Do speech registers differ in the predictability of words?

Previous research has demonstrated that language use can vary depending on the context of situation. The present paper extends this finding by comparing word predictability differences between 14 speech registers ranging from highly informal conversations to read-aloud books. We trained 14 statistical language models to compute register-specific word predictability and trained a register classifier on the perplexity score vector of the language models. The classifier distinguishes perfectly between samples from all speech registers and this result generalizes to unseen materials. We show that differences in vocabulary and sentence length cannot explain the speech register classifier's performance. The combined results show that speech registers differ in word predictability.

Keywords: speech registers, word predictability, text classification, statistical language modeling, register analysis

1. Introduction

People communicate in different situations and modalities, ranging from casual conversations between friends to formal lectures or public addresses. Many previous studies have shown that these different situations elicit different language use; see Biber and Conrad (2009) for an overview. The term 'register' is used to provide a link between a communicative act and the context of the situation it occurs in (Marco, 2000). Likewise, we will use the term register to refer to language variation in relation to the situation of use (see Lee, 2001 for a discussion). In this paper, we investigate register-specific differences in word predictability, defined as the conditional probability of a word given the preceding words. We conducted five experiments to test whether speech registers differ in word predictability.

To investigate register differences in word predictability, we use statistical language modelling, a technique widely used within the discipline of *natural language processing*. We compute register-specific word predictability scores with the aid of statistical language models (SLM) and use these scores to train a speech register classifier. The performance of the classifier shows, to what extent speech registers differ in word predictability.

In section 2, we will explain how this natural language processing approach complements register analysis. In section 3, we introduce the corpora we use for this study and outline our analysis approach. In the following sections, we describe the experiments we conducted. In Study 1, we investigate how to create SLMs that allow cross register comparison. In Study 2, we train and test register-specific SLMs to estimate register-specific word predictability. These word predictability scores are then used to train a speech register classifier. In Study 3, we validate the results from Study 2, by testing the speech register classifier on the validation corpus. In Study 4, we investigate the amount of data necessary for classification. Finally, in Study 5, we investigate the influence of average sentence length on word predictability. We end with a general discussion of our findings.

2 Characterizing text in register analysis and natural language processing

The most fundamental approach in register analysis is to count lexico-grammatical features (e.g. demonstrative pronouns), and compare their prevalence across registers. The studied materials may be written, or consist of orthographic transcription of speech samples. For example, Tottie (1991) investigated differences between spoken and written British English and found that negatives are twice as prevalent in spoken language as in written text. Van Gijsel et al. (2006) compared excerpts from different speech registers in Dutch and showed that word type-token ratio (TTR) is lower for informal dialogues than for formal monologues.

Biber (1988, 1995) developed an approach for register analysis known as multidimensional analysis, which aims at identifying co-occurring linguistic features and discovering underlying dimensions of language use by means of factor analysis. For example, Biber (1988) found that discourse particles, first and second person pronouns, and present tense verbs are typical of *involved* language. Conversely, a high frequency of nouns and prepositions and a high word type-token ratio are typical of *informational* language. The dimensions can be used to group or distinguish between different registers and give a functional interpretation to the patterns of lexico-grammatical features (Biber & Conrad 2009).

In contrast to the interpretative approach of register analysis, text classification methods as developed within the discipline of *natural language processing* characterize texts by a large set of (mostly) automatically generated features (see Killgariff 2001, for an overview). The feature set typically consists of sequences of one or more consecutive elements (e.g. words), called *n*-grams (e.g. unigrams, bigrams, trigrams). For example, *the big house* is a trigram, containing the bigrams *the big,* and *big house,* and the unigrams *the*, *big,* and *house*. Based on *n*-grams, a statistical language model (SLM) can be created that estimates the probability of a word given the preceding words. SLMs are a staple technology for applications such as machine translation, automatic speech recognition, and document retrieval (Jurafsky & Martin 2009).

Both approaches have advantages and disadvantages. For example, because register analysis uses a (small) set of handcrafted lexico-grammatical features to describe and interpret differences between registers (Biber and Conrad, 2009), it precludes data-driven research. That is, registers can only be characterized and compared with the features that are defined beforehand, based on previous research or on the researcher's intuitions. Statistical language modelling avoids this and opens up the possibility of a data-driven search of patterns in a corpus.

From the perspective of register analysis, there is a disadvantage to SLMs; because typically many features are used, the interpretation of patterns of textual differences is difficult. Nevertheless, the distribution of *n*-grams can provide valuable insight, hard to achieve with intuition alone. For example, Gries (2001) successfully used the statistics of word co-occurrences to disambiguate the meanings of near synonyms. Denoual (2006) used character n-grams (i.e. based on graphemes instead of words) to classify texts on a dimension ranging from literary to oral. Similarly, we propose that investigating the distribution of word *n*-grams across speech registers may reveal register differences not accessible with current register analysis tools.

In the current study, we investigated differences among registers using text classification techniques in the form of SLMs based on word *n*-grams. We used word *n*-grams because they are theory neutral; only minimal assumptions have to be made to count and compare *n*-grams of words (see also Gries & Ellis, 2015: 231). Moreover, previous research showed that listeners are sensitive to the statistics of word *n*-grams. We will discuss the relevant literature and define word predictability in the following section.

2.1 Word predictability

We define word predictability as the probability of a word given the previous context (i.e., the preceding words). For example, the predictability of the word *gun* given the context *The policeman pulled out his …* is high compared to a word like *socks*. Word predictability is thus the conditional probability P(*word*|*context*) of a word *word* given the preceding context *context*, which can be estimated with an SLM (e.g., Smith & Levy 2013).

Word predictability plays an important role in language comprehension (e.g. Kutas et al. 2010). Converging evidence from studies using different methodologies such as self-paced reading (e.g., Monsalve et al. 2012, Smith & Levy 2013), eye-tracking (e.g. Frisson et al. 2005), EEG (e.g. Van Berkum et al. 2005), and fMRI (e.g., Willems et al. 2016) show that that the processing of speech and text is influenced by the predictability of a word given the previous context. For an overview of frequency effects in language processing, see Ellis (2002).

Word predictability also plays a role in language production. For example, Bell et al. (1999) found that the pronunciation of English function words depends on word predictability, whereby less predictable words are pronounced in fuller form. Similarly, Pluymaekers et al. (2006) found that the duration and number of segments of Dutch suffixes are influenced by the predictability of the carrier word.

The widespread and converging evidence for the importance of word predictability in language comprehension and production led us to investigate to what extent word predictability differs across registers. One reason to suspect differences is the aforementioned finding that lexical richness differs across speech registers (e.g. Van Gijsel et al. 2006); more formal registers have higher word type-token ratios than more informal registers. If one register contains more word types compared to other registers, it is likely that this influences word predictability.

3. Methodology

3.1 Corpus

We used a subset of the Spoken Dutch Corpus (Oostdijk 2001). This corpus is ideally suited to investigate speech register differences, because it consists of *components* reflecting speech in different situations of use, ranging from spontaneous conversations to television news broadcasts and read aloud stories.

We used the orthographically transcribed recordings of adult native speakers in the Netherlands. We excluded the Flanders part (approximately one-third) of the corpus, because possible differences between Northern Dutch and Flemish Dutch speech styles are outside the scope of our study. In addition, we excluded one component (*masses and solemn speeches*) because it is comparatively small (fewer than 6,000 word tokens). This left 14 components for analysis (see Table 1). This subset consists of approximately five million word tokens of Netherlandic Dutch[2] speech.

Table 1. Overview of the 14 components in the Spoken Dutch Corpus used for Study 1 - 4.

ID	**Component description**
a	Spontaneous conversations (face-to-face)
b	Interviews with teachers of Dutch
c	Spontaneous telephone dialogues via a platform
d	Spontaneous telephone dialogues via a minidisc recorder
e	Business negotiations
f	Radio and television interviews and discussions
g	Debates, discussion and meetings (especially political)
h	Classes
i	Spontaneous radio and television commentaries (e.g. sports)
j	Radio and television newsroom and documentaries
k	News broadcast on radio and television
l	Reflections and commentaries broadcast on radio and television
n	Lectures and speeches
o	Read aloud speech

We also created a validation corpus to validate our findings and ensure they generalize beyond the materials in the Spoken Dutch Corpus. It consists of materials from three different corpora: two corpora of Dutch spontaneous speech, the *Institute of Phonetic Sciences Amsterdam Dialogue Video Corpus*, henceforth IFADV (Van Son et al. 2008), and the *Ernestus Corpus of Spontaneous Dutch*, henceforth ECSD (Ernestus 2000), and two components of the *STEVIN Dutch Reference Corpus*, henceforth SoNaR, (Oostdijk et al. 2013), namely a subset of Dutch teleprompt texts (news broadcasts) and Dutch books. We will refer to the combination of these new materials as the validation corpus, which consists of approximately 2.2 million word tokens.

The materials in the validation corpus were chosen because they correspond to three specific components in the Spoken Dutch Corpus. The two corpora of spontaneous speech (IFADV and ECSD) correspond to component *a* (spontaneous conversations), the set of Dutch teleprompt texts correspond to component *k* (news broadcasts for radio and television) and, finally, the Dutch books correspond to component *o* (read-aloud stories).

The SoNaR texts are not an orthographic transcription of speech, while this is the case for all other corpora that were used in this study. They are nevertheless similar to the respective components *k* and *o* in the Spoken Dutch Corpus, because news broadcasts (component *k*) are typically read from teleprompts and should conform to the teleprompt texts closely, and read-aloud stories (component *o*) are a collection of read-aloud audiobooks. Still differences could occur between the SoNaR materials and the orthographically transcribed texts, for instance, in the placement of sentence boundaries.

3.2 Analysis

We used statistical language models (SLM) to investigate whether speech registers influences word predictability. The reasoning is as follows. SLMs are sensitive to the difference between the language materials they are trained on and the materials they are tested on. The performance of a language model in terms of predicting the next word correctly on the basis of a sequence of previous words is known to suffer in general if the difference between the training and test set increases. We assert that this also is likely to apply to differences in speech register. For example, if an SLM is trained on spontaneous conversations and subsequently tested on read-aloud stories, the model’s predictive performance (i.e. its ability to assign the correct probability to the next word given the preceding context) is likely to be worse than in a test on an unseen set of spontaneous conversations. SLM performance can thus be utilized to assess the similarity of different registers to the register the model was trained on. We use this language model characteristic to determine word predictability differences between speech registers.

To test whether speech registers systematically differ in word predictability, we train a classifier[3] on the SLM performance measures. If word predictability differs between speech registers, the classifier should be able to differentiate these registers and achieve good register classification results. In addition, we investigate the amount of data necessary to achieve accurate classification of speech registers. Also we aim to rule out that our classifier results are driven by sentence length differences between speech registers. This is important because sentence length could influence the SLM results, as SLMs tend to assign higher likelihood scores to shorter sentences. Furthermore, Wiggers and Rothkrantz (2007) found that registers can differ in sentence length.

Because we aim to compare SLM word predictability scores between registers, the SLM vocabulary[1] deserves special consideration. An SLM's vocabulary is typically based on the texts it is trained on, referred to as a training set. The out-of-vocabulary words (i.e. words not part of the language model, also referred to as OOV words) are typically ignored in performance evaluation. However, we train SLMs on different registers and want to compare between them. If the number of OOV words differs between SLMs trained on different speech registers, this can influence test results of the SLM; for instance, if a register contains many OOV words, the SLM could attain an artificially boosted performance. Therefore, for a fair comparison between all register-specific SLMs, they should have the same register-insensitive vocabulary.

For the creation of the fixed SLM vocabulary, we need a corpus containing multiple registers and an approach for vocabulary word selection. Two extreme approaches are possible: Greedy selection, that is, selection of all or nearly all words occurring in the corpus; or robust selection, that is, selection of only those words that are most likely present if the corpus would be created again, regardless of register. For example, consider the word *gamble,* which can be used in many different registers, while the word *inning* typically occurs in sports commentaries. In this example, the word *gamble* is a good candidate for a robust vocabulary, while *inning* may not be.

The advantages of greedy selection are the maximum use of available data and a straightforward inclusion criterion, which typically consists of the selection of all words occurring above a certain frequency threshold (e.g. word frequency of 5) in the corpus. The disadvantage of greedy selection relates to the unreliability of the decision to include a word. For example, the burstiness of words, the phenomenon that a word's likelihood increases if it has been used recently (Church & Gale 1995), lead to an uneven distribution of tokens throughout a corpus. These findings make word frequency an unreliable measure to base word selection criteria on (Kilgariff 2001, Gries & Ellis 2015).

Robust selection addresses the word burstiness problem. Savický and Hlaváčová (2002) developed a metric called average reduced frequency (ARF), which adjusts word frequency based on the word’s dispersion in a corpus, whereby a word with low dispersion (i.e., with a bursty distribution) results in a lower ARF as compared to a word that is more evenly distributed (cf. section 4.1). If a word is used regularly throughout the corpus, it is more likely it will be found again in a newly sampled corpus, whereas a word that only occurs in local bursts may be an idiosyncratic (e.g. topical) characteristic of a specific corpus. Therefore, a vocabulary based on the highest scoring ARF words could improve word selection quality.

A potential disadvantage of robust selection is the reduction of the available data, because the resulting vocabulary will be significantly smaller than the vocabulary resulting from greedy selection. In addition, the word exclusion criteria are more complex and the quality of the vocabulary depends on the viability of these criteria. In sum, both approaches have their advantages and disadvantages, and it is unclear whether greedy or robust selection is the best way to create an SLM vocabulary for our purposes. Therefore, we test both approaches.

4. Study 1: SLM vocabulary selection

In study 1 we tested whether robust or greedy selection is better suited for the creation of a SLM vocabulary.

4.1 Procedure

We extracted the orthographic transcriptions from the Spoken Dutch Corpus and removed the special corpus-specific word codes (explicitly marking foreign words, dialectal words, regionally accented words, new words, interjections, onomatopoeia, hesitations and mispronunciations, see Goedertier et al. 2000). Further text normalization was not necessary because the orthographic transcriptions were already tokenized and normalized according to the protocol described in Goedertier et al. (2000).

We defined word type as the word surface form (i.e. *run* and *runs* are two different word types) and created the greedy vocabulary by selecting the 50,000 most frequent word types from the corpus. We created the robust vocabulary by ranking words types based on their average reduced frequencies (see below) and selected all word types with an average reduced frequency (ARF) of at least 50. This lower bound of the ARF was based on the trade-off between coverage and the constraint that word types should be present in most components of the corpus. This resulted in a list of 585 words types, covering 77.5% of all word tokens in the corpus.

To compute the ARF of each word in the corpus, we extracted the first 61,834 word tokens (i.e. the number of tokens in the smallest component) from each component, which ensures that the ARF scores are not influenced by the amount of materials of each component in the corpus. We then calculated the reduced frequency (RF) of each word (Savický & Hlaváčová 2002). The RF (Equation 1) equals a word's frequency if the word is evenly distributed throughout the corpus, and has a lower bound of one if the word is clustered in one location in the corpus (Hlaváčová & Rychly 1999). That is, words with the highest ARF are those words that occur evenly throughout the corpus and are therefore neither topic-specific nor register-specific.

To compute the RF for each word *w*, the corpus is divided into a number of intervals ($N_{intervals}$) equal to the frequency of word *w*. The RF is then computed as the number of intervals word *w* occurs in. Therefore, it is important to keep the original word order of texts and to group register-specific texts together.

$$\mathrm{RF} = \sum_{i=1}^{N_{intervals}} f_w(i)$$

$$\begin{cases} f_w(\mathrm{i}) = 1, & \mathit{if\ the\ word\ w\ occurs\ in\ the\ } i^{th} \mathit{\ interval} \\ f_w(\mathrm{i}) = 0, & \mathit{if\ the\ word\ w\ does\ not\ occur\ in\ the\ } i^{th} \mathit{\ interval} \end{cases} \tag{1}$$

The RF depends on the start and end points of the intervals and the start point of the first interval determines the start and end points of all other intervals. There are many possible starting points for the first interval. To avoid this arbitrariness, the RF is calculated for all non-redundant starting points in the corpus, that is, for the first word of the corpus, up to and including the word with number $v = \lfloor N_{words}/N_{intervals} \rfloor$, where v denotes the number of starting points, N_{words} the number of word tokens in the corpus and $N_{intervals}$ denotes the number intervals the corpus is divided into. We computed for each word the average reduced frequency by averaging over all RFs.

To compare the greedy and robust vocabularies, we created two versions of our corpus. All OOV words were mapped to the dummy string *unk.* In one version, we used the greedy vocabulary to determine the OOV words and in the other version we used the robust vocabulary.

We used frequency profiling, described in Rayson and Garside (2000), to discover those *n*-grams (restricted to unigrams, bigrams or trigrams) in each component that distinguish a given component from the other components, for both the greedy and robust corpus versions. Frequency profiling compares the frequency of a *n*-gram in different corpora by computing the log-likelihood (Equation 3) of the *n*-grams frequency in one corpus compared to the frequency in one or more other corpora. For the computation of the log-likelihood we used regular frequency (not the ARF) of a *n*-gram.

$$\mathrm{LL}_{n\text{-gram}} = 2\left(\sum_i O_i \ ln\left(\frac{O_i}{E_i}\right)\right) \qquad (3)$$

In Equation 3 O_i denotes the *n*-grams frequency in the *i*-th corpus. E_i denotes the expected value of the *n*-grams frequency in the i-th corpus and is computed according to Equation 4,

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \qquad (4)$$

where N_i refers to the total number of *n*-gram tokens in the *i*-th corpus.

To compute the log-likelihood statistic we used the Colibri-Core toolkit (Van Gompel & Van den Bosch 2016), which includes an implementation of frequency profiling. To investigate *n*-grams that are specific for a component compared to the rest of the corpus, we used the leave-one-out approach; we compared all *n*-grams in each component against the combination of the 13 other components. The log-likelihood statistic was calculated for all word unigrams, bigrams and trigrams, for both the robust and greedy corpus.

4.2 Results and discussion

The Colibri-Core toolkit returns *n*-gram lists ranked on log-likelihood, whereby the *n*-grams that distinguish a component most compared to the others are ranked at the top. We checked whether the greedy vocabulary resulted in a more uneven distribution of word forms across components compared to a robust vocabulary. We found that this was indeed the case. To illustrate this, we list the highest ranking unigrams for a sample of four components, in Table 2 for the greedy vocabulary version, and in Table 3 for the robust vocabulary version.

Table 2 Overview of most distinguishing word forms of four speech registers based on frequency profiling, Greedy vocabulary (50,000 most frequent word forms).

Casual conversation	**Interview Dutch teachers**	**Political debates**	**Sports commentary**
ja ‘yes’	*uh* ‘ehm’	*het* ‘it’	*bal* ‘ball’
nee ‘no’	*leerlingen* ‘pupils’	*de* ‘the’	*Kluivert* *
oh ‘oh’	*Nederlands* ‘Dutch’	*voorzitter* ‘chairman’	*Bergkamp* *
'k ‘I’	*lezen* ‘read’	*motie* ‘motion’	*Zenden* *
zo ‘later’	*onderwijs* ‘education	*u* ‘you’	*de the*
echt ‘really’	*literatuur* ‘literature’	*heer* ‘gentlemen’	*Davids* *
mmm ‘ehm’	*klas* ‘class’	*van* ‘of’	*balbezit* ‘ball possession’
wel ‘well’	*school* ‘school’	*mevrouw* ‘lady’	*Overmars* *
gewoon ‘just’	*vak* ‘course’	*minister* ‘minister’	*Boer* *
maar ‘but’	*ik* ‘I’	*vraag* ‘question’	*Cocu* *

** name of Dutch soccer player*

Table 3. Overview of most distinguishing words of four speech registers based on frequency profiling and a Robust vocabulary, (585 top ranking ARF words).

Casual conversation	**Interview Dutch teachers**	**Political debates**	**Sports commentary**
ja ‘yes’	*uh* ‘ehm’	*het* ‘it’	*<unk>* *
nee ‘no’	*lezen* ‘read’	*voorzitter* ‘chairman’	*de* ‘the’

oh ‘oh’	*school* ‘school’	*de* ‘the’	*speelt* ‘plays’
'k ‘I’	*ik* ‘I’	*u* ‘you’	*nul* ‘zero’
zo ‘later’	*vind* ‘think’	*heer* ‘gentlemen’	*nu* ‘now’
wel ‘well’	*mmm* ‘ehm’	*van* ‘of’	*helft* ‘half’
mmm ‘ehm’	*heel* ‘very’	*minister* ‘minister’	*voor* ‘before’
echt ‘really’	*dus* ‘so’	*<unk>* *	*meter* ‘meter’
maar ‘but’	*ben* ‘am’	*vraag* ‘question	*tweede* ‘second’
gewoon ‘just’	*kinderen* ‘childeren’	*om* ‘to’	*gaat* ‘go’

* <unk> is the dummy string that out-of-vocabulary words are mapped to.

We observe that the greedy selection approach produces a topicality confound (i.e. differences in *n*-gram frequency between components due to the topics discussed in the components). For example, the component containing interviews with teachers of Dutch contains many words specifically related to education (e.g. the Dutch equivalents of *pupils*, *school*, *class*), while the sports commentary component contains many proper names of Dutch soccer players (e.g. *Kluivert*, *Zenden*). A similar pattern is present in the higher order *n*-grams (i.e. bigrams and trigrams). Consequently, if we create SLMs based on a greedy vocabulary, it will not be possible to ascertain whether components are distinguished based on register or topic. The robust strategy, as illustrated in Table 3, attenuates the topicality confound. For example, the most distinguishing words for the sports commentary do not include proper names, and we see only few terms specifically related to education for the component containing interviews with teachers of Dutch. Note that, by using ARF to select words, we do not restrict the vocabulary to function words. As can be observed in Table 3; content words are also present in the robust vocabulary.

In sum, Study 1 showed that a greedy vocabulary introduces a topicality confound. Such a vocabulary contains many words that are specific for topics that happened to be discussed in one or several components of the Spoken Dutch Corpus. As a consequence, when we train the speech register classifier based on the SLM results obtained with the greedy vocabulary, we do not know whether speech registers are distinguished based on genuine register-specific word predictability or the coincidental distribution of topic-specific words. The robust vocabulary remedies this confound by excluding words that are not evenly distributed across the corpus.

5. Study 2: Training and testing of the speech register classifier

We created SLMs with the robust vocabulary to estimate register-specific word predictability. We tested to what extent speech registers can be distinguished based on word predictability differences.

5.1 Procedure

We used the same subset of the Spoken Dutch Corpus as described in Experiment 1 to train SLMs and create the speech register classifier. The Spoken Dutch Corpus was pre-processed as described in Study 1. We trained register-specific tri-gram models with the SRILM-toolkit[4] (Stolcke 2002), using the robust vocabulary created in Study 1. For smoothing, we used Witten-Bell discounting with interpolation (Witten & Bell 1991). We could not use the standard smoothing technique, that is, modified Kneser-Ney discounting (Chen & Goodman 1998), because of our small vocabulary of relatively frequent words. Kneser-Ney discounting needs counts of infrequent *n*-grams to asses the probability mass needed for unseen *n*-grams. Witten-Bell is able to deal with truncated count-of-count lists[5] because it uses the first occurrence of *n*-grams to asses the probability mass needed for unseen *n*-grams.

To create register-specific SLMs, we first mapped all OOV word tokens to the dummy string *unk*. The mapping was used to maintain the serial structure of the sentences. Next, we created training and test sets for each component in the Spoken Dutch Corpus by grouping all sentences of a given component into a single text file. Subsequently, the sentences of a given component were randomly assigned to one of ten equally-sized partitions to ensure a fair sampling of the register in all of the partitions.

For each component we ran a 10-fold cross-validation experiment on the partitions, using nine parts for training and one part for testing in a rotating fashion (see also Figure 1). The 10-fold cross-validation experiments yield perplexity scores for each of the 10 folds. Perplexity is a measure of how well a register-specific SLM predicts words (based on the preceding words) in new, unseen texts. Importantly for our study, registers similar to the SLM will generate lower perplexity scores than less similar registers.

The perplexity scores were computed with Equation 5, where *word* stands for a specific word token in the test file and *context* stands for the preceding words (maximally a bigram). N_{words} and $N_{sentences}$ represent the number of word tokens and sentences in the test set, respectively, and N_{OOV} represents the number of out-of-vocabulary words, which always equal 0 in our test sets, because all OOV words were mapped to the *unk* token.

$$perplexity = \frac{\sum_{all\ words} 10_{log}P(word|context)}{(N_{words} - N_{OOV} + N_{sentences})} \tag{5}$$

[Insert Figure 1 around here]

For each test file (ten from each of the 14 components), we created a 14-dimensional vector of perplexity scores (i.e. a list of 14 perplexity scores, one for each SLM) by applying all 14 trained language models to that test file. The resulting perplexity vector describes how well the test file is predicted by the 14 register-specific language models. The perplexity vectors for the 140 test files form a 140-by-14 similarity matrix, whereby each row describes the location of a test file in a 14-dimensional space, while the columns correspond to the register-specific SLMs in the Spoken Dutch Corpus. The perplexity similarity matrix shown in Figure 1 (step 3) is a subset of the complete similarity matrix we created based on the 14 components in the Spoken Dutch Corpus.

We used Linear Discriminant Analysis (LDA) to create a speech register classifier based on the similarity matrix. LDA finds a linear combination of features that maximizes class separation (see Equation 6).

$$\hat{x} = \underset{x}{argmax} \; \frac{x^t \sum_b x}{x^t \sum_w x} \tag{6}$$

The between-class and within-class scatter matrices are represented by $\sum_b$ and $\sum_w$ respectively. A vector of weights $\hat{x}$ is found that maximizes the coefficients of the between-class and within-class scatter matrices, which results in an optimal class separation when two assumptions hold about the data: homoscedasticity (identical within-class scatter matrices) and within-class multivariate Gaussian distributions.

Because our data do not conform to these assumptions, we validated our classifier, as will be discussed in Study 3.

5.2 Results and discussion

The speech register classifier was able to distinguish perfectly (accuracy 100%, on the held-out test sets) between all registers within the Spoken Dutch Corpus material. Compared to chance performance (accuracy 7.14%), the speech classifier performed considerably better. Performance metrics in terms of precision, recall and f1 can be found in Appendix 1.

6 Study 3. Validation of the speech register classifier

We showed that a classifier based on register-specific word predictability can distinguish between speech registers. However, the LDA assumptions do not hold for our dataset. To test whether our results are robust nevertheless, we performed two validation tests. First, we compared the results of the speech register classifier with a classifier trained on a random version of the corpus, and second, we tested the speech register classifier on materials from different corpora, to test whether the performance of the classifier generalizes to new data.

6.1 Procedure

We constructed 1,000 pseudorandom corpora with materials from the Spoken Dutch Corpus to validate the speech register classifier. For each pseudorandom corpus the sentences from the Spoken Dutch Corpus were randomly assigned to one of 14 components. The random components were made to contain as many word tokens as the original components in the corpus. We trained component-specific SLMs and tested these on held out test sets with 10-fold cross validation as in Study 2). Subsequently, we trained an LDA classifier for each pseudorandom corpus, also as we did in Study 2. If the speech register classifier based on the real corpus outperforms the classifiers based on the pseudorandom corpora, then the classification accuracy of the register classifier must be due to the grouping of sentence according to speech register.

The four components in the validation corpus were pre-processed individually. Since IFADV was annotated with the same protocol as used for the Spoken Dutch Corpus (Van Son et al. 2008, Goedertier et al. 2000), we used the same pre-processing steps as in Study 1. The ECSD used a slightly different annotation style with more elaborate punctuation. To approximate the annotation and tokenization of the Spoken Dutch Corpus, we created sentences by splitting the text materials on question marks, exclamation marks, commas and points. We replaced the capital letter at the start of each sentence with the lowercase equivalent, even if it was part of a proper name, since proper names were not included in the SLM vocabulary.

All sentences in the teleprompt texts and Dutch books from the SoNaR corpus already start with lower-case characters. We split on questions marks, exclamation marks, colons, commas and points and removed all remaining punctuation. For the set of teleprompt texts, we also removed special recording instructions (e.g. the Dutch equivalent of “start audio”).

The four components in the validation corpus were each split into ten equally sized partitions, equal to the 10-fold cross-validation structure we created for the Spoken Dutch Corpus. On each partition we applied the corresponding SLMs trained on the Spoken Dutch corpus. The resulting perplexity vectors were used as classifier test sets for the register classifier trained on the materials from the Spoken Dutch Corpus. Importantly, the validation materials did not influence the SLMs (which were exclusively trained on the Spoken Dutch Corpus) and did not influence the register classifier (which were trained only on the perplexity feature vectors from the Spoken Dutch Corpus). This validation therefore provides a strong test of whether our approach generalizes to new unseen data.

6.2 Results and discussion

The classifiers based on the pseudorandom corpora performed poorly, with a mean accuracy of 12% and a standard deviation of 5%. The performance is close to chance level performance (accuracy 7%). The result shows that a classifier based on perplexity scores cannot distinguish between random collections of sentences. The high performance of the classifier developed in Study 2 therefore indicates that the components of the Spoken Dutch Corpus are more homogeneous than those in the pseudorandom Corpora and that they differ in word predictability.

The speech register classifier developed in Study 2 yields an accuracy score of 93% on the validation corpus, compared to 100% accuracy on the held out classifier test sets of the Spoken Dutch Corpus. The classifier thus attained a high accuracy on materials from new corpora, which shows that the speech register classifier is not overfitted to idiosyncratic aspects of the Spoken Dutch Corpus. The accuracy score on the validation corpus was not perfect, however. The confusion matrix in Table 4 shows that all classification errors are made on the ECSD corpus of spontaneous speech. Interestingly, the ECSD is confusable with component *b*, interviews with Teachers of Dutch. There is considerable overlap between ECSD and component *b*, as both are unscripted dialogues, which suggests that the classification mistakes are not random.

Table 4. Confusion matrix of the speech register classifier test on the validation corpus.

corpora	**a**	**b**	**k**	**o**
SoNaR-books	0	0	0	10
ECSD	7	3	0	0
IFADV	10	0	0	0
SoNaR-teleprompt	0	0	10	0

In conclusion, a speech register classifier based on word predictability can distinguish between genuine speech registers, but not between randomly sampled sets of sentences. In addition, we showed that the register classifier cannot only classify materials from the training corpus (the Spoken Dutch Corpus), but also materials from the validation corpus. The combined results suggest that word predictability differs across speech registers.

7. Study 4. Investigation of the amount of text material necessary for speech register classification

We investigated the amount of text materials needed for creating a reliable register classifier. We divided registers into sentence sets containing a specific number of sentences. Subsequently, we trained and tested individual classifiers on all sets with a given number of sentences. We tested how many sentences are needed for accurate classification by comparing the performance of the classifiers.

Compared to the 10 partitions used in Study 2, dividing the text materials in small sentence sets increases the number of items to train and test a classifier on. The classifiers trained on small sentence sets are expected to confuse registers. They should especially confuse more similar registers as was seen with the ECSD in the validation corpus, which would provide further evidence that classification is based on register characteristics.

7.1 Procedure

We used materials from the Spoken Dutch Corpus and the validation corpus as described in section 3. We used a similar procedure as described in Study 2 except that we created perplexity vectors based on sets containing the following number of sentences from a specific speech register: 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024. We did this by dividing the text materials of each register from the Spoken Dutch Corpus into sets of a specific number of sentences. We computed the perplexity vectors for all sentence sets according to Equation 5 with the SLMs we created in Study 2. We trained and tested separate classifiers on the perplexity vectors for sentence sets with a given cardinality (i.e. 2, 4, … or 1024 sentences). For each register we randomly grouped half of the perplexity vectors for training and the other half for testing each classifier.

In addition, we used the text materials from the validation corpus obtained in Study 3. We divided each register into sentence sets containing the same number of sentences as before (2, 4, … 1024) and computed the perplexity vectors for all sentence sets. The register classifiers we trained on the Spoken Dutch Corpus materials were used to classify the sentence sets from the validation corpus. Again a classifier trained on sentence sets with a given cardinality (i.e. 2, 4, … or 1024) was used to test sentences sets with the same cardinality.

7.2 Results and discussion

The results, shown in Figure 2, show that the speech register classifier reaches ceiling performance (100%) when using sets of 512 sentences, while the classification of the validation corpus reaches its maximum performance (95%) with sets of 256 sentences. The accuracy results based on sets of 128 sentences are similar (92%) for the validation and Spoken Dutch Corpus. Larger sentence sets show slightly better performance for the Spoken Dutch Corpus, possibly a result of overfitting.

For the smaller sets of 2 – 64 sentences, the accuracy results for the validation corpus are higher than for the Spoken Dutch Corpus, which might come as a surprise. However, the components of the validation corpus belong to three very distinct speech registers, while the Spoken Dutch Corpus consists of 14 speech registers, including closely related registers (e.g. spontaneous conversations and telephone dialogues). This makes classification of the registers in the Spoken Dutch Corpus harder.

It is interesting to note that with small sets of sentences reasonably high accuracy is achieved. For the Spoken Dutch Corpus only 64 sentences are needed for 90% accuracy and for the validation corpus only 16 sentences are needed for a similar accuracy.

[Insert Figure 2 around here]

To investigate whether some speech registers are more similar in word predictability compared to others, we created a scatterplot based on the first two Linear Discriminants from the register classifier based on sets of 128 sentences (see Figure 4). The scatterplot shows that the four components of the validation corpus are located closely to the counterparts in the Spoken Dutch Corpus. Most registers are separated from all other registers except for the spontaneous dialogues (components *a, c, d),* which show considerable overlap.

[Insert Figure 3 around here]

Taken together the results show that it is possible to classify registers with a small amount of speech (i.e. 128 sentences) with high accuracy (92%). The scatterplot and the classification errors show that the spontaneous registers are similar, while all other components in the Spoken Dutch Corpus are more distinct.

8. Study 5: The sentence length confound

We addressed the potential sentence length confound, because the different components in the Spoken Dutch Corpus show a wide range in the average length of sentences, which could influence perplexity scores. The classifier may therefore be based on average sentence length rather than on word predictability. In Study 5 we investigated this by selecting a subset of our materials in such a way to reduce the difference in average sentence length between components. Furthermore, we created a classifier based on sentence length to test to what extent such a classifier can successfully distinguish between registers.

8.1 Procedure

We used the same materials as in Study 4, with the exception that, across all components, we only selected sentences containing 2 to 25 words. We excluded one-word sentences because they are mostly backchannels, which occur predominantly in more spontaneous speech registers and may therefore have a strong influence on overall perplexity score differences between speech registers. We excluded sentences longer than 25 words to restrict the range in average sentence length over all components.

To show the extent of average sentence length variability across registers, we tabulated, in Table 5, the average sentence length for the different speech registers in the Spoken Dutch Corpus and the validation corpus. The average sentence length differs quite extensively (range 6 – 28 words on average per sentence). The range was reduced to 7 – 15 words on average per sentence in the subset restricted by sentence length. Table 5 also shows that similar registers can differ in sentence length in different corpora. For example, component *k* and *o* from the Spoken Dutch Corpus have a high average sentence length, while the validation corpus equivalents (i.e. books and teleprompt texts) do not.

Table 5. Overview of the number of words, sentences and the average sentence length per component in the Spoken Dutch Corpus and the validation corpus. The table shows the number of word tokens, number of sentences and the average sentence length per component for all sentences (left column) and percentages for the subset of the sentences of 2 – 25 words (right column). The component column shows the letter identifier of each component in the Spoken Dutch Corpus. The names in capitals refer to the different corpora in the validation corpus.

All Sentences				Sentences with 2 - 25 words		
com-ponent	word tokens	sentences	average sentence length	% of total word tokens	% of total sentences	average sentence length
a	1,745,854	303,186	6	88	70	7
b	249,844	23,835	11	67	68	10
c	738,794	129,351	6	88	68	7
d	509,960	83,514	6	87	70	8
e	136,438	179,14	8	77	69	9
f	538,795	52,274	10	68	73	10
g	217,626	110,63	20	42	68	12
h	278,749	34,496	8	83	78	9
i	130,336	124,12	10	76	94	9
j	90,614	7,620	12	73	82	11
k	285,278	21,176	14	96	98	13
l	80,081	6,210	13	72	85	11
n	61,799	2,190	28	28	54	15
o	551,441	47,944	12	79	90	10
BOOKS	1,000,042	121,256	8	93	91	8
TP*	1,000,044	107,080	9	95	95	9
IFADV	70,170	12,203	6	92	74	7
ECSD	157,106	19,197	8	71	72	8

* TP = teleprompt texts

We trained the speech register classifiers using the same procedure and sentence sets as described in Study 4. In addition, we created a speech register classifier based solely on sentence length.

To create the latter classifier, we computed sentence length counts (counts of sentences with specific numbers of words) for each speech register in the Spoken Dutch Corpus. The histogram of sentence lengths per register represents a register-specific sentence length model analogous to the SLM used before. We created test sets for the sentence length model by computing sentence length counts for all sentence sets (of 2, 4, 8 … 1024 sentences) for both the Spoken Dutch Corpus and the validation corpus. We compared these test sets with the (speech register-specific) sentence length models using the Kullback-Leibler divergence (D_{KL}), presented in Equation 7. We used the D_{KL} as a similarity metric analogous to how we used perplexity scores.

$$D_{KL}(p||q) = \sum_i p(i) \log \frac{p(i)}{q(i)} \tag{7}$$

In Equation 7 q denotes the observed distribution (test set) and p the modelled distribution. The D_{KL} is a measure of the asymmetric difference between q and p. In our case the observed distribution q is the sentence length counts for a given set of sentences and the modelled distribution p is the sentence length counts of a given register (i.e. a component in de the spoken Dutch corpus).

We calculated the D_{KL} for each combination of a sentence set and speech register, similar to the approach used with the SLMs. We used the resulting D_{KL} similarity vectors for each sentence set to train and test register classifiers based on the Spoken Dutch Corpus. We validated these classifiers with sentence sets from the validation corpus. Classifiers for the smaller sentence sets (sets of 2,4,…,16 sentences) were not created, because of the prohibitively long computing time necessary for the calculation of all the D_{KL} values.

To quantify performance difference between word predictability and sentence length based classifiers, we calculated the average cross-entropy (ACE) for both the sentence length and word predictability classifiers (both LDA based). The cross-entropy reflects the difference between the probability the classifier assigns to each possible class (the fourteen different registers in this case) and the correct class. If a classifier assigns a high probability to the correct class, this results in a low cross-entropy. The cross-entropy is calculated according to Equation 8, where p denotes the probability of the class for the current test set (i.e. the correct class equals 1 and all other classes equal 0) and q denotes the probability for each class according to the classifier.

$$H(p, q) = - \sum_x p(x) \log q(x) \tag{8}$$

We computed the cross-entropy for all sentence sets for both the classifier based on word predictability and the one based on sentence length. Subsequently, we computed the ACE by averaging the cross-entropy across all sentence sets of specific cardinality for each classifier (based either on word predictability or sentence length) and compared the results.

8.2 Results and discussion

Figure 4 shows the results of the two different types of speech register classifiers, the one based on word predictability and the one based on sentence length. The results are provided for both the Spoken Dutch Corpus and for the validation corpus.

[Insert Figure 4 around here]

The speech register classifiers based on word predictability reach ceiling performance (accuracy 100%) with sets of 512 sentences. The validation corpus reaches maximum performance (accuracy 98%) with sets of 1024 sentences. The results are comparable to the results obtained in Study 3, which were based on all sentences (see Figure 2). This is a first indication that average sentence length differences across registers do not underlie the accuracy of our classifiers assumed to be based on word predictability, because differences in average sentence length were reduced in the current experiment.

The classifiers based just on sentence length were able to classify speech registers in the Spoken Dutch Corpus with reasonable accuracy. The classification performance does not generalize to the validation corpus. Furthermore, the ACE results also show that the word predictability based classifiers outperform the sentence length classifiers (Table 6): The comparison between classifier types shows a clear advantage for the word predictability classifier. We conclude that sentence length differences between registers cannot explain the results found with the classifiers based on word predictability.

Table 6. Performance comparison of the speech register classifiers based on Sentence Length and Word Predictability. ACE scores are based on the data from the Spoken Dutch Corpus. Lower scores indicate better performance.

	ACE scores for each register classifier	
Sentence set	Sentence length	Word predictability
32	1.74	0.43
64	1.49	0.23
128	1.23	0.08
256	0.94	0.01
512	0.68	0.0001
1024	0.38	< 0.0001

In conclusion, the results from Study 4 show that the performance of the speech register classifier based on word predictability cannot be attributed to sentence length differences between the components in the Spoken Dutch Corpus. When we restrict the corpus to sentences of 2 -25 words (to attenuate differences in sentence length between speech registers), the accuracy results are very similar to the results based on all sentences. Additionally, when we trained a classifier based on sentence length, the classifier performance did not generalize to the validation corpus and this classifier was also clearly outperformed by a classifier based on word predictability, as shown by the ACE comparison.

9. General Discussion and Conclusion

We investigated differences in word predictability between speech registers in Dutch. We created register-specific statistical language models (SLM) for each component in the Spoken Dutch Corpus. We tested the register-specific SLMs on unseen partitions of each register and used the resulting perplexity vectors to train a register classifier based on LDA. We found that the classifier was able to distinguish between speech registers accurately, and we were able to rule out the possibility that the classifier's high performance was based on confounds of topicality and sentence length with speech register.

In Study 1 we compared two approaches to create an SLM vocabulary. We found that there were substantial differences in word token frequency for some word types between speech registers. We used averaged reduced frequency (ARF) to filter out bursty words (i.e. words that only occur in concentrated bursts in the corpus). This approach was able to attenuate speech register vocabulary differences related to topic specificity.

An alternative solution would be to only include closed-class words. However, because we were specifically interested in word predictability differences between registers, we wanted to maintain as many word types as possible in our SLM vocabulary. Our results show that future studies that investigate differences in register or genre and that need to use a rich vocabulary best use a word selection criterion that penalizes topic-specific words.

Future research may further investigate the relation between word burstiness and the topic-specificity of words. For the current study we treated them as equivalent, but it may be possible to create a measure that more specifically targets topic-specificity, which may thus result in an improved inclusion criterion for a robust vocabulary.

In Study 2 we created register-specific SLMs and tested all SLMs using 10-fold cross-validation on unseen partitions from all components in the Spoken Dutch Corpus. We used half of the resulting perplexity vectors (which are lists of perplexity scores that relate to the ability of each register-specific SLM to predict an upcoming word based on preceding words for a register-specific text) to train a speech register classifier based on LDA and tested the classifier on the unseen other half of the perplexity vectors. We found that the speech register classifier was able to distinguish perfectly between all 14 speech registers in our subset of the Spoken Dutch Corpus.

In Study 3 we performed the same procedures on 1000 pseudo-random variants of the Spoken Dutch Corpus. These pseudorandom corpora contained similarly-sized components as the real corpus, but the sentences were randomly shuffled among components. The classifiers trained on pseudo-random variants of the Spoken Dutch Corpus performed poorly and could not distinguish the randomized components. This result shows that a classifier trained on perplexity scores cannot distinguish between random (heterogeneous) sets of sentences and that the accuracy results obtained with the speech register classifier are based on systematic differences between speech registers.

Furthermore, we created a validation corpus, with materials from other corpora. The validation of our register classifier is important in light of the finding by Miller & Biber (2015), who showed that the number of word types keeps growing with the addition of new texts to a corpus, even if they are from a restricted domain (i.e. psychology textbook). It is therefore important to test whether results hold across corpora. We tested the speech register classifier (trained on material from the Spoken Dutch Corpus) on the validation corpus and found that it can also accurately classify registers in this corpus. This shows that our speech register classifier is not overfitted to idiosyncratic aspects of the Spoken Dutch Corpus. The combined results support our hypothesis that word predictability differs across speech registers.

In Study 4 we investigated the amount of text needed to classify the register of a text based on word predictability. We split text materials of the Spoken Dutch Corpus into sets containing different number of sentences (i.e. 2,4,…, or 1024) and computed the perplexity vectors for each set. These were used to train and test classifiers for each sentence set size (i.e. a classifier for sentence set of size 2,4,…, or 1024). We found that a set of 128 sentences is sufficient to train a classifier with a classification accuracy of 92% and showed that this classifier performs similarly on materials from the validation corpus.

Figure 3 shows speech register differences captured by our classifier by means of a scatterplot based on the first two linear discriminants of the LDA. The plot shows that the components from the validation corpus are located closely to their counterparts in the Spoken Dutch Corpus. Furthermore, it shows that, compared to other registers, the different spontaneous registers cluster together closely. This is corroborated by the confusion matrices of the classifiers; most classification errors are made between the spontaneous conversations *a* and the two telephone dialogue components *c* and *d.* All other registers are well separated.

The clustering of spontaneous speech registers corresponds well with previous literature. Multiple factors contribute to the similarity of spontaneous speech registers (e.g., Leech 2000: 697-701, Ellis 2002: 156). For example, shared context between interlocutors reduces the need for specificity. Another contributing factor is the available processing time. Speakers only have limited time for processing and no possibility of editing, which typically results in a limited and reused repertoire. (i.e. the use of formulaic language to achieve a certain speech act; e.g., Schmitt 2010: 8-12). These factors work together to produce spontaneous speech registers that are similar, as is attested by the result from our study.

Previous research reported a distinction between informational and involved dimension in language use (Biber 1988, 1995) with factor analysis. Our cluster of spontaneous registers could be interpreted as registers that use involved language;

however, the other registers do not cluster together in an informational counterpart. This could be because instead of using comparatively small sets of lexico-grammatical features, we used large sets of *n*-grams width statistical language modelling. It is possible that a large feature set such as *n*-grams is sensitive to differences between registers that use informational language, which would explain why we did not find a cluster of informational registers. Our results suggest that register differences are not exclusively related to lexico-grammatical features, because word *n*-grams reveal subtle but robust differences across registers. We propose that register analysis based on lexico-grammatical features, could be fruitfully complemented by this new approach.

In Study 5 we investigated the influence of sentence length differences between components in the Spoken Dutch Corpus on the performance of the classifier. Sentence length differences can influence perplexity scores. To rule out that the classifier accuracy results are based on differences in sentence length instead of word predictability, we did the following. We used a subset of the corpus containing sentences of 2 – 25 words, which reduced differences in sentence length between components. We found results similar to those in Study 4, which suggests that sentence length differences cannot account for the performance of the speech register classifier.

Furthermore, we trained a register classifier based solely on sentence length. This classifier could distinguish between the components in the Spoken Dutch Corpus to some extent, but was clearly outperformed by the classifier based on word predictability. Additionally, the performance of the classifier based on sentence length did not generalize to the validation corpus, indicating that sentence length only is not a robust basis for a register classifier. This result is at odds with findings by Wiggers and Rothkrantz (2007), who reported that sentence length is a good indicator of speech register. However, our test was based on a subset of the materials that removed the most extreme sentence lengths and this could have influenced the results.

Our results have implications for studies investigating word predictability in relation to language comprehension. Given the sensitivity of readers and listeners to the predictability of words (e.g. Smith & Levy 2013), it is plausible that they are also sensitive to register-specific differences in word predictability. Future research has to show whether readers and listeners adapt their expectations based on the wider context of situation of use when comprehending written or spoken language.

Our results also raise important questions about the nature of lexical representations. For example, what type of lexical representation allows speakers to systematically adapt their word use to the appropriate register? Are different word predictabilities stored for every speech register and if so, how many registers are lexically represented? If listeners use register-specific word predictability to tune their anticipations of upcoming words, the question is again how these register-specific word predictabilities are mentally represented.

The study shows that the combination of register analysis and text classification with the aid of statistical language modelling provides important new insights about registers and the requirements needed for speech processing and the mental lexicon. Importantly, the study extends the finding that situation of use determines language variation, by reporting differences across speech registers in word predictability.

Notes

1 The SLM vocabulary lists all words that are used to train the SLM.

2 Netherlandic Dutch is a variety of Dutch spoken in the Netherlands.

3 An algorithm to determine the class of a new observation, based on a training set of observations that belong to various classes.

4 SRILM release 1.5.12, http://www.speech.sri.com/project

5 A count-of-count list lists the number of *n*-grams occurring a specific number of times (i.e. there are 15 unigrams that occur 3 times) in the training data.

References

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., & Gildea, D. (1999). Forms Of English Function Words-Effects Of Disfluencies, Turn Position, Age And Sex, And Predictability. In *Proceedings of ICPHS-99*.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443-467.

Biber, D. (1988). *Variation across speech and writing*. New York: Cambridge University Press.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. New York: Cambridge University Press.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. New York: Cambridge University Press.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, *13*(4), 359-393.

Church, K. W., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*(02), 163-190.

Denoual, E. (2006). A method to quantify corpus similarity and its application to quantifying the degree of literality in a document. *International Journal of Technology and Human Interaction*, *2*(1), 51-66.

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in second language acquisition*, *24*(2), 143-188.

Frisson, S., Rayner, K., & Pickering, M. J. (2005). Effects of contextual predictability and transitional probability on eye movements during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(5), 862-877.

Van Gijsel, S., Speelman, D., & Geeraerts, D. (2006). Locating lexical richness: a corpus linguistic, sociovariational analysis. In *Proceedings of the 8th International Conference on the statistical analysis of textual data*, 961-971.

Goedertier, W., Goddijn, S. M., & Martens, J. P. (2000). Orthographic Transcription of the Spoken Dutch Corpus. In *Proceedings of LREC-2000*.

Van Gompel, M., & Van den Bosch, A. (2016). Efficient n-gram, skipgram and flexgram modelling with Colibri Core. *Journal of Open Research Software*, *4*(1).

Gries, S. T. (2001). A corpus linguistic analysis of English -ic vs -ical adjectives. *ICAME Journal*, *25*, 65-108.

Gries, S. T., Ellis, N. C. (2015). Statistical Measures for Usage-Based Linguistics. *Language Learning, 65*(1), 228-255.

Hlaváčová J., Rychlý P. (1999). Dispersion of Words in a Language Corpus. In V. Matousek, P. Mautner, J. Ocelíková, P. Sojka (Eds.), *Text, Speech and Dialogue: Second International Workshop, TSD'99 Plzen, Czech Republic, September 13-17, 1999 Proceedings* (pp. 321-324). Berlin: Springer Berlin Heidelberg.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River: Pearson.

Kilgarriff, A. (2001). Comparing corpora. *International journal of corpus linguistics*, *6*(1), 97-133.

Lee, D. Y. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology,* 5(3), 37-72.

Leech, G. (2000). Grammars of Spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*(4), 675-724.

Marco, J. (2000). Register analysis in literary translation: A functional approach. *Fédération International des Traucteurs (FIT) Revue Babel 46*(1), 1-19.

Miller, D., Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, *20*(1), 30-53.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 398-408). Association for Computational Linguistics.

Oostdijk, N. (2001). The design of the spoken Dutch corpus. *Language and Computers*, *36*(1), 105-112.

Oostdijk, N., Reynaert, M., Hoste, V., Schuurman, I. (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In P. Spyns & J. Odijk (Eds.), *Essential speech and language technology for Dutch* (pp. 219-247). Berlin: Springer Berlin Heidelberg.

Pluymaekers, M., Ernestus, M., & Baayen, R. H. (2006). Effects of word frequency on the acoustic durations of affixes. In *Proceedings of the International Conference on Statistical Language Processing*. International Speech Communication Association.

Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora of ACL 2000* (pp. 1-6). Association for Computational Linguistics.

Savický, P., & Hlavácová, J. (2002). Measures of word commonness. *Journal of Quantitative Linguistics*, *9*(3), 215-231.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302-319.

Van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: a Free Dialog Video Corpus. In *LREC* (pp. 501-508).

Stolcke, A. (2002). SRILM-an extensible language modelling toolkit. In J. H. L. Hansen & B. L. Pellom (Eds.) *Proceedings of the International Conference on Statistical Language Processing*. International Speech Communication Association.

Tottie, G. (1991). *Negation in English speech and writing: A study in variation*. Academic Press.

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, 26(6), 2506-2516.

Appendix 1

In the tables below we report the precision, recall and f1 scores for different speech register classifiers discussed in the paper.

Table 1. Precision, recall and f1 scores for the speech register classifier in Study 2.

	precision	**recall**	**f1**
Spontaneous dialogues	1.00	1.00	1.00
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues via platform	1.00	1.00	1.00
Spontaneous telephone dialogues via a minidisc recorder	1.00	1.00	1.00
Business negotiations	1.00	1.00	1.00
Radio and television interviews and discussions	1.00	1.00	1.00
Debates, discussion and meetings (especially political)	1.00	1.00	1.00
Classes	1.00	1.00	1.00
Spontaneous radio and television commentaries (e.g. sports)	1.00	1.00	1.00
Radio and television newsroom and documentaries	1.00	1.00	1.00
News broadcast on radio and television	1.00	1.00	1.00
Reflections and commentaries broadcast	1.00	1.00	1.00
Lectures and speeches	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 2. Precision, recall and f1 scores for the speech register classifier tested on the validation materials in Study 3.

	precision	**recall**	**f1**
Spontaneous dialogues	1.00	0.85	0.92
News broadcast on radio and television	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 3. Precision, recall and f1 scores for the speech register classifier tested on the validation corpus materials in Study 4. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	1.00	0.61	0.76
News broadcast on radio and television	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 4. Precision, recall and f1 scores for the speech register classifier in Study 4 tested on the Spoken Dutch corpus materials. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	0.98	0.96	0.97
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues via platform	0.90	0.89	0.90
Spontaneous telephone dialogues via a minidisc recorder	0.81	0.90	0.85
Business negotiations	1.00	1.00	1.00
Radio and television interviews and discussions	1.00	1.00	1.00
Debates, discussion and meetings (especially political)	1.00	1.00	1.00
Classes	1.00	0.99	1.00
Spontaneous radio and television commentaries (e.g. sports)	1.00	1.00	1.00
Radio and television newsroom and documentaries	0.97	1.00	0.98
News broadcast on radio and television	1.00	1.00	1.00
Reflections and commentaries broadcast	1.00	1.00	1.00
Lectures and speeches	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 5. Precision, recall and f1 scores for the speech register classifier (based on word predictability scores) tested on the validation corpus materials in Study 5. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	1.00	0.73	0.84
News broadcast on radio and television	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 6. Precision, recall and f1 scores for the speech register classifier (based on sentence length) tested on the validation corpus materials in Study 5. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	0.75	0.19	0.30
News broadcast on radio and television	0.00	0.00	0.00
Read aloud speech	0.01	0.01	0.01

Table 7. Precision, recall and f1 scores for the speech register classifier (based on word predictability) tested on the validation corpus materials in Study 5. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	0.99	0.98	0.99
Interviews with teachers of Dutch	1.00	1.00	1.00
Spontaneous telephone dialogues via platform	0.94	0.91	0.93
Spontaneous telephone dialogues via a minidisc recorder	0.86	0.94	0.90
Business negotiations	1.00	1.00	1.00
Radio and television interviews and discussions	1.00	1.00	1.00
Debates, discussion and meetings (especially political)	1.00	1.00	1.00
Classes	1.00	1.00	1.00
Spontaneous radio and television commentaries (e.g. sports)	1.00	1.00	1.00
Radio and television newsroom and documentaries	1.00	1.00	1.00
News broadcast on radio and television	1.00	1.00	1.00
Reflections and commentaries broadcast	1.00	1.00	1.00
Lectures and speeches	1.00	1.00	1.00
Read aloud speech	1.00	1.00	1.00

Table 8. Precision, recall and f1 scores for the speech register classifier (based on sentence length) tested on the validation corpus materials in Study 5. The scores are provided for the classifier trained and tested on 128-sentence sets.

	precision	**recall**	**f1**
Spontaneous dialogues	0.77	0.56	0.65
Interviews with teachers of Dutch	0.36	0.38	0.37
Spontaneous telephone dialogues via platform	0.40	0.47	0.44
Spontaneous telephone dialogues via a minidisc recorder	0.23	0.29	0.26
Business negotiations	0.24	0.48	0.32
Radio and television interviews and discussions	0.55	0.47	0.51
Debates, discussion and meetings (especially political)	0.84	0.87	0.85
Classes	0.36	0.36	0.36
Spontaneous radio and television commentaries (e.g. sports)	0.18	0.51	0.27
Radio and television newsroom and documentaries	0.12	0.53	0.20
News broadcast on radio and television	1.00	0.99	0.99
Reflections and commentaries broadcast	0.15	0.30	0.20
Lectures and speeches	0.67	0.67	0.67
Read aloud speech	0.66	0.37	0.48