Variation between Different Discourse Types: Literate vs. Oral

Katrin Ortmann
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
ortmann@linguistics.rub.de

Stefanie Dipper
Department of Linguistics
Fakultät für Philologie
Ruhr-Universität Bochum
dipper@linguistics.rub.de

Abstract

This paper deals with the automatic identification of literate and oral discourse in German texts. A range of linguistic features is selected and their role in distinguishing between literate- and oral-oriented registers is investigated, using a decision-tree classifier. It turns out that all of the investigated features are related in some way to oral conceptuality. Especially simple measures of complexity (average sentence and word length) are prominent indicators of oral and literate discourse. In addition, features of reference and deixis (realized by different types of pronouns) also prove to be very useful in determining the degree of orality of different registers.

1 Introduction

Halliday distinguishes between two kinds of variation in language: social variation, which he calls *dialect*, and functional variation, which he calls *register* (e.g. Halliday, 1989, p. 44). VarDial's focus is on the first kind of variation, in particular diatopic variation, and addresses topics such as automatic identification of dialects but also includes topics like diachronic language variation. In this paper, we look at variation of the second kind, namely variation between literate/written and oral/spoken language (different registers, as Halliday would call it). However, we assume that the phenomenon of literate/written vs. oral/spoken language interacts with diachronic language change, which, in turn, interacts with diatopic variation (e.g. one dialect becomes more important than another one and has larger impact on the further development of the language). Hence, if we want to understand language change, we have to take into account different kinds of variation.

In general, human language is used in two major forms of representation: written and spoken. Both discourse modes place different demands on the language user. Spoken discourse has to be processed online by speakers and hearers and, hence, strongly depends on the capacity of the working memory. In contrast, written discourse proceeds independently of production and reading speed, and allows for a rather free and elaborate structuring of texts. This discrepancy can result in quite different utterances.

Moreover, as many linguists have noticed, there is also a high amount of variation *within* written and spoken language (Koch and Oesterreicher, 2007; Halliday, 1989; Biber and Conrad, 2009). For example, the language used in scientific presentations is rather similar to prototypical written language, despite its spoken realization. Chat communication on the other hand, although realized in the written medium, rather resembles spontaneous spoken speech. In other words, independently of their medial realization, language can show characteristics that are typical of the written or spoken mode. As Halliday (1989, p.32) puts it, "'written' and 'spoken' do not form a simple dichotomy; there are all sorts of writing and all sorts of speech, many of which display features characteristic of the other medium".

In the 1980s, Koch and Oesterreicher (1985) proposed to distinguish between *medial and conceptual orality and literacy*. On the medial dimension, an utterance can be realized either phonetically (spoken) or graphically (written), while the conceptual dimension forms a broad continuum between the extremes of conceptual orality and conceptual literacy. Example (1) from Halliday (1989, p.79) illustrates this continuum, from a clear conceptually-literate sentence in (a) to a clear conceptually-oral equivalent in (c).

(1) a. The use of this method of control unquestionably leads to safer and faster train run-

Proceedings of VarDial, pages 64–79
Minneapolis, MN, June 7, 2019

ning in the most adverse weather conditions.

b. If this method of control is used trains will unquestionably (be able to) run more safely and faster (even) when the weather conditions are most adverse.

c. You can control the trains this way and if you do that you can be quite sure that they'll be able to run more safely and more quickly than they would otherwise no matter how bad the weather gets.

The work reported here is part of a larger project which investigates syntactic change in German across a long period of time (1000 years). One of the working hypotheses of the project is that certain parts of syntactic change can be attributed to changes in discourse mode: Early writings showed many features of the oral mode. The dense, complex structure which is characteristic of many modern elaborate written texts is the product of a long development.

Interestingly, spoken language has also developed denser structures over time. It is commonly assumed that this is a reflex of the written language, and is due to the increasing amount of written language which became available after the invention of printing and since then has played a prominent role in the society. As Halliday (1989, p.45) argues, this feedback happens "particularly because of the prestige" of written registers.

The aim of the project is to trace these two strands of development, by investigating and comparing texts that are located at different positions of the orality scale. Of course, we do not have records of historical spoken language. Rather, we have to rely on written texts that are as close as possible to the spoken language. So we need to be able to identify conceptually-oral, i.e. spoken-like texts.

The present paper addresses the first step in this enterprise, namely to find means to automatically measure the conceptual orality of a given *modern* text. In particular, we investigate a range of linguistic features that can be automatically determined and seem useful for this task.

The remainder of this paper is structured as follows: Section 2 gives an overview of the related work. In Section 3, features of orality as proposed in the literature are presented, and the set of linguistic features used in the present study is specified. Section 4 introduces the data and describes their linguistic annotation as well as the way we determine expected orality. In Section 5, results from training a classifier on the linguistc features are discussed. Finally, Section 6 summarizes the results and gives an outlook at future investigations. An appendix provides further details of the analysis.

2 Related Work

Nowadays, the distinction between literate and oral language is widely recognized in linguistics. For instance, in a register analysis of typologically different languages Biber (1995) finds that the distinction between oral and literate language seems to be a dimension that plays a role in all these languages, although it can be expressed in different ways and he could not find "any absolute dichotomies between speech and writing" (p.236).

In the following, we focus on work that deals with features directly related to the difference between literate and oral language.

Koch and Oesterreicher (1985, 2007) list a number of universal characteristics, such as publicity vs. privacy, weak vs. strong emotional involvement, spatial and temporal distance vs. proximity, and monologicity vs. dialogicity. Combining these aspects in different ways results in different degrees of conceptual orality or literacy. Unfortunately, the characteristics are rather abstract and vague, and cannot be operationalized and applied to concrete texts.

To remedy this weakness, Ágel and Hennig (2006) extend the approach of Koch and Oesterreicher and create a framework that allows for objectively measuring the conceptual orality of a given text (in German). For this purpose, they consider a range of diverse linguistic features, e.g. deixis, ellipsis, interjections, number of complete sentences in the text, and compare the observed frequencies to a prototypical conceptually-oral text. The method as described by Ágel and Hennig (2006) requires careful manual inspection of every individual text, though, to determine a large number of linguistic features. Hence, it cannot be applied sensibly to a large amount of data.

A few approaches try to automate the process of feature identification: Rehm (2002) focusses on automatic identification of a small number of features in the restricted domain of computer-mediated communication (CMC) in German, such

as websites, emails, etc. The analyzed features include smileys, iterations, emphasis, isolated verb stems like *grins* 'smile', slang expressions or abbreviations, and a few other features like specific punctuation symbols and phonetic contractions marked with an apostrophe.

Following Biber (1995), Biber and Conrad (2009) conduct a register analysis based on automatically-identified co-ocurring linguistic features in English texts. In their analysis, the distinction of oral and literate language makes up the first dimension along which the analyzed registers differ. Biber (1995) showed that if this dimension is broken down, it turns out that it consists of fine-grained dimensions, e.g. dimensions concerning the degree of interactiveness (dialog vs. monolog), production circumstances (on-line vs. careful production), stance (overt marking of personal stance and involvement vs. non-personal/informational), and language-specific functions (e.g. abstract vs. non-abstract style in English, narrative vs. non-narrative in Korean).

3 Features of Orality

The aim of this paper is to identify linguistic features that (i) are useful predictors of the conceptual orality of a given text and (ii) can be recognized fully automatically in texts of any length. Previous work discusses a broad range of features that distinguish between written and spoken mode or literate and oral discourse. As explained above, the medium (written/spoken) and conceptuality (literate/oral) concern different aspects of language, and go hand in hand only in prototypical cases, e.g. edited news (written and literate) or spontaneous colloquial communication (spoken and oral). Researchers often investigate only one of the aspects in their work, and most of them focus on the medial distinction (written vs. spoken), e.g. Chafe (1982), Drieman (1962), Richter (1985), Tomczyk-Popińska (1987). Moreover, many of them consider prototypical cases. As a consequenc, for many features discussed in the literature it is not obvious whether they are indicative of the medium or of conceptuality.

The following presentation does not try to distinguish systematically between the two aspects, and, instead, makes a rough distinction between written/literate on the one hand, and spoken/oral on the other hand. Our study presented in Sec. 5 reveals which of the features correlate with oral conceptuality (whereas the medial aspect is not relevant to our purposes). The focus is on features proposed for English and German.

Reference/deixis As a consequence of the spatial and temporal co-presence of participants, spoken language shows an increased use of pronouns and demonstratives as compared to lexical nouns (Goody, 1987; Diamante and Morlicchio, 2015; Schwitalla and Tiittula, 2009; Tomczyk-Popińska, 1987). There are also some language-specific differences like the use of proper names with a definite article in German (Schwitalla and Tiittula, 2009) as in ***der*** *Peter* '(*the) Peter'. This construction is frequent in spoken (and oral) communication but disapproved in written (and literate) language.

Complexity As spoken language is produced and processed in real-time, it is largely dependent on the capacity of the working memory (Weiß, 2005). Therefore, spoken language is less complex than written language in many respects, e.g. it comes with shorter sentences and words (Bader, 2002; Richter, 1985; Tomczyk-Popińska, 1987; Drieman, 1962; Rehm, 2002), less complex noun phrases (Weiß, 2005), less subordination and more coordination (Ágel and Hennig, 2006; Bader, 2002; Müller, 1990; Richter, 1985; Schwitalla and Tiittula, 2009; Sieber, 1998; Speyer, 2013; Tomczyk-Popińska, 1987), which also leads to an increase of sentence-intial use of *and* and *but* (Chafe, 1982).

Moreover, written language shows a nominal style with a higher number of nouns and nominalizations, while spoken language shows a verbal style with a higher proportion of verbs (Bader, 2002; Chafe, 1982; Dürscheid, 2006; Goody, 1987; Halliday, 1989; Sieber, 1998). Finally, written and spoken language differ with respect to the information density, measured as lexical density, i.e. the ratio of lexical vs. functional words: written language uses more lexical words than spoken language (Halliday, 1989).

Syntax Further syntactic features that mark spoken language include a higher ratio of ellipsis (Ágel and Hennig, 2010; Bader, 2002; Fiehler, 2011; Müller, 1990; Richter, 1985; Schwitalla and Tiittula, 2009; Tomczyk-Popińska, 1987), and of parentheses and anacolutha (Müller, 1990; Richter, 1985). Similarly, spoken language shows a clear preference for active instead of passive

Feature	Description
mean_sent	Mean sentence length, without punctuation marks.
med_sent	Median sentence length, without punctuation marks.
mean_word	Mean word length.
med_word	Median word length.
subord	Ratio of subordinating conjunctions (tagged as KOUS or KOUI) to full verbs.
coordInit	Proportion of sentences beginning with a coordinating conjunction.
question	Proportion of interrogative sentences, based on the last punctuation mark of the sentence.
exclam	Proportion of exclamative sentences, based on the last punctuation mark of the sentence.
nomCmplx	Mean number of prenominal dependents for each noun in the dependency tree. This includes determiners but not punctuation marks, prepositions and contractions of prepositions and articles.
V:N	Ratio of full verbs to nouns.
lexDens	Ratio of lexical items (tagged as ADJ.*, ADV, N.*, VV.*) to all words.
PRONsubj	Proportion of subjects which are realized as personal pronouns, based on the head of the subject.
PRON1st	Ratio of 1st person sg. and pl. pronouns with lemmas *ich* 'I' and *wir* 'we' to all words.
DEM	Ratio of demonstrative pronouns (tagged as PDS) to all words.
DEMshort	Proportion of demonstrative pronouns (tagged as PDS) with lemmas *dies* 'this/that' or *der* 'the' which are realized as the short form (lemma *der* 'the').
PTC	Proportion of answer particles (*ja* 'yes', *nein* 'no', *bitte* 'please', *danke* 'thanks') to all words.
INTERJ	Proportion of primary, i.e. one-word interjections (e.g. *ach*, *oh*, *hallo*) to all words.

Table 1: Features used for classification. Tokens tagged as punctuation marks are not counted as words. The POS tags are from the STTS tagset.

structures (Chafe, 1982; Goody, 1987; Richter, 1985), and for analytic instead of synthetic verb forms (Müller, 1990; Richter, 1985; Sieber, 1998; Weiß, 2005) (e.g. past perfect instead of preterite). Finally, the *am*-progressive, as in *Er ist am Arbeiten* 'he is working', is a clear indicator of spoken language (Ágel and Hennig, 2010).

Lexicon A range of differences between written and spoken language can also be observed by inspecting individual words. Spoken language is characterized by frequent use of various particles, e.g. answer and modal particles in German (Diamante and Morlicchio, 2015; Fiehler, 2011; Müller, 1990; Richter, 1985; Schwitalla and Tiittula, 2009; Weiß, 2005), and interjections (Fiehler, 2011; Richter, 1985; Schwitalla and Tiittula, 2009). Furthermore, spoken language often contains vague expressions and hedges (Chafe, 1982).

Variation Since written texts can be carefully planned and revised, written language generally shows a high degree of grammatical and lexical variation, e.g. in the form of varying syntactic constructions and high type-token ratios (Drieman, 1962; Dürscheid, 2006; Müller, 1990; Sieber, 1998). In contrast, spoken language contains many repetitions (Diamante and Morlicchio, 2015; Green, 1982; Schwitalla and Tiittula, 2009). On the other hand, spoken language often exhibits a higher variation of sentence types, in that questions and exclamations are more frequent than in written language (Goody, 1987; Müller, 1990).

Graphical features Written language can express features of orality with specific graphical means, such as omission of characters, word contractions, or use of ellipsis dots, em dashes or apostrophes (Diamante and Morlicchio, 2015; Tomczyk-Popińska, 1987; Fiehler, 2011; Schwitalla and Tiittula, 2009; Richter, 1985; Rehm, 2002). Especially in the context of CMC, repetition of characters (*aaah*), and repetition of (combinations of) punctuation marks (*!!!, !?!?*), as well as capital letters or non-verbal symbols like smileys are clear indicators of orality (Rehm, 2002; Schwitalla and Tiittula, 2009).

Some of these features, such as use of specific particles, are language-dependent while others are language-independent, such as sentence or word length. This is also confirmed by Biber (1995), who shows that certain linguistic features fulfill the same functions in various languages while others are used with a specific function just in one language. In our analysis we mainly include language-independent features.

Not all of the features can be determined automatically. Some features require a detailed and reliable syntactic or semantic analysis, e.g. in the case of anacolutha or ellipsis. The present study only includes features that can be reliably identified based on automatically-created standard linguistic annotations.

Furthermore, it is to be expected that many

of the features correlate, which is precisely how Biber (1995) and Biber and Conrad (2009) identify the relevant features for their register analyses. In our study, we include various features of the different levels presented above, to allow for a broad coverage of features, and leave it to the classifier to determine the relevant ones. For an overview of the features used in the study, see Table 1.[1]

4 The Data

In order to evaluate the selected features for the task at hand, we compiled corpora from five different language registers, which differ with respect to their conceptual orality: newspaper articles (*News*), recited speeches (*Speech*), rehearsed talks (*TED*), chat communication (*Chat*), and spontaneous spoken communication (*Dialog*).

The News register includes various kinds of articles from two German newspapers.[2] In the Speech register, three different genres are considered: speeches and lectures[3] as well as modern Christian sermons.[4] The TED register consists of German transcripts of English TED talks.[5] For the Chat register, chat protocols were extracted from the Dortmunder Chatkorpus[6], including professional as well as informal chats. The texts of the Dialog register were taken from three sources: movie subtitles from the genres romance and drama,[7] subtitles of pranks filmed with a hidden camera from a German TV show[8], and work conversations.[9]

A random subset of texts with about 500,000 tokens was created for each of the five registers. Table 2 gives an overview of the data.

4.1 Preprocessing

To enable automatic identification of the described features, the data was automatically enriched with linguistic annotations. Except for the pre-tokenized texts, all corpora were automatically tokenized using the default NLTK tokenizers.[10] NLTK sentence tokenization was only applied within corpus-specific boundaries.[11]

After tokenization, the texts were tagged for part of speech (POS) with the spaCy tagger.[12] The German model uses the STTS-Tagset (Schiller et al., 1999) and overall achieves high accuracy scores.[13] All texts were automatically lemmatized using output from GermaLemma and the spaCy lemmatizer.[14] Finally, the texts were annotated with syntactic dependencies by the spaCy parser.[15]

[1] Besides syntactic features, which are excluded because they cannot be identified easily and reliably, the study also excludes graphical features, as our data includes transcriptions of spoken language which follow different notation conventions.

[2] We included articles from the Tüba-D/Z corpus (`http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-dz.html`) and the Tiger corpus (`http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/tiger.html`).

[3] The speeches and lectures were taken from Gutenberg-DE corpus, edition 14 (`http://gutenberg.spiegel.de/`), including only texts published after 1900, to allow the use of standard annotation tools for automatic processing of the orthographic surface forms.

[4] The sermons were automatically downloaded from the SermonOnline database (`http://www.sermon-online.de`).

[5] The transcripts were automatically downloaded from the official TED website at `https://www.ted.com/talks?language=de`.

[6] Release corpus from `http://www.chatkorpus.tu-dortmund.de/korpora.html`.

[7] The movie subtitles were downloaded from the OpenSubtitles database at `http://www.opensubtitles.org/de`.

[8] The subtitles were automatically downloaded from the YouTube channel of the show 'Verstehen Sie Spaß?' (`https://www.youtube.com/user/VSSpass`).

[9] From the Tüba-D/S corpus (`http://www.sfs.uni-tuebingen.de/ascl/ressourcen/corpora/tueba-ds.html`).

[10] Pre-tokenized texts are from Tiger, TüBa-D/Z and TüBa-D/S. NLTK tokenizer: `http://www.nltk.org/api/nltk.tokenize.html`. Some tokenizing errors were fixed by heuristic rules, which corrected the tokenization of repeated punctuation marks ('!!!!'), smileys and uses of the @-symbol.

[11] In particular: Movie subtitles were segmented across frames, chat protocols within messages, and lectures and speeches within lines, which usually correspond to paragraphs. In tokenizing TV subtitles, TED talks and sermons, frames or paragraph boundaries were ignored.

[12] `https://spacy.io/api/tagger` (v2.0). Certain tagging errors were automatically corrected, using word lists and regular expressions (e.g. 'ha+ll+o+', which matches all kinds of spellings of *Hallo* 'hello'). This concerned single-word interjections (ITJ), pronominal adverbs (PAV), and different punctuation types ($(, $, and $.).

[13] An evaluation of a random subset showed accuracy values of over 90% for all registers, except for the chat corpus with an accuracy of 85%. The most frequent confusions occur between nouns and proper names, and between adverbial adjectives, participles and adverbs.

[14] `https://github.com/WZBSocialScienceCenter/germalemma`, version from February 6, 2019, and `https://spacy.io/api/lemmatizer`, v2.0.
Words tagged as N.*, V.*, ADJ.* and ADV were lemmatized with GermaLemma. Pronouns (tagged as PPER, PRF and PPOS.*) were lemmatized using custom rules, to preserve information about 1st, 2nd and 3rd person. For all other words, the output of the spaCy lemmatizer was used.

[15] `https://spacy.io/api/dependencyparser` (v2.0).

Register	#Tokens	#Sentences	#Docs	Corpora
News	500,076	27,375	1,024	679 articles from the newspaper 'taz' (72%), 345 articles from the newspaper 'Frankfurter Rundschau' (28%)
Speech	500,475	18,833	31	11 (collections of) speeches (61%), 5 lectures (28%), 15 sermons (11%)
TED	500,035	30,809	224	224 talk subtitles (100%)
Chat	500,009	58,572	322	322 chat protocols (100%)
Dialog	500,622	66,815	140	30 movie subtitles (51%), 104 TV subtitles (26%), 6 work conversations (23%)

Table 2: Overview of the data. The numbers in brackets after each subcorpus provide the percentage of tokens in the register that stems from the respective subcorpus.

4.2 Expected orality

The features listed in Table 1 are designed for use by a classifier which locates the texts of the different registers on the continuum of conceptual orality. That means that we first have to assign an "index of orality" to each register. Admittedly, as Dürscheid (2006) points out, only individual texts can sensibly be located on the literate-to-oral continuum. However, it is possible to judge the prototypical conceptuality of a register based on its general characteristics. To this end, we establish four *situational characteristics* which allow us to manually determine the expected orality of the registers. The characteristics are based on features proposed by Koch and Oesterreicher (2007), Ágel and Hennig (2006) and Biber and Conrad (2009). The following paragraphs describe the characteristics in detail.

Participants: *many, few* The number of participants in the communication. We only distinguish between many (coded as -1) and few (1) participants, with few participants being an indicator of a higher conceptual orality. The value many refers to communications which usually involve hundreds or thousands of participants, such as public speeches or newspaper articles. In contrast, the value few refers to communications with usually less than ten participants. This characteristic is based on Koch and Oesterreicher (2007)'s distinction of private vs. public. We do not distinguish between addressor(s) and addresse(s), contrary to Biber and Conrad (2009).

Interactiveness: *monolog, dialog* The communication structure which can be either monologous (-1) or dialogous (1), with dialog being the indicator for conceptual orality. Dialogous registers show frequent changes of language producer(s) and recipient(s) while monologous registers are dominated by a single speaker. This characteristic has also been suggested by Koch and Oesterreicher (2007), and it is one of the "relations among participants" described by Biber and Conrad (2009) (the only one that can be determined rather easily and unambiguously).

Production circumstances: *synchronous, quasi-synchronous, asynchronous* The temporal circumstances of the production of utterances, also mentioned by Ágel and Hennig (2006) and Biber and Conrad (2009). Language production can be either synchronous, i.e. real-time production like in spontaneous communication, or asynchronous, i.e. planned production like in writing. As synchronous production is highly dependent on the working memory (Weiß, 2005), it is an indicator of higher conceptual orality. The intermediate value of quasi-synchronous language production was introduced by Dürscheid (2003) and refers to communication situations where the possibility of planning and revising one's utterances is given but possibly not exploited by the speaker, like, e.g., in chat communication or in a well-rehearsed but freely-performed presentation.

Reception circumstances: *synchronous, quasi-synchronous, asynchronous* The temporal circumstances of the reception of utterances, also emphasized by Ágel and Hennig (2006) and Biber and Conrad (2009). Like language production, reception can be either synchronous, when an utterance has to be processed in real time in the moment it is uttered, as in spontaneous communication, or asynchronous like in reading a book, where an utterance can be read multiple times and at any speed. Again, synchronous reception is an indicator of higher conceptual orality. The intermediate value of quasi-synchronous language reception is analogous to the production and refers to communication situations where the possibility of reading the speakers's utterances multiple times is given but possibly not exploited by the partici-

pants, like in chat communication, where participants usually want to answer immediately.

Table 3 shows the the five registers used in this study along with their situational characteristics. The characteristics locate the respective registers on a scale from highly literate (News) to highly oral (Dialog). The sum of the individual scores can be interpreted as an index of orality, with high scores indicating orally-oriented registers. It turns out that the two characteristics Participants and Interactiveness split the registers, as considered in the present work, in the same way so that we treat them as one property in the following section.

In order to validate our manual classification, we adapt the approach by Fankhauser et al. (2014), who compare American and British English from the 1960s and 1990s, based on unigram language models. We represent each register by POS unigram language models, which have been smoothed with Jelinek-Mercer smoothing (lambda = 0.005). We compute relative entropy (Kullback-Leiber Divergence, KLD) between each pair of registers as a measure of (dis)similarity of the two registers. In computing KLD, we can use one register as the reference register and compare it with the other four registers.[16] Fig. 1 shows the results for all registers. The plots arrange the registers according to their degree of orality (first bar: News, last bar: Dialog). When a reference register is compared with itself, (e.g. "N–N": News with News), KLD is zero and there is no column.

The plots show that the KLD scores of the orally-adjacent registers are systematically lower than KLD scores of distant registers. For instance, the first plot compares News with all other registers, and KLD is smallest with Speech (first bar) and highest with Dialog (last bar). The second plot compares Speech with all others, and, again, KLD is smallest with its immediate neighbors, News (left) and TED talks (right).[17]

[16] For probability distributions p, q, and an event space X, KLD is defined as: $\mathrm{KLD}(p||q) = \sum_{x \in X} p(x) log_2 \frac{p(x)}{q(x)}$. p represents the reference register and q is compared with it.

[17] As mentioned in the beginning of this section, a score of orality should be assigned to individual texts rather than registers. However, of the situational characteristics presented here, Interactiveness is the only one that can be observed in the data itself. All other characteristics would be part of metadata, which is not available. We therefore decided to pick registers with clear prototypical situational characteristics (e.g. TED talks aim at a large number of recipients, sermons are performed by just one speaker, etc.), so that we do not expect individual texts to diverge from the prototypical settings. There are some exceptions, though. For instance, some newspaper articles contain interviews, which are dialogous, whereas newspaper articles in general are monologous. Some movie sequence might feature a lecture, so that this part would have many participants, in contrast to other typical movie sequences (since we selected movies from the genres romance and drama, we expect such exceptional sequences to occur very rarely). Finally, chat data sometimes seem to involve many participants but looking at the data in detail shows that in fact communication takes place between small groups of people only. Hence, we assume that the vast majority of the texts exhibit the prototypical characteristics.

5 Results

As we have seen, the five registers we established in the previous section can be distinguished with regard to (expected) orality, by situational characteristics. The main question of this section is to determine in which way these registers also differ with regard to linguistics features, and which linguistic features can serve as indicators of specific registers and the degree of their conceptual orality.

In a first step, we plot the distribution of all linguistic features listed in Table 1 with regard to the different registers (cf. Fig. 2 in the appendix). The plots show that most of the features quite clearly distinguish between some of the registers. For instance, the feature mean sentence length (1st panel) clearly separates Chat and Dialog data from TED, Speech, and News.

We next train a classifier to determine the registers and their situational characteristics. We use J48 (Quinlan, 1993), a decision-tree classifier, which allows us to inspect the features most prominently used by the classifier.[18]

5.1 Classifying registers

The decision tree resulting from the full dataset is shown in the appendix in Fig. 3.

The major split distinguishes texts with sentences with a mean length of less or more than 10.5. It turns out that this split quite neatly separates oral-oriented registers, i.e. Dialogs and Chats (upper part, with shorter sentences in general), from literate-oriented registers, i.e. TED, Speech, and News (lower part, with longer sentences).

[18] We use J48 as implemented in Weka (Witten et al., 2011), combined with a filter that balances the size of the different classes in the training data. The minimum number of instances per leaf is set to 5, so that the options are set as follows:

```
weka.classifiers.meta.FilteredClassifier
-F "weka.filters.supervised.instance.
ClassBalancer -num-intervals 10" -S 1
-W weka.classifiers.trees.J48 -- -C 0.25
-M 5.
```

Register	Participants		Interactiveness		Production		Reception		Index of Orality
	value	score	value	score	value	score	value	score	score (sum)
News	many	-1	monolog	-1	asynchronous	-1	asynchronous	-1	-4
Speech	many	-1	monolog	-1	asynchronous	-1	synchronous	1	-2
TED	many	-1	monolog	-1	quasi-synchr.	0	synchronous	1	-1
Chat	few	1	dialog	1	quasi-synchr.	0	quasi-synchr.	0	2
Dialog	few	1	dialog	1	synchronous	1	synchronous	1	4

Table 3: Expected orality based on four situational characteristics of the registers. The characteristics rank the registers from highly literate (*News*) to highly oral (*Dialog*).

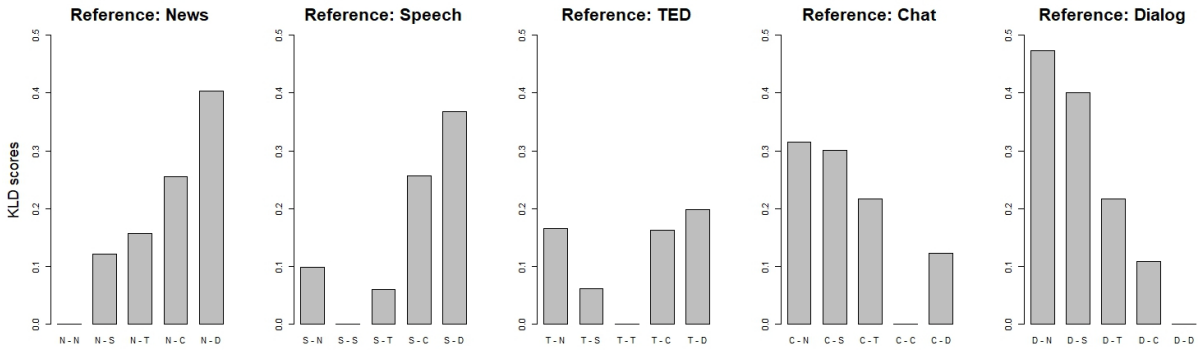

Figure 1: KLD scores of all register pairs.

Class	Precision	Recall	F-Measure
News	0.985	0.913	0.948
Speech	0.486	0.581	0.529
TED	0.731	0.848	0.785
Chat	0.817	0.857	0.836
Dialog	0.752	0.843	0.795
Weighted Avg.	0.894	0.883	0.886

Table 4: Results of classifying registers with the J48 decision-tree classifier.

In both partitions, the feature PRONsubj plays a prominent role: a low rate of pronominal subjects is indicative of News (in both partitions), and singles out certain chats (in the upper part).[19]

A 10-fold cross-validation results in an overall accuracy of 88.28%. Table 4 shows that the News register is classified with high accuracy whereas Speech data is classified with both low precision and low recall.

The confusion matrix in Table 5, which shows the confusions summed over all cross-validations, reveals that Speech data is often confused with News or TED, but very rarely or never with Chat or Dialog. Similarly, other confusions mainly occur between immediate neighbors, i.e. registers with similar levels of conceptual orality, e.g. Chat and Dialog.

classified as →	News	Speech	TED	Chat	Dialog
News	935	6	46	31	6
Speech	4	18	9	0	0
TED	5	11	190	12	6
Chat	5	0	14	276	27
Dialog	0	2	1	19	118

Table 5: Confusion matrix for the classification of registers.

Manual inspection of confusions shows that erroneous classifications of the Dialog and Chat registers mainly stem from errors in the data, e.g. missing punctuation marks, which result in long sentences or make it impossible to recognize questions automatically. Also, some features relevant to these registers, such as demonstratives, are not present in very short texts.

TED and News are mostly confused with Dialog or Chat data if they contain shorter sentences on average. This is also the main reason for the confusion of Speech data with TED talks.

Confusion of News with more orally-oriented registers (Dialog, Chat, TED) results from specific article types like interviews or literature excerpts, which contain more (first person) pronouns than is

[19]The relevance of the feature PRONsubj is also evidenced by the fact that this feature contributes the largest amount of information gain with respect to the class, as shown by Weka's "InfoGainAttributeEval", cf. Table 7 in the appendix.

typical for standard newspaper text.

5.2 Classifying situational characteristics

Since our project ultimately aims at investigating historical language data, we need classifiers that are based on functional, "timeless" features rather than features specific to modern-time registers. To this end, we trained classifiers for the different situational characteristics (see the resulting decision trees in the appendix, Fig. 4–Fig. 6).

Participants/Interactiveness As mentioned above, the registers used in this study only exhibit two combinations of these characteristics: either they are monologs with many participants or dialogs with few participants. Therefore, the resulting decision trees for the two characteristics are identical.

The most important feature for the classification of these characteristics is, again, mean sentence length. However, this time it does not introduce a clear distinction in the tree between oral- (few/dialog) and literate-oriented (many/monolog) characteristics, as we observed it for the registers.

Further relevant features are the ratio of first person pronouns (PRON1st) and questions. A large number of texts with long sentences can be classified by the (almost complete) absence of interjections (INTERJ).

The classifier achieves high scores of overall accuracy (97.13%) and average F-score (0.972, for details see Table 6 in the appendix).

Production and Reception The characteristics of the production and reception circumstances both have three possible values (asynchronous, quasi-synchronous, and synchronous), which are combined pairwise in five different ways by the five registers (see Table 3). Still, there are some interesting similarities between the two classifier trees. For both characteristics, features relating to pronouns (PRON1st for production, PRONsubj for reception) are used as the top-level split. In both cases, all synchronous instances fall into the lower part of the tree, which is marked by a larger number of these pronouns.

For reception, mean sentence length is the second most important feature while for production the mean word length is more discriminating.

It is interesting to note that with both characteristics, binary distinctions at the leaves almost never occur between the values asynchronous and synchronous. Instead, the two values are contrasted individually with quasi-synchronous. This seems to confirm the intermediate status of the quasi-synchronous value. Overall F-score of both characteristics is around 90% (see Table 6 in the appendix). In the case of production, confusions again occur mainly between neighbouring values.

6 Conclusion and Outlook

In this paper, we investigated a range of selected linguistic features, with the aim of automatically identifying conceptually oral and literate texts. It turned out that extremely simple measures of complexity, namely average sentence and word length, are prominent indicators of conceptuality. In addition, features of reference and deixis (realized by different types of pronouns) proved to be useful in determining the degree of orality of different registers.

Even though some of the features did not play major roles in the resulting decision tree, the distribution plots show that all of them are related in some way to oral conceptuality. This is confirmed by the fact that each feature is used at least once in some of the four decision trees. The features occurring least often in the decision trees are subord, exclam, and med_word.

Of course, when languages other than German are investigated, the set of linguistic features might have to be adapted, as features can be used with different functions in different languages (Biber, 1995).

When looking at diachronic data, one also has to consider that the relations between registers, their situational characteristics and the linguistic features might have changed over time. For instance, it is known that English scientific prose used to be closer to the oral mode than it is nowadays (Degaetano-Ortlieb et al., 2019).

Acknowledgments

We would like to thank the anonymous reviewers for very helpful comments. This work was supported by the German Research Foundation (DFG), SFB/CRC 1102 "Information density and linguistic encoding" (Project C6).

References

Vilmos Ágel and Mathilde Hennig. 2006. *Grammatik aus Nähe und Distanz: Theorie und Praxis am Beispiel von Nähetexten 1650-2000*. Niemeyer, Tübingen.

Vilmos Ágel and Mathilde Hennig. 2010. Einleitung. In *Nähe und Distanz im Kontext variationslinguistischer Forschung*, pages 1–22. de Gruyter, Berlin.

Jennifer Bader. 2002. Schriftlichkeit und Mündlichkeit in der Chat-Kommunikation. *NETWORX*, 29. Retrieved from `https://www.mediensprache.net/networx/networx-29.pdf`.

Douglas Biber. 1995. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

Wallace L. Chafe. 1982. Integration and involvement in speaking, writing, and oral literature. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, pages 35–53. Ablex Publishing Corporation, Norwood, NJ.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2019. An information-theoretic approach to modeling diachronic change in scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *From Data to Evidence in English Language Research*. Brill, Leiden, NL/Boston, MA.

Grazia Diamante and Elda Morlicchio. 2015. Authentische Dialoge im DaF-Unterricht? In Nicoletta Gagliardi, editor, *Die deutsche Sprache im Gespräch und in simulierter Mündlichkeit*, pages 91–114. Schneider Verlag Hohengehren, Baltmannsweiler.

G. H. J. Drieman. 1962. Differences between written and spoken language: An exploratory study. *Acta Psychologica*, 20:36–57. Doi:10.1016/0001-6918(62)90006-9.

Christa Dürscheid. 2003. Medienkommunikation im Kontinuum von Mündlichkeit und Schriftlichkeit: Theoretische und empirische Probleme. *Zeitschrift für angewandte Linguistik*, 38:37–56.

Christa Dürscheid. 2006. Äußerungsformen im Kontinuum von Mündlichkeit und Schriftlichkeit — Sprachwissenschaftliche und sprachdidaktische Aspekte. In Eva Neuland, editor, *Variation im heutigen Deutsch: Perspektiven für den Sprachunterricht*, pages 375–388. Peter Lang, Frankfurt a. M.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014. Exploring and visualizing variation in language resources. In *Proceedings of the 9th LREC*, pages 4125–4128.

Reinhard Fiehler. 2011. Mündliche Verständigung und gesprochene Sprache. In Sandro M. Moraldo, editor, *Deutsch aktuell: 2. Einführung in die Tendenzen der deutschen Gegenwartssprache*, pages 83–107. Carocci, Rome. Retrieved from `https://ids-pub.bsz-bw.de/frontdoor/deliver/index/docId/4338/file/Fiehler_Muendliche_Verstaendigung_und_gesprochene_Sprache_2011.pdf`.

Jack Goody. 1987. *The interface between the written and the oral*. Cambridge University Press.

Georgia M. Green. 1982. Some of my favorite writers are literate: The mingling of oral and literate strategies in written communication. In Deborah Tannen, editor, *Spoken and Written Language: Exploring Orality and Literacy*, pages 239–260. Ablex Publishing Corporation, Norwood, NJ.

Michael A. K. Halliday. 1989. *Spoken and written language*. Oxford University Press.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz: Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36:15–43.

Peter Koch and Wulf Oesterreicher. 2007. Schriftlichkeit und kommunikative Distanz. *Zeitschrift für germanistische Linguistik*, 35:246–275.

Karin Müller. 1990. *"Schreibe wie du sprichst!" Eine Maxime im Spannungsfeld von Mündlichkeit und Schriftlichkeit: Eine historische und systematische Untersuchung*. Lang, Frankfurt a. M.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Georg Rehm. 2002. Schriftliche Mündlichkeit in der Sprache des World Wide Web. In Arndt Ziegler and Christa Dürscheid, editors, *Kommunikationsform E-Mail*, pages 263–308. Stauffenburg, Tübingen. Retrieved from `http://www.georg-re.hm/pdf/Rehm-Muendlichkeit.pdf`.

Günther Richter. 1985. Einige Anmerkungen zur Norm und Struktur des gesprochenen Deutsch. *Deutsch als Fremdsprache*, 22(3):149–153.

Anne Schiller, Simone Teufel, Christine Stöckert, and Christine Thielen. 1999. *Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset)*. Retrieved from `http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf`.

Johannes Schwitalla and Liisa Tiittula. 2009. *Mündlichkeit in literarischen Erzählungen: Sprech- und Dialoggestaltung in modernen deutschen und finnischen Romanen und deren Übersetzungen*. Stauffenburg, Tübingen.

Peter Sieber. 1998. *Parlando in Texten: Zur Veränderung kommunikativer Grundmuster in der Schriftlichkeit*. Max Niemeyer, Tübingen.

Augustin Speyer. 2013. Performative Mündlichkeitsnähe als Faktor für die Objektstellung im Mittel- und Frühneuhochdeutschen. *Beiträge zur Geschichte der deutschen Sprache und Literatur*, 135(3):1–36.

Ewa Tomczyk-Popińska. 1987. Linguistische Merkmale der deutschen gesprochenen Standardsprache. *Deutsche Sprache: Zeitschrift für Theorie, Praxis, Dokumentation*, 15:336–375.

Helmut Weiß. 2005. Von den vier Lebensaltern einer Standardsprache. *Deutsche Sprache: Zeitschrift für Theorie, Praxis, Dokumentation*, 33:289–307.

Ian H. Witten, Eibe Frank, and Mark A. Hall. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*, 3rd edition. Morgan Kaufmann, Burlington, MA.

Appendix

Property	Values	#Texts	F-Score
Participants/ Interact.	many/monolog	1,279	0.980
	few/dialog	462	0.947
	weighted avg.		0.972
Production	asynchronous	1,055	0.942
	quasi-synchronous	546	0.838
	synchronous	140	0.795
	weighted avg.		0.898
Reception	asynchronous	1,024	0.951
	quasi-synchronous	322	0.852
	synchronous	395	0.849
	weighted avg.		0.909

Table 6: Results for classifying situational characteristics.

Information Gain	Feature
0.796	PRONsubj
0.738	V.N
0.732	PRON1st
0.69	question
0.676	PTC
0.673	mean_word
0.636	med_sent
0.633	mean_sent
0.508	lexDens
0.494	INTERJ
0.448	med_word
0.442	DEM
0.425	DEMshort
0.404	exclam
0.301	coordInit
0.206	nomCmplx
0.181	subord

Table 7: Ranking of the features according to their Information Gain with respect to the class of registers.

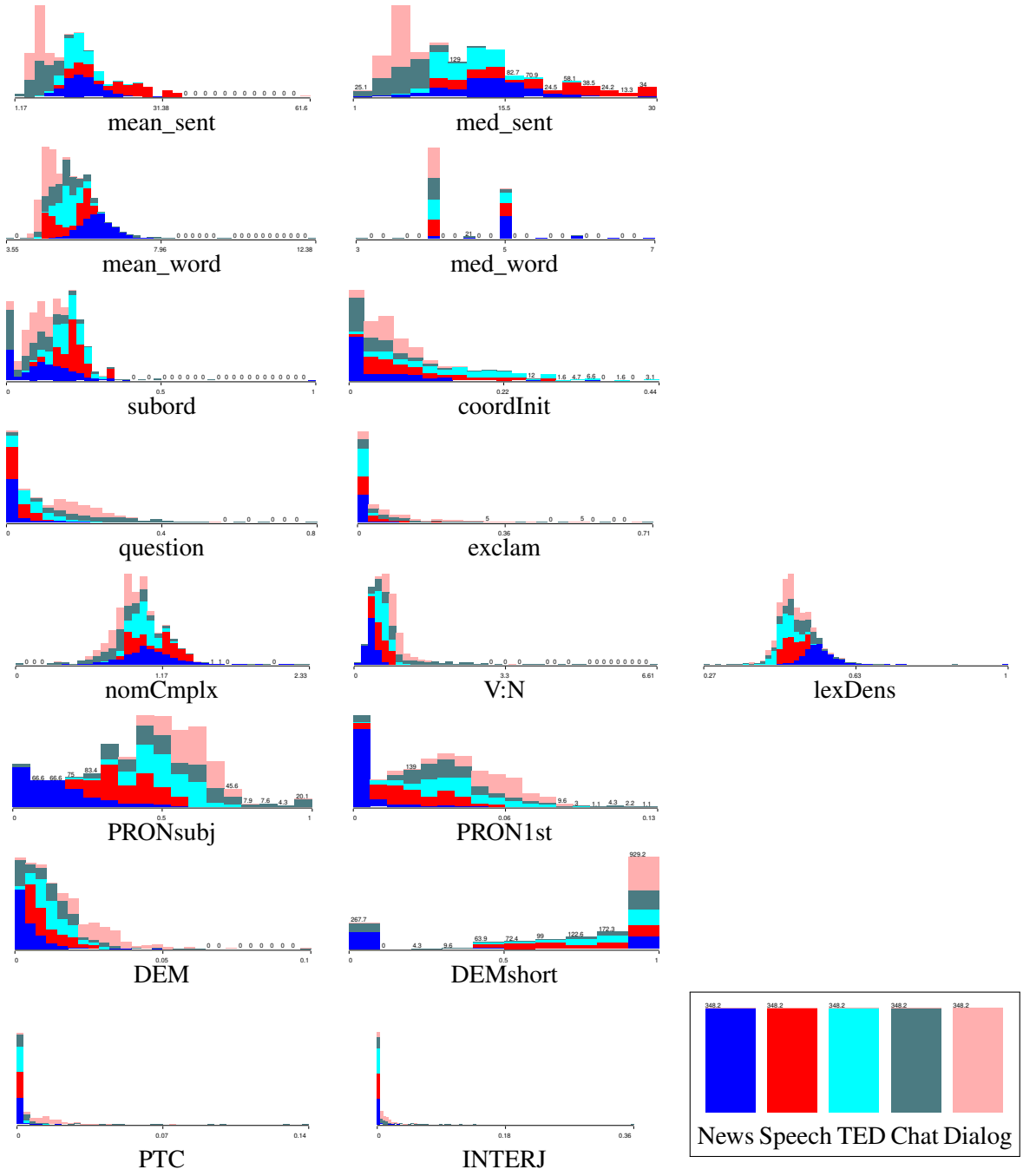

Figure 2: Weka plots for all 17 features investigated in the present study (see Table 1 for descriptions of the features). Registers are balanced and encoded by different colors (blue: *News*, red: *Speech*, cyan: *TED*, green: *Chat*, pink: *Dialog*, see the legend at the bottom right). The graphs plot the distributions of the respective features for each register. For example, the distribution of the feature PRON1st displays a large blue bar (*News*) on the left at value 0, as most newspaper articles do not contain any first person pronouns; the other registers show higher amounts of such pronouns, the pink bars (*Dialog*) achieve top values.

```
mean_sent <= 10.487395
|   PRONsubj <= 0.392857
|   |   question <= 0.066667
|   |   |   PRON1st <= 0.012245: News (34.13/2.16)
|   |   |   PRON1st > 0.012245: Chat (5.01/0.68)
|   |   question > 0.066667: Chat (59.01/1.7)
|   PRONsubj > 0.392857
|   |   DEMshort <= 0.942308: Chat (157.32/16.74)
|   |   DEMshort > 0.942308
|   |   |   V.N <= 1.659091
|   |   |   |   mean_word <= 5.123077
|   |   |   |   |   coordInit <= 0
|   |   |   |   |   |   mean_word <= 4.742631: Dialog (12.87/5.41)
|   |   |   |   |   |   mean_word > 4.742631: Chat (9.73)
|   |   |   |   |   coordInit > 0
|   |   |   |   |   |   question <= 0.277778: Dialog (319.51/13.59)
|   |   |   |   |   |   question > 0.277778
|   |   |   |   |   |   |   DEM <= 0.022727: Chat (9.73)
|   |   |   |   |   |   |   DEM > 0.022727: Dialog (9.95)
|   |   |   |   mean_word > 5.123077
|   |   |   |   |   question <= 0.142395
|   |   |   |   |   |   med_sent <= 5.5: Chat (10.81)
|   |   |   |   |   |   med_sent > 5.5
|   |   |   |   |   |   |   subord <= 0.192308: Dialog (10.63/0.68)
|   |   |   |   |   |   |   subord > 0.192308: Chat (5.41)
|   |   |   |   |   question > 0.142395: Chat (38.32/1.55)
|   |   |   V.N > 1.659091: Chat (30.28)
mean_sent > 10.487395
|   PRONsubj <= 0.232558: News (252.65)
|   PRONsubj > 0.232558
|   |   mean_sent <= 20.653846
|   |   |   lexDens <= 0.514085
|   |   |   |   exclam <= 0.019139
|   |   |   |   |   mean_word <= 4.755344: Speech (15.42/4.19)
|   |   |   |   |   mean_word > 4.755344
|   |   |   |   |   |   DEM <= 0.000745: News (6.58/2.16)
|   |   |   |   |   |   DEM > 0.000745
|   |   |   |   |   |   |   PRON1st <= 0.00905: News (6.32/1.55)
|   |   |   |   |   |   |   PRON1st > 0.00905: TED (298.92/16.01)
|   |   |   |   exclam > 0.019139
|   |   |   |   |   question <= 0.096045
|   |   |   |   |   |   mean_word <= 5.284568
|   |   |   |   |   |   |   lexDens <= 0.449857: TED (7.77)
|   |   |   |   |   |   |   lexDens > 0.449857: Speech (139.45/4.66)
|   |   |   |   |   |   mean_word > 5.284568: TED (10.83/3.06)
|   |   |   |   |   question > 0.096045
|   |   |   |   |   |   PTC <= 0.007194
|   |   |   |   |   |   |   coordInit <= 0.033113: Chat (6.49)
|   |   |   |   |   |   |   coordInit > 0.033113
|   |   |   |   |   |   |   |   lexDens <= 0.482862: TED (15.54)
|   |   |   |   |   |   |   |   lexDens > 0.482862: Chat (8.45/4.13)
|   |   |   |   |   |   PTC > 0.007194: Dialog (5.31/0.34)
|   |   |   lexDens > 0.514085
|   |   |   |   question <= 0.118103: News (28.76/1.55)
|   |   |   |   question > 0.118103: Chat (9.27/1.7)
|   |   mean_sent > 20.653846
|   |   |   question <= 0: Dialog (6.42/3.93)
|   |   |   question > 0
|   |   |   |   question <= 0.097561: Speech (205.1/2.91)
|   |   |   |   question > 0.097561: TED (5.0/0.34)
```

Figure 3: Weka decision tree for classifying registers. Class labels that have been assigned at the leaves are preceded by a colon. The first figure in parentheses states how many instances have been classified at this leaf (the figures do not correspond to actual instances but result from balancing the data, see Footnote 18). The second figure, after the slash, specifies how many instances were classified incorrectly, if any (because the data has missing attribute values, the algorithm used by Weka outputs fractional figures).

```
mean_sent <= 10.681818
|   PRON1st <= 0.008772
|   |   lexDens <= 0.486486: FEW/DIALOG (29.1/2.72)
|   |   lexDens > 0.486486
|   |   |   V.N <= 0.654762: MANY/MONOLOG (74.03/1.88)
|   |   |   V.N > 0.654762: FEW/DIALOG (16.43/1.36)
|   PRON1st > 0.008772
|   |   mean_sent <= 8.4: FEW/DIALOG (632.57/1.36)
|   |   mean_sent > 8.4
|   |   |   question <= 0.204724
|   |   |   |   V.N <= 0.969466
|   |   |   |   |   INTERJ <= 0.005882
|   |   |   |   |   |   med_word <= 4: MANY/MONOLOG (14.14/1.88)
|   |   |   |   |   |   med_word > 4: FEW/DIALOG (8.38/2.72)
|   |   |   |   |   INTERJ > 0.005882: FEW/DIALOG (15.07)
|   |   |   |   V.N > 0.969466: FEW/DIALOG (28.26)
|   |   |   question > 0.204724: FEW/DIALOG (101.75)
mean_sent > 10.681818
|   question <= 0.181024
|   |   INTERJ <= 0.004198: MANY/MONOLOG (744.95/3.77)
|   |   INTERJ > 0.004198
|   |   |   PRONsubj <= 0.473684: MANY/MONOLOG (20.42)
|   |   |   PRONsubj > 0.473684: FEW/DIALOG (5.13/1.36)
|   question > 0.181024
|   |   coordInit <= 0.086957
|   |   |   V.N <= 0.424528: MANY/MONOLOG (5.44)
|   |   |   V.N > 0.424528: FEW/DIALOG (37.16/1.36)
|   |   coordInit > 0.086957: MANY/MONOLOG (8.17)
```

Figure 4: Weka decision tree for classifying the situational characteristics participants or interactiveness. As the registers in the present study are either monologous with many participants or dialogous with few participants, the resulting decision trees for both properties are identical.

```
PRON1st <= 0.011905
|   question <= 0.193878
|   |   mean_sent <= 7.6
|   |   |   lexDens <= 0.501439: QUASI (5.31)
|   |   |   lexDens > 0.501439: ASYNC (12.58/2.13)
|   |   mean_sent > 7.6: ASYNC (499.84/5.31)
|   question > 0.193878
|   |   mean_sent <= 10.818182: QUASI (20.23/1.1)
|   |   mean_sent > 10.818182: ASYNC (5.46/1.06)
PRON1st > 0.011905
|   mean_word <= 5.072603
|   |   DEMshort <= 0.942308
|   |   |   PTC <= 0.034483: QUASI (77.66/2.2)
|   |   |   PTC > 0.034483: SYNC (11.48/3.19)
|   |   DEMshort > 0.942308
|   |   |   V.N <= 1.662281
|   |   |   |   mean_sent <= 9.375
|   |   |   |   |   coordInit <= 0
|   |   |   |   |   |   PRON1st <= 0.022489: SYNC (8.29)
|   |   |   |   |   |   PRON1st > 0.022489
|   |   |   |   |   |   |   PTC <= 0.045455: QUASI (15.94)
|   |   |   |   |   |   |   PTC > 0.045455: SYNC (5.21/1.06)
|   |   |   |   |   coordInit > 0
|   |   |   |   |   |   DEM <= 0.016575
|   |   |   |   |   |   |   question <= 0.263374: SYNC (38.51/5.35)
|   |   |   |   |   |   |   question > 0.263374: QUASI (8.5)
|   |   |   |   |   |   DEM > 0.016575: SYNC (489.24/4.25)
|   |   |   |   mean_sent > 9.375
|   |   |   |   |   PTC <= 0.0071
|   |   |   |   |   |   PRONsubj <= 0.52809
|   |   |   |   |   |   |   INTERJ <= 0.000617: ASYNC (5.46/1.06)
|   |   |   |   |   |   |   INTERJ > 0.000617: QUASI (5.35/1.1)
|   |   |   |   |   |   PRONsubj > 0.52809: QUASI (17.56/0.55)
|   |   |   |   |   PTC > 0.0071: SYNC (13.5/1.06)
|   |   |   V.N > 1.662281
|   |   |   |   med_sent <= 9.5: QUASI (24.45)
|   |   |   |   med_sent > 9.5: SYNC (5.21/1.06)
|   mean_word > 5.072603
|   |   PRONsubj <= 0.264706
|   |   |   PRON1st <= 0.020331: ASYNC (19.8)
|   |   |   PRON1st > 0.020331: QUASI (5.39/2.2)
|   |   PRONsubj > 0.264706
|   |   |   PTC <= 0.015456
|   |   |   |   nomCmplx <= 1.135593
|   |   |   |   |   mean_sent <= 7.429577
|   |   |   |   |   |   nomCmplx <= 1.03125: QUASI (72.83/0.55)
|   |   |   |   |   |   nomCmplx > 1.03125: SYNC (10.42/2.13)
|   |   |   |   |   mean_sent > 7.429577
|   |   |   |   |   |   question <= 0
|   |   |   |   |   |   |   lexDens <= 0.505458: QUASI (13.34/1.65)
|   |   |   |   |   |   |   lexDens > 0.505458: ASYNC (5.5)
|   |   |   |   |   |   question > 0: QUASI (260.38/14.85)
|   |   |   |   nomCmplx > 1.135593
|   |   |   |   |   med_sent <= 16
|   |   |   |   |   |   mean_word <= 5.38497: QUASI (18.07)
|   |   |   |   |   |   mean_word > 5.38497
|   |   |   |   |   |   |   DEMshort <= 0.844828: QUASI (5.86/0.55)
|   |   |   |   |   |   |   DEMshort > 0.844828: ASYNC (9.28/2.13)
|   |   |   |   |   med_sent > 16: ASYNC (8.25)
|   |   |   PTC > 0.015456
|   |   |   |   coordInit <= 0.069565
|   |   |   |   |   question <= 0.05: SYNC (5.21/1.06)
|   |   |   |   |   question > 0.05: QUASI (22.32)
|   |   |   |   coordInit > 0.069565: SYNC (14.56/2.13)
```

Figure 5: Weka decision tree for classifying production circumstances.

```
PRONsubj <= 0.232558
|   mean_sent <= 6.533333
|   |   med_word <= 4.5: QUASI (12.62)
|   |   med_word > 4.5: ASYNC (5.77/1.8)
|   mean_sent > 6.533333: ASYNC (467.55)
PRONsubj > 0.232558
|   mean_sent <= 11.309524
|   |   DEM <= 0.016575
|   |   |   question <= 0.054152
|   |   |   |   DEMshort <= 0.25
|   |   |   |   |   V.N <= 0.481481: ASYNC (9.63)
|   |   |   |   |   V.N > 0.481481: QUASI (24.8/3.17)
|   |   |   |   DEMshort > 0.25: SYNC (21.13/3.5)
|   |   |   question > 0.054152
|   |   |   |   mean_word <= 4.825082
|   |   |   |   |   V.N <= 1.360656
|   |   |   |   |   |   DEM <= 0.012663
|   |   |   |   |   |   |   question <= 0.266667
|   |   |   |   |   |   |   |   PRON1st <= 0.053691: SYNC (7.68/1.8)
|   |   |   |   |   |   |   |   PRON1st > 0.053691: QUASI (10.48/1.47)
|   |   |   |   |   |   |   question > 0.266667: QUASI (19.83)
|   |   |   |   |   |   DEM > 0.012663: SYNC (8.82)
|   |   |   |   |   V.N > 1.360656: QUASI (46.86)
|   |   |   |   mean_word > 4.825082
|   |   |   |   |   mean_sent <= 10.882353: QUASI (284.02/10.07)
|   |   |   |   |   mean_sent > 10.882353
|   |   |   |   |   |   V.N <= 0.726115: QUASI (13.75/1.13)
|   |   |   |   |   |   V.N > 0.726115: SYNC (5.88)
|   |   DEM > 0.016575
|   |   |   mean_word <= 5.072603
|   |   |   |   V.N <= 1.662281
|   |   |   |   |   coordInit <= 0: QUASI (29.97/2.94)
|   |   |   |   |   coordInit > 0: SYNC (194.7/9.58)
|   |   |   |   V.N > 1.662281: QUASI (27.03)
|   |   |   mean_word > 5.072603
|   |   |   |   mean_sent <= 10.3074: QUASI (101.06/7.35)
|   |   |   |   mean_sent > 10.3074: SYNC (6.21/1.8)
|   mean_sent > 11.309524
|   |   lexDens <= 0.511404
|   |   |   coordInit <= 0.003795
|   |   |   |   question <= 0.169811: ASYNC (9.97/1.47)
|   |   |   |   question > 0.169811: QUASI (5.41)
|   |   |   coordInit > 0.003795
|   |   |   |   question <= 0.189189
|   |   |   |   |   PRON1st <= 0.00905
|   |   |   |   |   |   mean_sent <= 23.821429: ASYNC (7.37)
|   |   |   |   |   |   mean_sent > 23.821429: SYNC (5.88)
|   |   |   |   |   PRON1st > 0.00905: SYNC (326.1/17.57)
|   |   |   |   question > 0.189189
|   |   |   |   |   lexDens <= 0.480859: SYNC (11.75)
|   |   |   |   |   lexDens > 0.480859: QUASI (5.97/0.57)
|   |   lexDens > 0.511404
|   |   |   exclam <= 0.047297
|   |   |   |   INTERJ <= 0.000251: ASYNC (44.54/1.47)
|   |   |   |   INTERJ > 0.000251
|   |   |   |   |   nomCmplx <= 1.109966: ASYNC (6.0/1.47)
|   |   |   |   |   nomCmplx > 1.109966: SYNC (5.54/1.13)
|   |   |   exclam > 0.047297
|   |   |   |   question <= 0.209459: ASYNC (7.47/1.8)
|   |   |   |   question > 0.209459: QUASI (7.21)
```

Figure 6: Weka decision tree for classifying reception circumstances.