

Diachronic Variation of Temporal Expressions in Scientific Writing through the Lens of Relative Entropy

Stefania Degaetano-Ortlieb

Saarland University
Saarbrücken
Germany

s.degaetano@mx.uni-saarland.de

Jannik Strötgen

Max Planck Institute for Informatics
Saarland Informatics Campus
Germany

jannik.stroetgen@mpi-inf.mpg.de

Abstract

The abundance of temporal information in documents has led to an increased interest in processing such information in the NLP community by considering temporal expressions. Besides domain-adaptation, acquiring knowledge on variation of temporal expressions according to time is relevant for improvement in automatic processing. So far, frequency-based accounts dominate in the investigation of specific temporal expressions. We present an approach to investigate diachronic changes of temporal expressions based on relative entropy – with the advantage of using conditioned probabilities rather than mere frequency. While we focus on scientific writing, our approach is generalizable to other domains and interesting not only in the field of NLP, but also in humanities.

1 Introduction

Many types of textual documents are rich in temporal information. A specific type of such information are temporal expressions, which again happen to occur in a wide variety of documents. Thus, during the last years, there has been a growing interest in temporal tagging within the NLP community. While variation of temporal expressions according to different domains has become a well established research area (Mazur and Dale, 2010; Strötgen and Gertz, 2012; Lee et al., 2014; Strötgen and Gertz, 2016; Tabassum et al., 2016), variation of temporal expressions according to time within a domain has received less attention so far.¹ Knowing how temporal expressions might have changed over time within a domain is interesting not only in the field

¹Note that here *domain* is defined as a group of documents sharing the same characteristics for the task of temporal tagging (cf. Strötgen and Gertz (2016)).

of NLP, e.g., for adaptation of temporal taggers to different time periods, but also in humanities studies in the fields of historical linguistics, sociolinguistics, and the like.

In this paper, we focus on temporal expressions in the scientific domain and study their diachronic development over a time frame of approx. 350 years (from the 1650s to the 2000s). While here we take an exploratory historical perspective, our findings have implications for improving temporal tagging, especially for recall.

Temporal expressions are related to situation-dependent reference (see notably, Biber et al. (1999)’s work), i.e., linguistic reference to a particular aspect of the text-external temporal context of an event (cf. Atkinson (1999, 120), Biber and Finegan (1989, 492)). While according to Biber et al.’s work, scientific writing has moved towards expressing less situation-dependent reference, to the best of our knowledge, there is no evidence of how this change has been manifested linguistically and whether the types of temporal expressions used in scientific writing have changed over time. To investigate this in more detail, we pose the following questions:

- Do the types of temporal expressions vary diachronically in scientific writing, and if so how is this manifested linguistically?
- What are typical temporal expressions of specific time periods and do these change over time?
- Are different types of temporal expressions, e.g., duration expressions and date expressions referring to points in time, equally affected by a potential change over time?

To process temporal information in scientific research articles, we use HeidelTime (Strötgen and Gertz, 2010), a domain-sensitive tagger to extract and normalize temporal expressions according to

the TimeML standard (Pustejovsky, 2005) for temporal annotation (see Section 4). To detect typical temporal expressions of specific time periods, we use relative entropy, more precisely Kullback-Leibler Divergence (KLD) (Dagan et al., 1999; Lafferty and Zhai, 2001). By KLD we measure how typical a temporal expression is for a time period vs. another time period (see Section 5). The methodology has been adopted from Fankhauser et al. (2016) and successfully used in Degaetano-Ortlieb and Teich (2016) to detect typical linguistic features in scientific writing, Degaetano and Teich (2017) to detect typical features of research article sections, and Degaetano (2017) to observe typical features of social variables.

In the analysis, we inspect general diachronic tendencies based on relative frequency and use relative entropy to investigate more fine-grained changes in the use of temporal expressions over time in scientific writing (see Section 6). On a more abstract level, we observe that the use of temporal information in scientific writing reflects the paradigm change from observational to experimental science (cf. Fankhauser et al. (2016), Glick (2010)) and moves further to descriptions of previous work (e.g., *in the last decades*) in contemporary scientific writing.

2 Related work

Temporal information has been often employed to improve information retrieval (IR) approaches (see Campos et al. (2014) and Kanhabua et al. (2015) for an overview). A prerequisite to exploit temporal information is temporal tagging, i.e., the identification, extraction, normalization, and annotation of temporal expressions based on an annotation standard such as the temporal markup language TimeML (Pustejovsky, 2005). While for quite a long time, temporal tagging was tailored towards processing news texts, in the last years, domain-sensitive approaches are being developed, as it has been shown that temporal information varies significantly across domains (Mazur and Dale, 2009; Strötgen and Gertz, 2016). Domain-sensitive temporal taggers are UWTime (Lee et al., 2014) and HeidelTime (Strötgen and Gertz, 2012). We choose HeidelTime as it is being reported to be much faster than UWTime (Agarwal and Strötgen, 2017).

Recently, there is also an increasing interest in temporal information in the field of digital humanities. An early approach to operationalize time in

narratology has been applied by Meister (2005). Strötgen et al. (2014) show how temporal taggers can be extended for temporal expressions referring to historical dates in the AncientTimes corpus. Fischer and Strötgen (2015) apply temporal taggers to analyze date accumulations in large literary corpora. An analysis of temporal expressions and whether they refer to the future or the past has also been performed on English and Japanese twitter data (Jatowt et al., 2015).

Considering the diachronic aspect of temporal information in scientific writing, it has been mainly investigated by considering temporal adverbs in the context of register studies. Biber and Finegan (1989) and Atkinson (1999), for example, have shown a decrease of temporal adverbs in scientific writing in terms of relative frequencies. Fischer and Strötgen (2015) also studied temporal expressions in a diachronic corpus, but only temporal expressions with explicit day and month information have been considered.

We use temporal tagging tailored at identifying temporal information in scientific writing to obtain a more comprehensive picture of possible diachronic changes. Moreover, besides considering changes in terms of relative frequency, we look at typical temporal expressions and patterns of temporal expressions of specific time periods.

3 Data

As a dataset, we use texts of scientific writing ranging from 1665 to 2007. The first time periods (1665 up to 1869) are covered by the Royal Society Corpus (Kermes et al., 2016a) build from the Proceedings and Transactions of the Royal Society of London – the first periodical of scientific writing – covering several topics within biological sciences, general science, and mathematics. For the later time periods (1966 to 2007), we also use scientific research articles from various disciplines (e.g., biology, linguistics, computer science) taken from the SciTex corpus (Degaetano-Ortlieb et al., 2013; Teich et al., 2013). For comparative purposes, we divide the corpus into fifty year time periods. Table 1 shows the time periods, their coverage and the sub-corpus sizes in number of tokens and documents.

The corpus has been pre-processed in terms of OCR correction, normalization, tokenization, lemmatization, sentence segmentation, and part-of-speech tagging (cf. Kermes et al. (2016b)).

period	coverage	tokens	documents
1650	1665-1699	2,589,536	1,326
1700	1700-1749	3,433,838	1,702
1750	1750-1799	6,759,764	1,831
1800	1800-1849	10,699,270	2,778
1850	1850-1869	11,676,281	2,176
1950	1966-1989	18,998,645	3,028
2000	2000-2007	20,201,053	2,111

Table 1: Corpus details.

4 Processing temporal information

4.1 Temporal expressions

Key characteristics Temporal expressions have three important key characteristics (cf. Alonso et al. (2011), Strötgen and Gertz (2016)). First, they can be *normalized*, i.e., expressions referring to the same semantics can be normalized to the same value. For example, *March 11, 2017* and *the 2nd Saturday in March of this year* point to the same point in time, even though both expressions are realized in different ways. Second, temporal expressions are *well-defined*, i.e., given two points in time X and Y , the relationship between these two points can always be determined, e.g., as X is before Y (cf. Allen (1983)). Third, they can be organized hierarchically on a granularity scale (from coarser to finer granularities and vice versa such as *day*, *month* or *year*). Relevant in our analysis are normalization and granularity. Normalized values are used to compare temporal expressions across time periods instead of considering only the single lexical realizations. In terms of granularity, we consider granularity scales to determine diachronic changes of temporal expressions.

Types According to the temporal markup language TimeML (cf. Pustejovsky et al. (2005)), there are four types of temporal expressions (cf. also Strötgen and Gertz (2016)):

- DATE expressions refer to a point in time of the granularity equal or coarser than ‘day’ (e.g., *March 11, 2017*, *March 2017* or *2017*).
- TIME expressions refer to a point in time of any granularity smaller than ‘day’ (e.g., *Saturday morning* or *10:30 am*).
- DURATION expressions refer to the length of a time interval and can be of different granularity (e.g., *two hours*, *three weeks*, *four years*).

- SET expressions refer to the periodical aspect of an event, describing set of times/dates (e.g., *every Saturday*) or a frequency within a time interval (e.g., *twice a day*).

In the analysis, we consider all these four types showing how their use has changed diachronically in scientific writing.

4.2 Temporal tagging

For temporal tagging we use HeidelbergTime (Strötgen and Gertz, 2010), a domain-sensitive temporal tagger. HeidelbergTime supports normalization strategies for four domains: news, narrative, colloquial, and autonomous. Although HeidelbergTime has been applied to process scientific documents using the autonomous domain, these scientific documents have been very specific, relatively short (biomedical abstracts) with many so-called autonomous expressions (i.e., expressions not referring to real points in time, but to references in a local time frame).

In contrast, our corpus is quite heterogeneous, containing letters and reports in the earlier time periods and full articles in the later time periods. Thus, we expect that most of the documents are written in such a way that the correct normalization of relative temporal expressions can be reached by using the document creation time as reference time. This makes the documents similar to news-style documents according to HeidelbergTime’s domain definitions. Thus, we apply HeidelbergTime with its news domain setting. Note, however, that in our analysis we use only normalized values of DURATION and SET expressions, which are normalized to the length and granularity of an expression but not to an exact point in time. Thus, our findings still hold if some of the occurring temporal expressions are not normalized correctly to a point in time.

HeidelbergTime uses TIMEX3 tags, which are based on TimeML (Pustejovsky et al., 2010), the most widely used annotation standard for temporal expressions. In the following, we briefly explain the *value* attribute of TIMEX3 annotations of DURATION and SET expressions, as we do consider their normalized values for a deeper analysis of the occurring temporal expressions. The value attribute of DURATION and SET expressions contains information about the length of the duration that is mentioned, starting with P (or PT in case of time level durations) followed by a number and an abbreviation of the granularity (e.g., years: Y, month: M, week: W, days: D; hours: H, minutes: M). In addi-

tion, fuzzy expressions are referred to by X instead of precise numbers, e.g., *several weeks* is normalized to P_XW, *monthly* is normalized to XXXX-XX and *annually* to XXXX.

4.3 Extraction quality

For meaningful analysis and substantiated conclusions of temporal expressions in our diachronic corpus, the extraction (and normalization) quality of the temporal tagger should be reliable. Although HeidelTime has been extensively evaluated before on a variety of corpora², our corpus is quite different from standard temporal tagging corpora as it contains scientific documents from multiple scientific fields published across several centuries. Creation of a proper gold standard with manual annotations covering all scientific fields across all time periods would not be feasible in an appropriate time frame. Instead, for a valuable statement of temporal tagging quality on our corpus, determining the correctness of expressions tagged by the temporal tagger would be meaningful.

For this, we use precision, i.e., we randomly sample 250 instances for each time period, and manually validate whether the automatically annotated temporal expressions are correctly extracted.³ Here, we consider correctly extracted instances (RIGHT) and wrongly extracted instances (WRONG). The latter are either cases of ambiguity (e.g., *spring* as ‘season’ or ‘water spring’ or *current* meaning ‘now’ or ‘electric current’) or wrongly assigned temporal expressions to numbers occurring in the text. On top we differentiate correctly assigned but not relevant instances (OTHER) due to noise in the data itself. These are, e.g., temporal expressions assigned to reference sections (especially in the 1950-2000 periods) or used within tables (mostly in the earlier time periods).

Table 2 presents precision information and the number of instances per assigned category of RIGHT, OTHER, and WRONG. We consider the OTHER instances to be correct in terms of precision of extraction. Across periods, precision achieves 0.89 to 0.96.

²E.g., on news articles as in the TempEval competitions (Verhagen et al., 2010; UzZaman et al., 2013) and on Wikipedia articles contained in the WikiWars corpus (Mazur and Dale, 2010).

³We chose to use an amount of instances per period rather than an amount of documents, due to possible sparsity of temporal expressions within documents. This also allows us to validate the same amount of instances across time periods, rather than varying amounts of instances across time periods.

period	RIGHT	OTHER	WRONG	precision
1650	219	13	18	0.928
1700	210	20	20	0.920
1750	218	21	11	0.956
1800	186	37	27	0.892
1850	181	48	22	0.912
1950	116	114	20	0.920
2000	145	96	9	0.964

Table 2: Precision across time periods.

5 Typicality of temporal expressions

To obtain temporal expressions typical of a time period, we use relative entropy, also known as Kullback-Leibler Divergence (KLD) (Kullback and Leibler, 1951) – a well-known measure of (dis)similarity between probability distributions used in NLP, speech processing, and information retrieval. In comparison to relative frequency, i.e., the unconditioned probability of, e.g., a word over all words in a corpus, relative entropy is based on conditioned probability.

In information-theoretic parlance, relative entropy measures the average number of additional bits per feature (here: temporal expressions) needed to encode a feature of a distribution A (e.g., the 1650 time period) by using an encoding optimized for a distribution B (e.g., the 1700 time period). The more additional bits needed, the more distant A and B are. This is formalized as:

$$D(A||B) = \sum_i p(feature_i|A) \log_2 \frac{p(feature_i|A)}{p(feature_i|B)} \quad (1)$$

where $p(feature_i|A)$ is the probability of a feature (i.e., a temporal expression) in a time period A , and $p(feature_i|B)$ the probability of that feature in a time period B . The $\log_2 \frac{p(feature_i|A)}{p(feature_i|B)}$ relates to the difference between both probability distributions ($\log_2 p(feature_i|A) - \log_2 p(feature_i|B)$), 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.

In terms of typicality, the more bits are used to encode a feature, the more typical that feature is for a given time period vs. another time period. Thus, in a comparison of two time periods (e.g., 1650 vs. 1700), the higher the KLD value of a feature for

one time period (e.g., 1650), the more typical that feature is for that given time period. 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).

To compare typical features across several time periods, the most high ranking features of each comparison are considered. For example, for 1650 we obtain six feature sets typical of 1650 as we have six comparisons of 1650 with each of the other six time periods (i.e., a feature set for features typical of 1650 vs. 1700, of 1650 vs. 1750, etc.). If features are shared across feature sets and are high ranking (e.g., in the top 5), these features are considered to be typical of 1650. In other words, these are features ranking high in terms of KLD, significant in terms of p-value, and typical of a time period across all/most comparisons with other time periods. As in our case we consider seven time periods, features are considered typical which rank high for one time period in 6 to 4 feature sets (i.e., typical in more than half of the comparisons).

6 Analysis

In the following, we analyze diachronic tendencies of temporal expressions from the period of 1650 to 2000 in terms of (1) relative frequency (i.e., unconditioned probabilities), and (2) typicality (i.e., conditioned probabilities of expressions in one vs. the other time periods as described in Section 5).

We show how the notion of typicality based on relative entropy leads to valuable insights on the change of temporal expressions in scientific writing w.r.t. more and less frequent expressions.

6.1 Frequency-based diachronic tendencies

Comparing temporal types across fifty years time periods in terms of frequency (see Figure 1 showing log of frequency per million (pM)), DATE is the most frequent type, followed by DURATION. SET and TIME expressions are less frequent. In addition, while DATE remains relatively stable over time, expressions of DURATION, SET and TIME drop quite a bit from 1850 onwards, getting relatively rare.

6.2 Diachronic tendencies of ‘typical’ temporal expressions

Inspecting diachronic change through the lens of relative entropy (as described in Section 5) allows

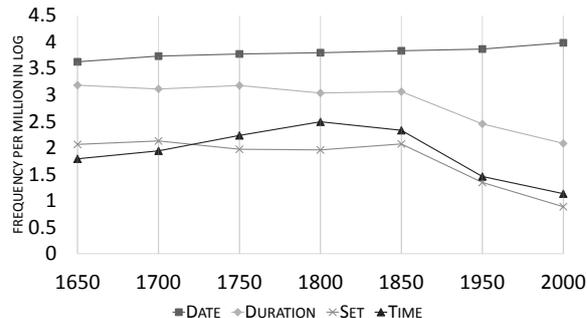


Figure 1: Diachronic tendencies of temporal expression types in scientific writing.

us to consider temporal expressions typical of one time period when compared to the other time periods. We study each type of temporal expression and carefully select the base of comparison.

DATE Considering DATE expressions, instead of comparing single dates (which mostly occur only once in the corpus, such as *June, 3, 1769*), we take a level of abstraction and consider part-of-speech (POS) sequences of annotated DATE expressions to better inspect the types of changes that might have occurred over time. For each DATE expression, we extract POS sequences and use relative entropy to detect typical POS sequences of temporal expressions for each time period.

period	POS sequence	example	comp.
1650	DT NN	<i>in the Spring</i>	5
	NP	<i>in Winter</i>	5
	RB	<i>now</i>	4
1700	NP	<i>in Summer</i>	5
	NP CD	<i>March 8</i>	5
	DT JJ IN NP	<i>the 6th of March</i>	4
1750	NP CD , CD	<i>June 3, 1769</i>	6
	NP CD	<i>April 19</i>	5
	CD NP	<i>2 June</i>	4
	DT NN	<i>the Spring</i>	4
1800	NP CD , CD	<i>June 18, 1784</i>	5
1850	CD	<i>in 1858</i>	4
1950	JJ	<i>current work</i>	5
	JJ JJ NN	<i>mid seventeenth century</i>	4
	DT JJ NNS	<i>the last decades</i>	5
2000	DT NNS	<i>the 1990s</i>	5
	JJ JJ NN	<i>late seventeenth century</i>	5

CD: cardinal number, DT: determiner, IN: preposition, JJ: adjective, NN: sing. common noun, NNS: pl. common noun, NP: sing. proper noun, RB: adverb

Table 3: Typical POS sequences of DATE.

Table 3 shows POS sequences typical of one time period vs. 6-4 other time periods (see column comp.)⁴. For example, for 1650 the POS sequences DETERMINER-NOUN (DT-NN) and PROPER NOUN (NP) are quite typical, which are all temporal expression referring to seasons in terms of lexical realizations (see Example 1 and 2). Both POS sequences are typical of 1650 vs. 1750 to 2000 (i.e., 5 comparisons). If we consider the POS sequences that are typical across time periods and their lexical realizations, there seems to be a development in terms of specificity and interval (see Figure 2).

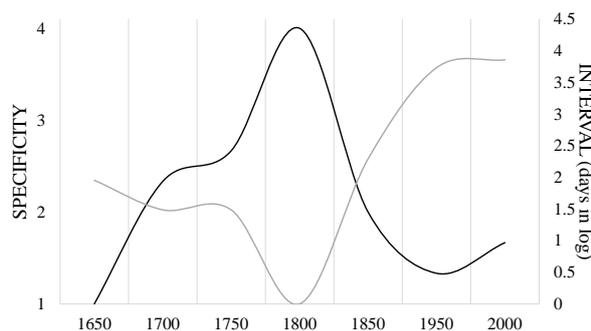


Figure 2: Specificity (black) and interval (gray) of typical DATE expressions.

To capture the notion of specificity, we consider how many pieces of temporal information are given by a POS sequence to make a temporal expression most specific, with a scale from 1 to 4, where 1 is least specific (e.g., NP denoting seasons such as *Winter* as we do not know of which year etc.) and 4 is most specific (e.g., NP CD, CD such as *June 3, 1769* which gives us an exact date)⁵. For comparison across time periods, Figure 2 shows the average of the specificity count over all typical POS sequences of a time period (black line). For the interval of typical DATE expressions, the amount

⁴Note that the examples in Table 3 show most frequent realizations for relatively generic expressions such as seasons (e.g., *in the Spring*) or examples taken randomly from the corpus for specific dates (e.g., *June 3, 1769*) as they occur only once.

⁵On this scale, 1 denotes the mere occurrence of a temporal expression (e.g., NP such as *Winter*), 4 denotes the mere expression plus the inclusion of day, month and year (e.g., NP CD, CD such as *June 3, 1769*, which is a temporal expression + day + month + year resulting in 4 points in total), and 2 and 3 denote an occurrence of a temporal expression plus either two or three combinations of day, month and year (e.g., CD such as *1858* for 2, i.e., temporal expression + year, and NP CD such as *March 8* for 3, i.e., temporal expression + month + day).

of days⁶ the expressions refer to is used (shown in log in Figure 2, gray line). The more specific an expression becomes, the smaller the interval it refers to and vice versa.

Figure 2 also shows how temporal expressions move from relatively unspecific (e.g., *in the Spring* in 1650) to very specific (*June 18, 1784* in 1800) and back to unspecific expressions (e.g., *the last decades* in 2000). The interval moves instead from a wider to a smaller span and back to a wider span in 2000. Investigating the contexts, in which these expressions arise, gives further insights. While in the early time periods, season mentioning is typical, from 1800 to 1850, temporal expressions are typical with exact date, year or month expressions. These expressions are used to present exact dates of observations made by a researcher at several points in time, especially in the field of astronomy (see Example 3). From the 1950 onwards, typical DATE expressions become less explicit, relating to broader (e.g., *the 1970s* in Example 4) and less specific (e.g., *current literature* in Example 5) temporal reference. These expressions are used, e.g., in the context of previous work descriptions in introduction sections of research papers.

- (1) *In Winter it will need longer infusion, than in the Spring or Autumn.* (1650)
- (2) *The difference between these two plants is this; the papaver corniculatum dies to the root in the winter, and sprouts again from its root in the spring;* (1750)
- (3) **March 4, 1783.** *With a 7-feet reflector, I viewed the nebula near the 5th Serpentis, discovered by Mr. MESSIER, in 1764.* (1750)
- (4) *In the 1970s, Rabin [38] and Solovay and Strassen [44] developed fast probabilistic algorithms for testing primality and other problems.* (2000)
- (5) *There is a significant confusion in the current literature on “cellular” or “tessellation arrays” concerning the concept of a “Garden-of-Eden configuration”.* (1950)

TIME To investigate typical TIME expressions, similarly to DATE expressions, we consider their POS sequences (see Table 4). It can be seen that only the intermediate time periods show typical expressions (1750 to 1850). In terms of granularity, in the period of 1750, expressions are less granular pointing to broader sections of a day (e.g., *morning, evening*) mostly used to describe observations made (see Example 6). In the 1850 period, expressions point to specific hours of a day (e.g., *9 A.M.*) mostly in descriptions of experiments.

⁶We use ‘day’ because DATE expressions refer to a point in time of the granularity ‘day’ or coarser, i.e., ‘day’ is the smallest unit.

period	POS sequence	example	comp.
1750	NP NN	<i>Sunday morning</i>	5
	JJ NN	<i>next morning</i>	5
1800	CD NN	<i>10 A.M.</i>	5
	CD NN	<i>7 A.M.</i>	5
1850	DT NN IN DT	<i>the evening of the</i>	4
	JJ IN NP	<i>28th of August</i>	
	IN CD NN	<i>about 8 A.M.</i>	4

CD: cardinal number, DT: determiner, IN: preposition, JJ: adjective, NN: sing. common noun, NP: sing. proper noun

Table 4: Typical POS sequences of TIME.

- (6) *Monday morning* she appeared well, her pulse was calm, and she had no particular pain. (1750)
- (7) *There being usually but one assistant*, it was impossible to observe during the whole twenty-four hours; the hours of observation selected were therefore from *3 A.M. to 9 P.M. inclusive*. (1850)

DURATION For DURATION, we consider their TIMEX3 value, as it directly encodes normalized information on the duration length and granularity of temporal expressions. Figure 3 shows typical TIMEX3 values (e.g., PID for expressions such as *one day*) of specific time periods⁷. The y-axis shows the duration length in seconds on a log scale. In general, duration length gets lower from 1750 to 1850 (with expressions of seconds and hours, which are more granular) and higher in 1950 and 2000 (with expressions of decades, which are less granular).

We then again consider the contextual environments of these typical expressions. In the earlier time periods (1650 and 1700), day and year expressions are typical, mostly relating to observations (see Example 8) or experiment descriptions (see Example 9).

- (8) *After the eleven Months, the Owner having a mind to try, how the Animal would do upon Italian Earth, it died three days after it had changed the Earth.* (1650)
- (9) *the Opium, being cut into very thin slices, [...] is to be put into, and well mixed with, the liquor; (first made luke-warm) and fermented with a moderate Heat for eight or ten Days, [...].* (1650)

From the period of 1750 to 1950, duration length is relatively low with expressions of seconds, minutes and hours being typical of these time periods.

⁷Note also that typical expressions can either be relatively explicit (e.g., PID for *24 hours*) or fuzzy (indicated by an X in the TIMEX3 value, e.g., *few hours* for the value PTXH).

These expressions are mainly related to observations (see Example 10) in the 1750 period and experiment descriptions from 1800 to 1950 (see Examples 11 to 13).

- (10) *June 4, the weather continued much the same, and about 9h 30 in the evening, we had a shock of an earthquake, which lasted about four seconds, and alarmed all the inhabitants of the island.* (1750)
- (11) [...] *the glass produced by this fusion was in about twelve hours dissolved, by boiling it in a proper quantity of muriatic acid.* (1800)
- (12) *In a few hours* a mass of fawn-coloured crystals was deposited; (1850)
- (13) *The patient is then switched to the re-breathing system containing 133 Xenon at 5 mCi/l for a period of one minute, and then returned to room air for a period of ten minutes.* (1950)

In the 1950, besides weeks and minutes, related to experiment descriptions (see Example 14), expressions of decades are typical. The latter is also true for the 2000. In both periods, expressions relating to decades refer to previous work (see Examples 16 and 15).

Thus, DURATION shifts from being used for purposes of observational to experimental science and finally to previous work references in the latest time periods.

- (14) *For each speaker, performance was observed across numerous repetitions of the vocabulary set within a single session, as well as across a 2-week time period.* (1950)
- (15) *It constitutes the usual drift-diffusion transport equation that has been successfully used in device modeling for the last two decades.* (1950)
- (16) *Provably correct and efficient algorithms for learning DNF from random examples would be a powerful tool for the design of learning systems, and over the past two decades many researchers have sought such algorithms.* (2000)

SET For SET expressions again their TIMEX3 value is considered. Figure 4 shows typical expressions with the times per year of a SET expression on the y-axis in log⁸, mirroring also less granular (*annually*) and more granular (*every day*) expressions. As we have seen from Figure 1, SET expressions are relatively rare in scientific writing and strongly decrease over time (see Section 6.1). This is also reflected in the few temporal expressions typical of each time period in Figure 4. In terms of granularity, there is a shift from day (see Example 17) to

⁸For example, *every day* corresponds to 365 times a year, while *annually* to once a year.

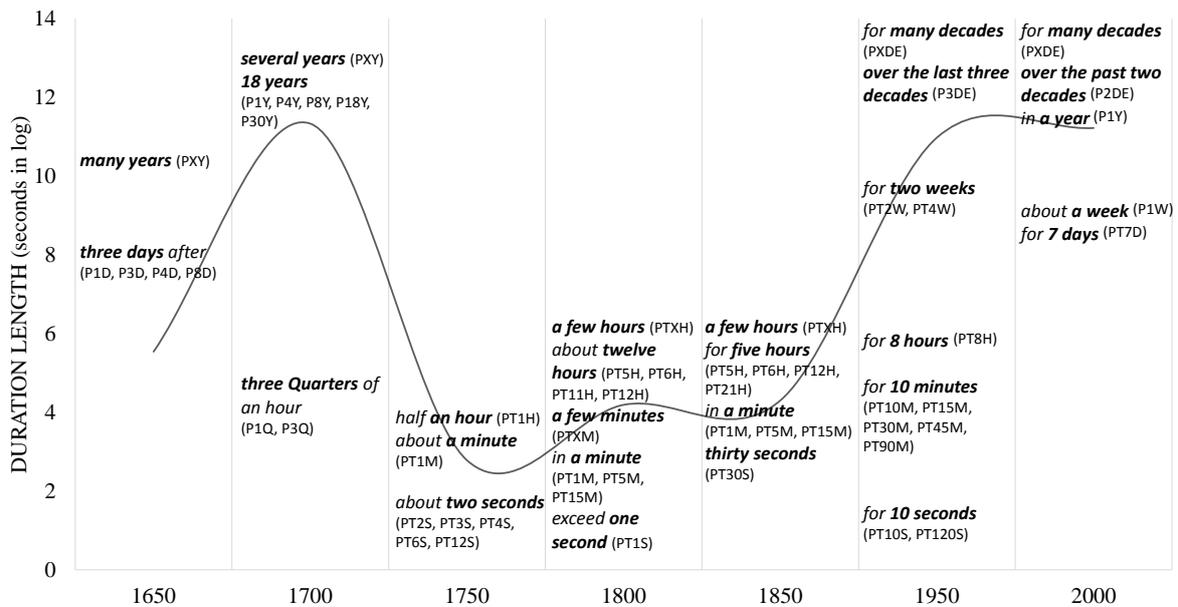


Figure 3: Diachronic tendencies of typical DURATION expressions.

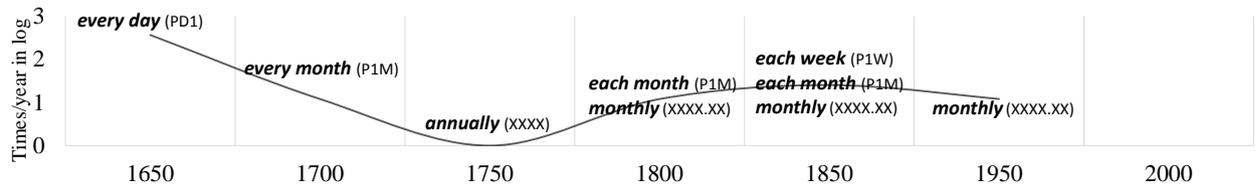


Figure 4: Diachronic tendencies of typical SET expressions in scientific writing.

month expressions (see Examples 18 to 19). Interestingly, for the latter, there has been a move from a noun phrase expression (*every/each month*) to an adverb expression (*monthly*). While, in the intermediate periods (1800 and 1850) both expressions are typical, in 1950 only *monthly* is typical. In 1750 to 1850, *every/each month* expressions relate to observations done on a monthly basis of which the mean or average is drawn (see Example 18) and the same applies for *monthly* used with *mean* as a term (see Example 19). In 1950, instead, *monthly* is solely used as an adverb. Thus, there is a replacement of longer noun phrase expressions (*every/each month*) by the shorter adverb expression *monthly*.

- (17) *Besides this, you may there see, that every day the Sun sensibly passes one degree from West to East , [...].* (1650)
- (18) *In order to determine the annual variations of the barometer, I have taken the mean of the observations in each month, [...].* (1800)
- (19) *The mean was then taken in every month of every lunar hour (attending to the signs), and the monthly means were collected into yearly means.* (1850)

- (20) *A disk resident file of all current recipient numbers is created monthly from the eligibility tape file supplied by Medical Services Administration.* (1950)

7 Discussion and Conclusion

We have presented an approach to investigate diachronic change in the usage of temporal expressions. First, we use temporal tagging to obtain a more comprehensive coverage of possible temporal expressions, rather than investigating specific expressions only, as was the case in previous work. Evaluation of the tagging results showed high precision (approx. 90%) across time periods.

Second, we use relative entropy to detect typical temporal expressions of specific time periods. A clear advantage to frequency-based accounts is that with relative entropy frequent as well as rare phenomena can be investigated in terms of their ‘typicality’ according to a variable (here: temporal expressions typical of specific time periods). Apart from gaining knowledge on diachronic changes specific to different types of temporal expressions, we also capture more abstract and more fine-grained

shifts. On a more abstract level, while our findings confirm the paradigm shift from the more observational to the more experimental character of scientific writing (cf. Fankhauser et al. (2016), Gleick (2010)) for DATE and DURATION expressions, we also show the tendency towards previous work descriptions for these two temporal types in contemporary scientific writing. On a more fine-grained level, for SET (a rarely used temporal type especially towards more contemporary time periods), there is a linguistic shift from longer noun-phrase to shorter adverb expressions.

These findings are not only interesting in historical linguistic terms, but are also relevant to improve adaptation of temporal taggers to different time periods. Especially for recall, gold-standard annotations are needed. Since this is a quite resource and time consuming task, our approach can help in gaining insights on the use of typical temporal expressions in specific contexts across periods. These contexts can then be further exploited in terms of possible temporal expression occurrences to achieve better recall. In addition, temporal expressions might change in terms of linguistic realization as with the SET type in our case. Accounting for shifts in linguistic realization will also improve recall. While this is true for diachronic variation, the approach also generalizes to domain-specific variation. In future work, we plan to work in this direction, further elaborating our methodology for diachronic and domain variation.

8 Acknowledgments

This work is partially funded by *Deutsche Forschungsgemeinschaft* (DFG) under grant SFB 1102: *Information Density and Linguistic Encoding* (www.sfb1102.uni-saarland.de). We are also indebted to Stefan Fischer for his contributions to corpus processing and Elke Teich for valuable improvement suggestions. Also, we thank the anonymous reviewers for their valuable comments.

References

Prabal Agarwal and Jannik Strötgen. 2017. Tiwiki: Searching Wikipedia with Temporal Constraints. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17), Companion Volume*, pages 1595–1600.

James F. Allen. 1983. Maintaining Knowledge about Temporal Intervals. *Communications of the ACM*, 26(11):832–843.

Omar Alonso, Jannik Strötgen, Ricardo Baeza-Yates, and Michael Gertz. 2011. Temporal Information Retrieval: Challenges and Opportunities. In *Proceedings of the 1st International Temporal Web Analytics Workshop (TAWAW '11)*, pages 1–8.

Dwight Atkinson. 1999. *Scientific Discourse in Socio-historical Context: The Philosophical Transactions of the Royal Society of London, 1675-1975*. Routledge, New York.

Douglas Biber and Edward Finegan. 1989. Drift and the Evolution of English Style: A History of Three Genres. *Language*, 65(3):487–517.

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

Ricardo Campos, Gaël Dias, Alípio M. Jorge, and Adam Jatowt. 2014. Survey of Temporal Information Retrieval and Related Applications. *ACM Computing Surveys*, 47(2):15:1–15:41.

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

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 (LaTeCH '16)*, pages 165–173. ACL.

Stefania Degaetano-Ortlieb and Elke Teich. 2017. Modeling Intra-textual Variation with Entropy and Surprisal: Topical vs. Stylistic Patterns. In Beatrice Alex, Stefania Degaetano-Ortlieb, Anna Feldman, Anna Kazantseva, Nils Reiter, and Stan Szpakowicz, editors, *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH and CLfL '17)*, pages 68–77. ACL.

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, Tübingen.

Stefania Degaetano-Ortlieb. 2017. Variation in Language Use across Social Variables: A Data-driven Approach. In *Proceedings of the Corpus and Language Variation in English Research Conference (CLAVIER '17)*.

- Peter Fankhauser, Jörg Knappen, and Elke Teich. 2016. Topical Diversification over Time in the Royal Society Corpus. In *Proceedings of Digital Humanities (DH '16)*.
- Frank Fischer and Jannik Strötgen. 2015. When Does (German) Literature Take Place? On the Analysis of Temporal Expressions in Large Corpora. In *Proceedings of Digital Humanities (DH '15)*.
- James Gleick, 2010. *Seeing Further. The Story of Science and The Royal Society*, chapter At the Beginning: More Things in Heaven and Earth, pages 17–36. Harper Press, New York.
- Adam Jatowt, Émilien Antoine, Yukiko Kawai, and Toyokazu Akiyama. 2015. Mapping Temporal Horizons: Analysis of Collective Future and Past Related Attention in Twitter. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15)*, pages 484–494.
- Nattiya Kanhabua, Roi Blanco, Kjetil Nørvåg, et al. 2015. Temporal Information Retrieval. *Foundations and Trends in Information Retrieval*, 9(2):91–208.
- Hannah Kermes, Stefania Degaetano-Ortlieb, Ashraf Khamis, Jörg Knappen, and Elke Teich. 2016a. The Royal Society Corpus: From Uncharted Data to Corpus. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC '16)*, pages 1928–1931. ELRA.
- Hannah Kermes, Jörg Knappen, Ashraf Khamis, Stefania Degaetano-Ortlieb, and Elke Teich. 2016b. The Royal Society Corpus: Towards a High-quality Corpus for Studying Diachronic Variation in Scientific Writing. In *Proceedings of Digital Humanities (DH '16)*.
- Solomon Kullback and Richard A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.
- 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 (SIGIR '01)*, pages 111–119. ACM.
- Kenton Lee, Yoav Artzi, Jesse Dodge, and Luke Zettlemoyer. 2014. Context-dependent Semantic Parsing for Time Expressions. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL '14)*, pages 1437–1447. ACL.
- Paweł Mazur and Robert Dale. 2009. The DANTE Temporal Expression Tagger. In *Proceedings of the 3rd Language and Technology Conference (LTC)*, pages 245–257. Springer.
- Paweł Mazur and Robert Dale. 2010. WikiWars: A New Corpus for Research on Temporal Expressions. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*, pages 913–922. ACL.
- Jan Christoph Meister. 2005. Tagging Time in Prolog: The Temporality Effect Project. *Literary and Linguistic Computing*, 20:107–124.
- James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An International Standard for Semantic Annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC '10)*, pages 394–397. ELRA.
- James Pustejovsky. 2005. Temporal and Event Information in Natural Language Text. *Language Resources and Evaluation*, 39(2-3):123–164.
- Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High Quality Rule-based Extraction and Normalization of Temporal Expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pages 321–324. ACL.
- Jannik Strötgen and Michael Gertz. 2012. Temporal Tagging on Different Domains: Challenges, Strategies, and Gold Standards. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)*, pages 3746–3753. ELRA.
- Jannik Strötgen and Michael Gertz. 2016. *Domain-sensitive Temporal Tagging*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.
- Jannik Strötgen, Thomas Bögel, Julian Zell, Ayser Armiti, Tran Van Canh, and Michael Gertz. 2014. Extending HeidelTime for Temporal Expressions Referring to Historic Dates. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC '14)*, pages 2390–2397. ELRA.
- Jeniya Tabassum, Alan Ritter, and Wei Xu. 2016. TweepTime: A Minimally Supervised Method for Recognizing and Normalizing Time Expressions in Twitter. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP '16)*, pages 307–318. ACL.
- Elke Teich, Stefania Degaetano-Ortlieb, Hannah Kermes, and Ekaterina Lapshinova-Koltunski. 2013. Scientific Registers and Disciplinary Diversification: A Comparable Corpus Approach. In *Proceedings of 6th Workshop on Building and Using Comparable Corpora (BUCC '13)*, pages 59–68. ACL.
- Naushad UzZaman, Hector Llorens, Leon Derczynski, James F. Allen, Marc Verhagen, and James Pustejovsky. 2013. SemEval-2013 Task 1: TempEval-3: Evaluating Time Expressions, Events, and Temporal Relations. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval '13)*, pages 1–9. ACL.

Marc Verhagen, Roser Saurí, Tommaso Caselli, and James Pustejovsky. 2010. SemEval-2010 Task 13: TempEval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval '10)*, pages 57–62. ACL.