Simple statistics for DH

Václav Cvrček

June 8, 2022

CZECH NATIONAL CORPUS

Introduction

CZECH NATIONAL CORPUS

Overview

1. Why quantification matters?
2. Types of frequency measures
3. Why dispersion matters too?
4. How to interpret frequencies properly
5. Comparing frequencies
6. Comparing texts/corpora
    - KWords
    - QuitaUp

Quantification in DH research

Empirical research = quantification

- implicit quantification in qualitative research
  - studies use vague quantifiers like “often”, ”rarely”, “regularly” etc. in order to establish wider relevance of the findings (ten Have 2007: 158)
- evaluating and comparing phenomena
- quantitative analysis is *not* just aggregating single instances (analysis “built on the back of qualitative single case analysis”, Schegloff 1993: 102)
  - quantitative analysis has its own exploratory potential and can uncover hidden phenomena

Frequency

CZECH NATIONAL CORPUS

Type and token

Difference: *token* (instance) and *type* (class)

> *one and one and one is three*

…7 tokens, 4 types

- types have overall characteristics that tokens lack (e.g. frequency)

Frequency in description

- frequency characteristics of phenomena in
  - lexicon (Čermák & Křen 2011)
  - grammar (Cvrček et al. 2010)
  - …
- psycholinguistics
- cognitive lingvistics
- text interpretation (DH)

Frequency and its distribution in texts

Zipf's laws – George Kingsley Zipf (1902–1950)

- statistical properties of texts as a structure (empirical)
- balance between language economy and distinctiveness (unifying and diversifying forces)
- three types of relations (laws):
  1. $f \times r = k$, where r is rank of a word with frequency f
  2. $a_f \times f^2 = k$, where a is number of words with frequency f
  3. $\frac{m}{\sqrt{f}} = k$, where m is number of meanings of a word and f is its frequency

How to measure frequency

CZECH NATIONAL CORPUS

Types of frequencies

- **raw** (absolute) frequency – *poison* in DraCor: 84 hits
  - useless if it is an isolated information
  - something to compare with, e.g. *venom*, the play, the whole corpus
- **relative** frequency (w.r.t. whole)
  - instead of count we get a concentration (chemistry)
  - number of instances per million, thousand, per cent (ipm, wpt, %…)
  - $ipm = fq/N \times 1000000$
  - comparable between texts/corpora
  - which play has the highest *poison* concentration?
- **adjusted** frequency (cf. Gries 2008)
  - frequencies do not take into account dispersion, sometimes words come in bursts
  - *Romeo* vs. *hour* – are they spread evenly?
  - measures of dispersion vs. adjusted frequencies
  - ARF (Savický-Hlaváčová, 2002): average number of corpus parts that contain a word

How to interpret frequency

CZECH NATIONAL CORPUS

How to interpret frequency?

- raw/relative frequency is a *point* estimate
- would the frequency be the same in a similarly compiled corpus of the same size?
- binomial model (coin flipping) – content vs. function words

> *O, teach me how I should forget to think!*

> *C C F F F F C F C …4 content, 5 function*

- *interval* estimate is more adequate ⇒ **binomial confidence intervals**
- korpus.cz/calc (poison: 84 hits in 1M corpus)

What a zero means?

Is there a difference between frequency of 1 (hapax legomenon) and 0?

- 0 = the phenomenon is not attested in the data (we cannot be sure it does not exist!)
- 1 and more = we have a proof that it exists but no generalizations are possible (might not be attested in other corpora)

Line selection in KonText and evaluation of groups in Calc.

Comparing frequencies

CZECH NATIONAL CORPUS

How to compare frequency?

- statistical tests: chi2, log-likelihood, fisher (exact) test
- significance is not relevance (effect size)
- Calc modules:
  - 2 words in 1 corpus
  - 2 words in 2 corpora
- *poison* vs. *venom* in DraCor
- lemmas: *father* vs. *son* in DraCor

Keywords – comparing texts/corpora

Keywords:

- prominent units showing what the text is about (topic, register)
- words with unexpectedly high frequency in a text in comparison to reference corpus
- based on statistical tests
- KWords app: https://kwords.korpus.cz/

Texts: lines by Romeo and Juliet from R&J

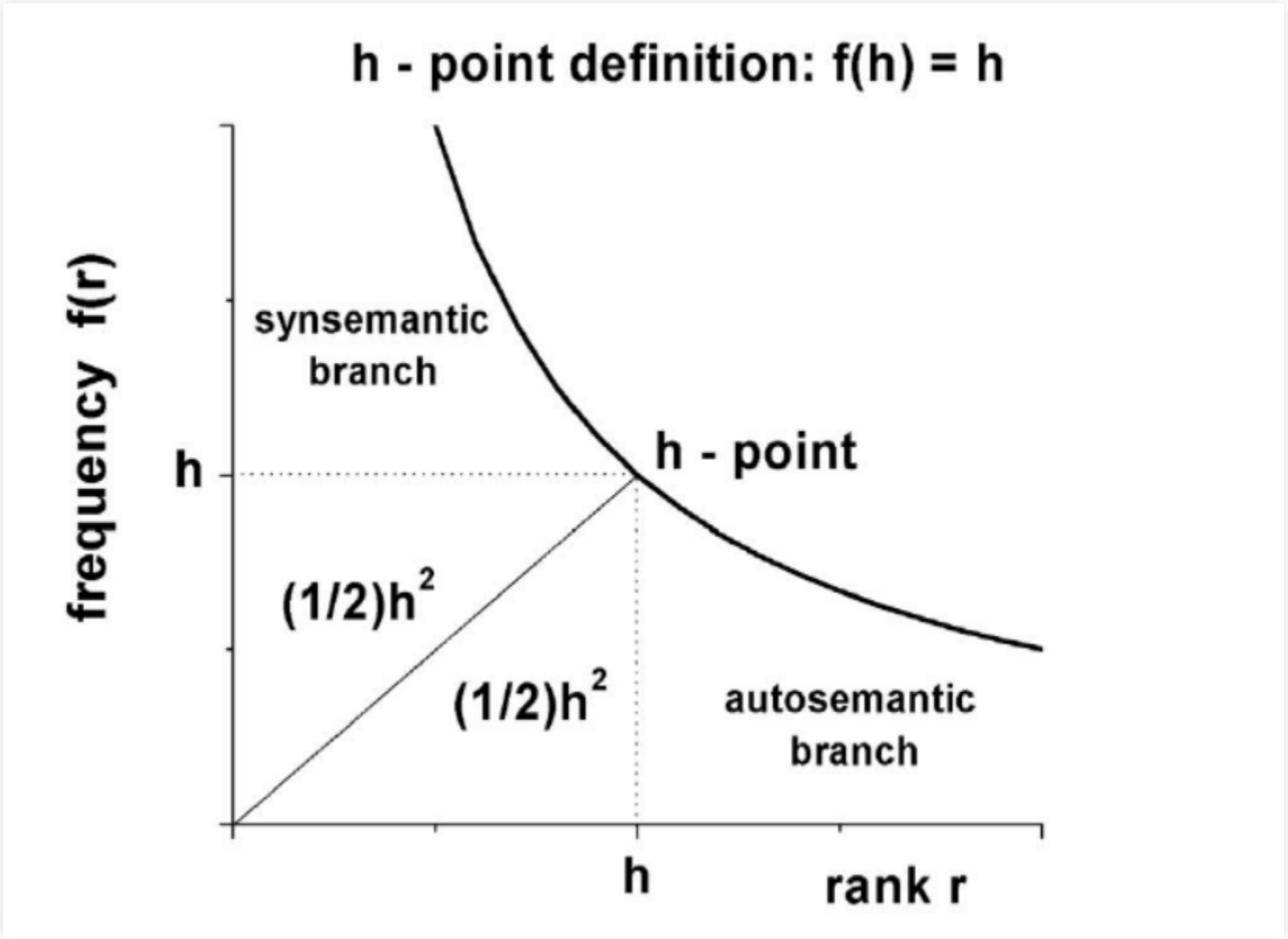
workshop cloud > Vaclav Cvrcek > `romeo.txt` and `juliet.txt`

Comparing texts

CZECH NATIONAL CORPUS

Thematic concentration

Different approach to prominent units (without a reference corpus)



h-point is sensitive to text length (Popescu et al. 2009)

Frequency-based indices

- TTR = type-token ration – sensitive to text length $TTR = \frac{V(N)}{N}$
- better alternatives:
  - Moving Average TTR: $MATTR = \frac{\sum_{i=1}^{N-L} V_i}{L(N-L+1)}$
  - Normalized TTR: $zTTR = \frac{TTR - Med_{TTR}}{IQR_{TTR}}$
- Entropy (H) – vocabulary diversity measure

$$H = \log_2 N - \frac{1}{N} \sum_{r=1}^{V} f_r \log_2 f_r$$

  - Activity (Q) – degree of action of a text $Q = \frac{V}{V+A}$
  - Descriptivity (D) – degree of description in a text $D = 1 - Q$

Application: QuitaUp

Texts: workshop cloud > Vaclav Cvrcek > `romeo.txt` and `juliet.txt`

References

- Cvrček, V. a kol.: Mluvnice současné češtiny. Nakladatelství Karolinum, Praha 2010.
- Čech, R., Kubát, M. (2018) Morphological Richness of Text. In Fidler, M., Cvrček, V. (eds.). Taming the Corpus. From Inflection and Lexis to Interpretation. Springer, 63-77.
- Čermák, F., Křen, M. (eds): A Frequency Dictionary of Czech: Core Vocabulary for Learners. Routledge, London 2011
- Gries, S. T. 2008. “Dispersions and adjusted frequencies in corpora”. International Journal of Corpus Linguistics, 13 (4), 403–437.
- ten Have, P. (2007). Doing Conversation Analysis. SAGE.
- Savický, P. & J. Hlaváčová: Measures of Word Commonness. In Journal of Quantitative Linguistics 9, 2002, 215–231. (https://www.tandfonline.com/doi/abs/10.1076/jqul.9.3.215.14124)
- Schegloff, E.A. (1993) Reflections on quantification in the study of conversation. Research on Language and Social Interaction 26: 99–128.
- Popescu, I. I. et al. (2009). Word frequency studies. Berlin / New York: Mouton de Gruyter.