Modeling intra-textual variation with entropy and surprisal: topical vs. stylistic patterns

Stefania Degaetano-Ortlieb
Saarland University
Campus A2.2
66123 Saarbrücken
s.degaetano@mx.uni-saarland.de

Elke Teich
Saarland University
Campus A2.2
66123 Saarbrücken
e.teich@mx.uni-saarland.de

Abstract

We present a data-driven approach to investigate intra-textual variation by combining entropy and surprisal. With this approach we detect linguistic variation based on phrasal lexico-grammatical patterns across sections of research articles. Entropy is used to detect patterns typical of specific sections. Surprisal is used to differentiate between more and less informationally-loaded patterns as well as types of information (topical vs. stylistic). While we here focus on research articles in biology/genetics, the methodology is especially interesting for digital humanities scholars, as it can be applied to any text type or domain and combined with additional variables (e.g. time, author or social group) to obtain insights on intra-textual variation.

1 Introduction

While there is an abundance of studies on linguistic variation according to domain, register and genre, text-internal variation, i.e. variation based on changing micro-purposes within a text (Biber and Finegan, 1994), has received much less attention. As such internal shifts occur in all kinds of discourse — be it in spoken (such as spontaneous conversation or speeches) or written mode (such as literary texts, written editorials, research articles) — there has been recently a growing interest in this type of variation. In general, knowledge on intra-textual variation leads to a more comprehensive understanding of the data underlying computational modeling, analysis, interpretation, etc.

In the field of NLP, there is a growing need in the development of applications that consider variation also at the textual level to improve performance. Considering research articles, approaches within BioNLP, for instance, have moved from focusing on abstracts as sources of text mining to using also full-text articles (Cohen et al., 2010), not least because this data is made available through repositories such as PubMedCentral (PMC). To obtain good performance, corpora created from such resources are highly annotated with linguistic as well as semantic categories characterizing e.g. gene names. From these, specific features are selected with a trade-off between ease of extraction and desired type of information. In the field of DH, intra-textual variation is considered especially in literary studies, computational stylistics, and authorship attribution. Hoover (2016) shows, for example, how knowledge about differences between text parts helps to improve computational stylistic approaches. In corpus linguistics, the common approach to intra-textual variation is to start with a set of pre-defined linguistic features (Biber and Finegan, 1994). While the choice of features is clearly linguistically informed, this initial step in analysis is manual and needs to be carried out anew for every new text type or register considered. Also, analysis is restricted to frequency (i.e. unconditioned probabilities).

We present a methodology for investigating intra-textual variation that is data-driven and based on conditional probabilities which are calculated using two information-theoretic measures, entropy and surprisal. Being data-driven, our approach can be applied to any text type or domain, avoiding extensive annotations and manual selection of features possibly involved in variation. Based on probabilities conditioned on ambient and extra-linguistic context, it allows to capture variation in a more fine-grained manner than by considering mere frequencies.

As a testbed for our approach, we use scientific research articles in genetics, as they clearly

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. Proceedings, pages 68–77, Vancouver, BC, August 4, 2017. ©2017 Association for Computational Linguistics

exhibit the typical IMRaD (Introduction, Methods, Results and Discussion) structure of scientific articles, with internal shifts in purpose (see e.g. Swales (1990)).

We use *relative entropy* (Kullback-Leibler Divergence) to detect features typical of specific sections. By considering *surprisal* (i.e. probabilities of features in their ambient context), we are able to detect the amount and type of information these typical features convey, e.g. more informationally-loaded expressions (e.g. terminology) vs. less informationally-loaded expressions (e.g. linguistic formula, such as *These results show that*). Thus, besides possible topical variation within articles across sections, we are able to detect also variation of stylistic lexico-grammatical patterns. While our focus is on research articles, the methodology can be applied to any text type or domain to detect (intra-textual) variation in a data-driven way.

2 Related work

Related work in (corpus) linguistics has mainly focused on variation across domains, registers or genres (represented by corpora) and less on variation within text. Among the few approaches to intra-textual variation is Swales' work on moves, discourse-structuring units with specific communicative purposes (Swales, 1990), which he applies to the analysis of research articles. A different approach is taken by Biber and colleagues (e.g. Biber et al. (2007)), who use multi-dimensional analysis considering detailed, pre-defined linguistic features to observe intra-textual variation across research article sections. Gray (2015) applies the same approach to observe features of 'elaborated' vs. 'compressed' grammatical structures (e.g. finite complement clauses such as *that*-clauses vs. adjectives as nominal pre-modifiers) across disciplines and research article sections. While quite detailed and linguistically informed, these approaches are clearly biased towards the pre-selection of features to be investigated.

In computational stylistics, there is related work on style variation of literary works, where it has been recently shown that knowledge on intra-textual variation among literary texts possibly improves computational stylistic tasks (Hoover, 2016). In terms of methods, similar work is done especially in the field of authorship attribution. These approaches aim to determine probable authors of disputed texts, ranging from considering frequencies of words, keywords and keyness to measures such as Burrow's Delta and Kullback-Leibler Divergence (see e.g. Burrows (2002); Hoover (2004); Jannidis et al. (2015); Pearl et al. (2016); Savoy (2016)). While we also use Kullback-Leibler Divergence to obtain typical features (here: of specific sections of research articles), in our approach we also account for the amount and type of information typical features provide, allowing a more fine-grained differentiation between topical vs. stylistic features.

In computational linguistics, a related problem is discourse segmentation. For an early approach see e.g. TEXTTILING (Hearst, 1997), a cohesion-driven approach for segmentation of multi-paragraph subtopic structure. More recently, topic modeling (notably LDA) has been applied to discourse segmentation as well (e.g. Misra et al. (2011); see also Riedl and Biemann (2012) for an overview). The dominant interest is on topical shifts in text as indicator of discourse structure, however topic modeling estimation is computationally expensive and needs domain-adaptation.

Recently, there is also an increasing interest in argumentative and rhetorical structure (e.g. Gou et al. (2011); Séaghdha and Teufel (2014)). While recent approaches in this field achieved promising results, they rely on highly annotated data and have to be adapted for different domains.

Further, there is work on intra-textual variation within the BioNLP community, motivated by the need to extract biomedical knowledge not only from abstracts, but also from full-text articles (Cohen et al., 2010). Besides the use of a pre-defined linguistic feature set, in BioNLP also ontologies are widely employed. This again involves a bias towards feature selection, use of highly annotated data combined with a restricted use to specific domains.

More recently, information-theoretic notions have been employed to analyze intra-textual variation. For example, Verspoor and colleagues employ Information Gain to measure the difference between conditional probabilities of tokens being part of a term within an ontology (Groza and Verspoor, 2015). The intuition behind this is to model the amount of information a token such as *activity* provides when being part of a term such as *alpha-1, 6-mannosyltransferase activity*. In this exam-

ple, *activity* provides a low amount of information, as it is also widely used within other entries (over 25,000) in the Gene Ontology. Others combine entropy with a Bayesian approach to unsupervised topic segmentation (Eisenstein and Barzilay, 2008).

We propose here to employ entropy and surprisal to model intra-textual variation. First, this allows us to detect linguistic features typical of specific sections (rather than using pre-defined ones), modeling intra-textual variation in a data-driven way. Second, by considering the amount of information (i.e. more or less informationally-loaded) and the type of information these typical features provide (i.e. topical vs. stylistic), we obtain a more comprehensive picture of the type of variation. Moreover, while the majority of approaches relies on lexical features, we take a step of abstraction, focusing also on grammatical patterns, which adds to the genericity of our approach.

3 Methodology

3.1 Data

As a dataset, we use a subsection of the SciTex corpus (Degaetano-Ortlieb et al., 2013) with research articles from genetics, amounting to approx. 2.5 million tokens (see Table 1), and covering the years 2004 to 2006. For tokenization, lemmatization and part-of-speech (POS) tagging, we use TreeTagger (Schmid, 1994) with an updated list of abbreviations specific to academic writing. Sentence splitting is based on labels of punctuations from POS information.

journal	tokens	texts
Gene	1,972,206	280
Nucleid Acids Research	612,988	71

Table 1: Journals with corpus size and number of texts

The two selected journals have the advantage of having a relatively systematic section labeling, which allows us to automatically detect sections by trigger words (e.g. Abstract, Introduction). The automatic annotation is revised manually to ensure a high quality section labeling. Table 2 shows the amount of tokens across sections[1].

[1]As the body of an article can be split into a variety of sections, rather than trying to match these into methods and result sections, we opted for putting this material into one Main part.

section	tokens
Abstract	33,577
Introduction	143,863
Main (Methods/Results)	2,136,679
Conclusion	271,075

Table 2: Section size

3.2 Methods

To observe differences in phrasal lexico-grammatical patterns across sections of research articles, we consider part-of-speech (POS) trigrams as features[2], as they have shown to perform best in inspecting lexico-grammatical patterns[3]. To consider whether a phrasal pattern transports more or less information, we also consider the amount of information in bits being transmitted by the lexical fillers of POS trigrams in a running text. For this, we use a model of *average surprisal* (AvS), i.e. the (negative log) probability of a given unit (e.g. a word) in context (e.g. its preceding words) for all its occurrences, measured in bits:[4]

$$AvS(w) = \frac{1}{|w|} \sum_i - \log_2 p(w_i|w_{i-1}w_{i-2}w_{i-3}) \quad (1)$$

where w_i is a word, w_{i-1} to w_{i-3} its three preceding words with $p(w_i|w_{i-1}w_{i-2}w_{i-3})$ being the probability of a word given its preceding three words. To obtain AvS values for POS trigrams, we take the mean of the AvS of the three lexical fillers:

$$AvS(trigram_i) = \frac{AvS(w_1) + AvS(w_2) + AvS(w_3)}{3} \quad (2)$$

This allows us to measure the amount of information in bits each instance i, i.e. each lexical realization of a POS trigram, conveys. The distribution of $AvS(trigram_i)$ is divided up into three quantiles, categorizing the data into low, middle and high AvS ranges, a methodology that already

[2]We exclude POS trigrams consisting of characters constituting sentence markers (e.g. fullstops, colons), brackets, and symbols (e.g. equal sign).

[3]Note that bi-grams proved to be too short to capture grammatical information (e.g. passives), four- and five-grams lead to sparse data.

[4]For a similar approach see Genzel and Charniak (2002).

phrase type	example trigram (POS.AvS)	example
AdjP mod	JJ.NN.NN.high	***paa2 gene cluster***
Citation	NP.CC.NP.high	*Indeed,* ***Wolner and Gralla*** *(12) showed that*
Compound	NP.NN.NP.high	***TbR-I inhibitor SB-431542***
Gerund	VVG.MD.VV.high	***silencing should prove***
NP demonstrative	DT.NNS.VHP.low	***these studies have***
Passive	NNS.VHP.VBN.high	*In plants,* ***polyamines have been*** *reported to play a crucial role in morphogenesis*
Past participle	VVN.IN.DT.low	***Based on the*** *data presented in Figure 5*
PP mod	NN.IN.JJ.middle	***use of alternative***
Semi-modal	VVP.TO.VB.low	*more detailed studies* ***need to be*** *done*
that-clause	IN.PP.MD.low	*but it was possible* ***that they could*** *be transcribed*
to-inf evaluative	JJ.TO.RB.middle	***useful to finally***
V coordination	NNS.CC.VV.high	*to functionally characterize the identified* ***mutations and distinguish*** *between polymorphisms*
Evaluative *it*-pattern	PP.VBZ.JJ.low	***it is remarkable*** *that*
VP existential	EX.VBP.JJ.low	***There are several*** *hypotheses about*
VP interactant	PP.VVP.IN.low	***we show that***
VP modal	MD.VV.DT.middle	***could explain the***
VP reporting	NNS.VVP.IN.low	***data suggest that***

Table 3: Typical phrase types with examples of POS trigrams with AvS range and examples

proved to be useful in capturing diachronic variationDegaetano-Ortlieb and Teich (2016)[5]. We then combine for each instance i information about the POS trigram and the AvS range it belongs to. At the same time, this also provides for each POS trigram the number of i with low, middle and high AvS, i.e. how many times a POS trigram occurs with low, middle or high AvS. These POS trigrams with AvS ranges serve then as features, providing a set of 19,776 features.

Detection of typical features from this feature set is based on Kullback-Leibler Divergence (KLD; or *relative entropy*), a well-known measure of (dis)similarity between probability distributions (Kullback and Leibler, 1951) used in NLP, speech processing, and information retrieval. Based on work by Fankhauser et al. (2014a,b), we create KLD models for each section (ABSTRACT, INTRODUCTION, MAIN, CONCLUSION), calculating the average amount of additional bits per feature (here: POS trigrams with AvS ranges) needed to encode features of a distribution A (e.g. ABSTRACT) by using an encoding optimized for a distribution I (e.g. INTRODUCTION). The more additional bits are needed, the more distinct or distant A and I are. This is formalized as:

$$D(A||I) = \sum_i p(feature_i|A) log_2 \frac{p(feature_i|A)}{p(feature_i|I)} \quad (3)$$

where $p(feature_i|A)$ is the probability of a feature in a section A (e.g. ABSTRACT) and $p(feature_i|I)$ is the probability of that feature in a section I (e.g. INTRODUCTION). The $log_2 \frac{p(feature_i|A)}{p(feature_i|I)}$ relates to the difference between the probability distributions $(log_2 p(feature_i|A) - log_2 p(feature_i|I))$, giving the number of additional bits. These are then weighted with the probability of $p(feature_i|A)$ so that the sum over all $feature_i$ gives the average number of additional bits per feature, i.e. the relative entropy. This allows us to determine whether any two sections are distinct or not and if they are, to what degree and by which features. For this, we inspect the ranking (based on KLD values) of features for one section vs. the other sections. In terms of typicality, the more additional bits are used to encode a feature, the more typical that feature is for a given section vs. another section. For instance, in a comparison between two sections (e.g. ABSTRACT vs. INTRODUCTION), the higher the KLD value of a features for a section (e.g. ABSTRACT), the more typical that feature is for that given section. In addition, we test for significance of a feature by an unpaired Welch's t-test. Thus, features considered typical are distinctive according to KLD and show a p-value below a given threshold (e.g. 0.05).

We thus obtain typical features for each section, i.e. typical POS trigrams combined with AvS ranges, allowing us to see whether a typical POS trigram carries more or less information (i.e. the amount of information) as defined by AvS.

For analysis, we then categorize typical POS trigrams into phrase types. Table 3 shows examples

[5]We also considered a division into quartiles, but it proved to be too narrow.

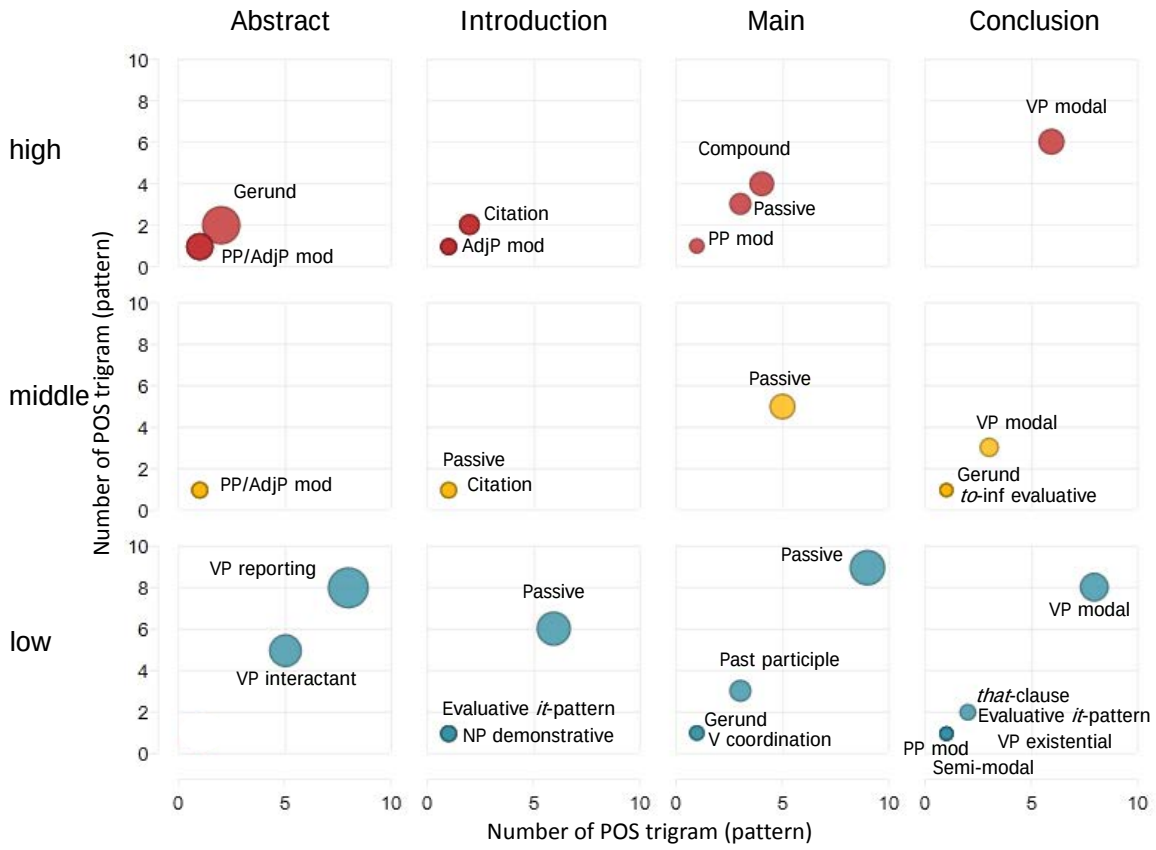

Figure 1: Typical phrase types across sections and AvS range

of POS trigrams with AvS range by their phrase type with examples of lexical realizations.

4 Analysis

In the analysis, we aim to explore intra-textual variation taking a variationist approach (rather than a text segmentation approach) and pursue the following questions:

(a) *Typical features*: Which phrasal lexico-grammatical patterns are typical of specific sections?

(b) *Amount of information*: How much information (by means of AvS) do phrasal lexico-grammatical patterns convey?

(c) *Type of information*: What type of information do phrasal lexico-grammatical patterns convey?

4.1 Typical phrase types across sections

For better comparison across sections, Figure 1 shows the number of POS trigrams (patterns) for a specific phrase type (on the x and y-axis) and the frequency per million (fpM) of the phrase type by circle size across sections with respect to high (red), middle (yellow) and low (blue) AvS values. For examples of each phrase type consider Table 3.

Considering ABSTRACT and low AvS (lower left part of Figure 1), it is strongly characterized by reporting patterns, mainly used with *that*-clauses and relatively general nouns (e.g. *data suggest that*, *analysis showed that*), and by interactant patterns (such as *we show that* and *we report here*). Considering the high AvS range (red), gerunds (see Example 4) as well as adjectival and prepositional modification are typical (see Examples 5 and 6, respectively).

(4) ***Considering*** *some severe limitations of viral systems [...] synthetic nonviral systems are highly desirable in the above applications.* (ABSTRACT; VVG.DT.JJ)

(5) *The T. maritima rpoA gene coding the subunit does not complement the* ***thermosensitive rpoA112 mutation*** *of E. coli.* (ABSTRACT; JJ.NN.NN)

(6) *The minichromosome maintenance (MCM) proteins are thought to function as the replicative helicases* ***in eukarya and archaea****.* (ABSTRACT; IN.NN.CC)

INTRODUCTION is characterized by passives (e.g. *been used with*), especially with low AvS, followed by citation with middle and high AvS (e.g. *Wolner and Gralla*). Also typical is the evaluative *it*-pattern (see Example 7) and a demonstrative pattern (e.g. *these studies/proteins have*) both in the context of presenting previous work/knowledge.

(7) ***It has become evident*** *in the last decade that many, if not the majority, of genes are regulated post-transcriptionally* [...]. (INTRODUCTION, low AvS; PP.VHZ.VVN)

MAIN is strongly characterized by passives (e.g. *analysis was performed*), especially with low AvS, but also with middle and high AvS. Also typical in the low AvS range are past participle patterns (e.g. *performed as described*, *based on the*), gerund (e.g. *purified by using*), and coordination (e.g. *and visualized with*). In addition, compound patterns are typical in the high AvS range, being clearly terminological (such as *TbR-I inhibitor SB-431542*, *SG parallel G-quadruplex*, *GC12/ GC3 correlation*).

In CONCLUSION modal verb patterns are most typical across all three AvS ranges (e.g. *units might result*, *could explain the*). In addition, with low AvS *that*-clauses are typical (e.g. *suggests that it may require*), evaluative *it*-patterns (e.g. *it is important to note*, *it is possible that*) as well as semi-modals (e.g. *seem/appear to be*), existentials (e.g. *there are several/other*) and prepositional post-modification (e.g. *present/useful in the*). Thus, modality and evaluation are quite typical for CONCLUSION sections in genetics.

Comparing typical phrase types across sections, we see that while for INTRODUCTION and MAIN passives are quite typical (especially with low AvS for both), ABSTRACT and CONCLUSION are marked by relatively unique typical phrase types (e.g. reporting verb phrases for ABSTRACT vs. modal verb phrases for CONCLUSION).

While this is in line with observations made by Biber and Finegan (1994), who have shown e.g. a preference of passives in the main part of articles as well as a common use of modal verbs in conclusions, besides other features (such as evaluative patterns) we also show the amount of information these features transmit (by AvS). Typical phrase types with high AvS values belong mostly to nominal groups (compounds and nouns modified by adjectives (AdjP mod) and prepositional phrases (PP mod)) conveying topical information, while those with low AvS values mostly to verb groups (passives and verb phrases with different functions such as reporting, evaluative, etc.) conveying a more stylistic type of information.

4.2 Amount of information and type of information of typical phrase types

Zooming into the most frequent lexical realizations of specific patterns, gives a clearer picture of the type of information conveyed by different ranges of AvS.

Here, we present two examples: First, we zoom into typical patterns of ABSTRACT, showing how the type of information differs from topical to stylistic based on the AvS range. Second, we look at CONCLUSION considering its typical modal verb phrase across AvS ranges.

Figure 2 shows lexical realizations of typical phrase types within ABSTRACT across AvS ranges (high: reddish, middle: yellowish, low: blueish) with the size relating to frequency for each range.

Typical reporting verb patterns (VP reporting) with low AvS values (blueish) make use of relatively general nouns (*data*, *analysis*, *results*) with verbs such as *suggest*, *show* and *indicate*. For VP interactional, the phrase *we show that* is the dominant lexical realization, followed, for example, by phrases such as *we characterized the/demonstrate that/report here*. The amount of information transmitted by these phrases is relatively low. The purpose of use of these phrases is more style-oriented rather than topic-oriented.

Comparing this to lexical realizations of high AvS values (reddish) for ABSTRACT (see again Figure 2), we see that these are clearly related to quite compact linguistic forms expressing either processes with the gerund form (*lining the gastrovacular*) or scientific terms with adjectival (e.g. *multiple gene cassette*) and prepositional modification (e.g. *helicases in eukarya and archaea*[6]). Clearly, the amount of information these phrases

[6]Note that for this pattern we have shown more context for better understanding, as the pattern would only show Preposition-Noun-Conjunction, which in the example is realized as *in eukarya and*.

Figure 2: Lexical realizations of typical patterns with high, middle and low AvS in ABSTRACT

transmit is relatively high and the type of information is topic-oriented.

Lexical realizations of middle AvS lie in between, i.e. terms seem to be more generic (e.g. by adjectival modification such as *positive regulatory factor* or prepositional modification such as *lack of regulatory*) and reporting verb phrases are used in a less confined ambient context (e.g. *showed a common* instead of *showed that*). Thus, these phrases transmit a relatively moderate amount of information and can be style- or topic-oriented.

In Figure 3 we zoom into CONCLUSION, showing how the same typical phrase type (here: VP modal; compare also with Figure 1) can differ in the information type it conveys depending on its AvS range. Here, lexical realizations of verb patterns with modal verbs are shown for high (reddish), middle (yellowish), and low (blueish) AvS values. With high AvS, the modal verb is used in combination with specific terms (e.g. *tlh genes*, *tRNA isodecoders*). From Examples 8 and 9, we can see how within the whole sentences assumptions are put forward about the two terms *tlh genes* and *tRNA isodecoders*. In the middle range, the modal verb patterns are used with a variety of verbs. Examples 10 and 11 show relatively generic preceding contexts (*the structure of the substrate* in 10 and *subtle changes* in 11), which are used with modal expressions of middle AvS. In the lower range, the modal verbs are used with a confined set of verbs (*suggest*, *result*), in relatively formalized lexical phrases (*may/might be due*), and in relational phrases (*may be an*, *might be the*). Example 12 to 14 show a quite vague preceding context realized by the use of referring expressions such as *this* and *there* for modal verbs used with low AvS.

Figure 3: Lexical realizations of modal verb patterns (VP modal) with high, middle and low AvS in CONCLUSION

(8) *A possibility suggested by the presence of transposable elements in the pericentromeric heterochromatin of higher eukaryotes is that the* ***tlh genes might*** *be parts of transposons that preferentially transpose to heterochromatic regions.* (CONCLUSION)

(9) *One can envision that* ***tRNA isodecoders may*** *be more harmful than useful in translation.* (CONCLUSION)

(10) *This assay is performed in the absence of dNTP, so that the structure of the substrate*

cannot be affected *by polymerization .* (CONCLUSION)

(11) *PAX 7 gene expression levels are highly controlled during tissue development and subtle* ***changes could lead*** *to important effects.* (CONCLUSION)

(12) *Our work does not suggest that gene expression contributes to the asymmetric evolution of paralogs that we observed but again* ***this may be due*** *to small sample size.* (CONCLUSION)

(13) ***This may be due*** *to the short length (11 bp) of the primer* [...]. (CONCLUSION)

(14) ***There may be a*** *few possible reasons for why hix-AG is not bound by Hin* [...]. (CONCLUSION)

Given that this is just one type of phrase, i.e. modal verb phrase being typical for CONCLUSION in genetics, by considering AvS we clearly see how it still differs in the type of information it transmits, depending on the ambient context it occurs with, being either topical or stylistic.

5 Section classification

While in the analysis we have taken a variationist approach, we also test how well sections can be distinguished by typical features obtained by our approach. Our baseline is a classifier using all POS trigrams without AvS ranges. In Table 5 we report the F-Measure of three classifiers (Naive Bayes, Support Vector Machine (SVM) and RandomForest (RF)). Adding AvS ranges improves classification for all classifiers. Using only typical POS trigrams obtained by our approach improves the model considerably. A further improvement is achieved by considering typical POS trigrams with AvS ranges. The random forest classifier achieving the best result with 86.0 of F-Measure.

set	**BL** (NaiveBayes)	SVM	RF
POS 3grams	76.6	78.2	72.9
POS 3grams+AvS	77.0	80.3	74.2
typPOS	80.3	82.5	85.6
typPOS+AvS	81.1	81.0	**86.0**

Table 4: Classification results with typical POS trigrams and AvS ranges.

Considering classification performance of sections with Random Forest, ABSTRACT and MAIN can be best predicted with 94.5 and 92.5 of F-Measure, followed by INTRODUCTION with 82.8. CONCLUSION is less well distinguishable, but still achieves a considerable improvement when considering typical POS trigrams (from 17.4 to 61.2 of F-Measure).

set	**ABS**	**INTRO**	**MAIN**	**CONC**
POS 3grams	84.1	71.2	88.2	17.4
POS 3grams+AvS	84.8	75.5	87.9	20.1
typPOS	93.3	81.0	92.7	61.2
typPOS+AvS	**94.5**	82.8	92.5	60.5

Table 5: Classification results by F-Measure for each section (RandomForest)

6 Conclusion

This paper has presented a novel data-driven approach to intra-textual variation. We have shown how sections of research articles from genetics differ with respect to the phrasal lexico-grammatical patterns used across sections (see Section 4.1). We used *relative entropy* to obtain typical lexico-grammatical patterns for each section. Moreover, we have modeled the amount and type of information these lexico-grammatical patterns convey (see Section 4.2) by using *average surprisal* (AvS), showing that sections vary in topical as well as stylistic type of information. In future work, we plan to model different scientific domains to investigate which of these lexico-grammatical patterns would generalize across domains and which are domain-specific.

Being data-driven and using part-of-speech information to generate features (see Section 3.2), our approach can be applied to any other domain, text type and even other languages (given a good quality POS annotation), since it is not biased by topical variation. While here we have modeled intra-textual variation, additional variables such as time, author, social group, production type, language etc. can be integrated into the model. For an application on diachronic data see Degaetano-Ortlieb et al. (2016) and Degaetano-Ortlieb and Teich (2016). As long as the variables are known (e.g. publication dates for time, author names for author, etc.), our approach allows to investigate variation at a more abstract linguistic level than topical variation. Thus, our approach is directly relevant to studies in sociolinguistics, historical linguistics and digital humanities in general.

Assessing the amount and type of information

of typical lexico-grammatical patterns is relevant for more sophisticated text analysis. For example, historical linguists might be interested in the whole AvS range, as specific linguistic features might move across time between high, middle and low AvS. A linguistic feature might have high AvS in one time period (e.g. when it enters language use its ambient context may be expected to vary a lot), and low AvS in a later time period (where the feature is well-established in language use and might be more confined to a specific ambient context). The transition period would be seen in the use of the feature in the middle AvS range. In information retrieval, instead, features with high AvS are more relevant as they convey more information and are topic/content-related. AvS ranges could also be more fine-grained in this scenario to distinguish relatively established from new terms. Considering more fine-grained ranges of high AvS combined with time as a variable might be a possible way to explore knowledge change.

7 Acknowledgments

This work is funded by *Deutsche Forschungsgemeinschaft* (DFG) under grants SFB 1102: Information Density and Linguistic Encoding (www.sfb1102.uni-saarland.de) and EXC 284: Multimodal Computing and Interaction (www.mmci.uni-saarland.de). We are also indebted to Stefan Fischer for his contributions to corpus processing and Peter Fankhauser (IdS Mannheim) for his support in statistical analysis. Also, we thank the anonymous reviewers for their valuable comments.

References

Douglas Biber, Ulla Connor, and Thomas A. Upton. 2007. *Discourse on the Move: Using Corpus Analysis to Describe Discourse Structure*. Benjamins, Amsterdam.

Douglas Biber and Edward Finegan. 1994. Intratextual Variation within Medical Research Articles. In Susan Conrad and Douglas Biber, editors, *Variation in English: Multi-dimensional Studies*, Routledge Taylor & Francis Group, pages 108–123.

John Burrows. 2002. Delta: a Measure of Stylistic Difference and a Guide to Likely Authorship. *Literary and Linguistic Computing* 17(3):267–287. https://doi.org/10.1093/llc/17.3.267.

K. Bretonnel Cohen, Helen L. Johnson, Karin Verspoor, Christophe Roeder, and Lawrence E. Hunter. 2010. The Structural and Content Aspects of Abstracts versus Bodies of Full Text Journal Articles are Different. *BMC Bioinformatics* 11(492):1–10.

Stefania Degaetano-Ortlieb, Hannah Kermes, Ashraf Khamis, and Elke Teich. 2016. An Information-Theoretic Approach to Modeling Diachronic Change in Scientific English. In Carla Suhr, Terttu Nevalainen, and Irma Taavitsainen, editors, *Selected Papers from Varieng - From Data to Evidence (d2e)*, Brill, Language and Computers.

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*, Narr, volume 3 of *Corpus Linguistics and Interdisciplinary Perspectives on Language - CLIP*, pages 93–104.

Stefania Degaetano-Ortlieb and Elke Teich. 2016. Information-based Modeling of Diachronic Linguistic Change: From Typicality to Productivity. In Nils Reiter, Beatrice Alex, and Kalliopi A. Zervanou, editors, *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Association for Computational Linguistics (ACL).

Jacob Eisenstein and Regina Barzilay. 2008. Bayesian Unsupervised Topic Segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '08, pages 334–343. http://dl.acm.org/citation.cfm?id=1613715.1613760.

Peter Fankhauser, Hannah Kermes, and Elke Teich. 2014a. Combining Macro- and Microanalysis for Exploring the Construal of Scientific Disciplinarity. In *Digital Humanities*. Lausanne, Switzerland. URL: http://dharchive.org/paper/DH2014/Poster-126.xml.

Peter Fankhauser, Jörg Knappen, and Elke Teich. 2014b. Exploring and Visualizing Variation in Language Resources. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*. European Language Resources Association (ELRA), Reykjavik, pages 4125–4128. http://nbn-resolving.de/urn:nbn:de:bsz:mh39-26224.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy Rate Constancy in Text. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Philadephia, Pennsylvania, USA, pages 199–206. http://dl.acm.org/citation.cfm?id=1073117.

Yufan Gou, Anna Korhonen, and Thierry Poibeau. 2011. A Weakly-supervised Approach to Argu-

mentative Zoning of Scientific Documents. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Edinburgh, UK, pages 273–283.

Bethany Gray. 2015. On the Complexity of Academic Writing. Disciplinary Variation and Structural Complexity. In Viviana Cortes and Eniko Csomay, editors, *Corpus-based Research in Applied Linguistics. Studies in Honor of Doug Biber*, John Benjamins Publishing Company, Amsterdam / Philadelphia, volume 66 of *Studies in Corpus Linguistics (SCL)*, pages 49–77.

Tudor Groza and Karin Verspoor. 2015. Assessing the Impact of Case Sensitivity and Term Information Gain on Biomedical Concept Recognition. *PLoS One* 10(3):1–22.

Marti A. Hearst. 1997. TextTiling: Segmenting Text into Multi-paragraph Subtopic Passages. *Computational Linguistics* 23(1):33–64.

David L. Hoover. 2004. Testing Burrows's Delta. *Literary and Linguistic Computing* 19(4):453–475. https://doi.org/10.1093/llc/19.4.453.

David L. Hoover. 2016. Textual Variation, Text-Randomization, and Microanalysis. In *Proceedings of Digital Humanities Conference (DH)*. Kraków, Poland, pages 223–225.

Fotis Jannidis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. Improving Burrows' Delta - An Empirical Evaluation of Text Distance Measures. In *Digital Humanities Conference (DH)*. Sydney, Australia.

Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22(1):79–86.

Hemant Misra, François Yvon, Olivier Cappé, and Joemon Jose. 2011. Text Segmentation: A Topic Modeling Perspective. *Information Processing & Management* 47(4):528 – 544. https://doi.org/http://dx.doi.org/10.1016/j.ipm.2010.11.008.

Lisa Pearl, Kristine Lu, and Anousheh Haghighi. 2016. The Character in the Letter: Epistolary Attribution in Samuel Richardsons Clarissa. *Digital Scholarship in the Humanities* https://doi.org/https://doi.org/10.1093/llc/fqw007.

Martin Riedl and Chris Biemann. 2012. Text Segmentation with Topic Models. *Journal for Language Technology and Computational Linguistics (JLCL)* 27(1):47–70.

Jacques Savoy. 2016. Estimating the Probability of an Authorship Attribution. *Journal of the Association for Information Science and Technology* 67(6):1462–1472. https://doi.org/10.1002/asi.23455.

Helmut Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *International Conference on New Methods in Language Processing*. Manchester, UK, pages 44–49.

Diarmuid Ó Séaghdha and Simone Teufel. 2014. Unsupervised Learning of Rhetorical Structure with Un-topic Models. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*. Dublin, Ireland, pages 2–13.

John M. Swales. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge Applied Linguistics. Cambridge University Press, Cambridge.