Semantics, Similarity, and Corpus Search in the *Deutsches Textarchiv*

Bryan Jurish
jurish@bbaw.de

2. DTA-/CLARIN-D-Konferenz und CLARIN-D-Workshop
Textkorpora in Infrastrukturen für die Geistes- und Sozialwissenschaften
Berlin-Brandenburgische Akademie der Wissenschaften
17th–18th November 2014

Overview

The Big Picture

- Semantics
- Similarity
- Corpus search

Tools

- Thesauri & WordNets
- DTA SemCloud

Use Cases

- Similarity-based recommendations in DTAQ
- Lexical ambiguity and semantic search filtering
- History of ideas: topic histograms

Summary & Outlook

The Big Picture

Semantics: study of ***meaning***

- **lexical** semantics: word meanings
- **compositional** semantics: phrase- & sentence-meanings *(ignored here)*
- **ontological** relations: synonymy, hypernymy, hyponymy, ...

(semantic) Similarity

- **thesauri/wordnets**: manually encoded relations
- **distributional** similarity: shared contexts ⇒ similar semantics

> *"You shall know a word by the company it keeps"* —J. R. Firth

Text-Corpus Search

- manual browsing ⇝ topic-based navigation
- linguistic/lexicographic research ⇝ semantic pruning
- "distant reading" ⇝ topic retrieval

Tools: Thesauri

Semantic Relations

- **Synonymy** (∼ *"is-equivalent"*, ...)
- **Hyponymy** (∼ *"is-subordinate"*, *"species-of"*, *"is-subsumed-by"*, ...)
- **Hypernymy** (∼ *"is-superordinate"*, *"genus-of"*, *"subsumes"*, ...)

GermaNet

(Lemnitzer & Kunze, 2002)

- http://www.sfs.uni-tuebingen.de/GermaNet
- developed at Universität Tübingen, 1997–present
- available at no charge for academic use
- **Example:** http://kaskade.dwds.de/germanet/?q=Schloss

OpenThesaurus

(Naber, 2005)

- http://www.openthesaurus.de
- crowd-sourced community project
- free open-source license (LGPL)
- **Example:** http://kaskade.dwds.de/openthesaurus/?q=Schloss

Tools: DTA SemCloud

Distributional Semantics *(Berry, Dumais, & O'Brien, 1995)*

- track "latent" word associations (shared contexts)
- model lexical semantics as high-dimensional vector space (bag-of-words)

... in the DTA

- modern lemmata via DTA::CAB *(Jurish, 2013)*
- content filter via PoS-tags (STTS)
- vector space modelling with Perl & PDL *(Glazebrook & Economou, 1997)*

DTA SemCloud http://kaskade.dwds.de/dtasem

- exploratory interface $\rightsquigarrow$ "tag-cloud" visualization
- k-nearest neighbor search over any of three levels:
  - terms (modern lemmata) `Sinn` *(term→pages)*
  - pages ("documents") `page=frege_sinn_1892.0030` *(page→books)*
  - books ("categories") `book=frege_sinn_1892` *(book→terms)*

Use Case: Similarity-based Recommendations

Idea: facilitate topic-based "lateral" (inter-work) browsing

- suggest other items of potential interest based on ***current item*** only
- straightforward use of SemCloud *(page→pages)* and *(book→books)* queries

Example: Pages http://deutschestextarchiv.de/dtaq/book/view/frege_sinn_1892/?p=30

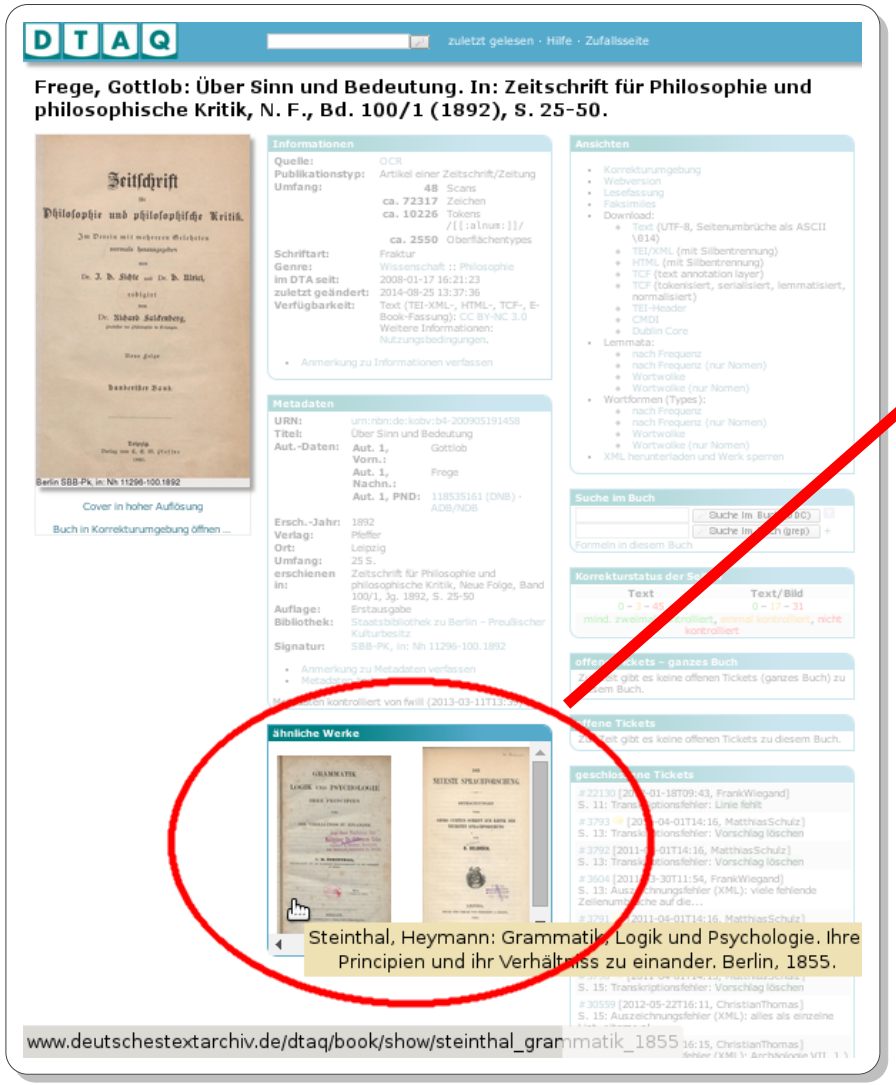
Frege, Gottlob: *Über Sinn und Bedeutung*. In: Zeitschrift für Philosophie und philosophische Kritik, N. F., Bd. 100/1 (1892), S. 25-50. (facs. #30)

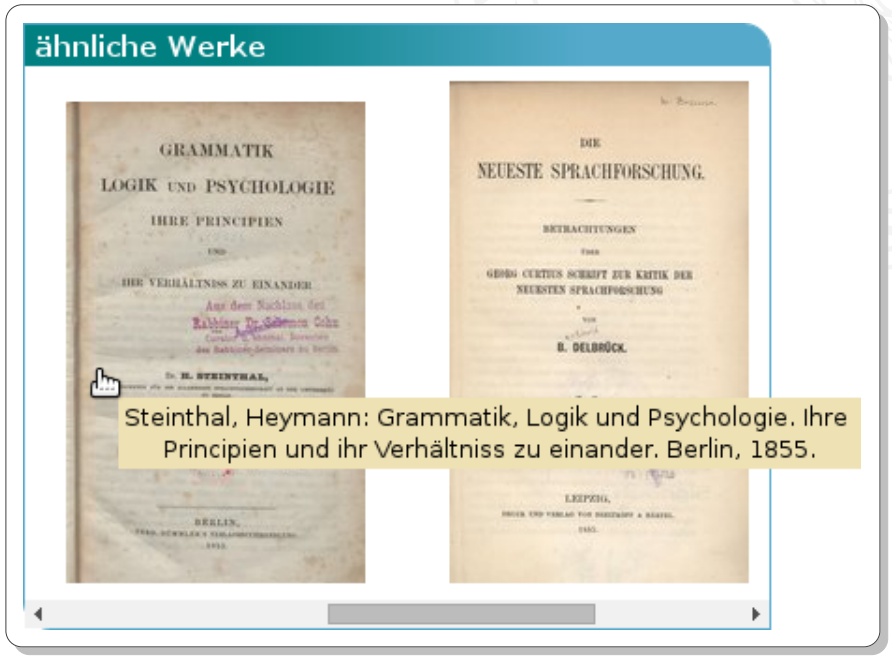
hegel_logik02_1816.0092

Hegel, Georg Wilhelm Friedrich: *Wissenschaft der Logik*. Bd. 2. Nürnberg, 1816. (facs. #92)

Use Case: Similarity-based Recommendations: by Book

http://deutschestextarchiv.de/dtaq/book/view/frege_sinn_1892/





- Schuchardt, H.: *Ueber dei Lautgesetze. Gegen die Junggrammatiker*. Berlin, 1885.
- Lambert, J. H.: *Neues Organon*. Bd. 2. Leipzig, 1764.
- Steinthal, H.: *Grammatik, Logik und Psychologie. Ihre Principien und ihr Verhältniss zu einander*. Berlin, 1855.
- ...

Use Case: Lexical Ambiguity Resolution

Idea: filter corpus search results by topic

- allow use of thesaurus relations and distributional similarity
- implementation uses DDC *external term expansion* mechanism

Examples: *Flügel* ["(bird's) wing"]

- naïve search: `Flügel`
- similar terms: `Flügel Vogel|sem`
- similar documents (books): `Flügel #has[docsim,'Vogel@100']`
- GermaNet hyponyms: `Flügel Vogel|gn-sub`
- OpenThesaurus hyponyms: `Flügel Federvieh|ot-sub`
- similar terms wrt. hyponyms: `Flügel Federvieh|ot-sub|sem`

Examples: *Flügel* ["grand piano"]

- similar terms: `Flügel Klavier|sem`
- similar books: `Flügel #has[docsim,'Musik@100']`
- proximity: `near(Flügel, Musikinstrument|gn-sub without Flügel, 5)`

Use Case: Topic Histograms

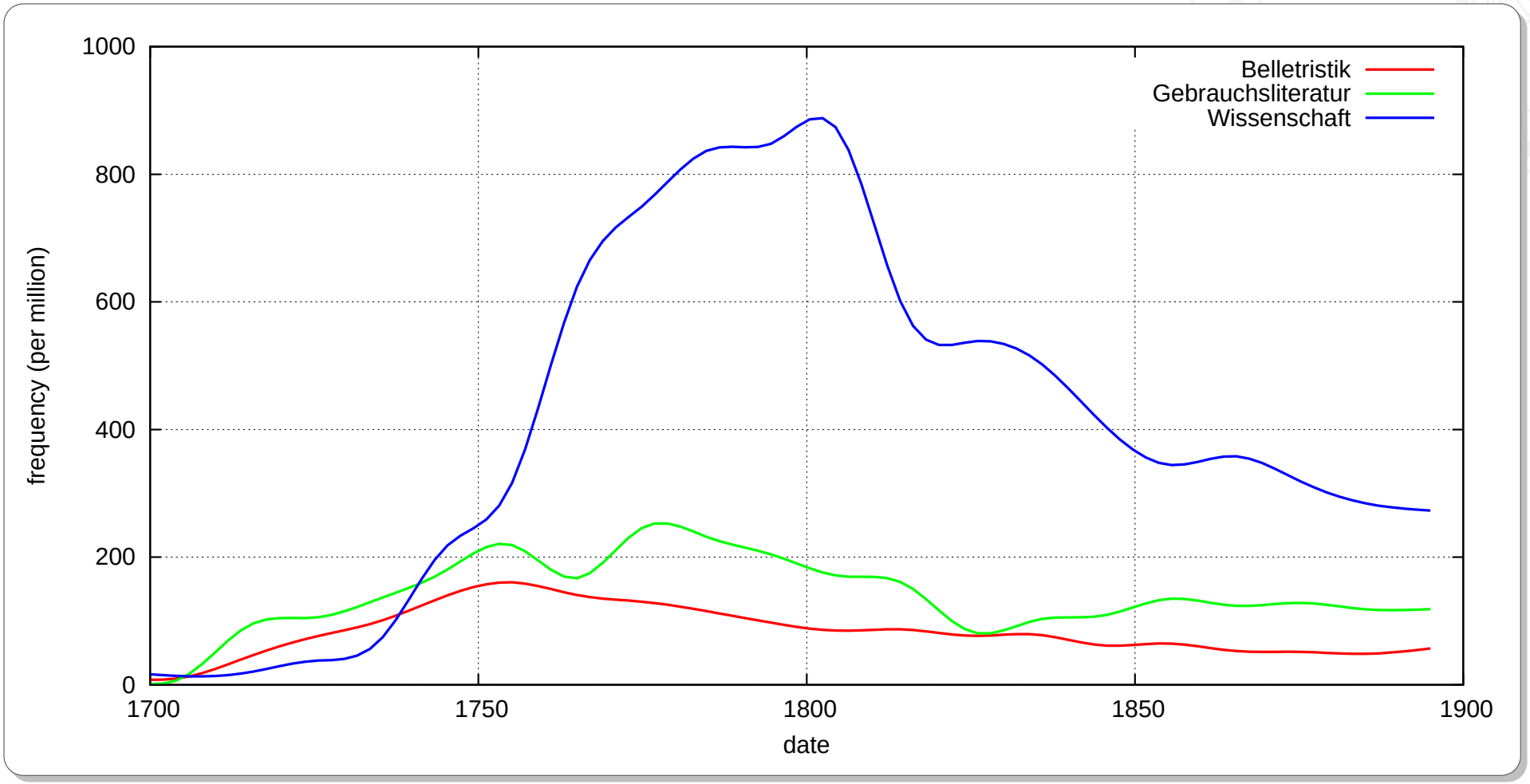
Idea: Exploit distributional similarity for "distant reading"

- approximate tracking of ***discourse topics*** over time
- implementation uses DDC histogram function (version $\geq$ 2.0.23)

Example: 25 characteristically Kantian terms `book=^kant_@25`

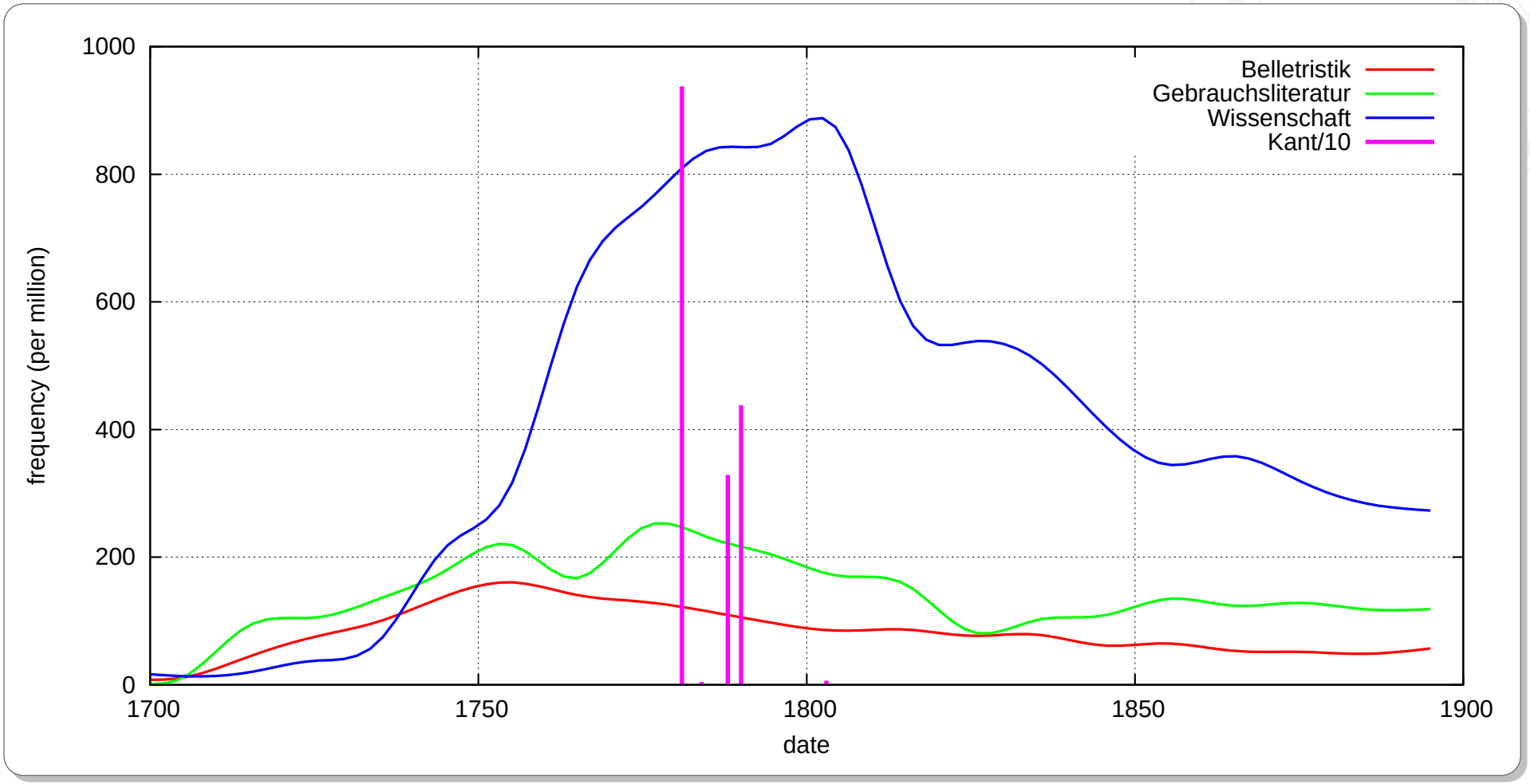
Example: Kantian terms

```
'book=^kant_@25'|sem
```



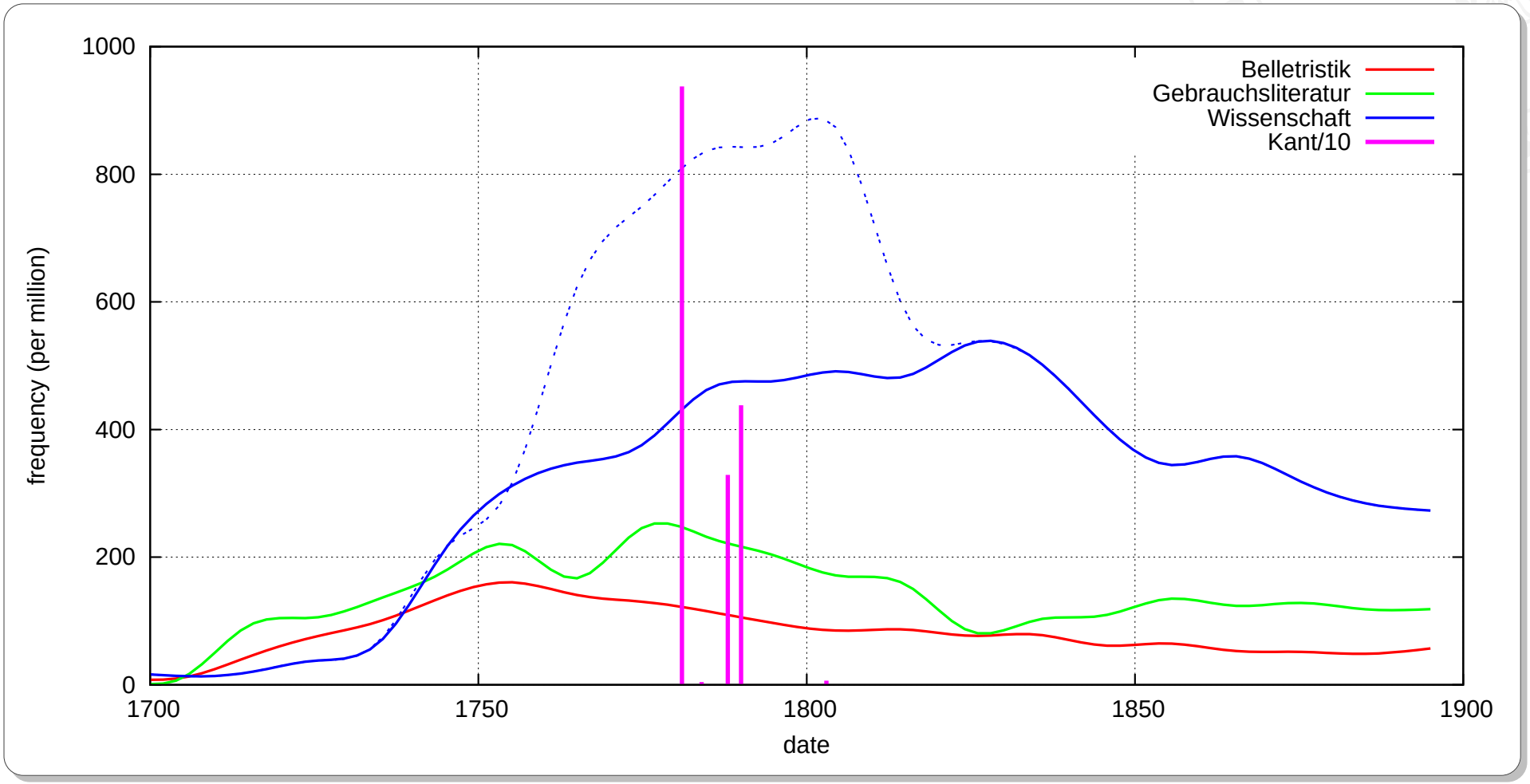
Example: Kantian terms ... in Kant

```
'book=^kant_@25'|sem #has[author,/Kant/]
```



Example: Kantian terms . . . in other authors' work

```
'book=^kant_@25'|sem !#has[author,/Kant/]
```



Summary & Outlook

"Semantic" Search

- initial exploration ⇝ *distributional tag clouds*
- ontological relations ⇝ *manual thesauri*
- corpus research ⇝ *term expansion*

Use Cases

- similarity-based recommendations ⇝ *lateral navigation*
- lexical ambiguity resolution ⇝ *semantic filtering*
- distant reading ⇝ *topic histograms*

Future Work

- document clustering & genre classification
- improve distributional models, scalability (e.g. Blei et al., 2003)
- improve accessibility for typical tasks ⇝ *which ones?*
- ...? ⇝ *suggestions?*

— The End —

Thank you for listening!