Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity

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

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

Abstract

We present a new approach for modeling diachronic linguistic change in grammatical usage. We illustrate the approach on English scientific writing in Late Modern English, focusing on grammatical patterns that are potentially indicative of shifts in register, genre and/or style. Commonly, diachronic change is characterized by the relative frequency of typical linguistic features over time. However, to fully capture changing linguistic usage, feature productivity needs to be taken into account as well. We introduce a data-driven approach for systematically detecting typical features and assessing their productivity over time, using information-theoretic measures of entropy and surprisal.

1 Introduction

The analysis of diachronic corpora is of great interest to linguistics, history and cultural studies alike. The challenges in dealing with diachronic material are manifold, ranging from corpus compilation and annotation to analysis. Here, we address questions of analysis, notably the data-driven detection and evaluation of linguistic features marking shifts in register, genre and/or style (Halliday, 1988; Halliday and Hasan, 1985). Specifically, we focus on the *productivity* of features over time, i.e. the property of a grammatical pattern to attract new lexical items and to spread to new contexts (cf. Barðdal (2008)).

In terms of methods, we propose a systematic approach to feature detection and evaluation based on information-theoretic measures such as entropy and surprisal. These measures are based on probability in context, where that context may be the ambient context (as in n-gram models) or the extra-linguistic context (here: time, register) (cf. Section 3 for details). While we investigate diachronic linguistic change in English scientific writing, our methodology can easily be applied to other scenarios analyzing differences/similarities across registers/genres/languages/time and the like in terms of typicality and productivity.

To detect features, we employ *relative entropy* or Kullback-Leibler Divergence (KLD), a well-known measure of similarity/dissimilarity between probability distributions used in natural language and speech processing and information retrieval (see e.g. Dagan et al. (1999); Lafferty and Zhai (2001)). Using KLD, we compare different time periods and obtain typical features of scientific texts for further analysis. As features, here, we use part-of-speech (POS) 3-grams to approximate grammatical patterns. To capture productivity, we apply the notion of *average surprisal* (AvS). Using surprisal, we compare differences in probabilities for selected units (here: parts-of-speech) and contexts across different time periods and registers (here: scientific vs. "general" language), which allows us to evaluate their contribution to change in terms of productivity. For example, passive voice is considered a typical feature of scientific writing (compared to other registers) (cf. Biber et al. (1999)). Diachronically, its productivity may have been low in the beginning and increasing later on or it may first have been high and then decreasing over time. For example, in scientific writing passive may have initially been used with only a few verbs (e.g. BE + *made/seen/found*) and in few contexts (e.g. *as/it may be seen*) and then extended to more verbs (e.g. BE + *made/seen/found/observed/determined/produced*) and spread to more contexts (e.g. *as/it/that may/will/must* + VERB), which would indicate a

Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH), pages 165–173,
Berlin, Germany, August 11, 2016.

shift from a lower to a higher productivity.

In the following, we describe related work (Section 2) as well as the data, methods and analytic procedures (Section 3), followed by selected analyses and results (Section 4). We conclude with a summary and envoi (Section 5).

2 Related Work

Existing work on diachronic change in scientific language typically focuses on short-term change (e.g. work on the ACL anthology corpus; (Hall et al., 2008)) and mostly investigates lexis-related change (e.g. topical shifts). Here, we address long-term change and grammatical patterns, focusing on their productivity.

Productivity has a long history in the field of derivational morphology, i.e. the word formation processes employed by speakers to generate new words. Different methods have been proposed to measure productivity of affixes (e.g. Baayen and Lieber (1991); Hay and Baayen (2002)). More recently, there is also some interest in modeling syntactic productivity, i.e. the combination of syntactic patterns or constructions with lexical items, with approaches ranging from simple measures such as proportional preference (Biber, 2012) to collostructions (Stefanowitsch and Gries, 2003) and distributional semantics (Perek, 2014, 2016).

In corpus linguistics, existing approaches to diachronic change are essentially frequency-based and work from predefined features known to be involved in linguistic change (e.g. Biber and Gray (2011); Gray and Biber (2012); Taavitsainen and Pahta (2012); Moskowich and Crespo (2012)). While frequency is clearly a major indicator of change, it does not provide a full picture. To investigate syntactic productivity, we clearly need an approach which accounts for context of use. Perek (2014), for instance, considers the semantic context of a specific lexical phrase (V *the hell out of* NP) in diachrony (from 1930 to 2009 in the COCA corpus (Davies, 2008)) by applying distributional semantic models. He shows how the different verbs filling the lexical phrase are semantically related and how visualization techniques and statistical modeling can be used to analyze the semantic development of a construction in terms of syntactic productivity.

We model the productivity of grammatical patterns that become increasingly typical over time by using the notion of *surprisal*. Surprisal is rooted in information theory (Shannon, 1949) and is widely applied in psycholinguistic studies (e.g. Hale (2001); Levy (2008); Demberg and Keller (2008)) to assess cognitive processing effort. We apply surprisal to calculate a unit's probability in context to analyze diachronic shifts in productivity considering a unit's probability in a given context as well as the probability of a context with a given unit (see more details in Section 3).

3 Data and methods

Data The corpus of scientific writing we use consists of the first two centuries of publication of the Royal Society of London (1665-1869; RSC), altogether 35 million tokens (Kermes et al., 2016). It is encoded for text type (article, abstract), author and date of publication. For analysis, the corpus can be flexibly chunked up in different time periods (e.g. decades). Linguistic annotation of the corpus has been obtained by using existing tools: VARD (Baron and Rayson, 2008) for normalization and TreeTagger (Schmid, 1994, 1995) for tokenization, lemmatization and part-of-speech (POS) tagging. For training and evaluation, we created a manually annotated (normalization, part-of-speech tags) subcorpus ($\sim$56.000 tokens). The trained model for VARD exhibited a 10% increase (61.8% to 72.8%) and double the recall (31.3% to 57.7%). For TreeTagger we obtained 95.1% on normalized word forms (Kermes et al., 2016). This procedure ensured a relatively reliable part-of-speech tagging of historical texts.

For comparative purposes, we employ a register-mixed corpus, the Corpus of Late Modern English Texts, version 3.0 (CLMET) (Diller et al., 2011), which has a similar size and roughly spans the same period (1710-1920) as the RSC.

Methods and analytic procedures For feature detection, we create KLD models for RSC on the basis of part-of-speech (POS) 3-grams[1]. Kullback-Leibler Divergence (or *relative entropy*) measures the difference between two probability distributions by calculating the difference in the number of bits between the cross-entropy of two data sets A and B and the entropy of A alone, i.e. $H(A;B) - H(A)$. The more additional bits are

[1]To further avoid possible POS tagging errors, in the extraction procedure nouns were restricted to a size of >2 characters. Furthermore, we exclude POS 3-grams consisting of characters constituting sentence markers (e.g. fullstops, colons), brackets, symbols (e.g. equal signs), and words tagged as foreign words.

needed for encoding a given unit (here: POS 3-gram), the more distinctive (and thus typical) that unit is for a given time period vs. another time period. On this basis, we compare the probabilities of 3-grams across the five time periods in RSC[2], aiming to obtain those 3-grams that become increasingly typical of scientific language over time. For this, we create four KLD models for each (fifty years) time period, starting with 1700 vs. its preceding time period based on 1184 POS 3-grams. We then inspect the ranking (based on KLD values) of 3-grams typical of one time period vs. a previous time period. Thus, we obtain the 3-grams typical of 1700 vs. 1650, 1750 vs. 1700, 1800 vs. 1750, 1850 vs. 1800.

We then further analyze selected typical 3-grams in terms of relative frequency, comparing their distributions across RSC. In addition, to confirm typicality within scientific language, we also compare the use of these 3-grams within a general language corpus (CLMET) (cf. Section 4.2).

For studying productivity we apply *surprisal*, a measure of information calculating the number of bits used to encode a message. The amount of bits being transmitted by a given linguistic unit in a running text depends on that unit's probability in context. Formally, surprisal is quantified as the negative log probability of a unit (e.g. a word) in context (e.g. its preceding words):

$$S(unit) = -\log_2 p(unit|context)$$

In corpus analysis, we are interested in the surprisal of all occurrences of a given linguistic unit, i.e. its *average surprisal*:

$$AvS(unit) = \frac{1}{|unit|} \sum_i -\log_2 p(unit_i|context_i)$$

For instance, using words (uni-grams) as units, we can inspect whether a given word is more "surprising" in one context vs. in another context. We create AvS models for RSC and CLMET on the basis of uni-grams in the context of three preceding tokens and compare the AvS of the selected 3-grams across RSC and CLMET. For assessing their productivity, we inspect the AvS ranges of their lexical heads in the preceding context of three tokens as well as their type-token ratios (cf. Section 4.3).

[2](1650: 1665–1699, 1700: 1700–1749, 1750: 1750–1799, 1800: 1800–1849, 1850: 1850–1869)

3-gram	example	type
DT.JJ.JJ	*the same general*	nominal
NN.TO.DT	*respect to the*	
TO.DT.JJ	*to the same*	prepositional
IN.VVG.DT	*for determining the*	gerund
DT.NN.VBZ	*the latter is*	verbal; BE
VV.DT.JJ	*produce the same*	verbal; base form
VV.IN.DT	*account for the*	
MD.VB.VVN	*will be found*	verbal; passive
VB.VVN.IN	*be considered as*	
VBD.VVN.IN	*were made with*	
VBZ.VVN.IN	*is composed of*	
VVN.TO.VV	*found to contain*	verbal; to-inf

DT: determiner, JJ: adjective, IN: preposition, MD: modal verb, NN: common noun, TO: to-particle/preposition, VB: verb *be*, VBD: verb *be* past, VBZ: verb *be* present, VV: verb base form, VVG: ing-verb, VVN: past tense verb

Table 1: List of 3-grams increasingly typical in RSC obtained from KLD ranking

4 Analysis and results

4.1 Typicality

From the KLD models (built as described in Section 3), we obtain altogether twelve 3-grams which become increasingly typical over time (see Table 1). A subset of these clearly reflect particular (sets of) grammatical patterns.

Consider, for example, the gerund 3-gram consisting of a preposition followed by an *ing*-verb and a determiner (IN.VVG.DT). According to previous historical linguistic studies (cf. De Smet (2008); Fanego (2004, 2006)), this grammatical pattern reflects the verbal gerund, which has been shown to have developed from Middle English onwards. By our method, we can show that it becomes increasingly typical in scientific writing over time, confirming also Gray and Biber (2012)'s frequency-based results. Another grammatical pattern that becomes increasingly typical in our data is passive voice (reflected by four 3-grams; see again Table 1), which is in line with standard reference works on English Grammar, such as Biber et al. (1999). In addition, there are also two nominal patterns which become increasingly typical over time (DT.JJ.JJ and NN.TO.DT) as well as a prepositional pattern (TO.DT.JJ) and other verbal patterns (DT.NN.VBZ, VV.DT.JJ, VV.IN.DT, and VVN.TO.VV).

In the following, we focus on the two grammatical patterns gerund and passive (overall five 3-grams).

4.2 Frequency-based diachronic changes

All five 3-grams increase in frequency up until 1800 in RSC (see Figure 1 showing frequencies per million in the five time periods). The gerund then drops from 1800 to 1850. The past passive decreases from 1800 to 1850, while the present passive increases. This may indicate a replacement of past tense with present tense for the passive in RSC. The modal passive and BE passive seem to develop a stable distribution from 1750 onwards.

Comparing RSC with CLMET (compare Figure 1 with Figure 2), while there is generally a frequency increase in RSC, CLMET shows a decreasing tendency. Nevertheless, the past passive increases both in RSC and CLMET from 1700 to 1750 to a similar level, but then while in RSC it keeps increasing till 1800, in CLMET it decreases.

Considering the gerund 3-gram, it shows similar frequencies across RSC and CLMET around 1700, while it clearly drops in use in CLMET compared to RSC around 1850. Passive 3-grams show a similar tendency: all 3-grams are less frequently used in CLMET than in RSC around 1850, even though

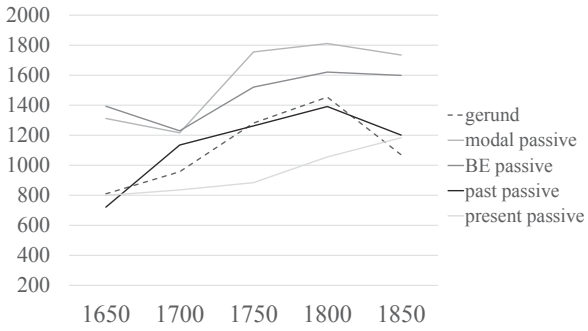

Figure 1: Diachronic frequency distribution of gerund (dashed line) and passive 3-grams in RSC

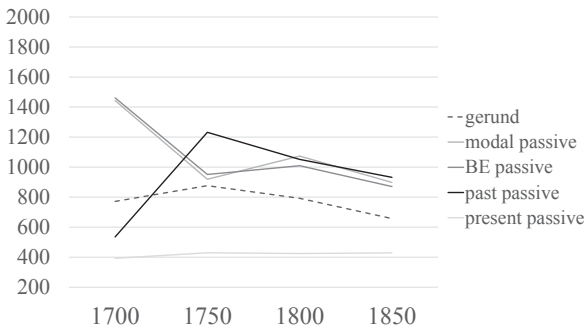

Figure 2: Diachronic frequency distribution of gerund (dashed line) and passive 3-grams in CLMET

around 1700 3-grams with the verb *be* in base form (modal passive, BE passive) were used to similar extents in RSC and CLMET. Finally, the present passive is less frequently used over time in both RSC and CLMET.

In summary, scientific writing and general language become increasingly distinct over the given time period: Overall, in RSC the gerund as well as the passive increase in frequency, in CLMET their frequencies decrease. This indicates an increasingly more formal, expository and abstract style of scientific written English in comparison to "general" English.

4.3 Productivity

In our discussion of productivity, we focus on the gerund (IN.VVG.DT) and the modal passive (MD.VB.VVN) 3-grams.

4.3.1 Number of types/tokens

To inspect degree of lexical variation, we consider how many types a 3-gram has per tokens over the time periods of RSC and CLMET. We observe that RSC uses fewer types over time, while in CLMET the number of types is fairly stable from 1750 onwards. See Figure 3 and Figure 4 showing the gerund and the modal passive 3-gram, respectively. Thus, in scientific writing the lexical variation of these typical 3-grams goes down over time, giving rise to a more conventionalized use (lower productivity). In general language, instead, lexical variation in these 3-grams increases. Note that overall, the variation is mainly due to the lexical units in the two 3-grams, i.e. VVG and VVN, since the other parts-of-speech are function words.

4.3.2 Preceding contexts

To inspect variation in context, we consider the average surprisal (AvS) of the individual verbs filling

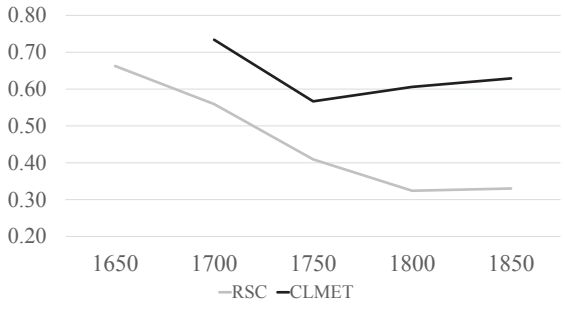

Figure 3: Types per tokens for the gerund 3-gram (IN.VVG.DT)

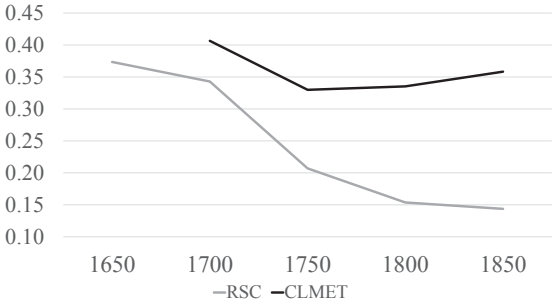

Figure 4: Types per tokens for the modal passive 3-gram (MD.VB.VVN)

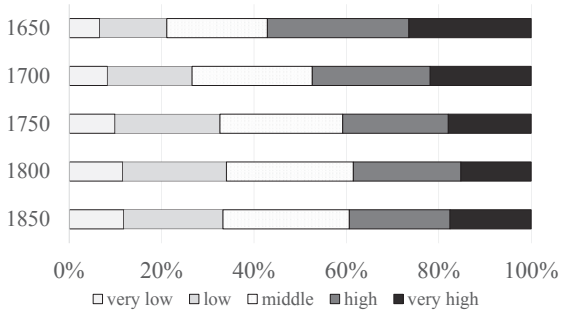

Figure 5: AvS values of lexical verb in the gerund 3-gram (RSC)

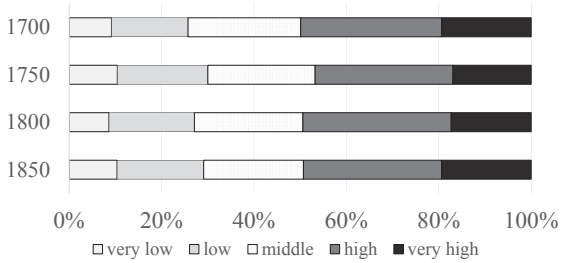

Figure 6: AvS values of lexical verb in the gerund 3-gram (CLMET)

VVG and VVN in their preceding contexts of three words:

$$AvS(verb) = \frac{1}{|verb|} \sum_{i} - \log_2 p(verb_i | w_{i-1} w_{i-2} w_{i-3})$$

Here we want to see whether these verbs obtain rather low or rather high AvS values. Low values would indicate a relatively conventionalized use of the verb in its context, i.e. a low degree of productivity, because based on its preceding words the verb is quite predictable. High AvS values would point to verbs which are hard to predict by their previous context (e.g. new verbs entering the vocabulary, which would indicate a higher degree of productivity).

The AvS values range from 0 to 22. For comparison, we define a scale based on five quantiles.

Gerund (VVG) Figure 5 shows the AvS distribution of the lexical verbs realizing the gerund 3-gram. Diachronically, in RSC an increasing number of verbs have very low to low AvS values (from ∼20% in 1650 to ∼30% in 1850, see light gray shades) but a decreasing number have high to very high AvS values (from ∼60% in 1650 to ∼40% in 1850, see dark shades). This seems to indicate that an increasing number of verbs are used over time in the same context pointing to lower productivity, while rare, untypical or new verbs become less frequent. The middle range (white shade) remains relatively stable over time. Comparing this to the AvS of the lexical verbs realizing the gerund in CLMET, a different tendency is observed (see Figure 6). In general, there is less variation in the distribution of the AvS values in CLMET in comparison to RSC, i.e. productivity does not seem to change diachronically.

To test whether AvS can really be a measure showing effects of productivity, we inspect the lexical realizations of the preceding context of the gerund in RSC considering again the distribution according to five quantiles and the number of types over tokens for each time period. Thus, a low AvS value of the full verb would be an indicator of low productivity in terms of preceding context and vice versa a high AvS value would be an indicator of high productivity. Figure 7 shows how very low to low AvS values (light gray shades) of the verb have also a low number of types over tokens in the preceding context (∼0.4–0.6) and high to very high AvS values (dark shades) have a high number of types over tokens (∼0.9–1.0). This relation is relatively stable over time. Thus, AvS can be used to distinguish higher vs. lower productivity.

We then inspect the concrete lexical items that have very low to low AvS, i.e. which are quite predictable given the previous context (preceding three words). This allows us to see which items are used in relatively fixed expressions and how this changes over time. Thus, we can inspect how the unit (gerund) changes over time as well as how the context changes (see Table 2). In 1650 and 1700 a relatively general verb, *making*, is used, while over time more specific verbs appear (*determining*, *examining*, *obtaining*). Moreover, the

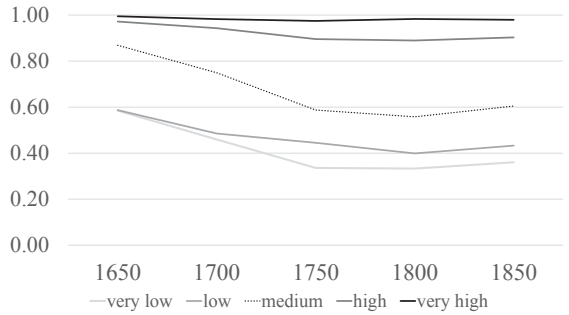

Figure 7: AvS of lexical verb and type-token ratio of preceding context of the gerund 3-gram (RSC)

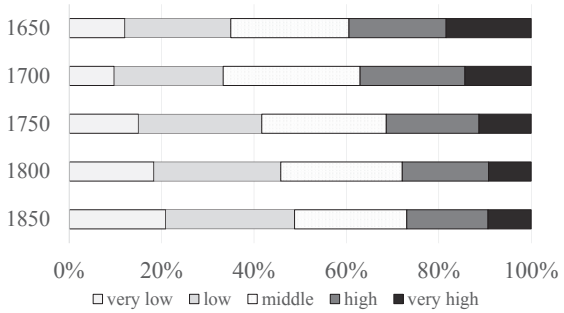

Figure 8: AvS values of lexical verbs in the modal passive 3-gram (RSC)

context becomes more restricted over time, i.e. the same lexical realization for the preceding context is used (e.g. *an opportunity of* and *the purpose of*) in combination with different verbs. Thus, diachronically lexical variation of the gerund may increase, while its context of use gets increasingly restricted.

Passive (VVN) Considering the modal passive 3-gram (MD.VB.VVN), we observe a similar pattern (see Figure 8). The very low to low AvS values for RSC of the past tense verb (VVN) rise up to around 50% in 1850, while the number of high AvS values decreases over time (to around 30%). Again, this indicates lower productivity over time in RSC. Comparing this to the distribution in CLMET (see Figure 9), it again remains fairly stable over time in comparison to RSC. Thus, also for the modal passive 3-gram, productivity in CLMET remains stable.

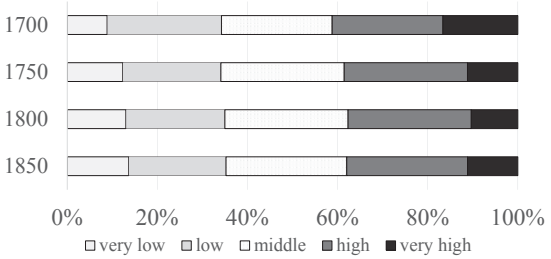

Figure 9: AvS values of lexical verbs in the modal passive 3-gram (CLMET)

From Figure 10, we again observe how low to high AvS values correlate with low to high number of types over tokens, respectively. This confirms that AvS is an indicator of productivity.

Further, we inspect the concrete lexical items that realize the full verb for the passive and which are relatively predictable given the previous context (low to very low AvS). We can see from Ta-

period	context + VVG	freq	%
1650	the way of **making**	12	3.40
	the opportunity of **making**	5	1.42
	the way of **measuring**	3	0.85
1700	be made by **multiplying**	11	1.45
	be capable of **producing**	4	0.53
	the pleasure of **seeing**	3	0.40
1750	an opportunity of **examining**	15	0.64
	be capable of **producing**	12	0.51
	the manner of **making**	11	0.47
1800	the purpose of **determining**	37	0.83
	the purpose of **ascertaining**	36	0.81
	an opportunity of **examining**	17	0.38
1850	the purpose of **ascertaining**	24	0.63
	the purpose of **determining**	23	0.61
	the purpose of **obtaining**	20	0.53

Table 2: Gerund verbs in the gerund 3-gram for very low to low AvS (RSC)

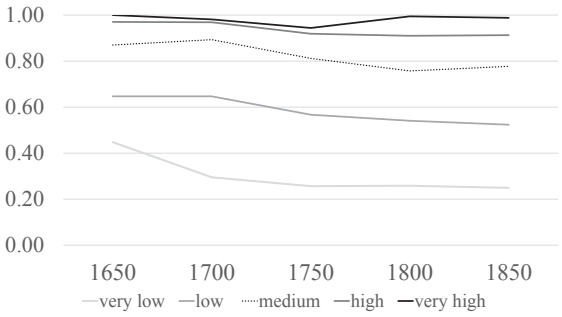

Figure 10: AvS of lexical verb and type-token ratio of preceding context of modal passive 3-gram (RSC)

ble 3 that the verbs used are basically the same diachronically (*found*, *observed*, *seen*). What changes is the context they appear in, which becomes progressively restricted over time and develops into a relatively fixed expression by 1850 (*it will be* VVN).

period	context + VVN	freq	%
1650	as may be **seen**	48	5.17
	that may be **made**	21	2.26
	it will be **found**	16	1.72
1700	as may be **seen**	48	4.56
	it will be **found**	26	2.47
	it may be **observed**	24	2.28
1750	it will be **found**	117	2.97
	it must be **observed**	93	2.36
	it may be **observed**	74	1.88
1800	it will be **seen**	351	4.76
	it will be **found**	257	3.48
	it may be **observed**	95	1.29
1850	it will be **seen**	494	5.68
	it will be **observed**	244	2.81
	it will be **found**	227	2.61

Table 3: Passive verbs in the modal passive 3-gram for very low to low AvS (RSC)

5 Conclusions

We have presented an approach to long-term diachronic change — here: in scientific writing — combining typicality and productivity of features involved in changing language use. While relative frequency is clearly a major indicator of change, also productivity, i.e. the lexical extensibility of a linguistic unit and the degree of variation in its immediate context, may change. To address productivity, we have suggested to employ the notion of (average) surprisal, which measures the predictability of a linguistic unit in context. Predictability in context is a function of frequency of a unit, variation of the unit and variation of its context. In a given context, the more frequent a given unit (e.g. a particular part-of-speech) and the less varied its realizations (e.g. lexical types) are, the less surprising that unit is (and vice versa). Also, the contexts in which a unit occurs may change over time, they may expand or become more restricted. More contextual variation makes the unit less predictable, less variation makes it more predictable.

We have investigated a set of POS 3-grams becoming increasingly typical of scientific writing diachronically (mid 17th to mid 19th century), as determined by KLD and feature ranking. We then inspected the relative frequency of the selected 3-grams as well as their productivity over time by means of AvS. Compared to "general language", the analyzed 3-grams become more frequent over time, while their productivity diminishes. Both gerund and passive (with modal verb) exhibit fewer types over time and the contexts in which their lexical heads are used become more restricted. Such restricted language use has been noted before as a property of specialized sublanguages and is confirmed by our analyses (cf. Biber and Gray (2013)).

As the feature detection approach using KLD is based on part-of-speech tags, it can be applied to various other scenarios of comparison (e.g. different languages, registers, modes, etc.). Moreover, depending on the goal of analysis, other kinds of units, at all linguistic levels, and contexts can form the basis of (average) surprisal modeling (see e.g. Asr and Demberg (2015) on the predictability of discourse markers or Schulz et al. (2016) on vowel space and surprisal). In our ongoing work, we analyze the 3-grams that have not been considered here to determine whether they show similar productivity patterns or not. Given that scientific language is said to become increasingly nominal over time (cf. Halliday (1988); Biber and Gray (2011)), we would predict that nominal patterns (e.g. DT.JJ.JJ, NN.TO.DT; cf. Table 1) become more productive over time to ensure a sufficient level of expressivity in scientific language.

References

Fatemeh Torabi Asr and Vera Demberg. 2015. Uniform Information Density at the Level of Discourse Relations: Negation Markers and Discourse Connective Omission. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS 2015)*. London, UK, page 118.

Harald Baayen and Rochelle Lieber. 1991. Productivity and English Derivation: A Corpus-based Study. *Linguistics* 29(5):801–844.

Jóhanna Barðdal, editor. 2008. *Productivity: Evidence from Case and Argument Structure in Icelandic*. John Benjamins, Amsterdam/Philadelphia.

Alistair Baron and Paul Rayson. 2008. VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics*. Aston University, Birmingham, UK.

Douglas Biber. 2012. Register as a Predictor of Linguistic Variation. *Corpus Linguistics and Linguistic Theory* 8(1):9–37.

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*, John Benjamins, Amsterdam, pages 11–24.

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, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. Longman, Harlow, UK.

Ido Dagan, Lillian Lee, and Fernando Pereira. 1999. Similarity-based Models of Word Cooccurrence Probabilities. *Machine Learning* 34(1-3):43–69.

Mark Davies. 2008. The Corpus of Contemporary American English: 520 Million Words, 1990-present. Available online at http://corpus.byu.edu/coca/.

Hendrik De Smet. 2008. Functional Motivations in the Development of Nominal and Verbal Gerunds in Middle and Early Modern English. *English Language and Linguistics* 12(1):55–102.

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

Hans-Jürgen Diller, Hendrik De Smet, and Jukka Tyrkkö. 2011. A European Database of Descriptors of English Electronic Texts. *The European English Messenger* 19:21–35.

Teresa Fanego. 2004. On Reanalysis and Actualization in Syntactic Change: The Rise and Development of English Verbal Gerunds. *Diachronica* 21(1):5–55.

Teresa Fanego. 2006. The Role of Language Standardization in the Loss of Hybrid Gerunds in Modern English. In Leiv Egil Breivik, Sandra Halverson, and Kari Haugland, editors, *'These things write I vnto thee...': Essays in Honour of Bjorg Bækken*, Novus Press, pages 93–110.

Bethany Gray and Douglas Biber. 2012. The Emergence and Evolution of the Pattern N + PREP + V-ing in Historical Scientific Texts. In Isabel Moskowich and Begoña Crespo, editors, *Astronomy 'playne and simple'. The Writing of Science between 1700 and 1900*, John Benjamins, Amsterdam, pages 181–198.

John Hale. 2001. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies*. Pittsburgh, volume 2 of *NAACL '01*, pages 159–166.

David Hall, Daniel Jurafsky, and Christopher D. Manning. 2008. Studying the History of Ideas Using Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '08, 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*, Pinter, London, pages 162–177.

M.A.K. Halliday and Ruqaiya Hasan. 1985. *Language, Context, and Text: Aspects of Language*

in a Social-semiotic Perspective. Oxford University Press, Oxford.

Jennifer Hay and Harald Baayen. 2002. Yearbook of Morphology 2001. Springer Netherlands, Dordrecht, chapter Parsing and Productivity, pages 203–235.

Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016. The Royal Society Corpus: From Uncharted Data to Corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

John Lafferty and Chengxiang Zhai. 2001. Document Language Models, Query Models, and Risk Minimization for Information Retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '01, pages 111–119.

Roger Levy. 2008. A Noisy-channel Model of Rational Human Sentence Comprehension under Uncertain Input. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Honolulu, EMNLP '08, pages 234–243.

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

Florent Perek. 2014. Vector Spaces for Historical Linguistics: Using Distributional Semantics to Study Syntactic Productivity in Diachrony. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Baltimore, Maryland, USA, ACL '14, pages 309–314.

Florent Perek. 2016. Using Distributional Semantics to Study Syntactic Productivity in Diachrony: A Case Study. *Linguistics* 54(1):149–188.

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.

Helmut Schmid. 1995. Improvements in Part-of-Speech Tagging with an Application to German. In *Proceedings of the ACL SIGDAT-Workshop*. Kyoto, Japan.

Erika Schulz, Yoon Mi Oh, Zofia Malisz, Bistra Andreeva, and Bernd Möbius. 2016. Impact of Prosodic Structure and Information Density on Vowel Space Size. In *Speech Prosody*. Boston, pages 350–354.

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

Anatol Stefanowitsch and Stefan Th. Gries. 2003. Collostructions: Investigating the Interaction between Words and Constructions. *International Journal of Corpus Linguistics* 8(2):209–243.

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