Using the Crowd for Readability Prediction

Orphée De Clercq [1,3] Véronique Hoste [1,2] Bart Desmet [1,3]

Philip van Oosten [1,3] Martine De Cock [3] Lieve Macken [1,3]

[1] *University College Ghent, Faculty of Applied Language Studies,*
Ghent, Belgium
{orphee.declercq,veronique.hoste,bart.desmet,lieve.macken}@hogent.be

[2] *Ghent University, Department of Linguistics,*
Ghent, Belgium

[3] *Ghent University, Department of Applied Mathematics and Computer Science,*
Ghent, Belgium
martine.decock@ugent.be

(*Received 24 October 2012*)

Abstract

While human annotation is crucial for many natural language processing tasks, it is often very expensive and time-consuming. Inspired by previous work on crowdsourcing we investigate the viability of using non-expert labels instead of gold standard annotations from experts for a machine learning approach to automatic readability prediction. In order to do so, we evaluate two different methodologies to assess the readability of a wide variety of text material: a more traditional set-up in which expert readers make readability judgments and a crowdsourcing set-up for users who are not necessarily experts. To this purpose two assessment tools were implemented: a tool where expert readers can rank a batch of texts based on readability and a lightweight crowdsourcing tool which invites users to provide pairwise comparisons.

To validate this approach, readability assessments for a corpus of written Dutch generic texts were gathered. By collecting multiple assessments per text, we explicitly wanted to level out a readers background knowledge and attitude. Our findings show that the assessments collected through both methodologies are highly consistent and that crowdsourcing is a viable alternative to expert labeling. This is good news as crowdsourcing is more lightweight to use and can have access to a much wider audience of potential annotators.

By performing a set of basic machine learning experiments using a feature set which mainly encodes basic lexical and morphosyntactic information, we further illustrate how the collected data can be used to perform text comparisons or to assign an absolute readability score to an individual text. We do not focus on optimizing the algorithms to achieve the best possible results for the learning tasks, but carry them out to illustrate the various possibilities of our data sets. The results on the different data sets, however, show that our system outperforms the readability formulas and a baseline language modeling approach.

We conclude that readability assessment by comparing texts is a polyvalent methodology, which can be adapted to specific domains and target audiences if required.

1 Introduction

What is it that makes a particular text easy or hard to read? This question is central to any automatic readability prediction system. Ideally, the task is based on a clear-cut definition of the concept "readability". Since the emergence of the domain, however, the terms "readability" and "clarity" have been defined in a wide variety of ways, typically dependent on the intentions of the author. If one wishes, for instance, to have a criterion to select reading material, the concept of readability will evidently include the reading proficiency needed for text comprehension. This was one of the underlying principles in earlier work on readability formulas (e.g. (Dale & Chall, 1948; Gunning, 1952; Kincaid *et al.*, 1975; Staphorsius, 1994) etcetera.). Another perspective on the concept of readability was proposed by DuBay (2004) who defined it as "what makes some texts easier to read than others".

Automatic readability prediction has a long and rich tradition. Research in the 20th century, fueled especially by educational purposes, has resulted in a large number of readability formulas. Typically, these yield either an absolute score (Flesch, 1948; Brouwer, 1963) or a grade level for which a text is deemed appropriate (Dale & Chall, 1948; Gunning, 1952; Kincaid *et al.*, 1975) and are based on shallow text characteristics such as average word and sentence length and word familiarity. Owing to the advances in the fields of natural language processing (NLP) and machine learning (ML), readability research has seen a revival in the past decade or so. In recent studies, readability has been linked with more complex lexical and syntactic text characteristics (Schwarm & Ostendorf, 2005; Petersen & Ostendorf, 2009) and more recently, discourse features capturing local and global coherence across text are also being scrutinized (Graesser *et al.*, 2004; Pitler & Nenkova, 2008; Feng *et al.*, 2010).

Simultaneous to the introduction of modern NLP techniques in readability research, there has been an increased interest in private and public organisations to produce readable documents. The impetus for this has been a mixture of consumer demands and mainly government and industry regulation. Notable initiatives are the Clarity Campaign[1], the U.S. Plain Writing Act of 2010, which aims at enhancing citizen access to government information and services, the European Clear Writing Campaign, etc. These efforts have led to numerous clear writing guidelines tailored to specific domains and text types. Most of these efforts provide guidelines for writing although there are some semi-automatic authoring tools as well (see for example (Hoste *et al.*, 2010)). Another field in which readability research has gained a lot of interest is the medical health domain because of the "vocabulary gap" between the domain and its audience (Zeng *et al.*, 2008; Leroy *et al.*, 2010; Leroy & Endicott, 2011). Readability applications are not only interesting to determine the readability level of a given text, but could also aid to manage the information redundancy current society is faced with. For example, given the redundancy of information present on the web, the readability of web documents could be a criterion to rank

[1] `http://www.clarity-international.net`

retrieved documents (Kanungo & Orr, 2009; Kate *et al.*, 2010), or to summarize or automatically translate on-line content.

One of the well-known bottlenecks in data-driven NLP research is the lack of sufficiently large data sets for which annotators provided labels with satisfying agreement. Readability research is no exception to this rule. Since readability prediction was initially designed to identify reading material suited to the reading competence of a given individual, most of the existing data sets are drawn from textbooks and other sources intended for different competence levels (François, 2009; Heilman *et al.*, 2008; Feng *et al.*, 2010). Although recent annotation efforts have also tackled other text types (e.g. (Kate *et al.*, 2010; Feng *et al.*, 2010)), we are not aware of any publicly available data sets with readability assessments for generic texts.

In this paper, we investigate two different methodologies to assess the readability of a wide variety of text material, starting from a corpus of which the readability was previously unlabeled. To this end, we have designed and implemented two web applications that can be used to easily collect readability assessments from human annotators over any corpus. The first is a more traditional labeling set-up in which language experts are asked to sort texts based on readability. In the second set-up, we investigated the viability of a more lightweight crowdsourcing application in which users are confronted with two texts and asked to compare them on a five-point scale. In order to validate this approach, readability assessments for a corpus of written Dutch texts were gathered. By collecting multiple assessments per text, we explicitly aimed to level out the reader's background knowledge and attitude. Our findings show that the assessments collected through both annotation tools are highly consistent, i.e. the language experts agree with the users of the crowdsourcing application on how the texts should be ranked in terms of readability. This is good news as the crowdsourcing application is more lightweight to use and can tap into a much wider audience of potential annotators.

By performing some basic machine learning experiments, we further illustrate how the collected data can be used to perform text comparisons (e.g. allowing to compare different text pairs) or to assign readability scores to an individual text (e.g. to a legal text, insurance policy, Patient Information Leaflet or to reading material for language learners). In order to account for the latter case, we defined a readability measure that can be estimated from the data. The main purpose of the machine learning experiments reported in this paper is to demonstrate the various possibilities of our data set, no optimization has been carried out.

The remainder of this paper is organized as follows. Section 2 motivates the present study and presents an overview of related work on automatic readability prediction. Section 3 describes the two annotation strategies we propose for collecting readability assessments. In Section 4, we give an overview of the different basic machine learning experiments, followed by a description of the results in Section 5. Section 6 ends with a discussion and prospects for future research.

2 Related work

Since the first half of the 20th century, readability formulas have been widely used to automatically predict the readability of an unseen text. Although they were primarily intended to select reading material for language learners, their being part of editing environments such as MS Word illustrates a widespread usage. A readability formula can be described as a mathematical formula, typically consisting of a number of variables (i.e. text characteristics) and constant weights, intended to grasp the difficulty of a text. For some formulas a higher and for others a lower score indicates a more difficult text. The initial goal of readability tests, as for example conceived by McCall and Crabbs (Flesch, 1948), was to determine the reading proficiency of an individual reader on the basis of a set of standardised reading tests and to use readability formulas to select appropriate reading material. As Bailin and Grafstein (2001) state, readability formulas appeal because they are believed to objectively and quantifiably evaluate the difficulty of written material without measuring characteristics of the readers. Text characteristics typically figuring as variables are the average word length in number of syllables or characters, the average sentence length in number of words, the type/token ratio, the percentage of words also found in a pre-assembled word list of frequently used words, etc. These overall simple lexical and syntactic features are obtained by processing a text automatically.

Many objections have been raised against the classical readability formulas: their lack of absolute value (Bailin & Grafstein, 2001), the fact that they are solely based on superficial text characteristics (DuBay, 2004; DuBay, 2007; Kraf & Pander Maat, 2009; Feng *et al.*, 2009; Alice Davinson and Robert N. Kantor, 1982), the underlying assumption of a regression between readability and the modeled text characteristics (Heilman *et al.*, 2008), etc. Furthermore, there even seems to be a remarkably strong correspondence between the readability formulas, even across different languages (van Oosten *et al.*, 2010).

Thanks to advancements in the field of NLP new features have been introduced. Recent research on readability prediction using machine learning started with adding statistical language modelling (Si & Callan, 2001; Collins-Thompson & Callan, 2005) and later also more complex syntactic features introducing a text's complexity based on parse trees (Schwarm & Ostendorf, 2005; Heilman *et al.*, 2008) were added. More recently, semantic (vor der Brück *et al.*, 2008) and discourse features (Pitler & Nenkova, 2008; Feng *et al.*, 2010; Leroy & Endicott, 2011) are also being scrutinized. These and other studies have shown that more complex linguistic features are useful, however, the discussion on which features are the best predictors remains open. While Pitler and Nenkova (2008) have clearly demonstrated the usefulness of discourse features, their predictive power was not corroborated by for example Feng (2010). Nevertheless, we can deduct from previous research that features which are lexical in nature, such as language modeling features, have a strong predictive power. Besides various features, more intricate prediction methods such as Naive Bayes classifiers (Collins-Thompson & Callan, 2004), logistic regression (François, 2009) and support vector machines (Schwarm & Ostendorf, 2005; Feng *et al.*, 2010; Tanaka-Ishii *et al.*, 2010) have come into use.

Since the main focus of readability research, until recently, has been on finding appropriate reading material for language learners, most of the existing data sets are built on underlying corpora of educational material, such as school textbooks and comparable corpora that have been collected and studied representing various reader levels (Petersen & Ostendorf, 2009; Feng *et al.*, 2010). For Dutch, for example, the only large-scale experimental readability research (Staphorsius & Krom, 1985; Staphorsius, 1994) is limited to texts for elementary school children[2]. For English, the situation is similar, viz. a predominant focus on educational corpora. Notable exceptions are those corpora explicitly designed to represent an actual difference in readability based on its envisaged end-users (i.e. people with intellectual disabilities (Feng *et al.*, 2010)) or text genre (i.e. medical domain (Leroy & Endicott, 2011)). Recently, a more general corpus was assembled which is not tailored to a specific audience, genre or domain by the Linguistic Data Consortium (LDC) in the framework of the DARPA Machine Reading Program (Kate *et al.*, 2010). However, these data have not been made publicly available.

When constructing data sets for readability prediction, the inherent subjectivity of the concept of readability cannot be ignored. Other labeling tasks in NLP, such as annotating part-of-speech or coreferential relations, are all based on a set of predefined guidelines to which the annotators have to adhere. The agreement between the annotators evidently depends on the complexity of the annotation task and can lead to a further clarification and refinement of the underlying annotation guidelines. Readability labeling requires a different approach. The ease with which a given reader can correctly identify the message conveyed in a text is, among other things, inextricably related to the reader's background knowledge of the subject at

[2] Moreover, (Staphorsius, 1994) based his research entirely on cloze-testing. A cloze-test is a reading comprehension test introduced by (Rankin, 1959) in which test subjects are required to fill in automatically deleted words in an unseen text. It is unclear whether such tasks are actually suitable to estimate the readability of a text.

hand (McNamara *et al.*, 1993). The construction of a corpus that is to serve as a gold standard against which new scoring or ranking systems can be tested, thus requires a multifaceted approach taking into account both the properties of the text under evaluation and those of the readers. In recent years, a tendency seems to have arisen to explicitly address this subjective aspect of readability. Pitler and Nenkova (2008), for example, base their readability prediction method exclusively on the extent to which readers found a text to be "well-written" and Kate *et al.* (2010) follow a similar approach by defining readability assessment as "a subjective judgment of how easily a reader can extract the information the writer or speaker intended to convey".

The traditional way to collect readability assessments is to let people assign absolute scores to each text and use the resulting mean readability score. Pitler and Nenkova (2008) and Kate *et al.* (2010), for example, average out results collected from different readers. In our approach, however, all texts in the corpus are compared to each other by different people and by using different comparison mechanisms, i.e. pairwise comparison and ranking. By collecting multiple assessments per text, we aim to level out the reader's background knowledge and attitude as much as possible. This annotation process results in sets of text pairs, accompanied with a relative readability assessment per pair. This is similar to the work of Tanaka-Ishii *et al.* (2010) who also used text pairs. However, Tanaka-Ishii *et al.* (2010) required only two general classes of text to make pairs: easy versus difficult. In our approach, we not only allow for a more fine-grained comparison, but also for a comparison of all texts under evaluation.

3 Assessing Readability

For the construction of a readability prediction system, roughly three steps can be distinguished. First of all, a *readability corpus* containing text material of which the readability will be assessed must be composed. Second, a *methodology* to acquire readability assessments should be defined. Finally, based on the readability corpus and the acquired assessments, *prediction tasks* can be performed. The first two steps are extensively discussed in this section.

3.1 Readability Corpus

The corpus that was used for the readability assessments contains 105 texts extracted from the Dutch LassyKlein corpus (van Noord, 2009)[3]. LassyKlein is a syntactically annotated corpus and recently its texts have been enriched with semantic information (Schuurman *et al.*, 2010), which makes it a particularly interesting data set for further readability research. From the selected texts, we extracted snippets of between 100 and 200 words for readability assessment[4]. Most of these snippets are extracted from a larger context but are meaningful by themselves. This was objectively measured by letting two trained linguists rank 12 full texts and their 12 snippets as more difficult, less difficult or equally difficult independent of each other with an interval of one week (i.e. 11 full text pairs and 11 snippet pairs). We opted for an interval of one week, based on the assumption that the annotators could

[3] LassyKlein is distributed by the Flemish-Dutch Human Language Technology Agency (TST-Centrale)

[4] This was necessary in order to properly visualize the texts in our envisaged assessment tools

still remember (part of) the contents of both full texts and the snippets after one week, but forget about how they ranked both groups of texts. This was confirmed by both annotators. The agreement between the ranking of full texts and snippets is 90.9% for the first annotator and 100% for the second; the snippets can thus be considered as viable alternatives. When we refer in the following sections to the texts that were assessed, we actually mean the snippets.

The corpus which served as the basis for the assessments consists of texts coming from four different genres, viz. administrative texts, manuals (e.g. also including patient information leaflets), news articles and a miscellaneous section. This proportion of texts per genre reflects the corpus design of the underlying LassyKlein corpus. In Table 1, an overview is given of the number of texts and tokens included in the readability corpus together with their average sentence length.

Genre	#Documents	#Tokens	Avg senl.
Administrative	21	3,463	19.95
Journalistic	65	8,950	17.91
Manuals	8	1,108	15.98
Miscellaneous	11	1,559	19.10
Total	105	15,080	18.29

Table 1. Readability corpus data statistics

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. Although we are fully aware of the shortcomings of the existing readability formulas, we confronted our intuitive selection with the output of two classical readability formulas designed for Dutch so as to objectify the selection as much as possible. The formulas we used, are the Flesch-Douma formula given in (1) (Douma, 1960) and the Leesindex Brouwer given in (2) (Brouwer, 1963). Both are adaptations of the well-known English Flesch reading ease formula given in (3) (Flesch, 1948). As one can see, these formulas are based on shallow text characteristics such as the average sentence length (*avgsentencelen*) and the average number of syllables per word (*avgnumsyl*) in a particular text. The latter was calculated for Dutch by implementing a classification-based syllabifier (van Oosten *et al.*, 2010).

$$Douma = 207 - 0.93 \cdot avgsentencelen - 77 \cdot avgnumsyl \tag{1}$$

$$Brouwer = 195 - 2 \cdot avgsentencelen - 67 \cdot avgnumsyl \tag{2}$$

$$Flesch = 206.835 - 1.015 \cdot avgsentencelen - 84.6 \cdot avgnumsyl \tag{3}$$

Fig. 1 presents an overview of the readability scores each formula assigns to the texts in each of the genres. It clearly illustrates that the median values are all quite close across the different genres, e.g. between the journalistic texts and manuals according to Douma and Brouwer. Furthermore, the large spread both in terms of difference between the minimum and maximum values and interquartile range (difference between third and first quartile) makes us confident that our corpus presents various levels of readability per genre.

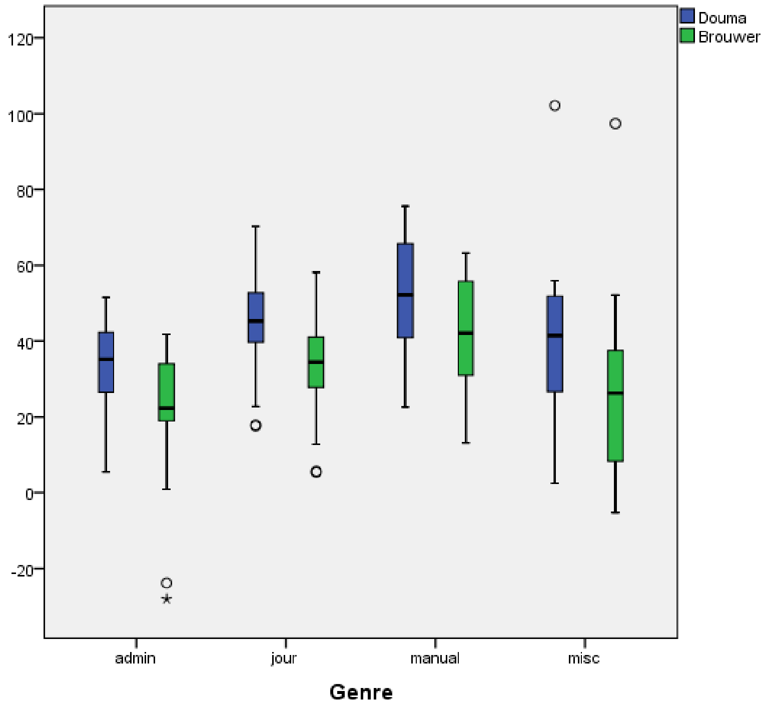


Fig. 1. Box-and-whisker plots representing the scores of the readability formulas on the different genres. The maximum and minimum values in the data set (except for the outliers) are represented as the upper end of the top whisker and the lower end of the bottom whisker. The first and third quartile are displayed as the bottom and the top of the box, the median as a horizontal stroke through the box. The outliers – indicated by a circle – are the scores which lie more than 3 times outside the interquartile range (IQR).

3.2 Two methodologies for readability assessment

While human expert annotation is crucial for many natural language processing tasks, it is often very expensive and time-consuming which explains the increasing success of cheaper and faster non-expert contributors over the Web (e.g. (Poesio *et al.*, 2008; Snow *et al.*, 2008; Finin *et al.*, 2010)). An important prerequisite, however, is that using non-expert labels for training machine learning algorithms is as effective as using gold standard annotations from experts.

Since we envisaged to collect multiple assessments per text in order to level out a reader's background knowledge and attitude as much as possible, we hypothesized that a crowdsourcing approach could be a viable alternative to expert labels. Snow *et al.* (2008), for example, demonstrated the effectiveness of using Amazon Mechanical Turk for a variety of NLP tasks, they found that for many tasks only a small number of non-expert annotations per item was necessary to equal the performance of an expert annotator. The task of assigning readability assessments to texts, however, is quite different from labeling tasks where a set of predefined guidelines have to be followed. Readability assessment remains largely intuitive, even in the case where annotators get instructions to pay attention to syntactic, lexical, etc. complexity when assigning a readability level to a given text. But then again, this lack of large sets of guidelines might be another motivation for the use of crowdsourcing.

This is why we decided to explore two different methodologies to collect readability assessments for our corpus, viz. a more classical expert labeling approach, in which we collected assessments of language professionals (i.e. teachers, linguists, writers), and a lightweight crowdsourcing approach.

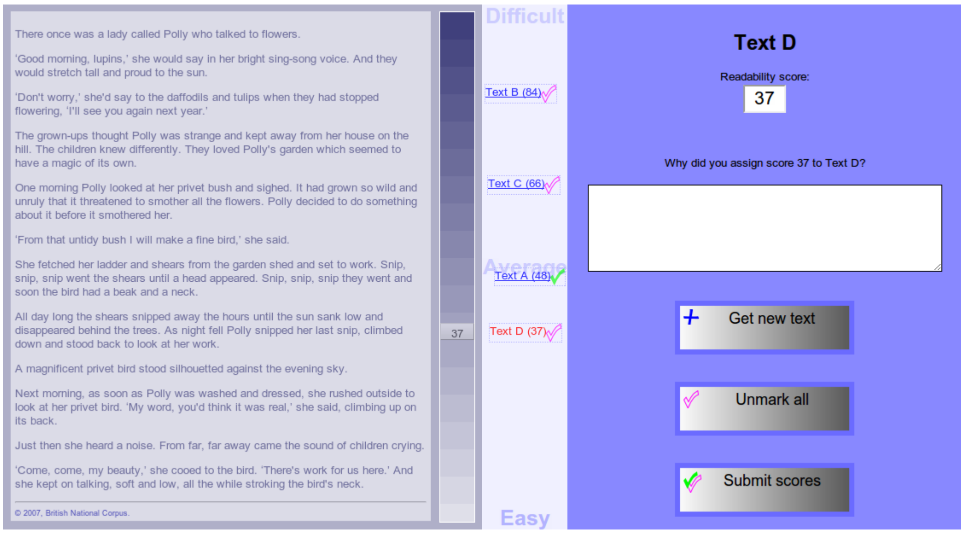
In designing both annotation methodologies, we started from DuBay's definition: "what makes some texts easier to read than others" (DuBay, 2004), a definition which specifically returns in the labels which can be assigned to text pairs in the crowdsourcing application. Since we wanted to level out individual background as much as possible, we added "to the community of language users" as an additional constraint.

3.2.1 Expert Readers

With the *Expert Readers* application[5] (see Fig. 2) we envisaged a more traditional approach to readability assessment in which experts decide on the level of readability of a given text. The interface allows the assessors to place each text in a correct position according to their readability perception, like in a slide sorter view in presentation software. Through this ranking set-up, the number of pairwise comparisons being performed grows quadratically with each assessed text. The experts are furthermore given the option to assign absolute scores to texts; this ordering of the scores defines the ranking. The benefit of the latter option is that it also allows users to easily assign the same score to multiple equally readable texts.

The application is only open to persons who are professionally involved with the

[5] The Expert Readers application is accessible at the password-protected url `http://lt3.hogent.be/tools/expert-readers-nl/`



There once was a lady called Polly who talked to flowers.

'Good morning, lupins,' she would say in her bright sing-song voice. And they would stretch tall and proud to the sun.

'Don't worry,' she'd say to the daffodils and tulips when they had stopped flowering, 'I'll see you again next year.'

The grown-ups thought Polly was strange and kept away from her house on the hill. The children knew differently. They loved Polly's garden which seemed to have a magic of its own.

One morning Polly looked at her privet bush and sighed. It had grown so wild and unruly that it threatened to smother all the flowers. Polly decided to do something about it before it smothered her.

'From that untidy bush I will make a fine bird,' she said.

She fetched her ladder and shears from the garden shed and set to work. Snip, snip, snip went the shears until a head appeared. Snip, snip, snip they went and soon the bird had a beak and a neck.

All day long the shears snipped away the hours until the sun sank low and disappeared behind the trees. As night fell Polly snipped her last snip, climbed down and stood back to look at her work.

A magnificent privet bird stood silhouetted against the evening sky.

Next morning, as soon as Polly was washed and dressed, she rushed outside to look at her privet bird. 'My word, you'd think it was real,' she said, climbing up on its back.

Just then she heard a noise. From far, far away came the sound of children crying.

'Come, come, my beauty,' she cooed to the bird. 'There's work for us here.' And she kept on talking, soft and low, all the while stroking the bird's neck.

© 2007, British National Corpus.

Text D

Readability score: 37

Why did you assign score 37 to Text D?

Fig. 2. A screenshot of the *Expert Readers* web application.

language under consideration, i.e. experts[6]. The experts can express their opinion by ranking texts on a scale of 0 (easy) to 100 (difficult), which allows them to compare texts while at the same time assigning absolute scores. Although it is unlikely that people can accurately assign scores on a 100 point scale (Cox, 1980), we decided to provide such an extensive scale because when many texts are compared at the same time, there should be a possibility to express differences, no matter how small these are. As a special feature, texts can be ticked as anchor points, which means that they are always kept in memory. As such, they can be used as a frame of reference which facilitates future rankings.

As experts, we specifically aimed at people who are professionally involved with language, thus following a more traditional data collection methodology. The experts are asked to assess the readability for language users in general. We delib-

[6] Ranking multiple texts is difficult and the interface requires users to first read through a user manual, which is provided along with an instruction movie.

erately do not ask more detailed questions about certain aspects of readability, because we want to avoid influencing the text properties experts pay attention to. Neither do we inform the experts in any way on how they should judge readability. Any presumption about which features are important readability indicators is thus avoided. However, the experts were offered the possibility to motivate or to comment on their assessments via a free text field in the interface. These comments, e.g. on the syntactic complexity of a given text fragment, will be used in future experiments integrating more high-level syntactic and semantic features.

Our current pool of active experts consists of 36 teachers, writers and linguists, who have contributed a total of 1,862 text rankings. As far as the ranking itself is concerned, we observed that every expert did mark some texts as anchor points, on average nine texts were kept in the frame of reference of each expert. This behavior confirms that the ranking task was well understood by the annotators, i.e. the expert rank the text relative to the other texts present in a batch instead of considering them all independently of each other. Per annotation session, the assessments were stored in a batch. We decided to only take into account submission batches where at least five texts were compared to each other. The sizes of the batches range from five to all available texts. As can be expected, different experts employ their own standards to assign readability scores on a scale from zero to 100. Fig. 3 clearly illustrates the large variability in the scores assigned to the individual texts (X-axis), viz. a large spread between the minimum and maximum scores.

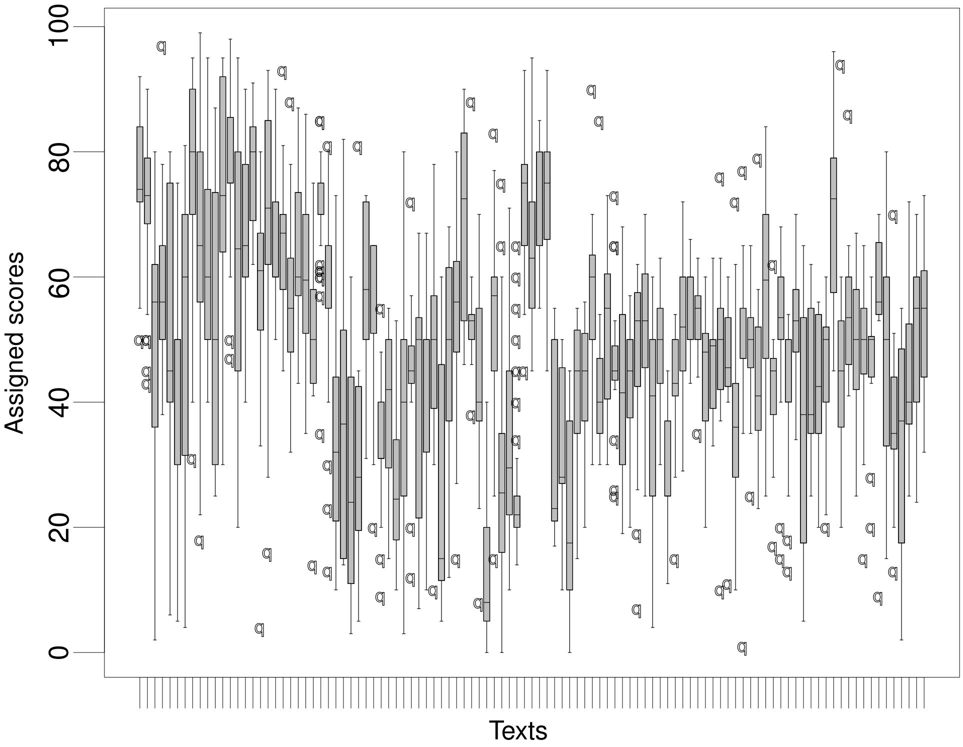
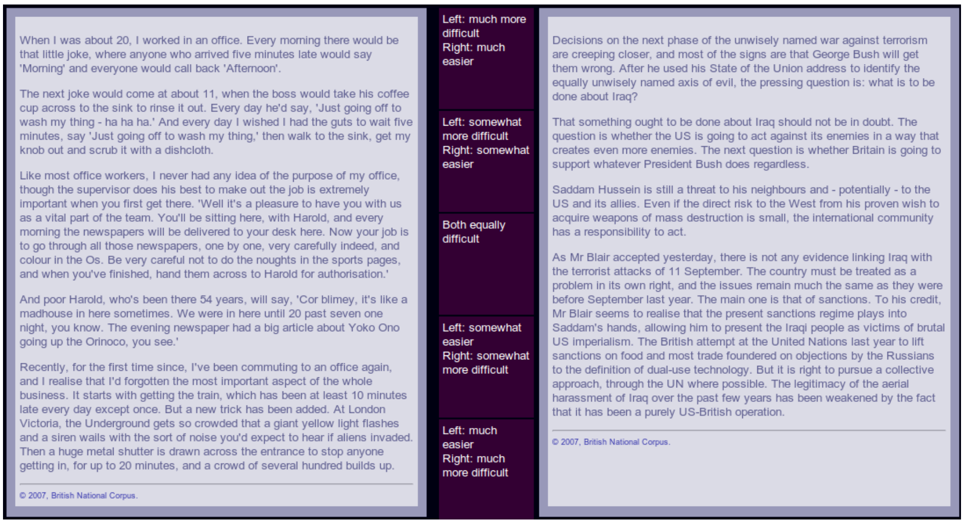


Fig. 3. Box-and-whisker plots showing the minimum, first quartile, median, third quartile, the maximum and the outliers for the scores assigned to each text in the Expert Readers application. A box plot is displayed for each text in which the minimum score that was assigned to each text is indicated as the lower end of the bottom whisker and the maximum score is the upper end of the top whisker, unless the distance from the minimum (maximum) value to the first (third) quartile is more than 1.5 times the IQR, in which case they are drawn as outliers. The first and third quartile are displayed as the bottom and top of the box, the median as a horizontal stroke through the box.



When I was about 20, I worked in an office. Every morning there would be that little joke, where anyone who arrived five minutes late would say 'Morning' and everyone would call back 'Afternoon'.

The next joke would come at about 11, when the boss would take his coffee cup across to the sink to rinse it out. Every day he'd say, 'Just going off to wash my thing - ha ha ha.' And every day I wished I had the guts to wait five minutes, say 'Just going off to wash my thing,' then walk to the sink, get my knob out and scrub it with a dishcloth.

Like most office workers, I never had any idea of the purpose of my office, though the supervisor does his best to make out the job is extremely important when you first get there. 'Well it's a pleasure to have you with us as a vital part of the team. You'll be sitting here, with Harold, and every morning the newspapers will be delivered to your desk here. Now your job is to go through all those newspapers, one by one, very carefully indeed, and colour in the Os. Be very careful not to do the noughts in the sports pages, and when you've finished, hand them across to Harold for authorisation.'

And poor Harold, who's been there 54 years, will say, 'Cor blimey, it's like a madhouse in here sometimes. We were in here until 20 past seven one night, you know. The evening newspaper had a big article about Yoko Ono going up the Orinoco, you see.'

Recently, for the first time since, I've been commuting to an office again, and I realise that I'd forgotten the most important aspect of the whole business. It starts with getting the train, which has been at least 10 minutes late every day except once. But a new trick has been added. At London Victoria, the Underground gets so crowded that a giant yellow light flashes and a siren wails with the sort of noise you'd expect to hear if aliens invaded. Then a huge metal shutter is drawn across the entrance to stop anyone getting in, for up to 20 minutes, and a crowd of several hundred builds up.

© 2007, British National Corpus.

Left: much more difficult Right: much easier
Left: somewhat more difficult Right: somewhat easier
Both equally difficult
Left: somewhat easier Right: somewhat more difficult
Left: much easier Right: much more difficult

Decisions on the next phase of the unwisely named war against terrorism are creeping closer, and most of the signs are that George Bush will get them wrong. After he used his State of the Union address to identify the equally unwisely named axis of evil, the pressing question is: what is to be done about Iraq?

That something ought to be done about Iraq should not be in doubt. The question is whether the US is going to act against its enemies in a way that creates even more enemies. The next question is whether Britain is going to support whatever President Bush does regardless.

Saddam Hussein is still a threat to his neighbours and - potentially - to the US and its allies. Even if the direct risk to the West from his proven wish to acquire weapons of mass destruction is small, the international community has a responsibility to act.

As Mr Blair accepted yesterday, there is not any evidence linking Iraq with the terrorist attacks of 11 September. The country must be treated as a problem in its own right, and the issues remain much the same as they were before September last year. The main one is that of sanctions. To his credit, Mr Blair seems to realise that the present sanctions regime plays into Saddam's hands, allowing him to present the Iraqi people as victims of brutal US imperialism. The British attempt at the United Nations last year to lift sanctions on food and most trade foundered on objections by the Russians to the definition of dual-use technology. But it is right to pursue a collective approach, through the UN where possible. The legitimacy of the aerial harassment of Iraq over the past few years has been weakened by the fact that it has been a purely US-British operation.

© 2007, British National Corpus.

Fig. 4. A screenshot of the *Sort by Readability* web application.

3.2.2 Crowdsourcing

Our crowdsourcing application, called *Sort by Readability*[7], is intuitive and user-friendly and is designed to be used by as many people as possible. The website is accessible to anyone having internet access. Users are not required to provide personal data.

A screenshot of the crowdsourcing application is shown in Fig. 4.

Two texts are displayed simultaneously and the user is asked to tick one of the statements in the middle column, corresponding to the five-point scale in Table 2.

After clicking one of the options, a text pair and its corresponding assessment are added to the database and two new randomly selected texts appear. To avoid that respondents click on one of the buttons too soon, i.e. without reading the texts, the

[7] The Sort by Readability application can be accessed through the following url: `http://lt3.hogent.be/tools/sort-by-readability-nl/`

Acronym	**Meaning**
LME	left text much easier
LSE	left text somewhat easier
ED	both texts equally difficult
RSE	right text somewhat easier
RME	right text much easier

Table 2. Five-point scale for the assessment of the difference in readability between two texts

buttons are disabled during a few seconds after the appearance of a new text pair. The only instructions respondents are given, are the following two sentences on the landing page of the application:

> "Using this tool, you can help us compose a readability corpus. You are shown two texts for which you can decide which is the more difficult and which is the easier one."

As was done for the experts, we gave no further instructions because we did not want to influence anyone on how to perceive readability. Since we deliberately chose to keep the crowdsourcing tool open to everyone, we do not know who performed the assessments. In the start-up phase, the tool was widely advertised among friends, family, students, etc., which might have caused a bias towards more educated labelers. But evidently, we do not have a clear view on the identity and background of the people who provided assessments. We can state with certainty, however, that the users of the crowdsourcing application differ from the experts selected for the first application (Section 3.2.2).

At the time of writing, 8,568 comparisons were performed and the number of assessments per text pair varies from 1 to 8. We also evaluated the number of times each button was pressed in the *Sort by Readability* application and found that the users are generally not biased towards finding texts on one side easier than on the other side.

3.3 From Assessments to Assessed Text Pairs

In order to compare the information collected with both applications, all data is converted to data points of the form (t_1, t_2, a), called *assessed text pairs*. An assessed text pair is a triplet in which t_1 and t_2 are texts from the readability corpus that are compared to each other and $a \in \{\mathit{LME, LSE, ED, RSE, RME}\}$ is the *readability assessment* for the text pair, i.e. the relative difference in readability between t_1 and t_2.[8]

Data collected through the crowdsourcing application is already in this format. Converting an expert's data to triplets is not as straightforward. At first sight, an intuitive way to work with the absolute expert scores would be to use the mean of all readability scores assigned to a text. Pitler and Nenkova (2008) and Kate *et al.* (2010), for example, average out results collected from different readers. Problems with this approach, however, immediately arise. The texts presented in each submission batch are selected randomly which implies that the annotator can be confronted with predominantly less or more readable texts, which may affect the

[8] Referring to t_1 as the "left text" and t_2 as the "right text". For the meaning of the acronyms we refer back to Table 2.

scoring. Moreover, different experts employ different standards to assign readability scores to texts. Being given the choice to label texts with marks between 0 (most readable) and 100 (least readable), some annotators decided to use a more coarse-grained labeling (e.g. by using multiples of 5, 10 or 20), whereas others used a fine-grained scoring. A similar conclusion was drawn by for example Anderson and Davison (1986) on a data set annotated with school grade levels by multiple annotators. Annotators may also use a different range of scores: while some may label a very readable text as 0 or 10, others might choose 30 as a lower bound, in case even more readable texts would later have to be scored. Some annotators used the maximum scoring range of 100 (between 0 and 100), while others scored between an upper and lower bound (e.g. 30-90, a scoring range of 60).

Fig. 3 clearly shows the large spread in scores per text. This variation can be explained by two factors: annotators' different perspectives on whether a text is hard to read, and their different scoring strategies. This prohibits a simple averaging of the scores: a good average should only reflect the former factor, and level out the latter as much as possible.

To normalize the differences between annotator scoring strategies, we propose to calculate the weighted normalized difference (WND) for each text pair coming from the experts. In a first step, we calculate the normalized difference (ND) of a text pair i, in which we account for the different scoring strategies of the experts:

$$ND_i = \frac{s_2 - s_1}{R} \tag{4}$$

, where s_1 and s_2 are the absolute scores of the first and the second text, and R is the scoring range of an individual expert over all batches. When dividing by R, differences in range are leveled out by rescaling every score between 0 (for the most readable text) and 1 (for the least readable text), while keeping the relative distances between scores intact.

Suppose the artificial first batches [9] listed in Table 3 (columns 2 and 3), coming from two experts. Note that the absolute scores differ between both batches, but the relative distances between scores are identical. If we would compare pair Text$_b$-Text$_e$ of the first expert with the same pair of the second expert simply by taking the average of the differences in scores, the difference in readability scored by the first expert (80-30=50) would be considered much more prominent than the second one (74-54=20). If, on the contrary, we also incorporate the range of each expert, the influence of scoring range is erased $\left(\frac{80-30}{60} = \frac{74-54}{24} = \frac{5}{6}\right)$.

Because the experts could submit their scores in different batches, they were sometimes confronted with texts which they already scored in earlier sessions, potentially leading to identical text pairs with different assessments (see for example the artificial Batch 2 from Expert 2 in Table 3). In order to account for these identical pairs with different scores, we group all the ND scores for that text pair from one expert using a weighted average. The canonical formula for a weighted average is given in Formula 5, where x_i is one element from a vector of values to be av-

[9] We would like to stress that these are artificial examples; because of place constraints, we were not able to include large batches

	Expert 1	**Expert 2**	**Expert 2**
	Batch 1	Batch 1	Batch 2
Text_a	20	50	51
Text_b	30	54	53
Text_c	50	62	
Text_d	60	66	
Text_e	80	74	60

Table 3. Three different artificial batches coming from 2 experts.

eraged, and w_i is the weight associated with that element. The weights determine how much a given element contributes to the weighted average.

$$\text{(5)} \qquad weighted\,average = \frac{\sum_{i=1}^{n} w_i.x_i}{\sum_{i=1}^{n} w_i}$$

In a weighted average, some data points contribute more to the average than others, depending on a specified weight. We specified weights W_i as a function of the batch size B_i in which a specific evaluation occurred:

$$\text{(6)} \qquad W_i = 1 - \exp\left(-\frac{B_i}{10}\right)$$

The weight ensures that assessments coming from small batches have a smaller influence on the average. We thus assign more confidence to assessments coming from bigger batches: we assume that the more texts an annotator compared in a batch, the more effort he/she made to position these texts. Moreover, the more texts a user is confronted with, the more likely it is that a particular batch will contain texts from various degrees of readability.

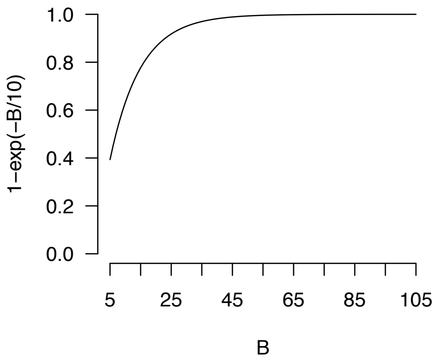


Fig. 5. The value of weight W_i as a function of batch size B. The minimum batch size in our corpus is 5, the maximum size is 105.

Fig. 5 shows how the weight function assigns mass to an evaluation based on batch size. Batches with the minimum size of 5 get a weight of approximately 0.4, after which the weight quickly increases with batch size towards 1.

We can thus define the weighted normalized difference (WND) of a text pair i by plugging our variables into the generic weighted average formula (Formula 5):

$$WND_i = \frac{\sum_{i=1}^{n} W_i.ND_i}{\sum_{i=1}^{n} W_i} \tag{7}$$

If we return to our example listed in Table 3 and more specifically to the text pair Text$_b$-Text$_e$, which was assessed twice by Expert 2, we obtain two different NDs. As exemplified in Table 4, normal averaging would result in a combined score

	Assessment One	**Assessment Two**
ND	$\frac{20}{24} \approx 0.83$	$\frac{7}{24} \approx 0.29$
weight	1-exp(- $\frac{5}{10}$) ≈ 0.39	1-exp(- $\frac{3}{10}$) ≈ 0.26
WND	$\frac{(0.83*0.39)+(0.29*0.26)}{0.39+0.26} \approx 0.614$	

Table 4. WND calculation example for 2 identical text pairs with different assessments from one single expert.

of 0.56, but by giving more importance to larger batches with a weighted average, the WND of 0.614 is closer to the assessment from the larger batch.

A WND can be calculated over different batches by a single annotator, but also over batches by multiple annotators. This is done by following the same procedure of calculating the ND first (taking into account the scoring range R of the annotator in question), and then averaging over all batches using their weights W. However, the ranking of texts may differ between batches, resulting in inconsistencies (e.g. when t_2 is ranked as more difficult than t_1 in one batch, and as easier in another).

In order to calculate the WND for text pair (t_1, t_2), we consider all NDs for that pair in that specific order. This may result in negative ND values, because (t_2, t_1, ND) (from a batch where t_1 is considered more difficult than t_2), can be rewritten as $(t_1, t_2, -ND)$ (following Formula 4). The same is true for any WND. If the WND of (t_1, t_2) is 0.5, the WND of (t_2, t_1) will be -0.5.

To finish, boundary values for each WND were determined to select an assessment from the five-point scale *LME, LSE, ED, RSE* or *RME* (Table 2) for each text pair (t_1, t_2). Two boundary values are required: if the absolute value of a WND is below the first value, the two texts in the text pair are regarded as equally difficult. Between the two boundaries, the readability of the texts is somewhat different and above the second boundary, there is much difference in readability between the two texts. The sign of the WND determines whether the first or the second text is the more readable one (a negative WND indicates that t_1 is harder to read than t_2). In other words, if the two boundary values are determined as b_1 and b_2, the five point scale *RME, RSE, ED, LSE, LME* corresponds to the following intervals: $[-1, -b_2],]-b_2, -b_1],]-b_1, b_1[, [b_1, b_2[, [b_2, 1]$. If b_2 is set at 0.5, a WND of -0.7 will fall in the $[-1, -0.5]$ interval, which corresponds to the *RME* label.

Different strategies can be followed to choose these boundary values. A simple approach is to choose ad hoc values. A somewhat more sophisticated method is to use values that lead to the same proportions of equally difficult, somewhat different or much different text pairs as in the crowdsourcing data set. In order to do so, we projected the number of times each button is pressed in the *Sort by Readability* application on the *Expert Readers* data set. This resulted in the boundary values 0.077 and 0.403.

3.4 Probabilistic Readability Measure

For some readability prediction tasks, a set of text comparisons is sufficient to learn the task. However, other tasks, such as predicting the reading difficulty of a document in isolation, require an actual readability score assigned to each individual text. In the latter case, it is important that the assigned readability score is meaningful. We define a readability score which is easy to estimate from the data sets. The readability of a text t_i can be identified as the probability $P_e(t_i)$ that t_i would be assessed as easier than any other text, by any assessor, i.e. the proportion of times that t_i was assessed as the easier text in a pair in which it occurs. Likewise, the unreadability (or difficulty) can be defined as the probability $P_m(t_i)$ that t_i would be assessed as more difficult than any other text.

$$\widehat{P_e}(t_i) = \frac{\#\{((t_i, t_j, a) \vee (t_j, t_i, a)) \| t_i \text{ is easier than } t_j\}}{\#\{(t_i, t_j, a) \vee (t_j, t_i, a)\}} \tag{8}$$

Likewise, the probability that t_0 is more difficult than any text in the corpus can be estimated by:

$$\widehat{P_m}(t_i) = \frac{\#\{((t_i, t_j, a) \vee (t_j, t_i, a)) \| t_i \text{ is more difficult than } t_j\}}{\#\{(t_i, t_j, a) \vee (t_j, t_i, a)\}} \tag{9}$$

We further use the notation $P_.(\cdot)$ for both the probability and its estimate. Note that in most cases $P_e(t_i) + P_m(t_i) < 1$, because for most texts, there are cases in which they are assessed as equally difficult as some other text in the corpus.

3.5 Correlations

In the previous sections we demonstrated how the output of both applications could be transformed into similar data sets, which now allows us to compare the readability assessments of the experts and the crowd and to verify whether crowdsourcing is indeed a viable alternative to expert labeling.

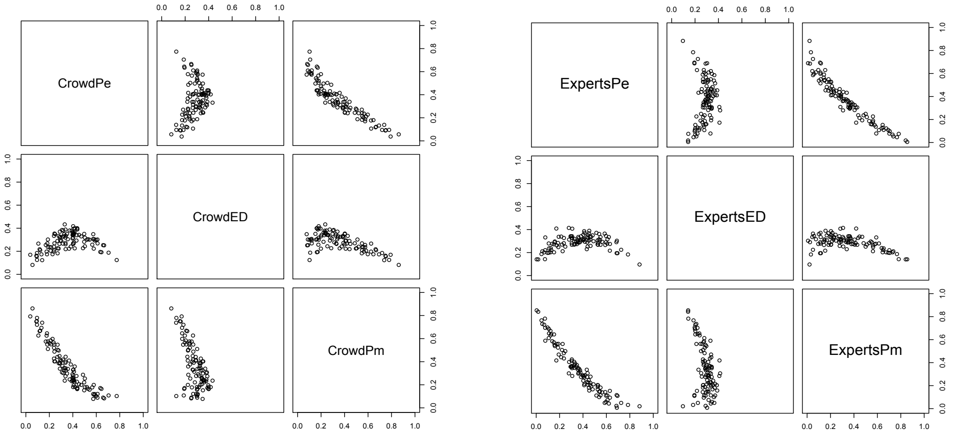
The data sets gathered through both applications, i.e. 8,568 text comparisons resulting from the *Sort by Readability* application (crowd) and 108 batches where at least five texts have been compared with the *Expert Readers* application (experts), were transformed to assessed text pairs (leading to 23,908 text comparisons), as described above in Section 3.3. This allowed us to know the actual probability measures (see Section 3.4) of each text according to both groups of assessors and discover some interesting correlations.

3.5.1 Experts versus crowd

The proportions with which each text has been assessed as easier (P_e), equally readable (ED) or more difficult (P_m) for both the experts and crowd data set are shown in Fig. 6. Since the lower left triangle and upper right triangle in both figures present the same information, we will limit the discussion to the lower left triangle. 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 6 in our crowd data set, this text has been assessed 0.77 times as easier, 0.12 times as equally difficult and 0.10 times as more difficult than any other text. These scores are visualised by the rightmost dot in the lower left plot of subfigure (a), corresponding to 0.77 on the lower left X-axis (easier) and a 0.12 on the lower right Y-axis (more difficult). The same text occupies the position 0.77 (easier) - 0.12 (equally difficult) in the left upper box and the position 0.12 (equally difficult) - 0.1 (more difficult) in the under middle box. We also observe that all plots show great similarity for both data sets.

In a next step we calculated the Pearson correlations between P_e and P_m, for both the *Expert Readers* data and the *Sort by Readability* data. We found that the correlation between Crowd P_e and Expert P_e is 86% and between Crowd P_m and

Expert P_m 90%. We can conclude that two very similar data sets are obtained from the applications, which means that experts rank the texts from the corpus in a very similar order as the crowd does by comparing these texts.



(a) (b)

Fig. 6. Proportion of times each text was assessed as easier, equally difficult or more difficult than any other text: (a) for the *Sort by Readability* data and (b) for the *Expert Readers* data. In both figures, the plots in the lower left triangle are transposed versions of those in the upper right triangle (the x and y axes are switched). They therefore present the same information.

3.5.2 Impact of the interface

Both groups were offered a different interface specifically tailored to their profile, i.e. experts used a rather complicated interface allowing them to enter (and comment on) fine-grained assessments whereas the crowdsourcing assessors were confronted with a very easy interface in which they had to compare two texts on a five-point scale. In order to verify that the assessments made by both groups were not biased by the different interface they used, we set up an experiment in which we asked five experts to assess twenty texts using both the *Expert Readers* and *Sort by Readability* application. To objectify the selection of these twenty texts as much as possible, we again relied on the Flesch-Douma formula given in (1) (see Section 3.1).

A comparison of the evaluations made in both applications is given in Fig. 7. On the x axis, three categories are plotted, taken from the crowdsourcing application evaluations: equally difficult (ED), somewhat different (SD) or very different (VD). We then looked up the WND scores (derived from the evaluations in the expert application) for all the text pairs present in a category, and plot these scores along the y axis. This results in one boxplot per category, showing the variability in (expert) WND scores assigned to pairs available in that (crowdsourcing) category.

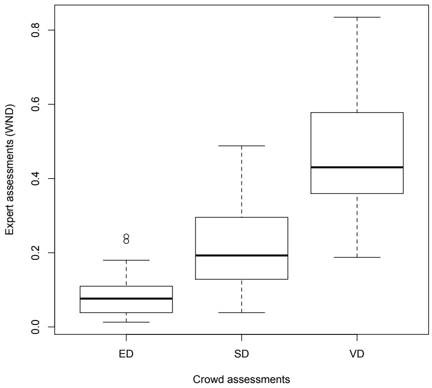


Fig. 7. Boxplot showing the correlation between the evaluations from one individual expert using both applications.

		Crowd tool		
		ED	SD	VD
	ED	22.6	7.5	0.0
Expert tool	SD	14.0	26.9	5.4
	VD	1.1	3.2	19.4

Table 5. Confusion matrix of the evaluations provided by one expert, taken from the crowdsourcing application and the sort-by-readability application (after conversion to pairwise assessments using the method described in Section 3.3). ED stands for equally difficult, SD for somewhat different and VD for very different. Values are in percent, over 93 pairs.

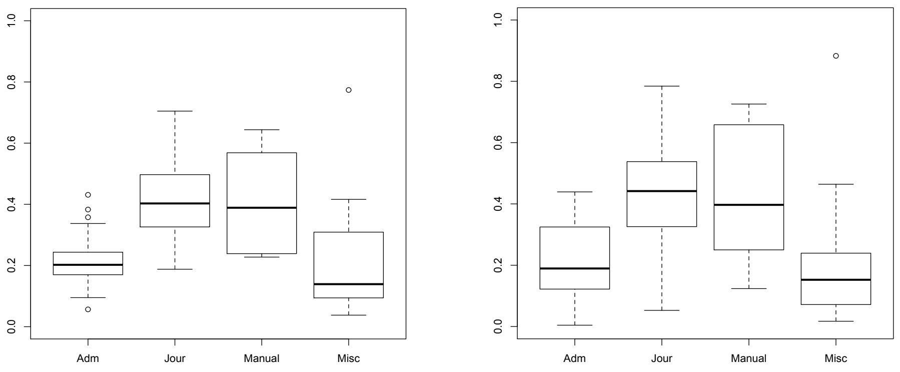
The boxplots of the three categories show some, but not much overlap. This indicates that the evaluations collected through one application are consistent with the ones collected through the other.

Another way of evaluating consistency across applications is presented in Table 5, which presents a confusion matrix between pairs assessed using the crowdsourcing application, and their equivalents from the expert data, converted and mapped to crowdsourcing evaluations using the method described in Section 3.3. Errors can be caused either by inconsistency on the part of the expert (evaluating text pairs differently in both applications), or due to the formula used for conversion from expert scores to pairwise comparisons.

Out of 93 evaluated pairs, 64 (69%) show no errors. Confusion is greatest between pairs that are assessed as equally difficult or somewhat different (20 errors on 66 pairs, or 30%) - the distinction between somewhat different and very different is easier (8 errors on 51 pairs, or 16%). Confusion between equally difficult and very different is only seen once (1 pair in 40, or 3%).

3.5.3 Genre versus readability

We also investigated how the experts and crowd assessed the texts from the different genres included in our readability corpus, in order to find out whether the different genres represent various readability levels. If one genre, for example, would predominantly consist of texts being assessed as difficult and another genre contains texts which are mainly labeled as easy to read, the task of readability prediction might boil down to the task of genre classification. The assessments for the different text genres are presented in Fig. 8 in which both the expert and crowd texts are ranked per genre according to the crowd P_e (a) and the experts P_e (b) (see Section 3.4. for the calculation of the probability measure P_e).



(a) (b)

Fig. 8. Ranking of the texts per genre according to the crowd P_e (a) and the experts P_e (b).

What strikes the eye immediately is the correspondence between the ranking of both the crowd (a) and the experts (b). Furthermore, within each genre we also perceive a spread in readability levels; every genre contains texts that have been assessed as more or less readable. Globally, however, their does seem to be a consistency between both groups, some texts that are perceived as less or more readable occur more in particular genres. We clearly see that the crowd perceives administrative text as more difficult than journalistic texts, the same is perceived by the experts. This is an interesting finding we would definitely like to investigate in closer detail in future work.

3.5.4 Experts and Crowd versus the classical readability formulas

Fig. 9 further illustrates that the readability formulas described in Section 3.1 (Flesch, 1948; Douma, 1960; Brouwer, 1963) strongly correlate with each other. The top left corner of Fig. 9, for example, compares the output of the Flesch Reading Ease formula with the predictions of Douma and reveals a very high correlation (reflected by the diagonal line). The bottom right corner, on the other hand, shows that the ranking of the texts according to their value for P_e correlates well between the crowdsourcing and the experts' data sets, which has also been demonstrated in Section 3.5.1. Furthermore, the correlation between each readability formula and each P_e value (represented in the six lower left plots and six upper right plots) is lower than the correlation between the P_e values themselves. These results confirm that both the experts and crowd assess the texts in a similar way, i.e. different than predicted by the various readability formulas. When designing a readability prediction system we should focus on predictors that provide better correlations, normally these will already outperform the readability formulas.

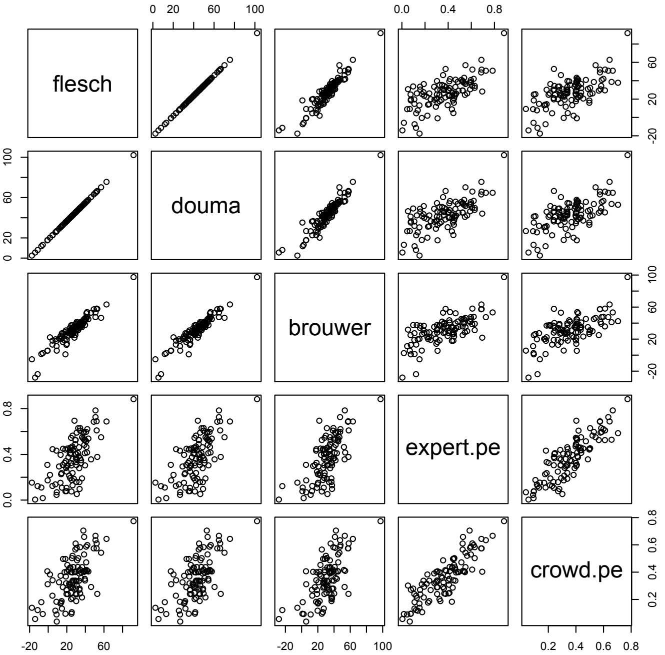


Fig. 9. Scatterplots showing the relations between three readability formulas and P_e for both data sets.

4 Predicting Readability

In the previous sections, we showed how the assessments collected with both web applications can be converted into *assessed text pairs* and how such a set can be used to estimate the readability of a text t, defined as the probability $P_e(t)$ that t would be assessed as easier than any other text. This approach allows us to construct a set of data points of the form $(t, P_e(t))$ in which each individual text t receives its own readability score. We came to the conclusion that the two readability assessment applications lead to two very similar data sets, which indicates that the users of the crowdsourcing application and the expert users rank the texts from the corpus in a very similar order. In the results displayed in Fig. 9, we saw that the correlation between the different readability formulas and each P_e value was lower than the correlation between the P_e (or P_m) values themselves. Based on this observation, we concluded that a predictor outperforms the readability formulas if it performs better on both (expert and crowd) P_e (or P_m) values than on the readability formulas.

In this section, we develop a basic readability prediction system and show how the data collected via both applications can be used in two different machine learning set-ups, one in which a given text receives a score/label and a second one in which texts are compared with each other .

4.1 Experimental set-ups

We experimented with two different experimental set-ups to show the possibilities of our data sets. In doing so, we believe that we cover both possible readability assessment scenarios. The first one allows to predict an absolute score to a given text, whereas the second one allows to compare two versions of a text. To this end, two different models were used: regression and classification. A schematic overview of all experiments is given in Table 6.

In the first experimental set-up, the task consists in **assigning an absolute readability score to a given text**, and more specifically, in predicting the value of $P_e(t)$ (see Section 3.4) for a text t[10]. In order to do so, we experimented with regression as a supervised learning model for readability prediction. *Regression* is a technique to predict a continuous value on the basis of a range of features.

[10] Note that a similar task can be defined for $P_m(t)$, but we further only focus on $P_e(t)$.

Experimental set-ups

Input	Output	Model
$t_1 ... \ t_n$	$P_e(t)$	regression
$(t_1, t_2) \ ... \ (t_{n-1}, t_n)$	LME, LSE, ED, RSE, RME (Table 2)	classification
	easier, more difficult	classification

Table 6. Schematic overview of the different readability experiments

Different algorithms to perform the regression task for readability prediction, i.e. logistic regression, SVM regression,... were recently used by vor der Brück *et al.* (2008), Heilman *et al.* (2008), Kanungo and Orr (2009), François (2009) and Kate *et al.* (2010).

The second experimental set-up aims at **comparing the readability of text pairs**. We defined two different subtasks.

The aim of the first experiment was to predict the correct assessment a from the five-point scale from Table 2 for the text pair (t_i, t_j). Recall however that the data sets of assessed text pairs collected with the annotation tools from Section 3.2 might contain inconsistencies, because different users might disagree in their assessment of the same text pair. For this reason, we considered two alternatives of the task, namely (1) a task in which we randomly sampled text pairs which occur more than once in the data, so as to avoid having perfect matches between the training and test data, and (2) a task in which we kept one assessment per text pair, calculated as the average over all available assessments for that text pair. The assessments for text pairs (t_i, t_j) and (t_j, t_i) are thereby merged. To calculate the average assessment, scores -2, -1, 0, 1 and 2 were associated with the LME, LSE, ED, RSE and RME classes, respectively. The mean of those scores per pair of texts was rounded and mapped back to a single assessment.

In a second experiment, the goal was to determine for a given text pair whether t_i was easier or more difficult than t_j, thus recasting the 5-class classification task as a binary choice. Text pairs which were assessed as equally difficult were discarded from the data set. Similar work on binary classification was for example done by Pitler and Nenkova (2008) and Tanaka-Ishii *et al.* (2010). A possible application would be, for example, to integrate a binary classifier into an editing environment, to detect whether edited versions of a text are more readable than the original. Again, we experimented on randomly sampled text pairs and on text pairs whose readability scores were averaged.

We experimented with different machine learning algorithms to perform the machine learning tasks. For the regression task, we used linear regression, whereas the classification tasks were performed using k-nn and random forests. For each algorithm, we used their implementation in GNU R [11].

Since the error measure is typically different in the context of supervised regression than for classification, we evaluated the performance of algorithms carrying out the regression task with the root mean squared error ($RMSE$) as the error to

[11] `http://www.r-project.org`

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, i.e. 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.

The evaluation of the classification experiments was done by measuring the classification accuracy (CA).

4.2 Features

A feature set has been extracted mainly capturing lexical and morpho-syntactic information. Previous research has shown that although more complex linguistic features are useful (Feng *et al.*, 2010), the most predictive ones are lexical in nature (Pitler & Nenkova, 2008; Zeng *et al.*, 2008; Kate *et al.*, 2010) and therefore constitute a viable base feature set. The following features were incorporated:

- Basic features representing text characteristics – which are also popular in the classical readability formulas – were extracted, including the average number of words per sentence, the average number of syllables[12]and characters per word, the proportion of words occurring in a list of the most frequently used words in Dutch and the proportion of words with three or more syllables.
- Character bigram and trigram frequencies were calculated. The underlying idea is that more deviation from the n-gram frequencies found in a reference corpus may result in a less natural looking text. This information can be

[12] By using a classification-based syllabifier (van Oosten *et al.*, 2010)

calculated as the sum of the average character bigram and trigram frequencies per word on the basis of a large reference corpus divided by the text length of the text under consideration. We used the Twente Nieuws Corpus[13] (TwNC) as reference corpus.

- Using the above-mentioned corpus, three additional statistical features were also extracted to detect those words that are specific to a text, based on corpus comparison: the mean TF-IDF value of all tokens in the texts, the average Log-Likelihood ratio and Mutual Information.

  TF-IDF originates from information retrieval and measures the relative importance or weight of a word in a document (Salton, 1989). We calculated TF-IDF for all terms in the readability corpus and to calculate the IDF we enlarged the readability corpus with all texts of the TwNC. Given a document collection D, a word w, and an individual document d in D,

$$W_d = f_{w,d} \cdot \log(|D|/f_{w,D})$$

  where $f_{w,d}$ equals the number of times w appears in d, $|D|$ is the size of the corpus and $f_{w,D}$ equals the number of documents in D in which w appears (Berger *et al.*, 2000). Calculating TF-IDF should thus enable us to extract those specific words in our texts that have much lower frequencies in the balanced background corpus. In short, the mean TF-IDF value of all the words in a large corpus estimates the mean importance of a word in any text.

  The Log-Likelihood ratio discovers keywords which differentiate between cor-

[13] `http://www.home.cs.utwente.nl/~druid/TwNC/TwNC-main.html`

	First Corpus	Second Corpus	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	c	c+d

Table 7. Contingency table to calculate Log-Likelihood

pora, in our case the TwNC corpus and the input text/corpus. We first produced a frequency list for each corpus and calculated the Log-Likelihood statistic for each word in the two lists. This was done by constructing a contingency table (see Table 7), where c represents the number of words in the TwNC corpus and d corresponds to the number of words in our corpus. The values a and b are known as the observed values (O).

Next, the expected value for each word is calculated as follows:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

where N corresponds to the total number of words in the corpus and i to the single words. The observed values correspond to the real frequency of a single word i in the corpus. So, for each $word_i$, the observed value O_i is used to calculate the expected value. Applying this formula to our contingency table (with N_1=c and N_2=d) results in:

$$E_1 = c \cdot (a + b)/(c + d)$$

$$E_2 = d \cdot (a+b)/(c+d)$$

Finally, the resulting expected values are used for the calculation of the Log-Likelihood (LL):

$$-2ln\lambda = 2\sum_i O_i ln\left(\frac{O_i}{E_i}\right)$$

which equates to:

$$LL = 2 \cdot \left(a \cdot \log\left(\frac{a}{E_1}\right)\right) + \left(b \cdot \log\left(\frac{b}{E_2}\right)\right)$$

More information about the calculation of the expected values and Log-Likelihood can be found in Rayson and Garside (2000). Since the reference corpus models usual everyday Dutch language, the intuition here is that texts with an overall unnatural use of words will be detected by the Log-Likelihood ratio.

Finally, *Mutual Information* attempts to indicate how informative the co-occurrence of two words close to each other in a text is, apart from the information of each individual word. If two points (words), *x* and *y*, have probabilities *P(x)* and *P(y)*, then their mutual information, *I(x,y)* is defined to be:

$$I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

We follow the implementation of Church and Hanks (1990) who estimated word probabilities *P(x)* and *P(y)* by counting the number of observations of *x* and *y* in a corpus, *f(x)* and *f(y)*, and normalizing by *N*, the size of the corpus (TwNC). Joint probabilities, *P(x,y)*, are estimated by counting the number of times that *x* is followed by *y* in a window of *w* words, f_w *(x,y)*, and normalizing by N. The window size parameter allows us to look at different scales. For our system, co-occurrences of words with 1 to 4 positions from each other were taken into account. The mean Mutual Information is a measure for the mutual information on the text level.

- Finally, to account for morpho-syntactic characteristics, another thirteen features were included. These correspond to the proportion of tokens referring to the thirteen main part of speech categories in Dutch[14] (i.e. all content words, articles, conjunctions, numerals, prepositions and interjections). Should this proportion of words with a certain part of speech differ from the expected value, the structure of the text and of the sentences in the text may differ from a reader's expectation.

All above-mentioned features consist of numeric values. For the task departing from a text t as input, the regression algorithms are given the numerical feature vector corresponding to t as described above. For the tasks which depart from a pair of texts (t_1, t_2) as input, an input feature vector is derived as the difference of the vectors corresponding to t_1 and t_2 respectively.

5 Results and Discussion

In order to evaluate the performance of the machine learning algorithms on the three different readability assessment tasks listed in Table 6, we calculated two different baselines. As our first baseline, a language modeling (LM) approach was used. A generic language model was generated from the Twente Nieuws Corpus

[14] To this end, we processed all texts with Tadpole (van den Bosch *et al.*, 2007).

using the SRILM package with standard settings (Stolcke, 2002) and by taking into account all n-grams up to order 5. Following Kate *et al.* (2010), we use the score assigned to a document by a generic language model and normalize this document-level score or perplexity by the number of words (to rule-out document length). We calculate this normalized document probability $NP(D)$ as follows:

$$NP(D) = (P(D|M))^{\frac{1}{|D|}}$$

,where M is our generic language model trained on the Twente Nieuws Corpus and $|D|$ is the number of words in document D. As a second baseline, we also considered the different readability formulas described in Section 3.1. In order to avoid a rescaling of the original readability scores, we only performed this second comparison for the binary classification experiments.

K-fold cross validation was used as the testing method. For our experiments, the readability data was split into 10 folds, so k was set to 10. In this paper, it is not our intention to optimize the algorithms to achieve the best possible results for the learning tasks at hand, but rather to illustrate the usefulness of the data collected with the tools from Section 3.2 in a machine learning set-up, and to provide a comparison of our approach with two baselines. Note that exactly because of the public unavailability of readability assessment data sets, such comparative evaluations are scarce in the literature.

The results for regression are shown in Table 8. They show that our model outperforms the generic LM baseline for both the crowd and experts data sets. Since our underlying corpus incorporates texts from different genres, we hypothesize

$t_1 \dots t_n$		
	Crowd	Experts
Baseline (LM)	0.156	0.1908
Our model	0.127	0.1334

Table 8. RMSEs for the regression task predicting the $P_e(t)$ value for a given text

that features which help to identify the genre of a given text, might improve the results of both approaches. A similar conclusion was drawn by Kate *et al.* (2010) when contrasting generic with genre-specific language models. Furthermore, we hypothesize that the usefulness of our model which currently incorporates rather shallow text characteristics will further increase if a syntactic, pragmatic and semantic layer will be added to the feature set.

An overview of the classification accuracies for the two experiments involving text pairs are presented in Tables 9 (multiclass) and 10 (binary). Since there seems to be no viable way to unambiguously transform the results of a readability formula for two texts to a five-point scale classification of a given text pair, we only calculated this baseline for the binary classification experiments. Overall, we can observe in Table 9 that the top accuracies obtained on both the experts and crowd data sets are quite similar. Furthermore, we see that the Random forest classifier outperforms the $k - nn$ classifier on this task, although these results could be subject to change

(t_1, t_2)... (t_{n-1}, t_n)			
	Averaged?	Crowd	Experts
Baseline (LM)			
$k - nn$	yes	0.3700	0.3963
Random forest		0.3700	0.3963
$k - nn$	no	0.4177	0.3700
Random forest		0.4177	0.3700
Our model			
$k - nn$	yes	0.3754	0.3395
Random forest		0.4564	0.3918
$k - nn$	no	0.3677	0.2854
Random forest		0.4136	0.4186

Table 9. Classification accuracy for the experiment in which a correct assessment from the five-point scale from Table 2 had to be assigned to a given text pair

in case parameter optimization were performed for both classifiers. The best result, i.e. 45.64%, is obtained by the Random forest classifier on the averaged crowd data. The rather low scores on this task can be explained by the fact that we used a rather strict evaluation in which each text pair had to be assigned to the exactly right class. Deviations of one class (e.g. ED being classified as RSE) were penalized equally severely as deviations of more classes (e.g. ED being classified as RME).

In the binary classification experiments, we contrasted both the language modeling approach and our model with the classical readability formulas. The results as displayed in Table 10 clearly show that the classical formulas perform poorly on this task. The baseline language modeling approach and our model, which currently also rely on rather shallow text characteristics and/or statistical information derived from general domain corpora, outperform the readability formulas. We expect that this gap will become even larger when incorporating deeper syntactic and semantic knowledge. A possible conclusion might be that the classical readability formulas, which are typically designed for selecting reading material for language learners, are clearly not fit for measuring the readability level of generic text. Furthermore, although averaging does not seem to help for the LM approach, it does so for our model on both the expert and crowd data sets. Similar to the 5-class classification task, the best performance is obtained by the Random forest classifier on the averaged crowd data, i.e. 77.51%.

Overall, we can conclude from the experiments that the data sets obtained through both methodologies yield similar results. Though our model already outperfoms the LM baseline, we are strongly convinced that additional syntactic, semantic and pragmatic knowledge should be incorporated to have a more diversified view on the complexity of a given text.

(t_1, t_2)... (t_{n-1}, t_n)			
	Averaged?	Crowd	Experts
Brouwer	yes	0.2442	0.2428
Douma		0.2482	0.2436
Flesch		0.2490	0.2434
Brouwer	no	0.2557	0.3358
Douma		0.4939	0.3315
Flesch		0.2542	0.3320
Baseline (LM)			
$k - nn$	yes	0.4952	0.4687
Random forest		0.4836	0.4028
$k - nn$	no	0.4865	0.4734
Random forest		0.5029	0.4020
Our model			
$k - nn$	yes	0.7347	0.5293
Random forest		0.7751	0.6378
$k - nn$	no	0.5495	0.5089
Random forest		0.7136	0.5722

Table 10. Classification accuracy for the binary classification experiment involving text pairs

6 Conclusions and Future Work

In this paper, we explored two different methodologies to collect readability assessments for texts in a selected corpus: a lightweight crowdsourcing approach and a more classical expert labeling approach. Since we intended to collect multiple assessments per text in order to level out a given reader's background knowledge and attitude as much as possible, we hypothesized that a crowdsourcing approach could be a viable alternative to expert labels. The corpus itself incorporates different genres.

As opposed to most assessment strategies which assign an absolute score to a given text, we opted for an approach in which all texts in the corpus are compared to each other by different people and by using different comparison mechanisms, i.e. pairwise comparison and ranking. To the best of our knowledge, only Tanaka-Ishii *et al.* (2010) used a similar, yet more coarse-grained pairwise comparison strategy. A comparison of the readability assessments collected with both methodologies revealed that the data sets are very similar, a similarity which was numerically confirmed by an analysis with Pearson's correlation coefficient. This allowed us to conclude that both the users of the crowdsourcing application and the experts rank the texts from the corpus in a very similar order.

Since the corpus itself incorporates different genres, we also investigated whether the various genres actually represented various readability levels. Our results show that this was indeed the case. However, when looking at the actual assessments we did notice a consistency between both groups that readability might be linked to text genre. In future work, we will closely investigate features representing the underlying differences between texts within a genre, i.e. genre-independent features, if these exist at all.

For some readability prediction tasks, e.g. in case an old text is compared to its easier, rewritten version, a set of text comparisons is sufficient to learn the task. However, other cases require a readability score assigned to each individual text, e.g. in case one wishes to determine the readability of a given patient information leaflet, manual, insurance policy, etc. In order to account for this type of tasks, we presented a novel definition of readability as a probability which is easy to estimate from the data sets. For a text t_0 in the readability corpus, the probability that it is easier/more difficult than any other text was estimated by the proportion of times that t_0 was assessed as easier/more difficult than any other text in the corpus.

Finally, we demonstrated how the data collected via both applications can be used in various machine learning set-ups to perform regression and classification. These experiments, however, should be considered as a basic set of experiments using rather shallow features. Adding a pragmatic and a semantic layer to the feature set, as well as performing a fine-grained analysis of the predictive power of the individual features, feature selection and feature construction are interesting topics for future research. Moreover, we would like to further combine and compare our approach with more advanced statistical language models which have turned out to be good predictors in previous research (Schwarm & Ostendorf, 2005; Feng *et al.*, 2010). In future work, we will also elaborate on how the readability assessments can be processed. Is it possible, for example, to weigh annotators, based on their assessments?

Note that we used a specific readability corpus and specific readability assessment applications. By including several genres in the data, however, we aimed to make our text collection as broad and as generic as possible. As with any other classification task, which is trained on a specific data set, the porting to another domain (say for example to legal texts or children's books) evidently necessitates domain adaptation of the classifier. To tailor readability prediction to specific settings, our approach is easily extensible to specific textual domains and to specific target audiences. To assess the readability of only legal texts, for example, a readability corpus containing such texts can be composed. The proposed methodology can also be adapted to a specific purpose, e.g. to promote safety by using readable instructions in manufacturing environments. In that case, the question asked to the experts should be to what degree the instructions promote safety, and not only how readable the instructions are.

We conclude that readability assessment by comparing texts is a polyvalent methodology, which can be adapted to specific domains and target audiences if required.

Acknowledgments

This research was funded by the University College Ghent Research Fund. We would like to thank the reviewers for their invaluable comments and insights. Furthermore, we thank everyone who contributed to our research by using our web applications.

References

Alice Davinson 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.

Anderson, Richard C., & Davison, Alice. 1986 (October). *Conceptual and Empirical Bases of Readability Formulas*. Tech. rept. 392. University of Illinois at Urbana-Champaign.

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

Berger, A., Caruana, R., Cohn, D., Freitag, D., & Mittal, V. (2000). Bridging the Lexical Chasm: Statistical Approaches to Answer Finding. *Pages 192–199 of: Proc. int. conf. research and development in information retrieval*.

Brouwer, R. H. M. (1963). Onderzoek naar de leesmoeilijkheden van Nederlands proza. *Pedagogische studiën*, **40**, 454–464.

Church, K., & Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational linguistics*, **16**(1), 22–29.

Collins-Thompson, Kevin, & Callan, Jamie. 2004 (May). A language modeling approach to predicting reading difficulty. *Proceedings of hlt / naacl 2004*.

Collins-Thompson, Kevyn, & Callan, Jamie. (2005). Predicting reading difficulty with statistical language models. *Journal of the american society for information science and technology*, **56**(November), 1448–1462.

Cox, Eli P. (1980). The optimal number of response alternatives for a scale: A review. *Journal of marketing research*, **17**(4), 407 – 422.

Dale, Edgar, & Chall, Jeanne S. (1948). A formula for predicting readability. *Educational research bulletin*, **27**, 11–20.

Douma, W.H. (1960). De leesbaarheid van landbouwbladen: een onderzoek naar en een toepassing van leesbaarheidsformules. *Bulletin*, **17**.

DuBay, W. H. (ed). (2007). *Unlocking Language: The Classic Readability Studies*. BookSurge.

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

Feng, L., Elhadad, N., & Huenerfauth, M. (2009). Cognitively Motivated Featurs for Readability Assessment. *Pages 229–237 of: Proceedings of the 12th conference of the european chapter of the acl*.

Feng, L., Jansche, M., Huenerfauth, M., & Elhadad, N. (2010). A Comparison of Features for Automatic Readability Assessment. *Pages 276–284 of: Proceedings of coling 2010 poster volume*.

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., & Dredze, M. (2010). Annotating Named Entities in Twitter Data with Crowdsourcing. *Page 8088 of: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazons Mechanical Turk*.

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. *Proceedings of the eacl 2009 student research workshop*.

Graesser, Arthur C., McNamara, Danielle S., Louwerse, Max M., & Cai, Zhiqiang. (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*. New York: McGraw-Hill.

Heilman, Michael, Collins-Thompson, Kevyn, & Eskenazi, Maxine. (2008). An Analysis of Statistical Models and Features for Reading Difficulty Prediction. *The third workshop on innovative use of nlp for building educational applications*.

Hoste, Véronique, Vanopstal, Klaar, Lefever, Els, & Delaere, Isabelle. (2010).

Classification-based scientific term detection in patient information. *Terminology*, **16**, 1–29.

Kanungo, Tapas, & Orr, David. (2009). Predicting the readability of short web summaries. *Pages 202–211 of: Proceedings of the second acm international conference on web search and data mining*. WSDM '09. New York, NY, USA: ACM.

Kate, Rohit J., Luo, Xiaoqiang, Patwardhan, Siddharth, Franz, Martin, Florian, Radu, Mooney, Raymond J., Roukos, Salim, & Welty, Chris. (2010). Learning to Predict Readability using Diverse Linguistic Features. *23rd international conference on computational linguistics*.

Kincaid, J. Peter, Jr., Robert P. Fishburne, Rogers, Richard L., & Chissom., Brad S. (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.

Kraf, Rogier, & Pander Maat, Henk. (2009). Leesbaarheidsonderzoek: oude problemen, nieuwe kansen. *Tijdschrift voor taalbeheersing*, **31**(2), 97–123.

Leroy, G., & Endicott, J.E. (2011). Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries. *International Conference on Asia-Pacific Digital Libraries (ICADL 2011)*.

Leroy, G., Helmreich, S., & Cowie, J. (2010). The influence of text characteristics on perceived and actual difficulty of health information. *International journal of medical informatics*, **79**(6), 438–449.

McNamara, Danielle S., Kintsch, Eileen, Songer, Nancy Butler, & Kintsch, Walter. (1993). *Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text*. Tech. rept. Institute of Cognitive Science, University of Colorado.

Petersen, S., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer speech & language*, **23**(1), 89–106.

Pitler, Emily, & Nenkova, Ani. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. *Pages 186–195 of: Emnlp.* ACL.

Poesio, M., Kruschwitz, U., & Chamberlain, J. (2008). ANAWIKI: Creating Anaphorically Annotated Resources through Web Cooperation. *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08).* European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Rankin, Earl F. (1959). The cloze procedure: its validity and utility. *Eighth yearbook of the national reading conference*, **8**, 131–144.

Rayson, Paul, & Garside, Roger. (2000). Comparing corpora using frequency profiling. *Pages 1–6 of: Proceedings of the workshop on Comparing corpora, 38th annual meeting of the Association for Computational Linguistics.*

Salton, G. (1989). *Automatic text processing: the transformation, analysis and retrieval of information by computer.* Addison Wesley.

Schuurman, Ineke, Hoste, Véronique, & Monachesi, Paola. (2010). Interacting semantic layers of annotation in SoNaR, a reference corpus of contemporary written Dutch. Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios, & Tapias, Daniel (eds), *Proceedings of the seventh international conference on language resources and evaluation (lrec'10).* Valletta, Malta: European Language Resources Association (ELRA).

Schwarm, Sarah E., & Ostendorf, Mari. 2005 (June). Reading Level Assessment Using Support Vector Machines and Statistical Language Models. *Pages 523–530 of: Proceedings of the 43rd annual meeting of the acl.* Association of Computational Linguistics, Ann Arbor.

Si, Luo, & Callan, Jamie. (2001). A Statistical Model for Scientific Readability. *Pages*

574–576 of: Proceedings of the tenth international conference on information knowledge management.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel, & Ng, Andrew Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. *Pages 254–263 of: Proceedings of the conference on empirical methods in natural language processing.* EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics.

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

Staphorsius, Gerrit, & Krom, Ronald S.H. (1985). *Cito leesbaarheidsindex voor het basisonderwijs: verslag van een leesbaarheidsonderzoek.* Specialistisch bulletin, no. 36. Cito Arnhem.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. *Proc. intl. conf. on spoken language processing.*

Tanaka-Ishii, Kumiko, Tezuka, Satoshi, & Terada, Hiroshi. (2010). Sorting Texts by Readability. *Computational linguistics*, **36**(2), 203–227.

van den Bosch, Antal, Busser, Bertjan, Daelemans, Walter, & Canisius, Sander. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. *Pages 191–206 of: Computational Linguistics in the Netherlands 2006.* Selected Papers of the 17th Computational Linguistics in the Netherlands Meeting.

van Noord, Gertjan J.M. 2009 (January). *Large Scale Syntactic Annotation of written Dutch (LASSY).*

van Oosten, Philip, Tanghe, Dries, & Hoste, Véronique. (2010). Towards an Improved Methodology for Automated Readability Prediction. Calzolari, Nicoletta, Choukri, Khalid, Maegaard, Bente, Mariani, Joseph, Odijk, Jan, Piperidis, Stelios, & Tapias, Daniel (eds), *Proceedings of the seventh international conference on language resources* *and evaluation (lrec'10)*. Valletta, Malta: European Language Resources Association (EL.

vor der Brück, Tim, Hartrumpf, Sven, & Helbig, Hermann. (2008). A Readability Checker with Supervised Learning Using Deep Indicators. *Informatica*, **4**, 429–435.

Zeng, Q., Goryachev, S., Tse, T., Keselman, A., & Boxwala, A. (2008). Estimating Consumer Familiarity with Health Terminology: A Context-based Approach. *Jamia journal of the american medical informatics association*, **15**(3), 349–356.