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

Stefania Degaetano-Ortlieb and Jannik Strötgen

Saarland University and Max Planck Institute for Informatics

Saarbrücken

IDeaL SFB 1102

UNIVERSITÄT DES SAARLANDES

max planck institut informatik

Background & Motivation

- Growing interest in NLP community for temporal tagging
- Awareness of domain variation of temporal expressions

(cf. Mazur&Dale 2010, Strötgen&Gertz 2016, Lee et al. 2014, Tabassum et al. 2016)

As language is a dynamic construct,

what about variation of temporal expressions according to time?

Background & Motivation

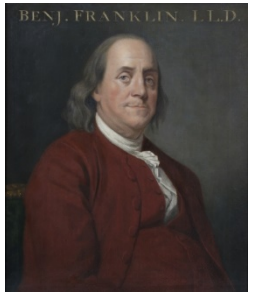
- Temporal expressions belong to situation-dependent reference (cf. Atkinson 1999, Biber&Finegan 1989, Biber et al. 1999)
- Scientific writing moves towards less use of situation-dependent reference (cf. Biber et al. 1999)
- BUT: So far, evidence based only on temporal adverbs for temporal expressions

→ Use temporal tagging for a more comprehensive coverage

Examples

*SIR, To perform **now** the promise I made you **the other day** […] by giving you an Account of what I tried on **Tuesday night last Octob. 29 1667** […] about the Relation between Air and Light, as this is to be found in some Bodies.* (**Robert Boyle**)

***Yesterday** I compared it with a fix foot Telescope, and found it not only to magnify more, but also more distinctly*. (**Isaac Newton**)



*Then I fully charged two six-gallon glass jars […] and I sent the united shock of these thro' the affected limb or limbs; repeating the stroke commonly **three times each day**.* (**Benjamin Franklin**)

*Germany, I have taken the opportunity of his absence to sweep in the neighbourhood of the sun , in search of comets ; and **last night**, the **1st of August**, **about 10 o'clock** , I found an object […] resembling in colour and brightness the 27th nebula of the Connoissance des Temps […].* (**Caroline Herschel**)

https://de.wikipedia.org/wiki/Caroline_Herschel
https://royalsociety.org/about-us/history/
https://de.wikipedia.org/wiki/Robert_Boyle

Research question

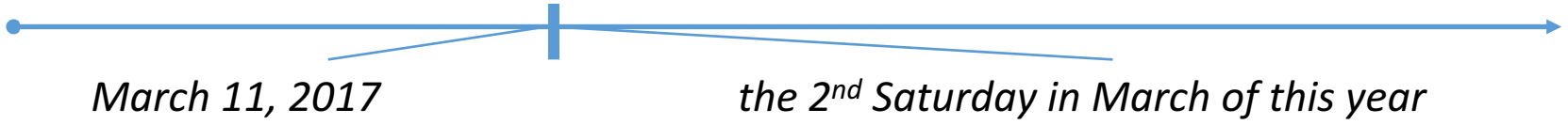
What are typical temporal expressions in scientific writing?

Do they change over time?

Temporal expressions

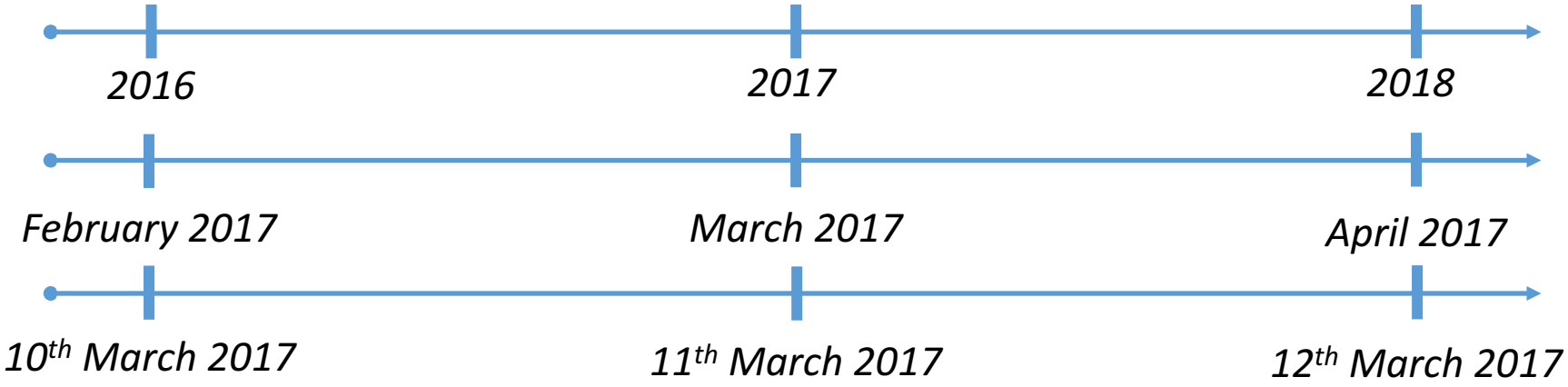
Key characteristics (cf. Alonso et al. 2011, Strötgen&Gertz 2016)

- Can be normalized to the same value



- Are well-defined

- Organized hierarchically on a granularity scale (from coarser to finer and vice versa)



Temporal expressions

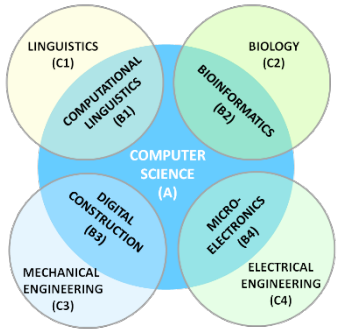
Types according to TimeML (temporal markup language; cf. Pustejovsky et al. 2005)

- DATE: >= day (e.g., *March 11, 2017*, *March 2017* or *2017*)
- TIME: < day (e.g., *Saturday morning* or *10:30 am*)
- DURATION: length of interval of different granularity (e.g., *two hours*, *three weeks*)
- SET: a set of times/dates (e.g., *every Saturday*) or frequency within a time interval (e.g., *twice a day*)

Data

Royal Society Corpus (Kermes et al. 2016) 1665-1869

SciTex Corpus (Degaetano-Ortlieb et al. 2013) 1966-2007



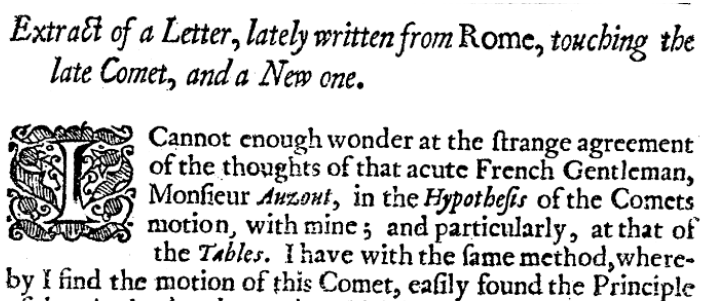
- Approx. 73 million tokens
- OCR corrected, normalized and linguistically pre-processed
- Different scientific fields
- Division into 50 year time periods

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

Methods: Annotation

HeidelTime (Strötgen&Gertz 2010)

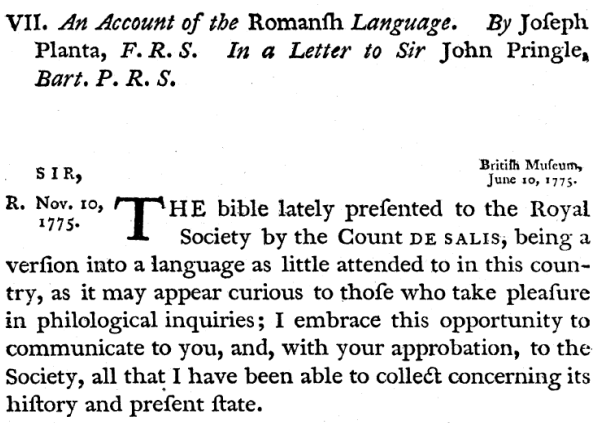
- Domain-sensitive temporal tagger (news, narrative, colloquial, autonomous)



Extract of a Letter, lately written from Rome, *touching the late Comet, and a New one.*

I Cannot enough wonder at the ſtrange agreement of the thoughts of that acute French Gentleman, Monſieur *Auzout*, in the *Hypotheſis* of the Comets motion, with mine; and particularly, at that of the *Tables*. I have with the ſame method, where-by I find the motion of this Comet, eaſily found the Principle

1665



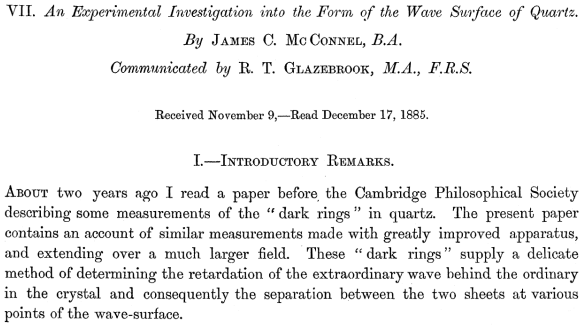
VII. *An Account of the* Romanſh *Language. By* Joſeph Planta, *F. R. S. In a Letter to Sir* John Pringle, *Bart. P. R. S.*

SIR, British Muſeum, June 10, 1775.

R. Nov. 10, 1775.

THE bible lately preſented to the Royal Society by the Count DE SALIS, being a verſion into a language as little attended to in this country, as it may appear curious to thoſe who take pleaſure in philological inquiries; I embrace this opportunity to communicate to you, and, with your approbation, to the Society, all that I have been able to collect concerning its hiſtory and preſent ſtate.

1775



VII. *An Experimental Investigation into the Form of the Wave Surface of Quartz.*

By JAMES C. MCCONNEL, *B.A.*

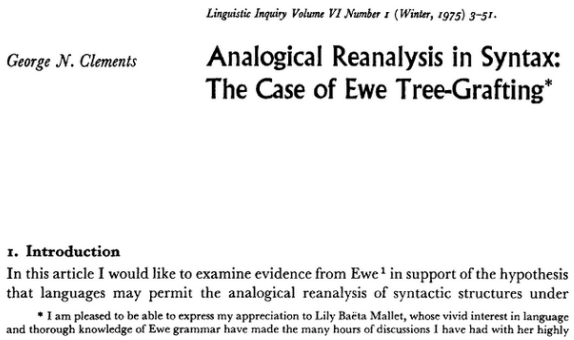
Communicated by R. T. GLAZEBROOK, *M.A., F.R.S.*

Received November 9,—Read December 17, 1885.

I.—INTRODUCTORY REMARKS.

ABOUT two years ago I read a paper before the Cambridge Philosophical Society describing some measurements of the "dark rings" in quartz. The present paper contains an account of similar measurements made with greatly improved apparatus, and extending over a much larger field. These "dark rings" supply a delicate method of determining the retardation of the extraordinary wave behind the ordinary in the crystal and consequently the separation between the two sheets at various points of the wave-surface.

1875



Linguistic Inquiry Volume VI Number 1 (Winter, 1975) 3–51.

George N. Clements

Analogical Reanalysis in Syntax: The Case of Ewe Tree-Grafting*

1. Introduction

In this article I would like to examine evidence from Ewe[1] in support of the hypothesis that languages may permit the analogical reanalysis of syntactic structures under

* I am pleased to be able to express my appreciation to Lily Baëta Mallet, whose vivid interest in language and thorough knowledge of Ewe grammar have made the many hours of discussions I have had with her highly

1975

- Providing normalization and TIMEX3 value tagging

Methods: Basis of comparison

- DATE and TIME:
  Part-of-speech realization due to rel. uniqueness of expressions
  (e.g., `CD`: *1985*, `DT-JJ-IN-NP`: *the 6th of March*)

- DURATION and SET: TIMEX3 tagging value
  - `P`: length of duration + number + abbreviation of different granularity
    (e.g., `P1D`: *1 day*, `P1W`: *1 week*)
  - `PT`: for time level durations (e.g., `PT12H`: *12 hours)*

- For all four consider also linguistic realization in context

Methods: Extraction quality

- Creation of gold-standard by manual annotation not feasible in reasonable time and with appropriate coverage
- Evaluation based on precision on a random sample of 1750 instances (250 per period)

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

WRONG:

- ambiguity *spring* as *season* or *water spring*
  current as now or electric current
- wrongly assigned numbers

OTHER:

- correctly assigned but not relevant due to noise in the data (e.g. references, tables)

Methods: Typicality

Difference by relative entropy (Kullback-Leibler Divergence)

$$D_{KL}(A||B) = -\sum_i p(unit_i|A) \log_2 \frac{p(unit_i|A)}{p(unit_i|B)}$$

- $unit$ = any linguistic unit (e.g. word, parts of speech, temporal expression)
- Probability distributions A and B = time periods
  A = 1650 B = 1700
  A = 1650 B = 1750
  …

Typicality of a $unit$ by feature ranking based on KLD

- High KLD = more typical
- Significance testing by unpaired Welch's t-test
- High ranking in 6-4 comparisons

Analyses

A1: General diachronic tendency

Does the amount of situation-dependent reference drop over time considering temporal expressions? (frequency-based)

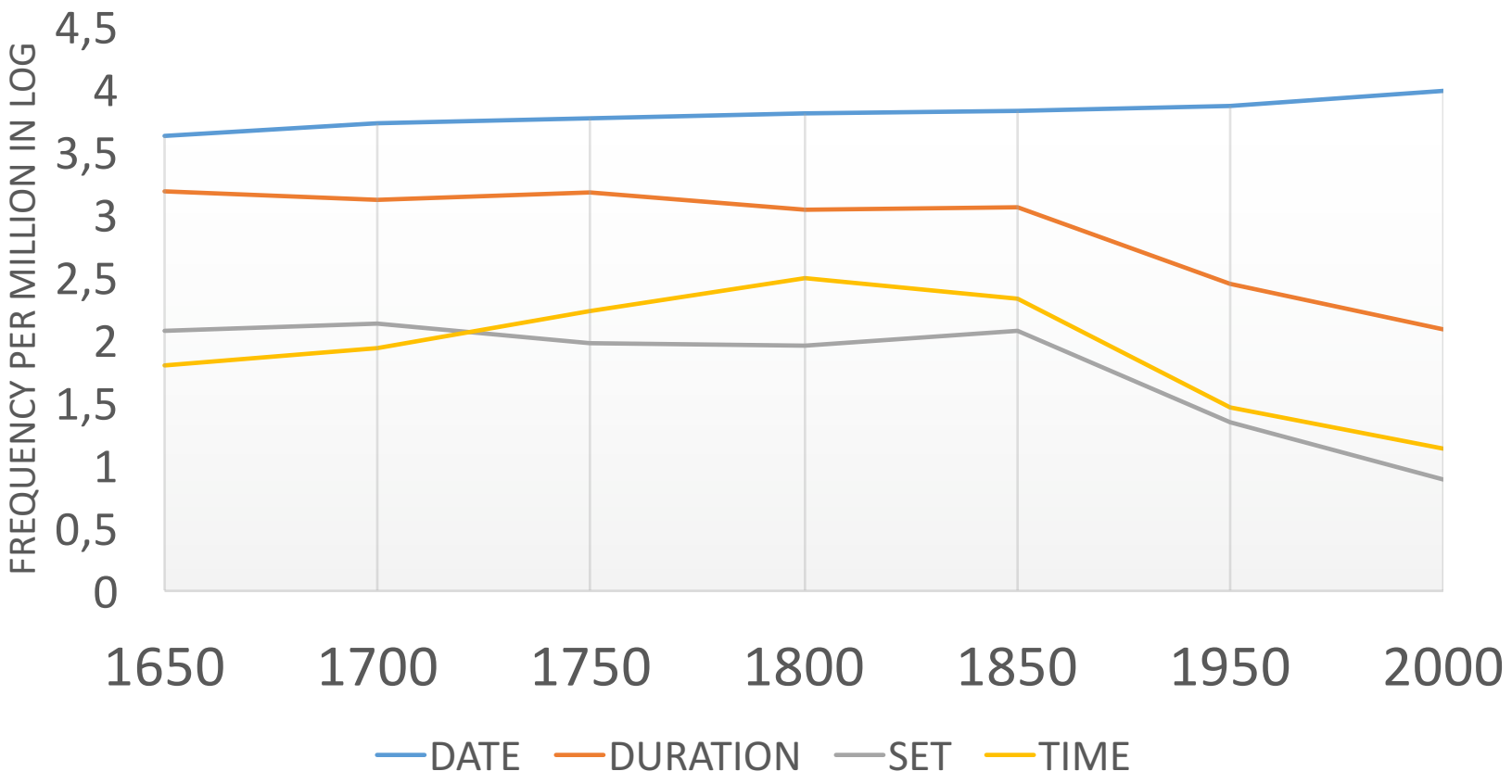
A2: Typical temporal expressions over time

Are there specific temp. expressions typical of particular periods?

Do these change over time?

If so, how do they change and are they equally affected by a potential change?

Frequency distribution of temp. expressions



DATE remains relatively stable

DURATION, SET, TIME decrease

DATE expressions

Frequent POS sequences across periods (top 5)

1650	1700	1750	1800	1850	1950	2000
CD	CD	CD	CD	CD	CD	CD
RB	NP	NP CD	RB	RB	JJ	RB
NP	RB	RB	NP CD	NP	RB	JJ
NP CD	NP CD	NP	NP	NP CD	NP CD	NP CD
DT NN	DT NN	NP CD , CD	NP CD , CD	JJ	NP	NP

- Specific years (CD; *1667*, *1795*, *1996*)
- Adverbs (RB; *now, recently*)
- Month and day (NP CD; *March 6*, *April 2*)

→ Used across periods

- Seasons (NP, DT NN; *Winter*, *in the Spring*)
- Full date in 1800 and 1850 (NP CD , CD; *June 3, 1769*)
- Adjectives from 1850 onwards (JJ): *current, recent*

→ Some preference in particular time periods

DATE expressions

Typicality of POS sequences across periods

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

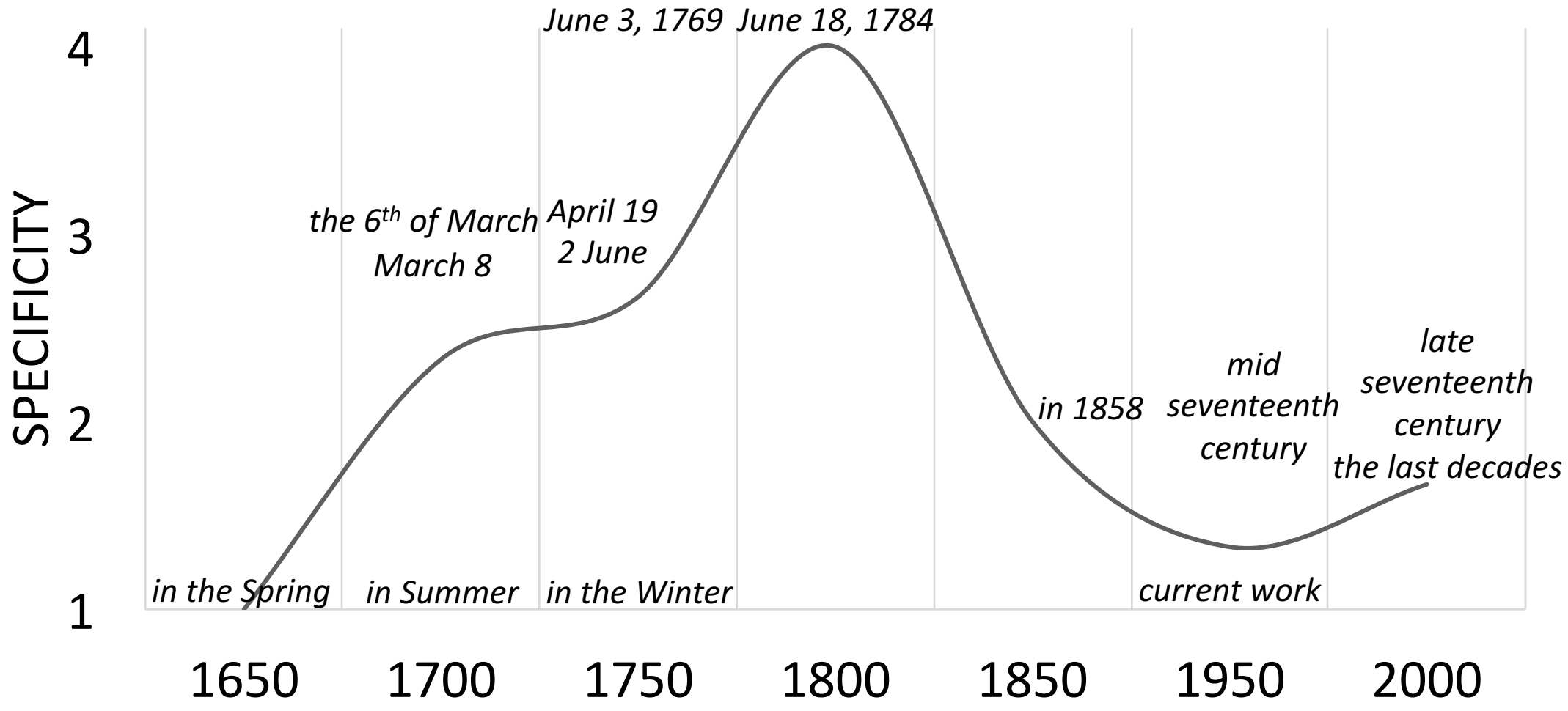
→ Very specific POS sequences

→ Not limited to high frequency items

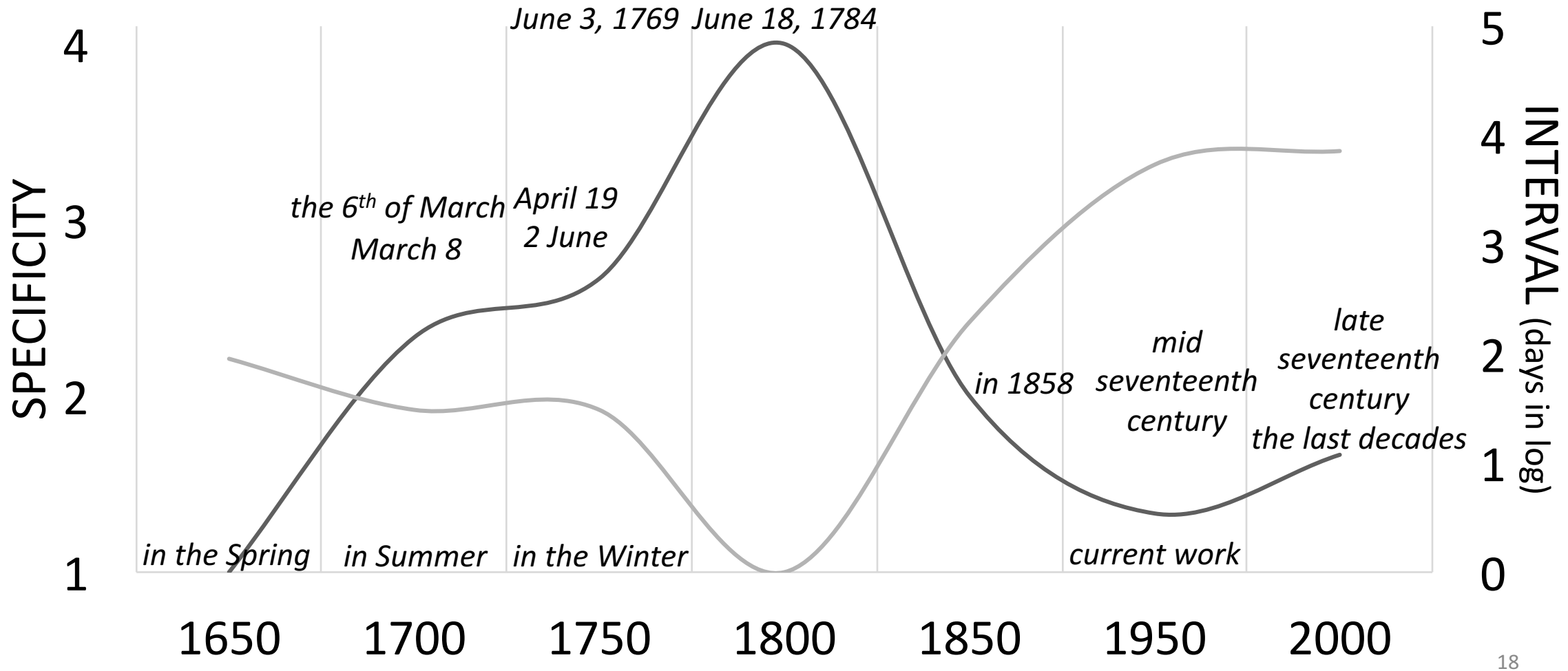
→ Changes seem related to

- Specificity (higher vs. lower)
- Interval (bigger vs. smaller)
- Contextual usage of temporal expressions

Specificity of DATE over time



Specificity vs. *Interval* of DATE over time



Contextual usage of DATE

Observational

Reference to seasons

*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)

Exact dates of observations made by a researcher (mostly astronomy)

***March 4, 1783**. With a 7-feet reflector, I viewed the nebula near the 5th Serpentis, discovered by Mr. MESSIER, in 1764*. (1750)

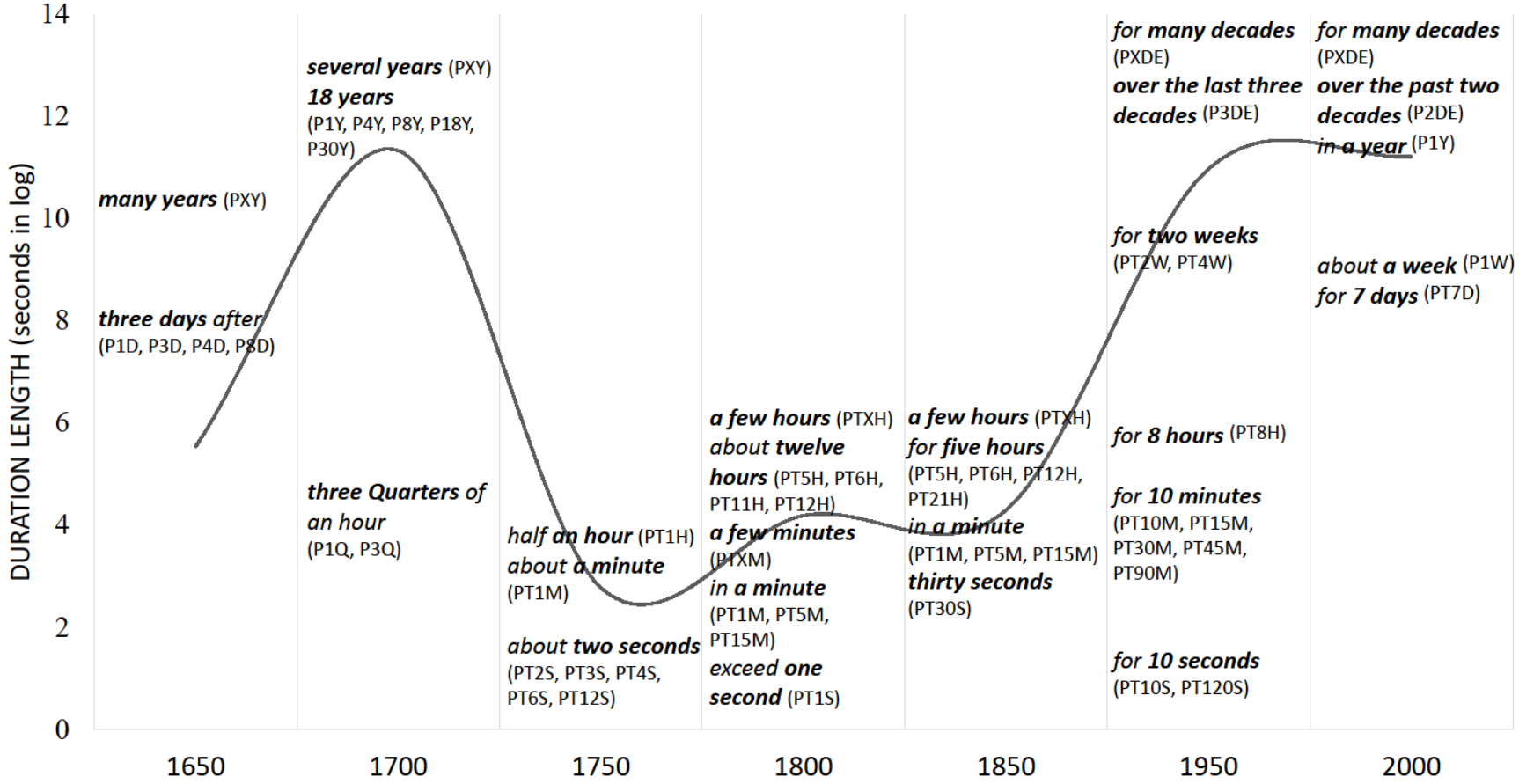
Retrospective

Previous work in introductions

*In **the 1970s**, Rabin [38] and Solovay and Strassen [44] developed fast probabilistic algorithms ….* (2000)

DURATION expressions

Typical TIMEX3 values across periods



- Duration length changes from shorter to longer
- Change in contextual usage

Contextual usage of DURATION

Observational / Experimental

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)

The Opium, [...] is to be put into […] the liquor, (first made luke-warm) and fermented with a moderate Heat for ***eight or ten Days****,* [...]. (1650)

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)

In a few hours *a mass of fawn-coloured crystals was deposited;* (1850)

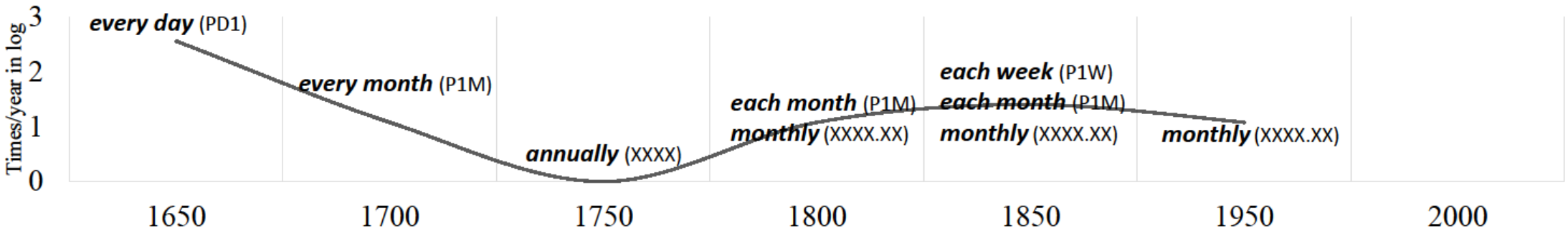
Retrospective

It constitutes the usual drift-diffusion transport equation that has been successfully used in device modeling for ***the last two decades****.* (1950)

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 expressions

Typical TIMEX3 values across time



- Few typical expressions (rarely used in scientific writing)
- Month expressions become most typical over time
- Lexical realization changes towards a more compact form

Conclusion

Temporal tagging

- Application to diachronic data with promising results (high precision ~90%)
- Wider coverage of temporal information in scientific writing (beyond adverbs)

Relative entropy vs. frequency

- Changes in *typical* features across time
- Application to different levels of abstraction (POS, TIMEX3 values)
- Perception of different kinds of changes

1. Contextual usage change: observational to experimental to retrospective
2. Fine-grained linguistic changes: nominal to adverbial style

Envoi

- Zoning of temporal expressions in (structured) documents

  i.e., based on typical expressions and where they occur in e.g. scientific articles
  choose material for gold-standard creation to improve recall
- Capture (more/less conventionalized) linguistic realizations by considering $\mathrm{P}(\mathrm{tempexp}|\mathrm{context})$
- Application to domain-specific variation

THANK YOU!

QUESTIONS, COMMENTS, HINTS?

TIME expressions

Typical POS sequences

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

→ Only for intermediate time periods (rarely used in scientific writing)

→ Some trends towards change of granularity from high to low

Monday morning *she appeared well, her pulse was calm, and she had no particular pain.* (1750)

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)

Most frequent DATE expressions over time

1650		1700		1750		1800		1850		1950		2000	
pos	freq	pos	freq	pos	freq	pos	freq	pos	freq	pos	freq	pos	freq
CD	3234	CD	5359	CD	10824	CD	27478	CD	38601	CD	103486	CD	170279
RB	2178	NP	3070	NP CD	5887	RB	7791	RB	7725	JJ	10946	RB	8673
NP	1763	RB	2983	RB	5647	NP CD	7736	NP	7100	RB	10709	JJ	7622
NP CD	841	NP CD	2821	NP	3946	NP	5540	NP CD	6334	NP CD	5393	NP CD	2921
DT NN	445	DT NN	526	NP CD , CD	3439	NP CD , CD	4577	JJ	5967	NP	2407	NP	1328

- Specific years (CD) prevail across time: *1667*, *1795*, *1996*
- Adverbs (RB) in top 3 across time: *now, recently*
- Month and day (NP CD) in top 5 across time: *March 6*, *April 2*

→ Used across periods

- Months or seasons (NP, DT NN): *June*, *in the Spring*
- Full date (NP CD , CD) in 1800 and 1850: *June 3, 1769*
- Adjectives (JJ) from 1850 onwards: *current, recent*

→ Some preference in particular time periods