causal verbs (Graesser et al. 2004). A similar approach is to compute the actual word overlap. This word overlap was introduced without further investigations in the Coh-Metrix in three ways: noun overlap, argument overlap, and stem overlap (Graesser et al. 2004). Subsequent readability research by Crossley, Greenfield, and McNamara (2008) looked only at content overlap and showed it to be a significant feature. However, similar work by Pitler and Nenkova (2008) did not lead to the same conclusion. The first study to actually investigate the validity of the Coh-Metrix as a readability metric concluded that noun overlap can be indicative of causal and nominal coreference cohesion, which in turn allows to distinguish between coherent and incoherent text (McNamara et al. 2010).

More intricate methods are also available based on various techniques. A first technique is to use latent semantic analysis (LSA).This technique was first introduced in readability research by Graesser et al. (2004) under the form of local and global LSA in the Coh-Metrix but not further investigated. The first to measure the impact of modeling local LSA for readability prediction were Pitler and Nenkova (2008); they found that the average cosine similarity between adjacent sentences was not a significant variable. Also, the validity of LSA as implemented in the Coh-Metrix could not be corroborated in the previously mentioned study by McNamara et al. (2010). François (2011) was the first to study LSA in greater detail, which seemed very helpful for his readability research for second language learners, but in more recent work his approach was criticized because of the specificity of the corpus used (Todirascu et al. 2013).

An alternative to LSA was introduced by Barzilay and Lapata (2005). They define three linguistic dimensions that are essential for accurate prediction: entity extraction, grammatical function, and salience. These three dimensions are combined in the entity-grid model they propose in which all entities can be defined in a text on a sentence-to-sentence basis and where the transitions are checked for each sentence. Their main claim is that salient entities prefer prominent over non-prominent syntactic positions within a clause and are more likely to be introduced in a main clause than in a subordinate clause. Though originally devised for other research purposes, they found that the proportion of transitions in this entity grid model results in predicting the readability of a text in combination with the syntactic features as introduced by Schwarm and Ostendorf (2005). Subsequent work by Pitler and Nenkova (2008) compared this entity grid model with the added value of discourse relations as annotated in the Penn Treebank (Prasad et al. 2008). They treat each text as a bag of relations rather than a bag of words and compute the log likelihood of a text based on its discourse relations and text length compared to the overall treebank. They found that these discourse relations are indeed good in distinguishing texts, especially when combined with the

entity grid model. Because these discourse relations were only based on gold standard information whereas, in the end, a readability prediction system should be able to function automatically, Feng et al. (2010) proposed an alternative that should be able to compute this type of information. Besides entity-density and entity-grid features, they introduced features based on lexical chains that try to find relations between entities (such as synonym, hypernym, hyponym, coordinate terms [siblings], etc. [Galley and Mckeown 2003]). Moreover, they incorporated coreferential inference features in order to study the actual coherence between entities. However, this study did not come to a positive conclusion for incorporating these types of features. In a follow-up study, Feng et al. (2010) found that enlarging the corpus, which exclusively consisted of texts for primary school children, with more diverse text material allowed for an overall better performance. However, the added value of the discourse relations to the system was still not significant.

We can conclude that the introduction of more complex linguistic features has indeed proven useful. However, the discussion on which features are the best predictors remains open. Although Pitler and Nenkova (2008) have clearly demonstrated the usefulness of discourse relations, the predictive power of these was not corroborated by, for example, Feng et al. (2010). Nevertheless, we can deduce from previous research that features that are lexical in nature, such as language modeling features, have a strong predictive power. Many studies are also difficult to compare because they all use their own definition of readability and corpora to measure readability. Furthermore, we see that most studies focus on human judgments by, for example, people with specific disabilities, or that they work with corpora of texts targeting a specific audience (mostly language learners). The work of Feng et al. (2010), for example, is very valuable thanks to its focus on discourse features while including features from previous work, but their main focus is on texts aimed at primary school students. A similar observation can be made about the work of François (2011), who investigated a wide variety of current state-of-the-art readability features, but focused on second language learners. We envisaged from the beginning building a corpus that consists of texts adult language users are all confronted with on a daily basis.

3. Data Collection

In order to build an *unbiased* readability system, one which is not targeted towards a specific audience or trained on highly specific text material only, we needed to select texts that adult language users are all confronted with on a regular, daily basis. To this purpose, we collected comparable English and Dutch text snippets taken from reference corpora. For English, we selected snippets from the British National Corpus (Aston and Burnard 1998), the English part of the Dutch Parallel Corpus (Macken, De Clercq, and Paulussen 2011), and Wikipedia.[2] For Dutch, we used the corpus collected by De Clercq et al. (2013) that incorporates texts from the SoNaR corpus (Oostdijk et al. 2013), which has recently been enriched with semantic information (De Clercq, Monachesi, and Hoste 2012). Some data statistics are presented in Table 1. Both data sets consist of 105 texts each, contain data from different genres in order to represent a variety of text material and presumably also various readability levels. The *administrative* genre comprises reports and survey or policy documents written within companies or institutions.

2 www.wikipedia.org.

Table 1
Data statistics of the English (En) and Dutch (Du) part of the readability corpus.

Genre	# En docs	# En tokens	# Du docs	# Du tokens
Administrative	21	6,466	21	3,463
Informative	64	17,090	65	8,950
Instructive	9	2,011	8	1,108
Miscellaneous	11	2,311	11	1,559
Total	105	27,878	105	15,080

The texts falling under the *informative* genre can be described as current affairs articles in newspaper or magazines and encyclopedic information such as Wikipedia entries. The *instructive* genre consists of user manuals and guidelines. Finally, the *miscellaneous* genre covers other text genres such as very technical texts and children's literature. We acknowledge that including multiple genres might influence our final training system in that it only learns to distinguish between various genres instead of various readability levels. To account for this as much as possible, we carefully tried to select texts of varying difficulty for each text genre (see De Clercq et al. [2014] for more information).

For the actual assessment, we were inspired by DuBay's (2004) vision on readability, notably, "what is it that makes a particular text easier or more difficult to read than any other text," which means that we assessed readability by comparing texts with each other.

Deciding how readability will be assessed is not a trivial task and there exists no consensus on how this should be done. In modern readability research, we see that most readability data sets consist of graded passages, that is, the texts have received a grade level or absolute difficulty score typically assigned by experts (Collins-Thompson 2014). Consulting these experts or language professionals is both time- and money-consuming, which might explain the increasing success of using cheaper and non-expert contributors over the Web, also known as crowdsourcing (Sabou, Bontcheva, and Scharl 2012).

The task of assigning readability assessments to texts, however, is quite different from annotation tasks where a set of predefined guidelines have to be followed. Readability assessment remains largely intuitive, even in cases where annotators are instructed to pay attention to syntactic, lexical, or other levels of complexity. But then again, this lack of large sets of guidelines might be another motivation to use crowdsourcing instead. This is why we explored two different methodologies to collect readability assessments for our corpora—namely, a more classical expert labeling approach, in which we collect assessments of language professionals, and a lightweight crowdsourcing approach. For more details we refer readers to De Clercq et al. (2014).

The experts are language professionals (language teachers, linguists) that were asked to rank the texts on a scale from 0 (easy) to 100 (difficult). These experts were asked to assess the readability for language users in general. We deliberately did not ask more detailed questions about certain aspects of readability because we wanted to avoid influencing the text properties experts pay attention to. Neither did we inform the experts in any way on how they should judge readability. Any presumption about which features should be regarded as important readability indicators was thus avoided. However, in order to have some idea about their assessment rationale the

experts were offered the possibility to motivate or to comment on their assessments via a free text field. Our pools consisted of 23 English and 36 Dutch experts who ranked 3,736 and 2,564 texts, respectively.

The crowd, on the other hand, consisted of nonprofessionals who were asked to sort text pairs using a five-point scale (see Table 2). As was done for the experts, we gave no further instructions because we did not want to influence anyone on how to perceive readability. Everyone participating in the crowd assessments remained anonymous. In the start-up phase, the crowdsourcing was widely advertised among friends, family, and so forth, which might have caused a bias towards more educated labelers, but we can nevertheless state that the assessors participating in the crowd differ from the experts. In total, 8.297 English and 11,038 Dutch text pairs were assessed.

Using the same techniques as described in De Clercq et al. (2014), the information collected through both assessor groups was converted into assessed text pairs, resulting in 27,323 English and 23,908 Dutch assessed expert text pairs and the above-mentioned numbers of assessed crowd pairs. A comparison of the English data sets reveals some interesting similarities, as illustrated in Figure 1.

In this figure, the proportions with which each text has been assessed as easier, equally readable, or harder for both the experts and crowd data set is shown. Each dot in the figures represents one text, so every plot in both figures represents the 105 assessed texts. If we take, for example, text 105, we see that this text has been assessed in our Experts data set 0.63 times as easier, 0.29 times as equally difficult, and 0.07 times as more difficult than any other text. In our Crowd data set the same text has been assessed 0.62 times as easier, 0.28 times as equally difficult, and 0.09 times as more difficult than any other text. Overall, we observe that all plots show great similarity for both data sets.

If we calculate the Pearson correlation, we find that the correlation between both groups regarding the easier texts is 90.9% and 89.7% when we look at the number

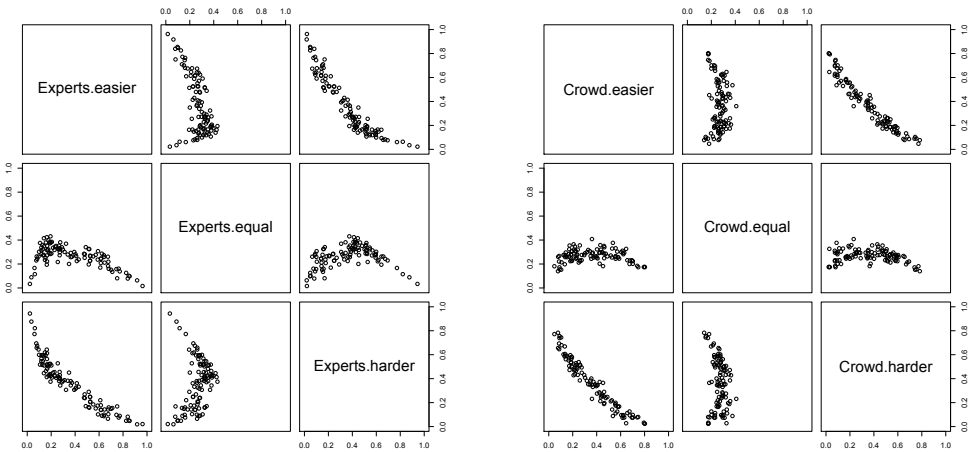

Figure 1
Scatter plots of the English data sets in which the proportion of times is marked each text was assessed as easier, equally difficult, or harder than any other text: (a) for the *Experts* and (b) *Crowd* data. The plots in the lower left triangle are transposed versions of those in the upper right triangle (the x and y axes are switched) and thus present the same information.

Table 2
Total amount of text pairs for each of the five scales.

Acronym	Meaning	Value	#EN pairs	#DU pairs
LME	left text much easier	100	310	260
LSE	left text somewhat easier	50	2,836	2,782
ED	both texts equally difficult	0	4,615	4,836
RSE	right text somewhat easier	−50	2,836	2,782
RME	right text much easier	−100	310	260

of times a text was considered harder.[3] The strong correlations between our Experts and Crowd data sets made us confident that we could combine both data sets for the experiments. This led to an English data set comprising 27,323 and a Dutch one comprising 34,946 assessed text pairs. Considering that for each language we had 105 texts as input corpus, the maximum number of assessed text pairs that can exist within a data set is 10,920 pairs (i.e., every text in the corpus being compared to every other text, viz. 105×104). To this purpose, we averaged all text pairs that were assessed multiple times, and this resulted in 10,907 English and 10,920 Dutch text pairs, as presented in Table 2. In order to be able to calculate such an average value, every assessment label was assigned a corresponding value. The assessment label LME, for example, means that the left text is much easier than the right text which corresponds to this pair receiving the value 100 (left text minus right text, i.e., 100 – 0). Because every text pair has been included in both directions, the experimental corpus shows an even distribution.

4. Experiments

We performed two learning tasks reflecting the two possible readability prediction set-ups: a regression task in which an absolute score is predicted for a given text and a classification task in which two text are compared to each other.

In this section, we will discuss the types of text characteristics we implemented and how we assessed their added value by exploiting a wrapper-based approach to feature selection using genetic algorithms. Finally, we will give an overview of the full experimental pipeline.

4.1 Information Sources

In an attempt to determine the optimal mix of readability predictors, we implemented different types of text characteristics, ranging from traditional to semantic and discourse features. We selected the features to be implemented in our readability prediction system on the basis of the existing literature on the topic (see Section 2) and the various comments left by our expert assessors. The scrutiny of these comments allowed us to discover some interesting tendencies with respect to which text characteristics guided their assessments most. Although the experts did not receive any guidelines on which characteristics to take into consideration when assessing readability, most

3 In De Clercq et al. (2014) we revealed similar results for the Dutch data sets, that is, correlations of respectively 86% and 90%.

assessors commented on their assessments in a similar manner. These comments can be categorized into four groups, as illustrated in Figure 2. The first class includes all comments relating to **Vocabulary** in some way or another, including comments relating to lexical familiarity ("text is full of difficult economics words which might be unknown to a layman") or the level of concreteness ("too many abstract words"). A second class, **Structure**, includes comments relating to syntactic constructs ranging from superficial characteristics ("The sentences are way too long, they should be divided into smaller parts") to complaints about more complex structures ("The complex grammatical structure hinders reading"). The third class groups all comments that relate to the **Coherence** of the overall discourse and again ranges from simple ("The reasoning in this text is not logical; where are the linking words?") to more complex issues ("Every sentence refers to an element of the previous sentence which causes confusion"). Finally, the **Other** class contains all those comments that could not be grouped under a certain linguistic category ("I had to read the text twice").

We observe that in both languages vocabulary is the most important obstructor or facilitator of text readability: It accounts for almost half of all comments, indicating that lexical features are indeed crucial when trying to predict readability (i.e., 47% and 41% of the English and Dutch comments, respectively). However, the syntactic (17% and 18%) and semantic (11% and 14%) aspects of a text should not be ignored either. What also draws the attention is the rather elaborate Other category, accounting for 25% of the comments in both languages. It is difficult to attribute these comments to one particular characteristic; sometimes they hint at layout problems, sometimes at the cognitive load. At this point of our research, we focus on linguistic characteristics. We implemented various lexical, syntactic, and semantic features in our readability prediction system. Furthermore, we also decided to integrate more "traditional" lexical and syntactic features—those that are used in the classical readability formulas—as a separate group because they have proven good predictors of readability in addition to the NLP-inspired features (Pitler and Nenkova 2008; François 2011). In total, we encoded no fewer than 87

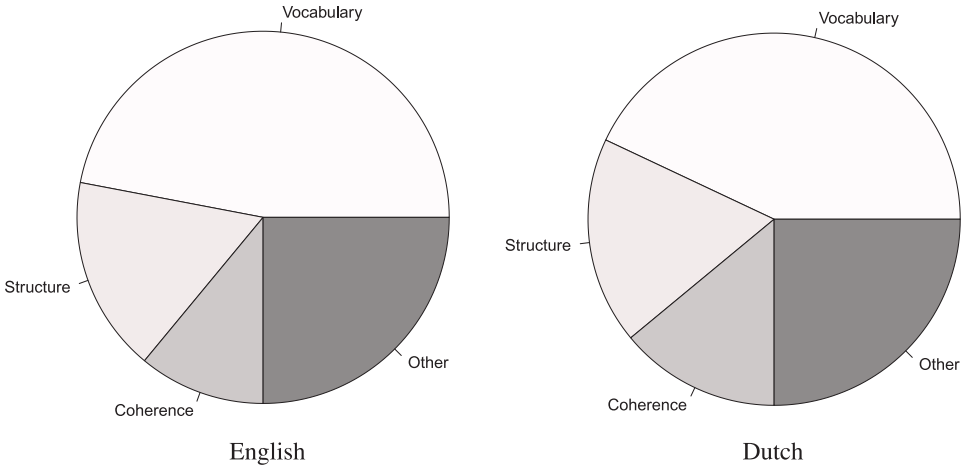

Figure 2
Pie charts representing the importance added to the various feature groups by the expert assessors based on their individual comments for both the English (left) and Dutch (right) data set.

Table 3
Overview of all features that were implemented for the readability prediction tasks divided into various subgroups.

Traditional	*tradlen*	4	**Semantic**	*shallowsem*	12
	tradlex	2		*ner*	7
				coref	5
Lexical	*lexlm*	2		*srl*	20
	lexterm	2			
Syntactic	*shallowsynt*	27			
	deepsynt	6			

distinct features, which were all computed on the document level using state-of-the-art text processing tools. A schematic overview can be found in Table 3.

– **Traditional features**: We included four length-related features (*tradlen*) that have proven successful in previous work (Feng et al. 2010; Nenkova et al. 2010; François and Miltsakaki 2012): the average word and sentence length, the ratio of long words in a text (i.e., words containing more than three syllables), and the percentage of polysyllable words. We also incorporated two traditional lexical features (*tradlex*): the percentage of words that can be found in the Chall and Dale list (1995) for the English texts or in the CLIB list (Staphorsius 1994) for the Dutch texts.[4] We also calculated the type token ratio to measure the level of lexical complexity within a text. All these features were obtained after processing the text with a state-of-the-art English (LeTs; Van de Kauter et al. 2013) and Dutch (Frog; van den Bosch et al. 2007) preprocessor and a designated classification-based syllabifier (van Oosten, Tanghe, and Hoste 2010).

– **Lexical features**: Because we envisaged having no presupposition on the various levels of complexity in our corpus, we decided to build two generic language models, one for English based on the written part of the BNC corpus (Aston and Burnard 1998) and one for Dutch based on a subset of the SoNaR corpus (Oostdijk et al. 2013) containing only newspaper, magazine, and Wikipedia material. These language models were built up to an order of 5 ($n = 5$) with Kneser-Ney smoothing using the SRILM toolkit (Stolcke 2002). As features (*lexlm*), we calculated the perplexity of a given text when compared with this reference data and also normalized this score by including the document length, as seen in Kate et al. (2010). Besides these n-gram models, which have proven strong predictors of readability in previous work (Feng et al. 2010; Kate et al. 2010; François 2011), we also introduced two other metrics that were calculated using the same reference corpora (*lexterm*). Inspired by terminological work, we included the Term Frequency-Inverse Document Frequency, aka tf-idf (Salton 1989) and the Log Likelihood (Rayson and Garside 2000) ratio of all terms included in a particular text.

– **Syntactic features**: We incorporated two types of syntactic features: a shallow level where all features are computed based on PoS-tags (*shallowsynt*) and a deeper level based on dependency parsing (*deepsynt*). We included 25 shallow features, inspired by Feng et al. (2010), relating to the five main part-of-speech classes: nouns, adjectives, verbs, adverbs, and prepositions. For each class, we indicated their absolute and relative frequency in the text, in the sentence and the average type per sentence. In addition, we calculated two additional features, the average number of content and function words

4 Both lists contain words that are particularly frequent in the respective languages.

within a text (Leroy et al. 2008). For these calculations, the same preprocessor tools were used as mentioned above. For the deep syntactic features, we incorporated the parse tree features as first introduced by Schwarm and Ostendorf (2005) that have proven successful in many other studies (Pitler and Nenkova 2008; Petersen and Ostendorf 2009; Feng et al. 2010; Nenkova et al. 2010). We calculated the parse tree height, the number of subordinating conjunctions, and the ratio of the noun, verb, and prepositional phrases. We also included the average number of passive constructions in a text. The parsers underlying these features were the Stanford parser (de Marneffe, MacCartney, and Manning 2006) for English and the Alpino parser (van Noord et al. 2013) for Dutch.

– **Semantic features**: Because connectives serve as an important indication of textual cohesion in a text (Halliday and Hasan 1976; Graesser et al. 2004), we integrated several features based on a list look-up of connectives (*shallowsem*). The English and Dutch lists were drawn up by linguistic experts. As features, we counted the average number of connectives within a text and the average amount of causal, temporal, additive, contrastive, and concessive connectives on both the sentence and document level. As named entity information provides us with a good estimation of the amount of world knowledge required to read and understand a particular text, we calculated the number of entities and unique entities and the number of entities on the sentence level, and we made a comparison between predicted named entities (that is, recognized by a NER system) and shallow entities (based on PoS-tags [*ner*]). For English, we used the Stanford NER (Finkel, Grenager, and Manning 2005) and for Dutch the NERD system (Desmet and Hoste 2013). Coreferential relations, then, might indicate how structured and thus how coherent a particular text is. We represented as features the number of coreferential chains present in a text, the average length of a chain, the average number of coreferring expressions and unique mentions, and we also count how many chains span more than half of the text (*coref*). To this purpose, we used the Stanford Coreference Resolver (Lee et al. 2013) for English and COREA (De Clercq, Hendrickx, and Hoste 2011) for Dutch. In order to determine how many agents or modifiers a particular text contains, we also calculated the average number of arguments and modifiers and the average occurrence of every possible PropBank label (Palmer, Gildea, and Kingsbury 2005) (*srl*). For the construction of these features, we used the English semantic role labeler (SRL) as part of the Mate-Tools (Björkelund, Hafdell, and Nugues 2009) and for Dutch the SoNaR SRL (De Clercq, Monachesi, and Hoste 2012).

Both the entity and coreference features were tested before in the work of Feng et al. (2010). They found that none of these features possesses a high predictive power for readability research which was mainly because of the low performance of the individual tools used for making them. As the text material that was selected for our Dutch data set was drawn from the SoNaR corpus, which was enriched with manual dependency tree, named entity, coreference, and semantic role semantic information, we are able to work with gold-standard information and can thus assess for Dutch the upper bound impact of including these different types of information.

4.2 Two Prediction Tasks

For our experiments, we considered two readability prediction tasks: regression and classification.

- In the case of **regression**, the task consists in assigning an absolute readability score to a given text. For the regression task, the text pairs from Table 2 are turned into individual texts which receive an absolute score by

calculating how many times each particular text is labeled as much or somewhat easier in comparison to other texts and by dividing this by the total number of times this text appears as part of a text pair.

- In the **classification** set-up, we defined two subtasks: a binary classification task in which we determine for a given text pair whether text *a* is easier or more difficult than text *b* and a multiclass classification task where multiple classes have to be predicted representing the five possible readability values between two texts. The 10,908 English text pairs and 10,920 Dutch text pairs can be used as such for the multiclass classification. For the binary experiments, we excluded all equally difficult pairs and put together the much and slightly easier or more difficult text pairs, leading to reduced data sets of 6,922 English and 6,084 Dutch text pairs.

All experiments were conducted using support vector machines (SVMs), and more specifically the LibSVM[5] implementation which supports both support vector regression and support vector classification. In preliminary experiments, we also tested two other machine learning methods, CRF and TiMBL, but SVMs were found superior.

We evaluated the performance of our regression experiments with the root mean squared error (*RMSE*) as the error to be optimized:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - x_i)^2}$$

in which X_i is the prediction and x_i the response value, that is, the correct value, for the regression task at hand, and m is the number of texts for which a prediction is made. The lower the RMSE value, the better.

Given the even distribution of the data, the classification tasks were evaluated in terms of accuracy:

$$accuracy = \frac{\text{true positives} + \text{true negatives}}{\text{total number of instances}}$$

4.3 Exploring the Optimal Feature Mix

The selection of relevant features and the elimination of the irrelevant features is an important problem in machine learning. Most inductive methods incorporate some type of feature selection or feature weighting to distinguish between the informativeness of the features and to measure their relevance in a given learning task, in our case readability prediction. Apart from assigning weights or degrees of informativeness to the different features, it is also possible to eliminate the non-informative features, thus creating a feature subset of the most informative features. There are two main types of feature selection techniques, namely, filter and wrapper approaches (Aha and Bankert 1996). The **filter approach** uses an evaluation function (e.g., mutual information or Pearson correlation) for determining feature relevance and selects the best features

5 `http://www.csie.ntu.edu.tw/cjlin/libsvm`.

independently of the performance of the learning algorithm. The assumption is that features should have a strong correlation with the target class. It is common practice in readability research to measure the correlation between textual features and the human assessments (Pitler and Nenkova 2008; François 2011). It has also been shown, however, that the features considered most predictive in classification experiments do not necessarily overlap with those having the highest correlation (Pitler and Nenkova 2008). We will come back to this observation in Section 5.

In a **wrapper approach**, on the other hand, feature informativeness is determined while running some induction algorithm on a training data set and the best features are selected in relation to the problem (e.g., readability prediction) to be solved. Finding a good subset of features requires searching the space of feature subsets. However, as an exhaustive or greedy search of this space is often practically impossible—because this implies searching 2^n possible subsets for n attributes, other more realistic approaches have been explored to search the space of possible feature combinations. Techniques such as forward selection, backward elimination (John, Kohavi, and Pfleger 1994), and bidirectional hillclimbing (Caruana and Freitag 1994) differ in the point where they start their search, but all share the potential problem of convergence to a local optimum. In the case of genetic algorithms (GAs) search does not start from a local search point, but from a population of individuals, thus exploring different areas of the search space in parallel (and it also allows multiple optima). Genetic algorithms for feature selection in readability prediction have, for example, been used by Falkenjack and Jonsson (2014) to determine the added value of syntax features for Swedish readability prediction.

Because, besides feature selection, changing the hyperparameters of an algorithm can also have a dramatic effect on classifier performance (Hoste 2005; Desmet 2014) and should be determined experimentally, we chose to use GAs as a computationally feasible way to tackle this optimization problem, which involves searching the space of all possible feature subsets and parameter settings to identify the combination that is optimal or near-optimal.

Genetic algorithms (see Goldberg [1989] and Mitchell [1996] for more information) are search methods based on the mechanics of natural selection and genetics. They require two things: fitness-based selection and diversity. Central principles in genetic algorithms are selection, recombination, and mutation. As illustrated in Figure 3, the principle behind GAs is quite simple: search starts from a population of individuals, which all represent a candidate solution to the optimization problem to be solved. These individuals are typically represented as a bit string of fixed length, called a "chromosome" or "genome." In our experiments, the individuals are represented as bit strings. Each individual contains particular values for all algorithm parameters (e.g., RBF) and for the selection of the features (0 or 1). A possible value of a bit is called an "allele." The population of chromosomes has a predefined size. Larger population sizes increase the amount of variation present in the population at the expense of requiring more fitness function evaluations. To decide which individuals will survive into the next generation, a selection criterion is applied defining how good the individual is at solving the problem—its fitness. For our experiments, we run 10-fold cross-validation on the training data and use the resulting performance values, RMSE for regression and accuracy for classification, as the fitness scores to be optimized. After the fitness assignment, a selection method determines which individuals in the parent generation will survive and produce offspring for the next generation. We used the common technique of tournament-based selection (Goldberg and Deb 1991). Here, a fixed number of individuals is randomly picked from the population to compete in a tournament,

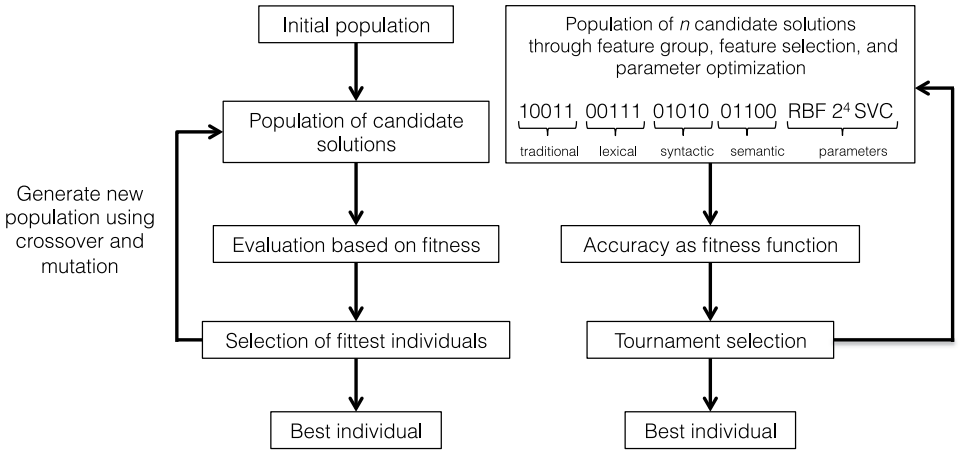

Figure 3
Feature selection using a genetic algorithm approach. The left-hand side of the figure illustrates the general procedure, and the right-hand side translates this GA search to our task of readability prediction.

where an individual's probability of winning is proportionate to its fitness. The winner is selected as parent. This process is repeated as many times as there are individuals to be selected. Unless the stopping criterion is reached at an earlier stage, optimization stops after a predefined set of generations. In order to combine effective solutions and maintain diversity in the population, chromosomes are combined or mutated to breed new individuals. The mutation operator forms a new chromosome by making alterations to the information contained in the genome of a parent according to a given probability distribution, expressed in the mutation rate. Crossover is an operator which creates an offspring's chromosome by joining segments chosen alternately from each of two parents' chromosomes which are of fixed length. This crossover reproduction is performed with a certain probability: the crossover rate which can vary between 0 (no crossover) and 1 (crossover always applies).

4.4 Experimental Set-up

After setting our baseline, in which we use all available features and the default hyperparameter settings of LibSVM for both the regression and classification readability prediction tasks, we performed two rounds of optimization experiments. In both optimization set-ups, we allowed 100 generations and set the stopping criterion to a best fitness score that remained the same during the last five generations. The mutation rate was set to 0.3 and we applied single-point crossover with a probability of 0.9.

- *Round 1: feature selection*, allowing variation between the features in two different setups, while relying on LibSVM's default hyperparameters.
  - In the first set-up, we perform *feature group selection* by splitting the feature set in ten feature groups (i.e., *tradlen*, *tradlex*, *lexlm*, *lexterm*, *shallowsynt*, *deepsynt*, *shallowsem*, *ner*, *coref*, and *srl*). Here we start from a population of 100 individuals.

  - In the second set-up, we freeze the features within the *tradlen, tradlex, lexlm, lexterm, shallowsynt*, and *shallowsem* groups and allow *individual feature selection* among the features requiring deeper linguistic processing (*deepsynt, ner, coref*, and *srl*). Here, our search space starts from a population of 300 individuals to allow sufficient variation.

- *Round 2: combined hyperparameter and feature selection*, in which we again discern two different set-ups: one focusing on feature groups and starting from 100 individuals and one where we allow individual feature selection and start from a population of 300 individuals.
  Three different LibSVM types were chosen for our two prediction tasks: For the classification we worked with C-SVC and for the regression we allowed both epsilon-SVR and nu-SVR. As to the hyperparameter optimization, when using SVMs, much depends on which kernel you decide to use to weigh the training instances in the new feature space (see Cristianini and Shawe-Taylor [2000] for an in-depth discussion). In LibSVM, four different kernels can be used: the default Gaussian radial basis function (RBF) or a linear, polynomial, or sigmoid kernel. For the linear kernel, no additional kernel-specific parameters have to be set; the ones that were varied for the other three kernel functions are summarized in Table 4 together with how they were configured for our purposes. Besides these kernel-specific settings, we configured the other hyperparameters as follows:

  - We used the soft margin method to allow training errors when constructing the decision boundary, and vary the associated cost parameter C between 2^{-6} and 2^{12}, stepping by a factor of 4 (*default = 1*).

  - Shrinking heuristics are always used, which is also the *default* option. Shrinking is a technique to reduce the training time: By identifying and removing some bounded elements in the optimization problem, it becomes smaller and can be solved in less time.

  - The stopping criterion or ϵ is set to the *default of 0.001*. Because the optimization method only asymptotically approaches an optimum, it is terminated after satisfying this stopping condition.

  - For epsilon-SVR the epsilon in the loss function was allowed to vary between 0.1 and 1.0, in steps of 0.1 (*default = 0.1*).

All optimization experiments are performed using the Gallop toolbox (Desmet and Hoste 2013). Gallop provides the functionality to wrap a complex optimization problem as a genome and to distribute the computational load of the GA run over multiple processors or to a computing cluster. It is specifically aimed at problems involving natural language.

5. Results

In this section, we present the results of our experiments for the regression and classification tasks. For each task, we first performed a baseline experiment (Section 5.1),

Table 4
Hyperparameters for the RBF, polynomial, and sigmoid kernels.

	RBF	polynomial	sigmoid
Function	$exp(-\gamma\|\|x_i - x_j\|\|^2)$	$\left(\gamma {x_i}^{\mathrm{T}} x_j + c\right)^{\mathrm{d}}$	$tanh\left(\gamma {x_i}^{\mathrm{T}} x_j + c\right)$
Parameters	free parameter γ: vary between 2^{-14} and 2^4, stepping by factor 4 (*default = 3*)		
	d: vary between 2 and 5 (*default = 1/number of features*)		
	c (constant trading off): fix to *default of 0*		

followed by two different rounds of optimization experiments. In the discussion of our results, we make a distinction between the readability prediction experiments performed on our two languages under consideration using only automatically derived features (Section 5.2) and the experiments where the fully automatic Dutch readability prediction system is compared with a system where gold-standard features have been derived (Section 5.3). We start each time by presenting the optimal results after which we discuss in close detail which features contributed most to the readability predictions.

5.1 Baseline Results for English versus Dutch Readability Prediction

In Table 5, we present the baseline results using LibSVM in a 10-fold cross validation set-up for our two readability prediction tasks. For both tasks, the default learner options were set and all available features were fed to the learners.

For the regression task, we achieve a better result on the English data set, whereas the opposite seems to hold for the classification experiments—that is, both the binary and multiclass experiments on the Dutch data set achieve a superior accuracy score. As expected, the performance on the binary data sets is much higher than on the multiclass data sets.

5.2 Capturing the Complex Interplay between Various Aspects of Readability

5.2.1 Round 1 and 2 Experimental Results. Table 6 gives an overview of the results of the two different rounds of optimization experiments that were conducted. On the left-hand side we present the results on the regression task, and on the right-hand side those of

Table 5
Baseline results for English versus Dutch. The regression task is evaluated with RMSE and the classification task with accuracy.

		Regression		**Classification**			
				BINARY		*MULTI*	
		EN	DU	EN	DU	EN	DU
Baseline	Default, all features	0.1489	0.1813	85.31	92.83	57.35	59.49

Table 6
Results of the optimization experiments on the English and Dutch data sets for our two readability prediction tasks in a 10-fold cross validation set-up. The regression task is evaluated with RMSE and the classification task with accuracy. **Boldface** represents the best results.

		Regression		**Classification**			
				BINARY		*MULTI*	
		EN	DU	EN	DU	EN	DU
Round 1	Feature groups	0.1242	0.1492	85.60	93.16	57.38	60.87
	Individual features	0.0985	0.1470	86.28	93.61	58.14	61.31
Round 2	Joint feature groups	0.0060	**0.0003**	96.27	98.01	70.35	73.35
	Joint individual features	**0.0059**	0.0004	**96.88**	**98.24**	**71.00**	**73.62**

the binary and multiclass classification tasks. The results of these two different rounds will be discussed separately.

In the *Round 1* experiments, LibSVM's hyperparameters were set to the default options and the focus was on selecting the optimal features for readability prediction in both languages. In a first set-up, variation between the ten different feature groups was allowed, and in the second set-up those features requiring deep processing were optimized individually. We observe a similar tendency in both prediction tasks. Compared with the baselines (Table 5), better results are always achieved when performing feature selection. We also observe that for both tasks the best results are achieved with the individual feature selection optimization experiments, though the performance increase is moderate, which is not that remarkable given the inherent feature weighting in the greedy type of learning that SVMs perform.

In *Round 2* similar experiments were performed, but this time LibSVM's hyperparameters were jointly optimized while selecting the optimal features. We observe that this setting results in the best results (indicated in bold) for both prediction tasks. If we have a closer look at the differences between both set-ups, joint feature groups versus joint individual features, we see that the differences in performance are moderate. For the regression task, we observe for both languages a minimal difference of 0.001 points. For the classification tasks, these differences are more outspoken: For the English data set we achieve an increase of 0.61 points for the binary and 0.65 points for the multiclass experiments. For the Dutch data set, we achieve a performance increase of 0.23 and 0.27 points, respectively.

As the latter experiments led to the best results, we will now discuss which features and which hyperparameters were selected in the fittest individuals.

5.2.2 Feature (Group) Informativeness. Because, at the end of a GA optimization run, the highest fitness score may be shared by multiple individuals having different optimal feature combinations or parameter settings, we also considered runner-up individuals to that elite as valuable solutions to the search problem. When discussing the results of the GA experiments, we therefore refer to the *k*-nearest fitness solution set; these are the individuals that obtained one of the top *k* fitness scores, given an arithmetic precision (e.g., by rounding the scores to four decimal places). Following Desmet (2014), we used a precision of four significant figures and set *k* to three.

We will discuss which hyperparameters, and especially which features groups, were selected in both languages. The features are visualized using a color range: The closer to blue, the more this feature group was turned on and the closer to red, the less

important the feature group was for reaching the optimal solution. The numbers within the cells represent the same information but percentagewise. In Figure 4, we illustrate which feature groups were considered important using this color range.

What immediately draws our attention is the discrepancy between the regression and classification tasks in both languages. Apparently, the optimal regression results can be achieved with far fewer features: for both languages only the lexical (i.e., the *tradlex* and *lexterm* for English and the *lexterm* and *lexlm* for Dutch) and semantic role features (*srl*) seem crucial. For both the binary and multiclass classification tasks it is better to have more feature information available, especially for the multiclass experiments.

Regarding those features requiring more complicated linguistic processing (the *deepsynt*, *ner*, *coref* and *srl* features), we observe that these feature groups are always selected for the classification tasks in both languages. Because the best results for the classification experiments were achieved when performing an individual selection of those features we made an additional analysis of the individual features that were or were not retained in those optimal set-ups. These are presented in Figure 5, in which a black box refers to a selected feature, and a white box refers to a feature that was not selected. For both languages, we observe that more than 50% of the features in each of the four groups requiring deep processing was selected, which also explains why these feature groups were retained (see Figure 4). When comparing our two languages under consideration, we observe that similar features are selected. For the binary classification task all deep syntactic features (6/6) are selected in both languages, as well as most of the deep semantic features (4 versus 5 out the 7 *ner*, 5 versus 3 of the *coref* and 18 out of the 20 *srl* features). The multiclass experiments reveal a similar tendency though here the coref features seem to beat to deep syntactic features when it comes to being selected in both languages. Also, most of the *ner* (5 versus 6 out of 7) and *srl* (15 and 14 out of 20) features are selected in both languages. This confirms that for the classification task the features requiring deep linguistic processing are important to achieve optimal performance.

For the regression experiments, we perform a similar analysis but go one step further in that we also analyze text correlates. These findings are presented in the next section.

	ENGLISH			DUTCH		
	REG	BIN	MULTI	REG	BIN	MULTI
tradlen	31.25	83.33	100	0	100	100
tradlex	56.25	100	100	42.86	0	100
lexterm	81.25	100	50	100	100	100
lexlm	43.75	83.33	100	100	100	100
shallowsynt	0	100	100	0	100	100
deepsynt	18.75	100	100	0	100	100
shallowsem	0	100	100	0	100	100
ner	0	100	100	0	100	100
coref	12.5	100	100	0	100	100
srl	100	100	100	100	100	100

Figure 4
Illustrating which feature groups were selected in the joint optimization set-ups for the regression, binary, and multiclass classification tasks for both languages under consideration. A blue cell means that a feature group was selected more often whereas a red cell implies the opposite. The numbers within the cells represent this information percentagewise.

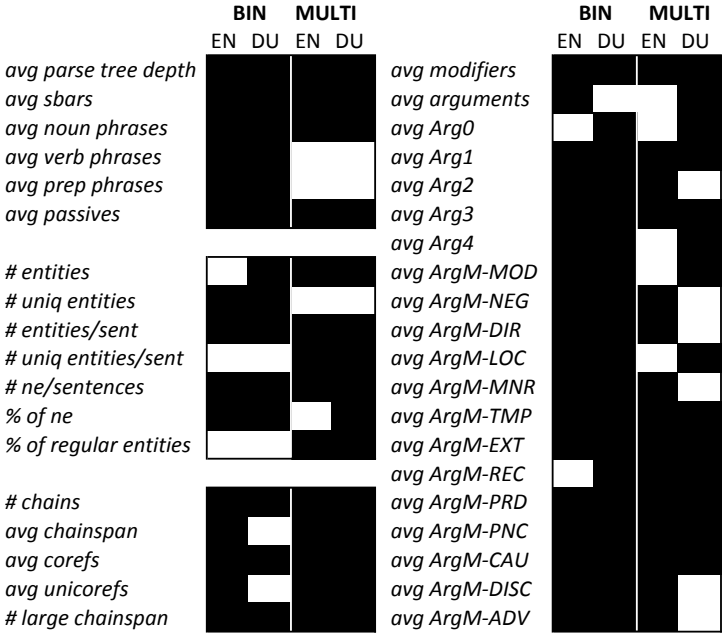

Figure 5
Illustrating which individual deep syntactic and semantic features were selected (= black) or not (= white) in the joint optimization classification experiments for English (EN) and Dutch (DU).

5.2.3 Identifying Text Correlates. When it comes to selecting the best features for readability prediction, there seems to be the consensus that first the correlation between the features and human assessments is measured (Pitler and Nenkova 2008; François 2011). The next step, if included at all, is then to see which features come out as good predictors when performing machine learning experiments such as regression (Pitler and Nenkova 2008), or classification (Feng et al. 2010) by including or excluding features or feature groups from the prediction task. Interestingly, the most predictive features often do not overlap with those having the highest correlation (Pitler and Nenkova 2008).

We compute the Pearson correlation coefficient between all individual features and our regression data set, in which we have an absolute score for each individual text. As we observed in our experiments, the optimal settings for regression did not require the activation of many feature groups in both languages (see Figure 4). We hope to shed more light on this by identifying text correlates. In our discussion we only report on features with a significant correlation coefficient (i.e., with p-values less than 0.05).[6]

Regarding the *traditional features*, we found that in both languages the four length-related features (*tradlen*) correlate with our regression data set; the features related to word-length show an especially stronger correlation. Regarding the two traditional lexical features (*tradlex*), only for English does the percentage of words that can be found in the Chall and Dale list (1995) correlate significantly ($r = -0.53$).

This brings us to the *lexical features*. For the Dutch data set, the perplexity of a given text when compared with our reference corpus (i.e., a subset of the SoNaR corpus [Oostdijk et al. 2013]) was found to correlate ($r = 0.36$), although when perplexity was

6 The individual correlations are presented in Tables 9 and 10.

averaged over text length this was not the case. For English, these language modeling features (*lexlm*) do not correlate. Looking at the terminological metrics (*lexterm*), however, we found that the tf-idf value correlates in both languages ($r = 0.38$ for English and $r = 0.21$ for Dutch).

At the level of **syntactic features**, we make a division between shallow features computed based on PoS-tags (*shallowsynt*) and a deeper level based on dependency parsing (*deepsynt*). For the PoS-related features, we observe a clear difference between the English and Dutch data sets in that 78% of the English features versus only 48% of the Dutch features correlate (i.e., 21 versus 13 out of 27 to be exact). However, for both languages at least one feature representing the five main part-of-speech classes (nouns, adjectives, verbs, adverbs, and prepositions) does correlate. For English, the average amount of function and content words also correlates. From the group of deep syntactic features, we see that for Dutch all six features correlate significantly and for English they all correlate but one.

This brings us to our final group of features, the *semantic features*. The lists of connectives (*shallowsem*) do not correlate much; for English, only the number of temporals per sentence ($r = 0.26$) do and for Dutch only the amount of concessive connectives per document ($r = 0.22$). As to the named entity features (*ner*), we again observe some differences between English and Dutch. Whereas for English especially the average amounts of entities and named entities correlate, for Dutch the overall percentages of entities and named entities in a document correlate more. The added value of the coreference features (*coref*) seems trivial in both languages: For English none of the features correlate whereas for Dutch only the average length of a chain does ($r = -0.24$). Finally, we considered the semantic role features (*srl*). For English, these seem obsolete; only one out of 20 features correlates, that is, the average amount of modifiers of direction ($r = 0.29$). For Dutch, on the other hand, the total number of arguments and the Arg1 and Arg3 arguments correlate significantly together with three modifiers.

If we extrapolate these individual feature correlates to the group level, we find that for English we have six feature groups of which 50% or more of the features correlate whereas for Dutch we only have three. In Figure 6, we compare these results with the analysis of the feature groups coming from the optimal regression set-ups. A black

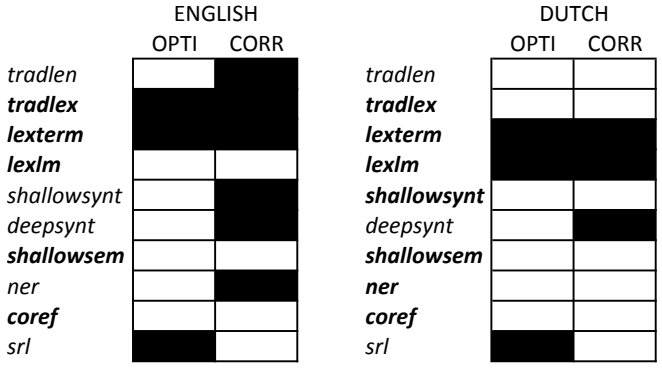

Figure 6
Comparison on the regression data set between those feature groups that were (= black) or were not (= white) selected in the optimal setting (OPTI) and the feature groups where more (= black) or less than (= white box) 50% of the features were found to correlate (CORR). **Boldface** marks those feature groups revealing a similar tendency.

cell means that a feature group was either selected in the optimal setting or found to correlate. Those feature groups revealing similar tendencies (two black or two white cells) have been indicated in bold. For English, we observe that only five out of the ten feature groups show a similar tendency, whereas for Dutch seven out of the ten feature groups do. This implies that for our English data set there is a less outspoken link between features correlating and them being selected in the optimal regression experiments, which is in line with the results presented by Pitler and Nenkova (2008). What is especially striking is that the feature group containing the strongest correlations in the English data set, the *tradlen* group where three correlations of more than $r = -0.5$ were found, was not selected in the optimal setting. The same is true for both languages considering the *deepsynt* group; in both languages the significant correlation coefficients are above $r = -0.3$ but this feature group was never selected in the optimal settings.

Given that the optimal results were achieved while jointly optimizing both features and hyperparameters, we briefly list which hyperparameters were selected. For the regression task, there was each time a preference for the nu-SVR LibSVM type. For both languages a linear kernel was chosen and the cost-value ranges from 2^{12} to 2^{13}. For the classification tasks we observe that for the binary task a linear kernel is preferred whereas for the multiclass task the default more complex RBF kernel. C-values are slightly lower: 2^{11} to 2^{12}. The free parameter γ for the RBF kernels was very small or zero.

5.3 Impact of Dutch Fully Automatic versus Dutch Gold-Standard Deep Syntax and Semantic Features

Another aspect of this research was to investigate in closer detail the contribution of those features requiring deep linguistic processing. Though many advances have been made in NLP, the more difficult text-understanding tasks such as coreference resolution or semantic role labeling still achieve moderate performance rates. Implementing such features in a readability prediction system is thus risky as the automatically derived features might not truly represent the information at hand. Because we have gold-standard deep syntactic and semantic information available for our Dutch readability data set, we were able to investigate in close detail their added value in predicting readability.

5.4 Baseline Results for Dutch Fully Automatic versus Dutch Gold-Standard Readability Prediction

In Table 7, we present the baseline results using LibSVM in a 10-fold cross validation set-up for our two readability prediction tasks. For both tasks, the default learner options were set and all available features were fed to the learners.

For the regression task we observe that relying on a feature space with gold-standard deep syntax and semantic features harms performance whereas for the classification tasks, especially for the multiclass experiments (i.e., from an accuracy of 59.49 to one of 62.58), it proves beneficial.

5.4.1 Optimization Results. Table 8 gives an overview of the results of the two different optimization rounds. On the left-hand side, we present the results on the regression task, and on the right-hand side those of the binary and multiclass classification tasks. The best individual results for the Dutch language are indicated in bold. We see that for

Table 7
Baseline results for Dutch fully automatic versus Dutch gold standard. The regression task is evaluated with RMSE and the classification task with accuracy.

		Regression		Classification			
				BINARY		*MULTI*	
		Auto	Gold	Auto	Gold	Auto	Gold
Baseline	Default, all features	0.1813	0.1965	92.83	92.92	59.49	62.58

Table 8
Results of the optimization experiments on the Dutch automatic and gold-standard data sets for our two readability prediction tasks running 10-fold cross validation experiments. The regression task is evaluated with RMSE and the classification task with accuracy. **Boldface** represents the best individual results.

		Regression		Classification			
				BINARY		*MULTI*	
		Auto	Gold	Auto	Gold	Auto	Gold
Round 1	Feature groups	0.1492	0.1437	93.16	93.34	60.87	63.55
	Individual features	0.1470	0.1585	93.61	94.08	61.31	63.73
Round 2	Joint feature groups	**0.0003**	0.0080	98.01	97.68	73.35	72.78
	Joint individual features	0.0004	0.0689	**98.24**	98.06	**73.62**	72.95

both tasks these best results are achieved with the Dutch fully automatic feature space. We will start by discussing the results of the two different optimization rounds.

In the *Round 1* experiments, we observe a different tendency in both prediction tasks. For the regression task, a set-up with gold-standard features never outperforms the results achieved with the fully automatic features. In the classification tasks, however, and especially in the multiclass experiments, relying on gold-standard deep syntactic and semantic features seems beneficial (an increase of 2.68 points in the first and one of 2.42 in the second set-up), which is in line with our baseline results. In the *second round*, counterintuitively, we notice that the best results for both languages and both tasks are achieved with the fully automatic features. Because the only difference between the two data sets are the feature values of the deep syntactic and deep semantic feature groups, we had a close inspection of these particular features.

5.4.2 Feature (Group) Informativeness. Figure 7 gives an overview of the feature groups which were considered important in the optimization. Again, the groups are visualized using the previously mentioned color range (see Section 5.2.2).

When relying on gold-standard deep syntactic and semantic information we observe that more feature groups are considered important for the regression task, 8 out of the 10 groups (including *deepsynt*, *ner*, *coref*, and *srl*) become selected versus 3 in the experiments where automatically derived features were used. For the classification tasks the situation alters less, in the binary experiments one feature group appears more important (*tradlen*), and in the multiclass experiments one semantic feature group even gets turned off (*coref*) in the gold standard.

We make an additional analysis of the individual features that were or were not retained in the optimal set-ups, this comparison is presented in Figure 8. In the

	AUTOMATIC			GOLD		
	REG	BIN	MULTI	REG	BIN	MULTI
tradlen	0	100	100	100	100	100
tradlex	42.86	0	100	75	53.85	100
lexterm	100	100	100	100	61.54	100
lexlm	100	100	100	25	100	100
shallowsynt	0	100	100	0	100	100
deepsynt	0	100	100	75	100	100
shallowsem	0	100	100	100	100	100
ner	0	100	100	75	100	100
coref	0	100	100	75	100	0
srl	100	100	100	100	100	100

Figure 7
Illustrating which feature groups were selected in the joint optimization set-ups for the regression, binary, and multiclass classification tasks with Dutch fully automatic and Dutch gold standard. A blue cell means that a feature group was selected more often whereas a red cell implies the opposite. The numbers within the cells represent this information percentagewise.

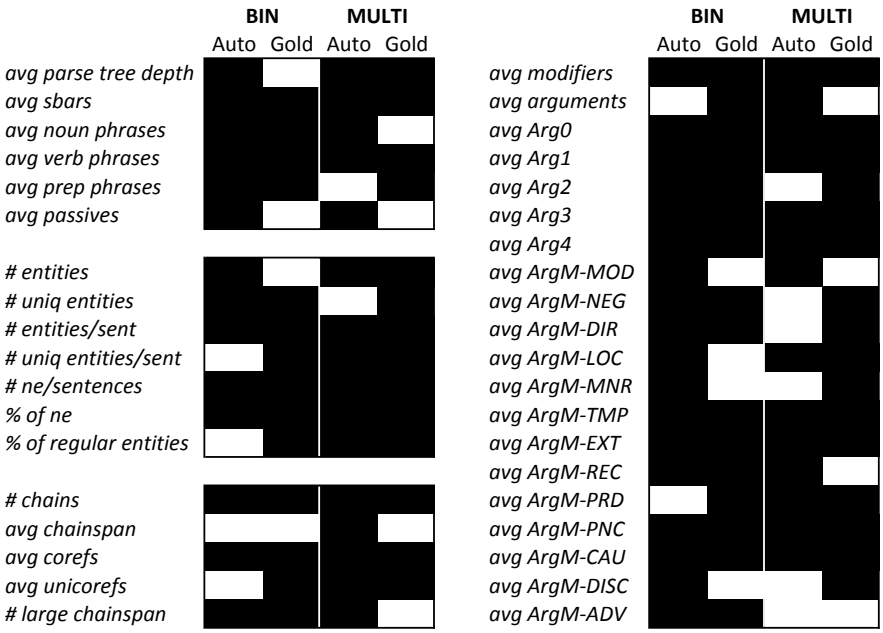

Figure 8
Illustrating which individual deep syntactic and semantic features were selected (= black) or not (= white) in the joint optimization classification experiments when relying on fully automatic (Auto) or gold-standard (Gold) Dutch information.

remainder of this section we zoom in on the classification experiments and in the next section we do the same for the regression experiments. When comparing the fully automatic with the gold-standard features we see that for the binary task fewer deep syntactic and semantic role features are chosen, whereas the named entities and

coreference features are selected more. For the multiclass classification task we also observe that fewer deep syntactic features are selected, but here also the coreference features get selected less often. This final finding explains why only the *coref* feature group as a whole was not selected in Figure 7. Overall, we see that for the binary classification task more fully automatic deep syntactic and semantic features are selected (32 versus 30), whereas for the multiclass task the opposite is true (30 versus 33). In total, we have 38 individual deep syntactic and semantic features; we can thus conclude that for both classification tasks including this type of information is important, regardless of whether it was obtained automatically or from gold-standard information. For the regression experiments, we perform a similar analysis but go one step further in that we also analyze text correlates. These findings are presented in the next section.

5.4.3 Identifying Text Correlates. As we observed in the experiments, the optimal settings for regression differed when relying on automatic versus gold-standard features. The optimal result was achieved in the fully automatic setting (RMSE of 0.0003) when relying less on those feature groups requiring deep linguistic processing (see Figure 7 where only the *srl* group is blue). We hope to shed more light on this by identifying text correlations.[7] We limit our discussion to the features requiring deep syntactic and semantic information.

Looking at the **syntactic features** based on dependency parsing, we observe that all six fully automatic features correlate with our regression data set, whereas the number of verb phrases is the only feature not correlating when relying on gold-standard information. This brings us to the *deep semantic features*. Here, we observe that for all the groups, more features correlate when relying on gold-standard information than when relying on fully automatic information; this is especially the case for the named entities (*ner*) where six out of the seven features correlate versus only three.

If we extrapolate these individual feature correlates to the group level, we find that only the deep syntactic group correlates both in fully automatic or gold-standard form with our regression data sets. For the semantic features, only the named entities correlate in the gold standard. In Figure 9, we compare these results with the analysis of the feature groups coming out of our optimal regression set-ups. A black cell again means that a feature group was either selected in the optimal setting or found to correlate. Those feature groups revealing similar tendencies (two black or two white cells) have been indicated in bold.

We observe that in both set-ups two of the feature groups were or were not selected or found to correlate. Again, it draws the attention that all feature groups requiring deep semantic processing were selected in the gold-standard set-up whereas only two of these contain features that correlate most of the time with our regression data set. In order to gain more insights into this, we performed a final analysis where we compare the individual feature correlates with the optimization experiments where both the hyperparameters and individual deep syntactic and semantic features were jointly optimized. The comparison is presented in Figure 10.

Regarding the **syntactic features**, we observe that although all these features were found to correlate when derived automatically, these were not selected in the optimal setting. When deep syntactic information derived from gold-standard dependency trees

7 The actual correlation coefficients can be found in Table 10.

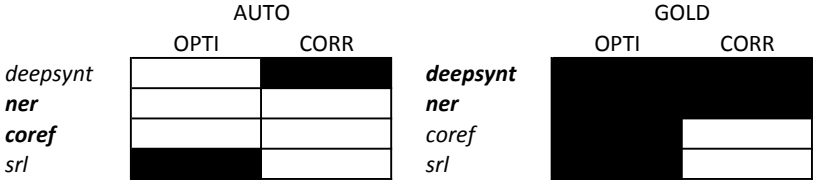

Figure 9
Comparison on the regression data set between those deep syntactic and semantic feature groups that were (= black) or were not (= white) selected in the optimal setting (OPTI) and the feature groups where more (= black) or less than (= white box) 50% of the features were found to correlate (CORR). **Boldface** indicates feature groups revealing similar tendencies.

was used we see that only the number of verb phrases did not correlate, surprisingly this feature was selected in the optimal setting whereas the other two features, the average parse tree depth revealing a high correlation ($r = -0.5$) and the number of passives ($r = -0.34$) were not selected in the optimal setting.

Having a closer look at the *named entity features*, we see that not many features correlate when derived automatically, as a result they are also not often selected, the percentages of named entities and regular entities present in a text are important. In their gold-standard form, we observe that more of these features reveal a correlation with our data set. However, in our optimal setting only the previously mentioned percentages together with the total number of entities present in a text seems important for the prediction.

This brings us to the *coreference features*. The performance of most automated coreference resolvers is moderate, which might explain that only one automatically derived feature, the average chainspan, was found to correlate. In the optimal setting we see that the number of coreferential relations and the number of chains with a large chainspan were selected. The same two features were selected in the gold-standard setting, but when it comes to the correlations we see that relying on gold-standard coreferential information only shows correlations with the average number of coreferential relations (*coref* and *unicorefs*).

Finally, the *semantic role features*. Though only few automatic semantic role features correlate with our data set, many of them were retained in the optimal settings. The same holds when relying on gold-standard features.

Overall, if we compare the fully automatic with the gold-standard set-up we observe that in the gold-standard set-up there are more similarities between features being selected or not in the optimal setting and their correlation with the data set, that is, in total 16 features (those indicated in italics). In the fully automatic setting this number is less outspoken, only 13 features. Nevertheless, our results reveal that the best individual results for the Dutch language are achieved when relying on fully automatic deep syntactic and semantic features.

Again, we finish this discussion by briefly listing which hyperparameters were selected in the optimal settings. For the regression task, each time there was a preference for the nu-SVR LibSVM type. Whereas for the fully automatic features a linear kernel was chosen, our system preferred a sigmoid kernel for the set-up with gold-standard features. The cost-value ranges from 2^{12} to 2^{13}. For the classification tasks we observe that for the binary task a linear kernel is preferred, whereas for the multiclass task the default more complex RBF kernel is preferred. C-values are slightly lower: 2^{11} to 2^{12}. The free parameter γ for the RBF kernels was very small or zero.

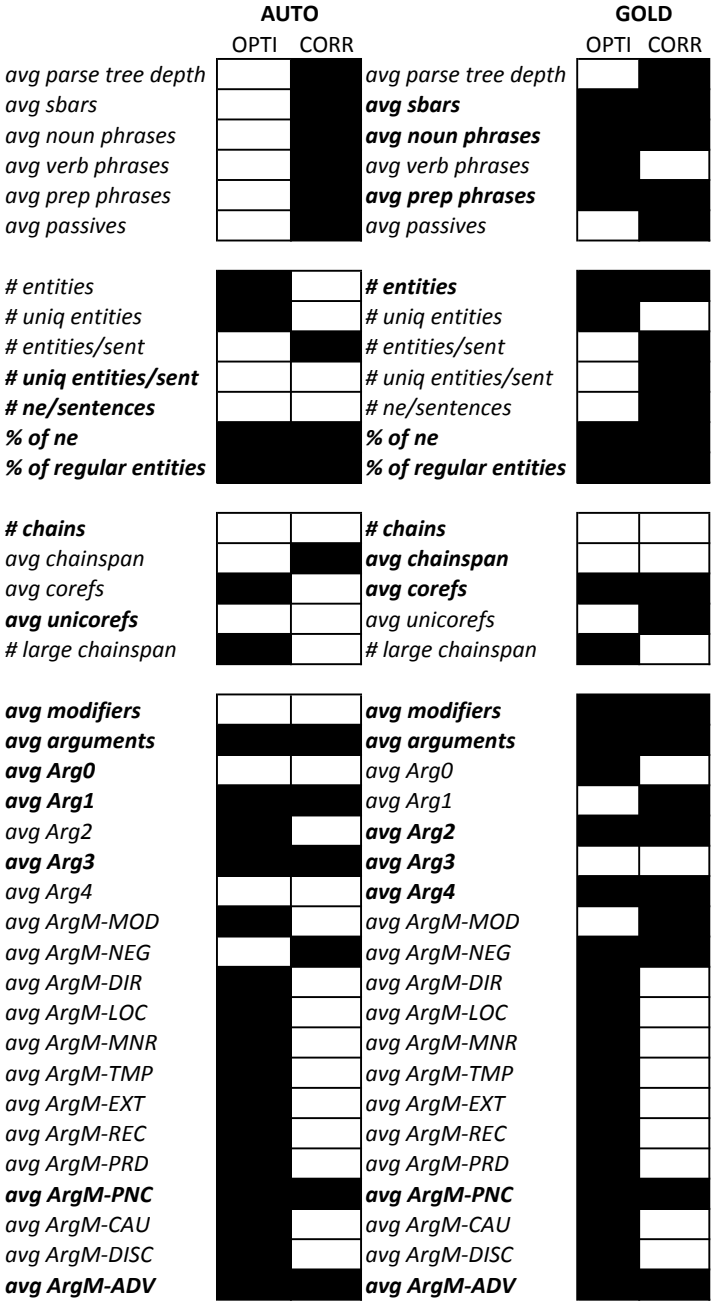

Figure 10
Comparison on the regression data set between those deep syntactic and semantic individual features that were (= black) or were not (= white) selected in the optimal setting (OPTI) and that did (= black) or did not (= white) (CORR). **Boldface** indicates feature groups revealing similar tendencies.

Table 9
Pearson correlation coefficients of the *tradlen*, *tradlex*, *lexlm*, *lexterm*, *shallowsynt*, and *shallowsem* English and Dutch automatically derived features with the English and Dutch regression data sets. The coefficients in **bold** indicate significance (with $p < 0.05$).

EN	DU	FEATURE	GROUP
−0.55	**−0.58**	average word length	*tradlen*
−0.40	**−0.37**	average sentence length	
−0.53	**−0.60**	ratio long words	
−0.52	**−0.58**	% of polysyllable words	
−0.53	0.07	% in frequency lis	*tradlex*
−0.13	0.15	type token ratio	
0.05	**0.36**	perplexity	*lexlm*
0.08	0.11	normalized perplexity	
0.38	**0.21**	TF-IDF	*lexterm*
−0.07	−0.03	Log Likelihood	
−0.27	0.16	average content words	*shallowsynt*
0.27	−0.16	average function words	
−0.30	**0.28**	average nouns	
−0.28	**0.21**	average type nouns	
−0.43	**−0.30**	average nouns/sentence	
−0.38	**−0.25**	average type nouns/sentence	
−0.30	0.16	average noun types	
−0.34	**−0.23**	average adjectives	
−0.29	**−0.20**	average type adjective	
−0.47	**−0.32**	average adjective/sentence	
−0.45	**−0.34**	average type adjectives/sentence	
−0.26	**−0.26**	average adjective types	
0.34	−0.09	average verb	
0.26	−0.11	average type verb	
−0.13	**−0.38**	average verb/sentence	
−0.11	**−0.41**	average type verb/sentence	
0.34	−0.15	average verb types	
0.24	0.06	average adverb	
0.21	0.03	average type adverb	
−0.01	**−0.22**	average adverb/sentence	
−0.04	**−0.26**	average type adverb/sentence	
0.24	0.00	average adverb types	
−0.36	−0.13	average prepositions	
−0.04	0.08	average type prepositions	
−0.44	**−0.37**	average prepositions/sentence	
−0.27	−0.28	average type preposition/sentence	
0.02	0.03	average preposition types	
0.07	−0.04	average connectives/document	*shallowsem*
−0.05	−0.15	average connectives/sentence	
−0.16	−0.11	average causal/document	
−0.16	−0.15	average causal/sentence	
0.26	−0.07	average temporals/document	
0.16	−0.67	average temporals/sentence	
n/a	−0.06	average additives/document	
n/a	−0.16	average additives/sentence	
−0.01	**0.22**	average contestive/document	
−0.08	0.17	average contestive/sentence	
n/a	−0.11	average concessives/document	
n/a	−0.11	average concessives/sentence	

Table 10
Pearson correlation coefficients of the *deepsynt*, *ner*, *coref*, and *srl* English and Dutch automatically derived and Dutch gold-standard features with the English and Dutch regression data sets. The coefficients in **bold** indicate significance (with $p < 0.05$).

EN	DU auto	DU gold	FEATURE	GROUP
−0.35	**−0.48**	**−0.50**	average parse tree depth	*deepsynt*
−0.07	**−0.30**	**−0.31**	average sbars	
−0.37	**−0.41**	**−0.41**	average noun phrases	
−0.44	**−0.32**	−0.17	average verb phrases	
−0.44	**−0.40**	**−0.41**	average prepositional phrases	
−0.30	**−0.38**	**−0.34**	average passives	
−0.30	−0.00	**−0.21**	number of entities	*ner*
−0.32	0.02	−0.14	number of uniq entities	
−0.43	**−0.24**	**−0.22**	number of entities/sentence	
−0.37	−0.19	**−0.23**	number of uniq entities/sentence	
−0.19	0.16	**0.28**	number of ne/sentences	
−0.11	**0.23**	**0.44**	perc of ne	
0.11	**−0.23**	**−0.44**	perc of regular entities	
0.02	0.01	−0.08	number of chains	*coref*
−0.04	**−0.26**	0.02	average chainspan	
0.15	−0.04	**0.29**	average corefs	
0.07	−0.02	**0.30**	average unicorefs	
0.07	−0.11	0.09	number large chainspan	
0.04	−0.18	**−0.27**	average modifiers	*srl*
−0.07	**−0.24**	**−0.25**	average arguments	
0.07	−0.06	−0.04	average Arg 0	
−0.14	**−0.33**	**−0.26**	average Arg 1	
−0.10	−0.09	**−0.28**	average Arg 2	
−0.05	**−0.22**	−0.01	average Arg 3	
0.06	0.14	**−0.20**	average Arg4	
0.09	−0.08	**−0.33**	average ArgM-MOD	
0.06	**−0.24**	**−0.25**	average ArgM-NEG	
0.29	−0.10	−0.10	average ArgM-DIR	
−0.02	−0.11	−0.12	average ArgM-LOC	
−0.00	−0.01	0.03	average ArgM-MNR	
−0.01	0.08	0.03	average ArgM-TMP	
−0.01	−0.06	−0.17	average ArgM-EXT	
n/a	0.09	0.08	average ArgM-REC	
−0.08	−0.07	−0.01	average ArgM-PRD	
−0.07	**−0.24**	**−0.37**	average ArgM-PNC	
−0.09	0.01	−0.04	average ArgM-CAU	
−0.15	−0.13	−0.19	average ArgM-DIS	
0.11	**−0.22**	**0.23**	average ArgM-ADV	

6. Conclusion

The aims of the research presented here were twofold. On the one hand we wished to identify whether it is possible to build an automatic readability prediction system that can score and compare the readability of English and Dutch generic text. On the other hand, we wanted to investigate which information sources optimally contributed to this readability prediction performance and determine if these features remained consistent in both languages. For Dutch, we could also investigate whether having gold-standard information available for those features requiring a deep linguistic processing is beneficial for the overall performance.

To this purpose, texts from various text genres were collected in both languages and these data were assessed by two user groups: experts and a crowdsource. Based on the correlations between those two assessor groups, we combined our data sets for performing experiments, reflecting the two possible readability prediction set-ups: predicting an absolute value (regression) or comparing two texts (classification). Based on the assessors' comments and a thorough literature overview, we included various feature groups representing both superficial features and text characteristics requiring deep linguistic processing. This resulted in instances with no less than 87 distinct features divided over ten feature groups. We used a wrapper-based approach using a genetic algorithm to perform combined hyperparameter optimization and feature selection for readability prediction.

Based on our results, we can state that we have succeeded in building a fully automatic readability prediction system for both English and Dutch generic text. The best results for both tasks were achieved while jointly optimizing LibSVM's hyperparameters and all our features. When comparing both readability prediction tasks we observed that in both languages the optimal regression result was achieved with fewer activated features. When these activated features were compared with their correlations with our regression data sets, we found that for English there is a less outspoken link. This is in line with previous research (Pitler and Nenkova 2008). Regarding the classification tasks, we observed that both languages selected the similar features in their optimal settings and that they rely on a large feature space including deep syntactic and semantic information.

Considering those features requiring deep linguistic processing, we observed that for both readability prediction tasks the best individual results on our Dutch data set were achieved when these features had been derived automatically. An analysis of which of these features were retained in the optimal classification settings revealed that including this type of deep linguistic information is important for both classification tasks, regardless of whether it was obtained automatically or from gold-standard information. For the regression task, we noticed that in the gold-standard set-up there are more similarities between features being selected or not in the optimal setting and their correlation with our data set. Nevertheless the best individual result on our Dutch data set was achieved while relying on deep syntactic and semantic features that have been derived automatically.

This research has sparked many ideas for future work. A next logical step in our research is to investigate how the current readability assessments can be used to pinpoint problematic passages in texts, which might probably also lead to redefining the readability scores at a sentence level or paragraph level. Based on the observation that 25% of the remarks given by the expert readers during the assessments could not be categorized in some linguistic category, we wish to further explore this category of comments and also include other methodologies, such as eye tracking, to measure

reading ease. Another interesting line of research could be to see if and how we need to adapt our system when dealing with more specific text genres such as legal texts. Lastly, the difference between readability and translatability is something which we would like to investigate in future research.

References

Aha, David W. and Richard L. Bankert. 1996. A comparative evaluation of sequential feature selection algorithms. In D. Fischer and J.-H. Lenz, editors, *Artificial Intelligence and Statistics V*. New York: Springer Verlag, pages 199–206.

Alderson, J. Charles. 1984. *Reading in a Foreign Language: A Reading Problem or a Language Problem*. Longman.

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh University Press.

Bailin, Alan and Ann Grafstein. 2001. The linguistic assumptions underlying readability formulae: A critique. *Language & Communication*, 21(3):285–301.

Barzilay, Regina and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 141–148, Ann Arbor, MI.

Barzilay, Regina and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Benjamin, Rebekah George. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24(1):63–88.

Björkelund, Anders, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL 2009): Shared Task*, pages 43–48, Boulder, CO.

Caruana, Rich and Dayne Freitag. 1994. Greedy Attribute Selection. In *Proceedings of the International Conference on Machine Learning (ICML-1994)*, pages 28–36, New Brunswick, NJ.

Chall, Jeanne and Edgar Dale. 1995. *Readability Revisited: The New Dale-Chall Readability Formula*. Brookline Books.

Collins-Thompson, Kevin. 2014. Computational assessment of text readability: A survey of current and future research. *Special Issue of the International Journal of Applied Linguistics*, 165(2):97–135.

Collins-Thompson, Kevin and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT - NAACL-2004)*, pages 193–200, Boston, MA.

Collins-Thompson, Kevyn and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56:1448–1462.

Cristianini, Nello and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Crossley, Scott A., Jerry Greenfield, and Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly*, 43(3):475–493.

Davison, Alice and Robert N. Kantor. 1982. On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2):187–209.

De Clercq, Orphée, Bart Desmet, Sarah Schulz, Els Lefever, and Veronique Hoste. 2013. Normalization of Dutch user-generated content. In *Proceedings of Recent Advances in Natural Language Processing*, pages 179–188.

De Clercq, Orphée, Iris Hendrickx, and Véronique Hoste. 2011. Cross-domain Dutch coreference resolution. In *Proceedings of the 8th International Conference on Recent Advances in Natural Language Processing (RANLP-2011)*, pages 186–193, Hissar.

De Clercq, Orphée, Paola Monachesi, and Véronique Hoste. 2012. Evaluating automatic cross-domain semantic role annotation. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, pages 88–93, Istanbul.

De Clercq, Orpheée, Véronique Hoste, Bart Desmet, Philip van Oosten, Martine

De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering, Cambridge Journals Online*, 20(3):293–325.

de Marneffe, Marie-Catherine, Bill MacCartney, and Christopher D. Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-2006)*, pages 449–454, Genoa.

Desmet, Bart. 2014. *Finding the Online Cry for Help: Automatic Text Classification for Suicide Prevention*. Ph.D. thesis, Ghent University.

Desmet, Bart and Véronique Hoste. 2013. Fine-grained Dutch named entity recognition. *Language Resources and Evaluation*, 48(2):307–343.

DuBay, William H. 2004. *The Principles of Readability*. Impact Information.

DuBay, William H., editor. 2007. *Unlocking Language: the Classic Readability Studies*. BookSurge.

Falkenjack, Johan and Arne Jonsson. 2014. Classifying easy-to-read texts without parsing. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 114–122, Gothenburg.

Feng, Lijun, Noémie Elhadad, and Matt Huenerfauth. 2009. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2009)*, pages 229–237, Athens.

Feng, Lijun, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pages 276–284, Beijing.

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 363–370, Ann Arbor, MI.

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

François, Thomas. 2009. Combining a statistical language model with logistic regression to predict the lexical and syntactic difficulty of texts for FFL. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop (EACL-2009)*, pages 19–27, Athens.

François, Thomas. 2011. *Les apports du traitement automatique du language à la lisibilité du français langue étrangère*. Ph.D. thesis, Université catholique de Louvain.

François, Thomas and Eleni Miltsakaki. 2012. Do NLP and machine learning improve traditional readability formulas? In *Proceedings of the 1st Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR2012)*, pages 49–57, Montreal.

Galley, Michel and Kathleen Mckeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence (IJCAI-2003)*, pages 1486–1488, Acapulco.

Goldberg, David. 1989. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison Wesley.

Goldberg, David and Kalyanmoy Deb. 1991. A comparative analysis of selection schemes used in genetic algorithms. In W. N. Martin and W. M. Spears, editors, *Foundations of Genetic Algorithms*. Morgan Kaufmann Publishers, pages 69–93.

Graesser, Arthur C., Danielle S. McNamara, Max M. Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments and Computers*, 36:193–202.

Gunning, Robert. 1952. *The Technique of Clear Writing*. McGraw-Hill, New York.

Halliday, Michael and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman Group Ltd.

Heilman, Michael, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd ACL Workshop on Innovative Use of NLP for Building Educational Applications (EANL-2008)*, pages 71–79, Columbus, OH.

Heilman, Michael J., Kevyi Collins-Thompson, Jamie Callan, and Maxine 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 and the North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2007)*, pages 460–467, Rochester, NY.

Hoste, Véronique. 2005. *Optimization Issues in Machine Learning of Coreference Resolution*. Ph.D. thesis, Antwerp University.

John, George H., Ron Kohavi, and Karl Pfleger. 1994. Irrelevant Features and the Subset Selection Problem. In *International Conference on Machine Learning*, pages 121–129, New Brunswick, NJ.

Kate, Rohit J., Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J. Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING-2010)*, pages 546–554, Beijing.

Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel. Research Branch Report RBR-8-75, Naval Technical Training Command Millington Tenn Research Branch, Springfield, Virginia.

Klare, George R. 1976. A second look at the validity of readability formulas. *Journal of Literacy Research*, 8:129–152.

Kraf, Rogier and Henk Pander Maat. 2009. Leesbaarheidsonderzoek: Oude problemen, nieuwe kansen. *Tijdschrift voor Taalbeheersing*, 31(2):97–123.

Lee, Heeyoung, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916.

Leroy, Gondy and James E. Endicott. 2011. Term familiarity to indicate perceived and actual difficulty of text in medical digital libraries. In *International Conference on Asia-Pacific Digital Libraries (ICADL-2011)*, pages 307–310, Beijing.

Leroy, Gondy, Steven Helmreich, James R. Cowie, Trudi Miller, and Wei Zheng. 2008. Evaluating online health information: Beyond readability formulas. In *Proceedings of the American Medical Informatics Association Annual Symposium (AMIA-2008)*, pages 394–398.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. Dutch Parallel Corpus: A balanced copyright-cleared parallel corpus. *Meta*, 56(2).

McNamara, Danielle S., Max M. Louwerse, Philip M. McCarthy, and Arthur C. Graesser. 2010. Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, 47(4):292–330.

Mitchell, Melanie. 1996. *An Introduction to Genetic Algorithms*. MIT Press.

Nenkova, Ani, Jieun Chae, Annie Louis, and Emily Pitler. 2010. Structural features for predicting the linguistic quality of text: Applications to machine translation, automatic summarization and human-authored text. *Empirical Methods in NLG, Lecture Notes in Artificial Intelligence*, 5790:222–241.

OECD. 2013. OECD Skills Outlook 2013. Technical report, OECD Publishing.

Oostdijk, Nelleke, Martin Reynaert, Véronique Hoste, and Ineke Schuurman. 2013. The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns and J. Odijk, editors, *Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing*. Springer, pages 219–247.

Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1):71–106.

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. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008)*, pages 186–195, Honolulu, HI.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, pages 2961–2968, Marrakech.

Rayson, Paul and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics Workshop on Comparing Corpora, (ACL-2000)*, pages 1–6, Hong Kong.

Sabou, Marta, Kalina Bontcheva, and Arno Scharl. 2012. Crowdsourcing research opportunities: Lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies (i-Know-2012)*, pages 17:1–17:8, Graz.

Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co.

Schwarm, Sarah E. and Mari Ostendorf. 2005. Reading level assessment using support vector machines and statistical language models. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, pages 523–530, Ann Arbor, MI.

Si, Luo and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the 10th International Conference on Information Knowledge Management (ICKM-2001)*, pages 574–576, Atlanta, GA.

Staphorsius, Gerrit. 1994. *Leesbaarheid en leesvaardigheid. De ontwikkeling van een domeingericht meetinstrument*. Cito.

Stolcke, Andreas. 2002. SRILM—an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP-2002)*, pages 901–904, Denver, CO.

Tanaka-Ishii, Kumiko, Satoshi Tezuka, and Hiroshi Terada. 2010. Sorting texts by readability. *Computational Linguistics*, 36(2):203–227.

Todirascu, Amalia, Thomas François, Nuria Gala, Cédric Fairon, Anne-Laure Ligozat, and Delphine Bernhard. 2013. Coherence and cohesion for the assessment of text readability. In *Proceedings of 10th International Workshop on Natural Language Processing and Cognitive Science (NLPCS-2013)*, pages 11–19, Marseille.

van Boom, Willem H. 2014. Begrijpelijke hypotheekvoorwaarden en consumentengedrag. In T.M. Berkhout en A.A. van Velten, editors, *Perspectieven voor vastgoedfinanciering (Congresbundel Stichting Fundatie Bachiene)*. Stichting Fundatie Bachiene, pages 45–80.

Van de Kauter, Marjan, Geert Coorman, Els Lefever, Bart Desmet, Lieve Macken, and Véronique Hoste. 2013. LeTs Preprocess: The multilingual LT3 linguistic preprocessing toolkit. *Computational Linguistics in the Netherlands Journal*, 3:103–120.

van den Bosch, Antal, Bertjan Busser, Walter Daelemans, and Sander Canisius. 2007. An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Proceedings of the Seventeenth Computational Linguistics in the Netherlands (CLIN)*, pages 191–206, Nijmegen.

van Noord, Gertjan J. M., Gosse Bouma, Frank van Eynde, Daniel de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: LASSY. In P. Spyns and J. Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing. Springer, pages 231–254.

van Oosten, Philip, Dries Tanghe, and Véronique Hoste. 2010. Towards an improved methodology for automated readability prediction. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC-2010)*, pages 775–782, Valletta.

Wolf, Allison. 2005. Basic skills in the workplace: Opening doors to learning. Technical report 3053, Chartered Institute of Personnel and Development.