Modeling Diachronic Change in Scientific Writing with Information Density

Raphael Rubino[†‡] **Stefania Degaetano-Ortlieb**[†] **Elke Teich**[†] **Josef van Genabith**[†‡]
[†]Universität des Saarlandes, [‡]DFKI
Saarbrücken, Germany
{s.degaetano,e.teich}@mx.uni-saarland.de
{raphael.rubino,josef.vangenabith}@uni-saarland.de

Abstract

Previous linguistic research on scientific writing has shown that language use in the scientific domain varies considerably in register and style over time. In this paper we investigate the introduction of information theory inspired features to study long term diachronic change on three levels: lexis, part-of-speech and syntax. Our approach is based on distinguishing between sentences from 19th and 20th century scientific abstracts using supervised classification models. To the best of our knowledge, the introduction of information theoretic features to this task is novel. We show that these features outperform more traditional features, such as token or character n-grams, while leading to more compact models. We present a detailed analysis of feature informativeness in order to gain a better understanding of diachronic change on different linguistic levels.

1 Introduction

Supervised classification has been applied to various natural language processing tasks over the past decades. To date, however, distinguishing between time periods has not received extensive attention. Early research on classifying time periods is presented in de Jong et al. (2005) for Dutch. Dalli and Wilks (2006) and Kumar et al. (2011) use word frequencies for temporal classification of documents, while Sagi et al. (2009) and Kim et al. (2014) predict semantic changes over time. While lexical features are commonly used for classification approaches of time periods, features based on more abstract linguistic levels have not yet been widely investigated.

In our study, we use supervised classification to distinguish scientific abstracts written in the 19th and 20th century at the sentence-level. From previous work, we know that in the scientific domain, shared expertise among authors and audience affects their language use. Over a longer time period, it drives the evolution of domain-specific language with respect to lexis (Halliday, 1988; Teich et al., 2016) and a more standardized and convention-driven style with respect to grammar (Biber and Gray, 2011; Biber and Gray, 2016; Banks, 2005).

Considering that language variation affects *all* linguistic levels — from sounds and words to syntactic structure — we investigate a set of features extracted at the lexis, part-of-speech and syntactic levels to test how well they act as predictors of time period-specific language use. Moreover, based on psycholinguistic evidence it has been shown that language users choose those linguistic options that they know to be relatively predictable in a specific context to optimize communication (Hale, 2001; Levy, 2008; Demberg and Keller, 2008). To model communication in this sense, in our research we employ features based on the information-theoretic notion of *surprisal* or *information density*.

Specifically, we make use of information theory inspired features on the linguistic levels of lexis, part-of-speech and syntax. In addition, these features allow an unlexicalized dense-vector representation, which enormously reduces the amount of features used for classification. Besides achieving high performance in classification, we are particularly interested in insights on long-term diachronic linguistic change, which are important to historical linguistics, sociolinguistics and the like. We do this by inspecting classification results and discriminative features more closely.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 750–761, Osaka, Japan, December 11-17 2016.

Our analyses are driven by the following assumptions:

1. *Lexical diversification*: On the lexical level, scientific abstracts from the 19th and 20th century will be well distinguished from one another, due to topical changes in the scientific domain.
2. *Grammatical consolidation*: On more abstract linguistic levels, scientific abstracts from the 19th and 20th centuries will be less well distinguished from one another, as grammatical changes develop rather slowly over time, but we expect a tendency towards denser grammatical encodings in the 20th century texts.

The remainder of the paper is structured as follows. In Section 2, we present previous work on classification of time periods, diachronic change and information density. Section 3 describes the experimental setup up followed by the results and detailed analysis in Section 4. Finally, Section 5 provides a short summary and conclusions.

2 Related work

2.1 Classification of time periods

Classification of time periods has been less investigated so far in comparison to other classification tasks. Most of the existing work is based on lexical features and the classification of documents rather than individual sentences. In the study conducted by de Jong et al. (2005), the authors classify Dutch texts according to time (considering the time span 1999 to 2005) using uni-gram language models achieving around 65% accuracy. Dalli and Wilks (2006) (considering weekly to yearly levels, with an accuracy of the yearly classifier of $\sim$88%) and Kumar et al. (2011) (yearly classification) use classification methods based on word frequencies to determine the time period a text was written.

Other approaches – also based on lexical features – investigate semantic change over time. For instance, Sagi et al. (2009) focus on specific words to identify their semantic change from Early to Modern English. Mihalcea and Nastase (2012) use supervised learning to predict a word's time period given the context it occurs in. More recently, Kim et al. (2014) use neural language models to identify words that have changed semantically from 1900 to 2009. So far, only few studies have used features other than lexical ones for time period classification. Štajner and Zampieri (2013) have used stylistic features (such as average word and sentence length, pos tag n-grams, etc.) for classification of Portuguese texts into centuries achieving an F-measure of 0.92.

Besides the fact that most approaches use lexical features to predict time periods, the common experimental setup involves document-level classification of texts. To the best of our knowledge, there has been no work on sentence-based classification of time periods. Classifying sentences rather than texts allows us to build finer-grained classification models. In our approach, we classify sentences according to time periods going beyond lexis-based representations by using information theory inspired features, which inherently account for the context of use.

2.2 Information Density (ID)

Assuming that language users strive for efficient communication, they will tend to encode their message using an approximately uniform information density that exploits channel capacity while avoiding to overload the recipient or being uninformative. Information theory (Shannon, 1949) measures the amount of information conveyed by a unit in a given context in *bits* (Shannon, 1949). This notion is also known as *surprisal* (Levy, 2008) and is formulated as the negative log probability of a unit (e.g. a word) in context (e.g. its preceding words): $S(unit_i) = -\log p(unit_i|Context)$. Based on a limited context of size n words, the surprisal value of the following word w_{n+1} corresponds to the negative log-probability: $S(w_{n+1}) = -\log P(w_{n+1}|w_1 \dots w_n)$.

There are two properties inherent to surprisal: (1) units with low probability convey more information than those with high probability, and (2) information conveyed by a unit is crucially dependent on its context. Thus, linguistic units that are highly predictable in a given context convey less information with troughs in surprisal, while less predictable units convey more information with peaks in surprisal.

Over a longer period of time, the predictability of a word will change according to its use in specific contexts. In the scientific domain, shared expertise among researchers, for example, will affect language

use and give rise to domain-specific language. Particular words (e.g. terminology) will become more predictable over time (showing lower surprisal values) and may result in shorter encodings (consider e.g. acronym use in scientific fields such as genetics). Among researchers this will optimize communication. A more conventionalized use of scientific language will result in changes of surprisal values over time with conventionalized expressions (e.g. formulaic expressions) showing lower surprisal.

However, not only changes in lexis will be reflected in changes of surprisal values. From studies on language change, we know that diachronically there has been, for example, a shift from a more verbal towards a more nominal style (cf. notably Biber and Gray (2011)). This will have an impact on surprisal values with respect to grammatical units (such as parts of speech or syntactic units), motivating the use of information theory inspired features to classify between time periods.

So far, these kinds of features have been successfully used in classification of Gospels (see Islam and Dundia (2015) being able to identify the Greek Gospel as the original text and the American and Georgian ones as translations) and classification of human translated texts (see Rubino et al. (2016) distinguishing original from manually translated texts of different levels of expertise).

2.3 Language Change

Previous computational work on diachronic change in scientific language mostly discusses short-term change (see e.g. Blei and Lafferty (2006; 2007) on changes in scientific topics and Hall et al. (2008) on the ACL anthology corpus, both using topic models) rather than long-term change and is mostly concerned with change related to lexis (such as topical shifts) rather than change on more abstract linguistic levels.

In corpus-linguistic work on language change, approaches are typically frequency-based (e.g. Biber and Gray (2011; Biber and Gray (2013; Biber and Gray (2016), Taavitsainen and Pahta (2012), Moskowich and Crespo (2012)) and do not inherently account for context – diachronic change being observed through the lens of unconditioned probabilities. In contrast, information density measures as we apply them here, are based on conditional probabilities and thus inherently take context into account. Based on our previous work on long-term change using information-theoretic features (Degaetano-Ortlieb and Teich, 2016), we have shown how these features help model diachronic change, further motivating their use to classify different time periods.

3 Experimental Setup

The experiments presented in this paper focus on the use of sentence-level information density measures — in particular n-gram log-probabilities according to a language model and n-gram distribution according to frequency quartiles — to classify texts from different time periods. In this section, we present the supervised classification setup and the set of features as well as the data used.

3.1 Supervised Classification

A linear Support Vector Machine (SVM) (Cortes and Vapnik, 1995) is used to train our time-period classification model based on a feature representation of sentences which aims at capturing the density of information. All training, development and test sentences are represented as feature vectors $\mathbf{x}_i$, and the two corpora (19th century and 20th century) are associated with a class y_i, resulting in instance-label pairs $(\mathbf{x}_i, y_i)$ with $\mathbf{x}_i \in R^n$, and $y \in \{0,1\}^l$ as a binary classification task. We use the $L2$-regularized $L2$-loss SVC implementation of LIBLINEAR (Fan et al., 2008) to solve the following optimization problem:

$$min_w w^T \frac{w}{2} + C \sum_{i=1}^{l} max(0, 1 - y_i w^T x_i)^2 \quad (1)$$

The cost parameter C is selected with grid-search using the accuracy obtained on the held-out development set. Finally, the model is evaluated using the precision, recall, f-measure per class and general accuracy obtained on the test set.

Corpus	Sentences		Tokens		Types	
	19cA	20cA	19cA	20cA	19cA	20cA
Train	20.0	20.0	890.5	495.2	28.8	31.5
Development	1.5	1.5	66.3	36.7	7.7	7.2
Test	1.5	1.5	64.2	37.0	7.7	7.3

(a) Corpora used as training, development and test sets.

Corpus	Sentences	Tokens	Types
19cLM	623.2	16,541.7	320.0
20cLM	423.4	11,137.8	261.9

(b) Corpora used as resources to train the language models and to extract n-gram frequencies.

Table 1: Statistics (in thousands) of the corpora used in our experiments.

3.2 Datasets

Four corpora are used in our experiments, two for each time period (early 19th century: 1800-1850; late 20th century: 1970-2007). Two corpora compose our training, development and test sets (henceforth: 19cA and 20cA) while two others allow us to train language models and extract n-gram frequencies (henceforth 19cLM and 20cLM). Statistics about these corpora are presented in Table 1a and Table 1b.

For the 19th century time period, we use a corpus of research articles from the Royal Society of London (Kermes et al., 2016). Abstracts are taken from this corpus to form the 19cA classification subset. For feature extraction full research articles (19cLM) are taken from the same corpus, filtering out articles with abstracts included in 19cA. For the 20th century time period, abstracts are taken from a corpus of research articles (Degaetano-Ortlieb et al., 2013) covering several disciplines[1] as our 20cA classification subset. For feature extraction, we collected abstracts from several fields (20cLM) matching those of 20cA. The main difference between 19cLM and 20cLM is the type of document used to extract them, the former being composed of full articles due to research abstract scarcity for this time period, while the latter is composed of abstracts. The classification subsets (19cA and 20cA) are pre-processed by means of regular expressions and manually verified in order to remove headlines preceding abstracts, dates, formulas and mathematical expressions, etc.

3.3 Feature Sets

We consider three sets of features: shallow base-line features, n-gram frequency features, and information density features. Both n-gram and features specifically referred to as information density features capture aspects of information density and rely on the external resources presented in Table 1b.

Shallow features Here we consider popular lexical features such as bags of character and token n-grams as a baseline, as well as bags of part-of-speech (POS) n-grams ($n \in [1; 3]$). For POS tagging and syntactic parsing, we use the Stanford NLP toolkit (Manning et al., 2014).[2] For bags of token n-grams, three feature sets are built: one taking into account all n-grams, one considering n-grams appearing at least 200 times in the training corpus and one keeping only n-grams appearing at least 500 times, noted *Tokens All*, *Tokens 200* and *Tokens 500* respectively. The two latter sets allow for more compact models and less sparsity in the feature vectors. Additionally, 13 surface features are used, extracted from the surface-level of each sentence, which aim to capture meta representations of sentences' lexical form including sentence and average word lengths, the number of punctuation marks, letter and word casing, binary values encoding whether the sentence ends with a period and starts with an uppercase letter, etc.

N-gram Frequency Features To capture the rarity of n-grams used in the sentences to classify, the percentage of n-grams in frequency quartiles are extracted ($n \in [1; 5]$). The corpora used to model the frequency quartiles are the same resources as the ones used for the language models (19cLM and

[1] computer science, computational linguistics, bioinformatics, computer-aided design, microelectronics, mechanical engineering, electrical engineering, biology, linguistics

[2] We use the bidirectional maximum entropy POS tagger with a pre-trained English model based on the WSJ sections 0-18, including word shape and distributional similarity features. The probabilistic context free grammar lexicalized parser is used to obtain syntactic information from text (Manning et al., 2014).

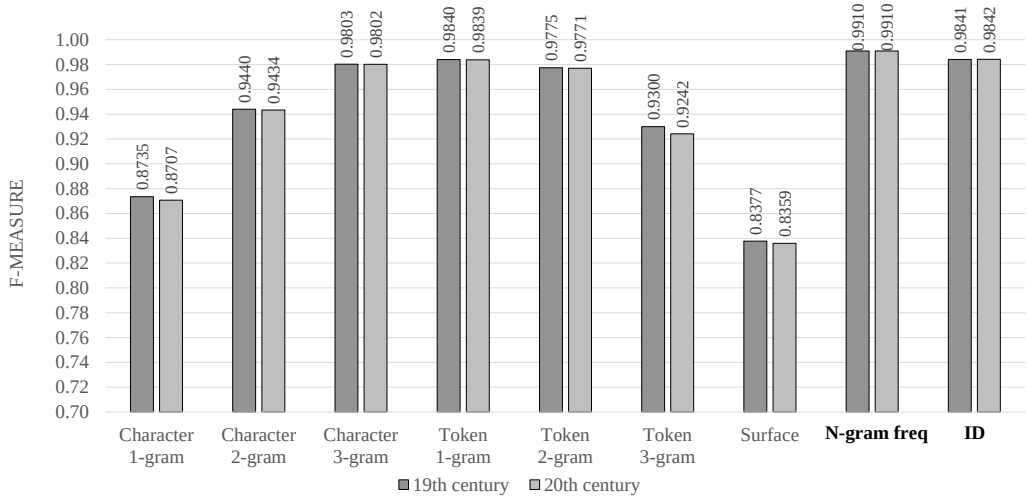

Figure 1: Classification results of 19th century and 20th century abstracts for lexis (LEX).

20cLM). Frequencies of word, part-of-speech and delexicalized flattened syntactic sequences are averaged at the sentence level, leading to 4 features per sentence (one per quartile) given one value of n, for each of the external resource used to model the quartiles (the corpora 19cLM and 20cLM). This approach leads to a dense representation of the information encoded in a sentence based on lexical, POS and syntactic information, without encoding raw word sequence n-gram features.

Information Density Features Using language models trained on sentences, delexicalized part-of-speech sequences and delexicalized flattened syntactic trees, a set of 120 sentence-level features are extracted: 15 features per individual LM resource (presented in Table 1b) and type of language model (lexical, POS and syntactic). We extract n-gram ($n \in [1; 5]$) log-probabilities (surprisal) as well as perplexities, with and without the tags indicating the beginning and ending of sentences, using the SRILM toolkit (Stolcke et al., 2011).

4 Results and Analysis

In the following, we present classification results of 19th vs. 20th century abstracts based on shallow as well as n-gram and information density features on three linguistic levels: lexis (LEX), part-of-speech (POS), and syntax (SYN). Moreover, by considering feature rankings obtained by the classification results, we analyze diachronic changes on these three linguistic levels.

4.1 Classification Results

Classification results on the lexical level (LEX) are shown in Figure 1. The bags of token n-grams features are unpruned (*Tokens All*). The best performing features are n-gram frequency (F-measure of 0.991) and ID features ranking second (0.984), both outperforming shallow features (bags of character and token n-grams, and surface features). Considering classification results on the part-of-speech level (POS), Figure 2 shows that POS 3-grams work best (0.9224) in classifying 19th c. and 20th c. abstracts, followed by POS 2-grams (0.9222) and ID features (0.9112).

Regarding classification at the syntactic level (SYN), Figure 3 shows that 19th c. and 20th c. abstracts are less well distinguished from one another in comparison to the lexical and part-of-speech level. Nevertheless, ID features work best on this task, achieving an F-measure of 0.88. Overall, ID features work relatively well on all three linguistic levels targeted in this study in comparison to other features, which work well on some levels (e.g. n-gram frequencies for lexis or POS 3-grams on the part-of-speech level), but less well on the other linguistic levels.

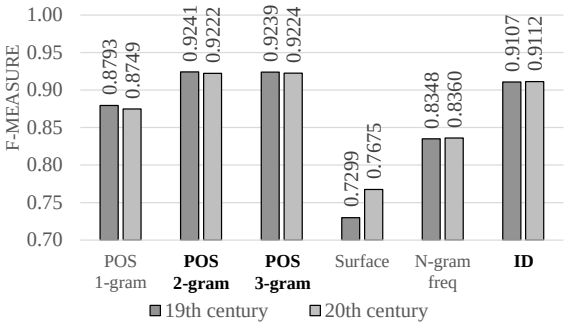

Figure 2: Classification results of 19th century and 20th century abstracts for part-of-speech (POS).

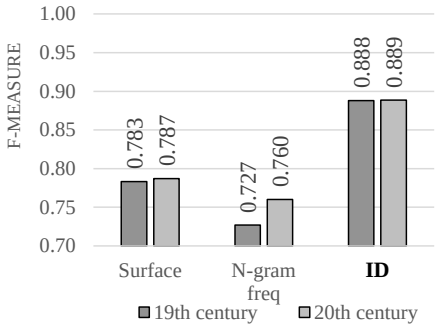

Figure 3: Classification results of 19th century and 20th century abstracts for syntax (SYN).

By considering combinations of feature sets, the best classification performance of 99.18 accuracy is obtained by a combination of n-gram frequency and ID features at all three linguistic levels (LEX, POS and SYN) (see Table 2). Moreover, looking at the number of features used for classification, better results can be obtained with a very small number of features (216) using n-gram frequency and ID features in comparison to the high number of features when using shallow features (e.g. more than 1 million features for surface tokens; see again Table 2). Pruning the low frequency n-grams considered in the bags of tokens features does not lead to accuracy improvement, but the resulting models are more compact with $3,081$ and 996 features for the *Tokens 200* and *Tokens 500* sets respectively.

			19cA			20cA		
Feature set	**Number of features**	**Accuracy**	**P**	**R**	**F**	**P**	**R**	**F**
Characters (LEX)	37,367	98.05	98.14	98.86	98.50	98.85	98.13	98.49
Tokens All (LEX)	1,896,263	98.12	97.43	98.84	98.13	98.82	97.39	98.10
Tokens 200 (LEX)	3,081	94.93	93.94	96.05	94.98	95.96	93.80	94.87
Tokens 500 (LEX)	996	90.38	89.68	91.26	90.46	91.10	89.50	90.29
POS-Tags (POS)	14,227	93.18	93.20	91.73	92.46	91.86	93.31	92.58
n-gram freq. + ID (LEX)	72	99.10	98.98	99.24	99.11	99.24	98.97	99.11
n-gram freq. + ID (POS)	72	88.85	93.20	91.73	92.46	91.86	93.31	92.58
n-gram freq. + ID (SYN)	72	88.67	91.38	89.10	90.22	89.37	91.59	90.47
n-gram freq. + ID (LEX, POS, SYN)	216	**99.18**	99.04	99.31	99.18	99.31	99.04	99.17

Table 2: Feature sets used in our experiments, number of features per set, accuracy obtained on the test set, as well as per class Precision (P), Recall (R) and F-measure (F).

4.2 Diachronic Changes at the Lexical Level

To investigate changes between 19th and 20th c. abstracts and evaluate the performance of the different feature types, we conduct a non-linear feature selection using the forest of randomized trees approach (Geurts et al., 2006) and describe the results for the top n-gram frequency and ID features in the paragraphs below. These two types of features are the focus of our study and lead to the best classification results as shown in Figure 1.

Lexis and N-gram frequency Inspecting the n-gram frequencies in detail based on feature ranking, we can observe general tendencies in lexis related change across the 19th and the 20th century. The highest ranking n-gram frequency features are based on 20cLM, comprising among the top 10 features 1- to 4-grams of very high (quartile 4) and low (quartile 1) frequency. This might be an indicator of conventionalized/formulaic language use with respect to high frequency high-order n-grams (4-grams quartile 4), on the one hand, and diversified language use with respect to low frequency low-order n-grams (1-grams quartile 1), on the other. Figure 4 shows the n-gram frequency distribution for 1- to

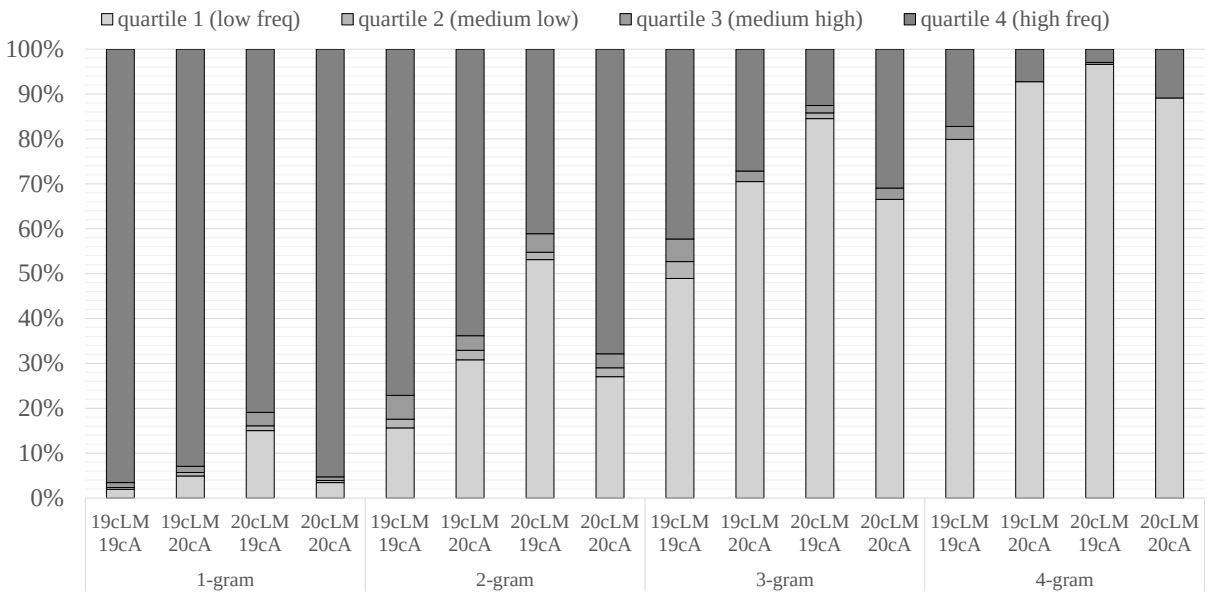

Figure 4: N-gram frequency distribution for 1- to 4-grams based on four frequency quartiles. 19cLM: 19th c. LM resource, 19cA: 19th c. abstracts, 20cLM: 20th c. LM resource, 20cA: 20th c. abstracts.

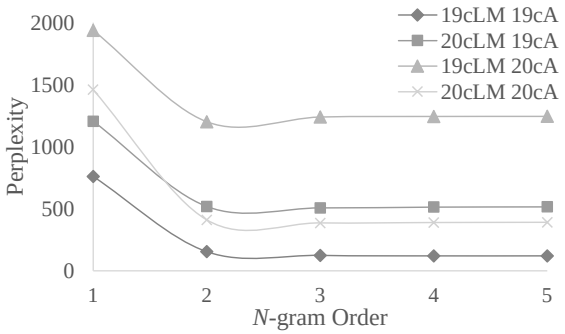

Figure 5: Averaged perplexity values obtained for 1- to 5-grams (LEX).

4-grams based on four quartiles (from high to low frequency) and on the resources used for language modeling (19cLM and 20cLM). Obviously, the higher the n-gram order, the higher the percentage of low frequency n-grams, i.e. rare n-grams. More specifically, Figure 4 shows that the 19cLM covers the 19th c. abstracts (19cA) quite well in terms of lexis, as the percentage of high frequency 1-grams (quartile 4) is high ($\sim$97%). The same can be observed for the 20cLM and the 20th c. abstracts (20cA) ($\sim$95%). However, while the 20cLM also covers relatively well the 19cA ($\sim$93% of high frequency 1-grams, quartile 4), the 19cLM on the 20cA shows a higher amount of low frequency 1-grams (quartile 1) covering high frequency 1-grams only by $\sim$80%. This indicates that while the vocabulary of 19th c. abstracts is relatively well covered by the 20th c. LM resource, the 20th c. abstracts make use of new words not covered by the 19th c. LM resource.

Considering 2-grams, which rank highest in classification, a similar but even more pronounced tendency can be observed, i.e. while the percentage of high frequency 2-grams is still relatively high for 20cLM and 19cA ($\sim$64%), 19cLM and 20cA show a relatively high amount of low frequency 2-grams ($\sim$53%). Nevertheless, the percentage of 20cLM and 19cA high frequency (quartile 4) n-grams remains higher than for 19cLM and 20cA. Thus, 20th c. abstracts have more diverse lexical n-grams than those written in the 19th c.

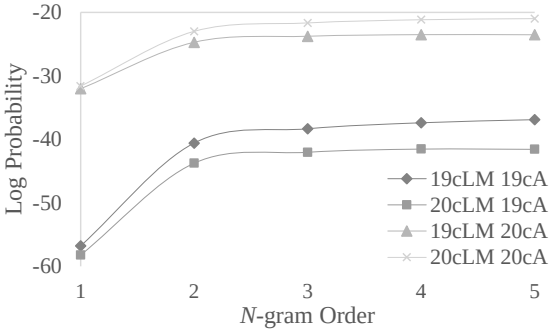

Figure 6: Averaged log probability values obtained for 1- to 5-grams (POS).

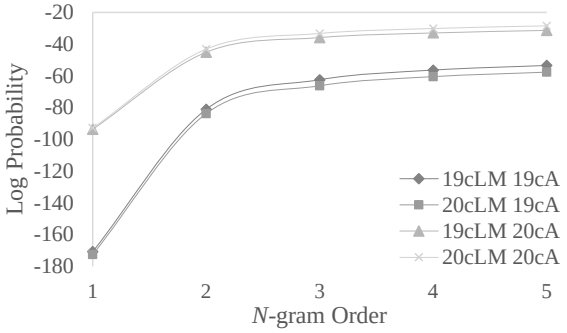

Figure 7: Averaged log probability values obtained for 1- to 5-grams (SYN).

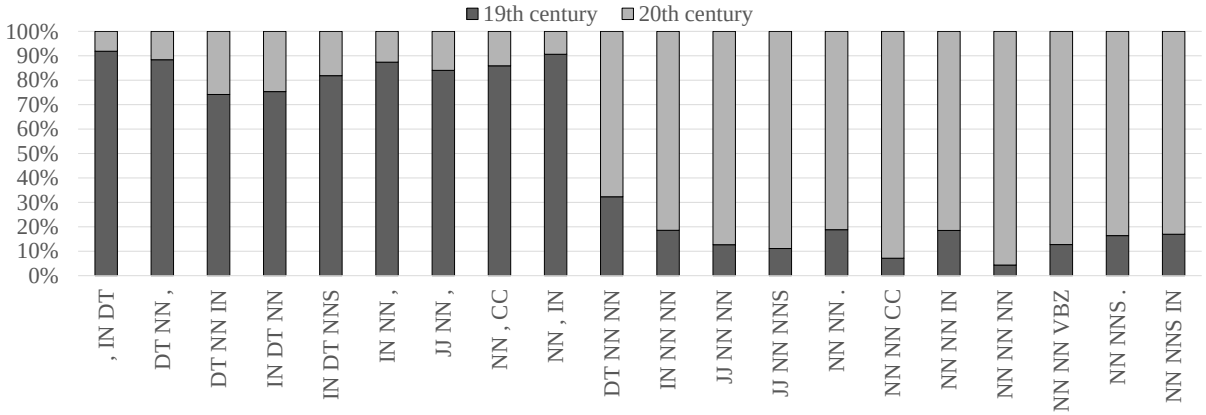

Figure 8: Distribution of top 20 POS 3-grams for the 19th c. and 20th c. abstracts. DT: determiner, CC: conjunction, JJ: adjective, IN: preposition, NN: singular common noun, NNS: plural common noun, VBZ: verb *be* present

In addition, the amount of high frequency (quartile 4) n-grams of the 19cLM and 19cA is higher than the ones for the 20cLM and 20cA. This might indicate a process of diversification in scientific writing, i.e. 19th c. texts in science are lexically closer than texts written in the 20th c.

Lexis and ID Feature ranking shows that perplexity values from 2- to 5-grams are the most discriminative ID features. Inspecting the perplexity values more closely (see Figure 5), we observe that from 2- to 5-grams 19cLM has relatively low average perplexity values for 19cA compared to the ones obtained for 20cA. While 19cLM is relatively close to 19cA, i.e. the abstracts' lexis is relatively predictable and obtains low perplexities according to the language model trained on 19cLM, lexis of 20th c. abstracts is less well predictable. This observation matches the results obtained with n-gram frequencies presented in Figure 4.

Considering 20cLM (see again Figure 5), it shows lower perplexity values for 20cA than for 19cA. However, the difference is relatively small in comparison to the difference observed for 19cLM on 19th and 20th c. abstracts. Thus, 20cLM is better in predicting lexical choices in both 19th c. and 20th c. abstracts compared to 19cLM. In terms of diachronic changes, this reflects how new lexical choices have entered scientific language, which were not present in the 19th century, while 19th c. language can still be understood by a contemporary language model. These results support the assumption of lexical diversification over time.

4.3 Diachronic Changes at More Abstract Linguistic Levels

	POS 3-gram	examples
19th century	, IN DT	*, that the*, *, in the*, *, on the*
	IN DT NN	*by the author*, *in this paper*, *of the earth*
	NN , CC	*water , and*, *acid , and*, *light , and*
	DT NN ,	*the author, , this paper, , the earth ,*
	DT NN IN	*the action of*, *the quantity of*, *the surface of*
20th century	JJ NN NN	*superficial gas velocity*, *natural language processing*, *optimal control problem*
	DT NN NN	*a computer program*, *the heat transfer*, *the nucleotide sequence*
	NN NN IN	*nucleotide sequence of*, *sequence analysis of*, *gene expression in*
	JJ NN NNS	*partial differential equations*, *open reading frames*, *linear matrix inequalities*
	NN NN .	*gene expression .*, *power consumption .*, *control system .*

Table 3: Most frequent lexical realizations of top 5 POS 3-grams for 20th c. and 19th c. abstracts

POS Sequences To investigate diachronic changes at the POS level, we consider POS 3-gram sequences, which perform best in POS-based classification (see again Figure 2). We inspect the top 20 features of the POS 3-gram sequences obtained by feature ranking and look at their frequency distribution in the training data of the 19th and 20th c. abstracts. Figure 8 shows how complex nominal structures (consisting of compounds with a at least two nouns, e.g. DT NN NN such as *the heat transfer*) are discriminative for the 20th c. abstracts, while shorter nominal structures (consisting of POS sequences with one noun, e.g. followed by a comma (DT NN , such as *the heat,*)) and prepositional phrases (e.g. IN DT NN such as *on the eye*) are discriminative for the 19th c. abstracts. This clearly reflects a shift towards a denser linguistic encoding in 20th c. abstracts, where information is more densely packed into longer nominal structures. Table 3 shows the top 5 POS 3-gram sequences of both periods with examples. We can see from the examples that while in the 19th c. there are relatively general and short nouns (such as *author*, *water*, *action*), in the 20th century more specific compound nouns are used. Inspecting these examples in their sentential context confirms the use of quite complex nominal phrases in abstracts of the 20th century (see examples (1) and (2)) vs. shorter, less complex ones in the 19th century (see examples (3) and (4)).

(1) *We have determined* ***the complete DNA nucleotide sequence of the carp Cyprinus carpio fast skeletal myosin heavy chain (MYH) gene***. (20th century)

(2) ***Nitric oxide generation rate and concentration distribution in combustors containing regions of recirculating flow*** *are calculated using* ***a computer program developed for two-dimensional elliptic compressible flows***. (20th century)

(3) *Those who cultivate* ***chemistry*** *with* ***any degree of ardour****, will be gratified to see in* ***this paper the pains*** *taken by* ***the author****, and* ***the various modes*** *he has devised, to produce* ***this compound metal*** *in its most perfect state of combination.* (19th century)

(4) ***The substance*** *here examined by* ***the author****, we are told, was first made known by* ***the celebrated Klaproth***. (19th century)

POS and ID We also inspect ID features as they achieve an F-measure above 0.90 (see Figure 2), indicating that 19th c. and 20th c. abstracts differ with respect to ID. The top three features refer to the log probabilities of 3- to 5-grams. In Figure 6, for both time periods the log probabilities increase with higher n-gram order, indicating lower surprisal values for POS n-grams of higher order. However, 19th c. abstracts differ from 20th c. abstracts as the log probabilities for the earlier period are lower (wrt both LM resources) than the ones for the 20th c. This indicates that POS sequences of 20th c. abstracts are

more predictable than those of the 19th c. abstracts, thus pointing to a more conventionalized use of POS sequences in 20th c. abstracts. These findings support our hypothesis of grammatical consolidation over time.

Syntax and ID The top three features on the syntactic level are again log probabilities of 3- to 5-grams. By inspecting the log probability distribution (see Figure 7), we see a very similar tendency to the results on POS. Thus, for both time periods the log probabilities increase with higher n-gram order, i.e. syntactic n-grams of higher order can be better predicted, indicating also on the syntactic level a more conventionalized use in 20th c. abstracts, which supports again our hypothesis of grammatical consolidation.

5 Conclusion

We have presented a sentence-based classification approach of time periods based on information theory inspired features. Our classification task focused on distinguishing 19th century and 20th century research abstracts. For this, we used features at three linguistic levels: lexis, part of speech, and flattened syntactic structure. This allows us to model not only lexical but also grammatical/stylistic long-term change in scientific writing.

Regarding classification, we show that while shallow features such as character and token n-grams achieve good results at the lexical level, applying features based on information density measures (log probability, perplexity) – achieves similar results and even outperforms shallow features at different linguistic levels. Furthermore, the best classification results were obtained by a combination of information density features considering all three linguistic levels with only a minimum number of features as we use unlexicalised dense feature-vector representations.

By a deeper analysis of the classification results, we obtained insights on long-term diachronic change with respect to our assumptions of *lexical diversification* and *grammatical consolidation*. Considering lexical diversification, new lexical choices have entered scientific writing from 19th to 20th century. Considering grammatical consolidation, a trend towards a denser linguistic encoding in terms of compact nominal structures was observed. Beyond lexical variation, we assume the methodology to be domain- and language independent. This will be pursued in future work with application on other genres/registers and languages.

Acknowledgments

This research is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding[3] and EXC-MMCI: Multimodal Computing and Interaction[4]. We would like to thank the anonymous reviewers for their insightful comments.

References

David Banks. 2005. On the Historical Origins of Nominalized Process in Scientific Text. In *English for Specific Purposes 24 (3), 347-357*.

Douglas Biber and Bethany Gray. 2011. The Historical Shift of Scientific Academic Prose in English towards Less Explicit Styles of Expression: Writing without Verbs. In Vijay Bathia, Purificación Sánchez, and Pascual Pérez-Paredes, editors, *Researching Specialized Languages*, pages 11–24. John Benjamins.

Douglas Biber and Bethany Gray. 2013. Being Specific about Historical Change: The Influence of Sub-register. *Journal of English Linguistics*, 41:104–134.

Douglas Biber and Bethany Gray, editors. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Cambridge University Press.

David M. Blei and John D. Lafferty. 2006. Dynamic Topic Models. In *Proceedings of ICML*, pages 113–120.

[3] IDEAL – http://www.sfb1102.uni-saarland.de/

[4] http://www.mmci.uni-saarland.de

David M. Blei and John D. Lafferty. 2007. A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1(1):17–35.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector Networks. *Machine learning*, 20(3):273–297.

Angelo Dalli and Yorick Wilks. 2006. Automatic Dating of Documents and Temporal Text Classification. In *Proceedings of the ACL Workshop on Annotating and Reasoning about Time and Events*, pages 17–22.

Franciska de Jong, Henning Rode, and Djoerd Hiemstra. 2005. Temporal Language Models for the Disclosure of Historical Text. In *Humanities, Computers and Cultural Heritage: Proceedings of the XVIth International Conference of the Association for History and Computing (AHC 2005)*, pages 161–168.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 165–173.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ekaterina Lapshinova-Koltunski, and Elke Teich. 2013. SciTex - A Diachronic Corpus for Analyzing the Development of Scientific Registers. In Paul Bennett, Martin Durrell, Silke Scheible, and Richard J. Whitt, editors, *New Methods in Historical Corpus Linguistics*, volume 3 of *Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*, pages 93–104. Narr.

Vera Demberg and Frank Keller. 2008. Data from Eye-tracking Corpora as Evidence for Theories of Syntactic Processing Complexity. *Cognition*, 109(2):193–210.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *The Journal of Machine Learning Research*, 9:1871–1874.

Pierre Geurts, Damien Ernst, and Louis Wehenkel. 2006. Extremely Randomized Trees. *Machine learning*, 63(1):3–42.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of NAACL*, volume 2, pages 159–166.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of EMLNP*, pages 363–371.

M.A.K. Halliday. 1988. On the Language of Physical Science. In Mohsen Ghadessy, editor, *Registers of Written English: Situational Factors and Linguistic Features*, pages 162–177. Pinter.

Zahurul Islam and Natia Dundia. 2015. Finding the Origin of a Translated Historical Document. In *Proceedings of PACLIC*, pages 96–105.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of LREC*, pages 1928–1931.

Yoon Kim, Yi-I Chiu, Kentaro Hanaki, Darshan Hegde, and Slav Petrov. 2014. Temporal Analysis of Language through Neural Language Models. In *Proceedings of the ACL Workshop on Language Technologies and Computational Social Science*, pages 61–65.

Abhimanu Kumar, Matthew Lease, and Jason Baldridge. 2011. Supervised Language Modeling for Temporal Resolution of Texts. In *Proceedings of CIKM*, pages 2069–2072.

Roger Levy. 2008. A Noisy-channel Model of Rational Human Sentence Comprehension under Uncertain Input. In *Proceedings of EMNLP*, pages 234–243.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of ACL: System Demonstrations*, pages 55–60.

Rada Mihalcea and Vivi Nastase. 2012. Word Epoch Disambiguation: Finding How Words Change Over Time. In *Proceedings of ACL*, pages 259–263.

Isabel Moskowich and Begoña Crespo, editors. 2012. *Astronomy 'playne and simple'. The Writing of Science between 1700 and 1900*. John Benjamins.

Raphael Rubino, Ekaterina Lapshinova-Koltunski, and Josef van Genabith. 2016. Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification. In *Proceedings of NAACL*, pages 960–970.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic Density Analysis: Comparing Word Meaning across Time and Phonetic Space. In *Proceedings of the EACL Workshop on GEMS: GEometical Models of Natural Language Semantics*, pages 104–111.

Claude E. Shannon. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, 1983 edition.

Sanja Štajner and Marcos Zampieri. 2013. Stylistic Changes for Temporal Text Classification. In *Proceedings of the International Conference on Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence (8082)*, pages 519–526. Springer.

Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, volume 5.

Irma Taavitsainen and Päivi Pahta, editors. 2012. *Early Modern English Medical Texts. Corpus Description and Studies*. John Benjamins.

Elke Teich, Stefania Degaetano-Ortlieb, Peter Fankhauser, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2016. The Linguistic Construal of Disciplinarity: A Data Mining Approach Using Register Features. *Journal of the Association for Information Science and Technology (JASIST)*, 67(7):1668–1678.