



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
- **ontological** relations: synonymy, hypernymy, hyponymy, ...

(ignored here)

(semantic) Similarity

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

“You shall know a word by the company it keeps” —J. R. Firth

Text-Corpus Search

- manual browsing
- linguistic/lexicographic research
- “distant reading”

\rightsquigarrow topic-based navigation

\rightsquigarrow semantic pruning

\rightsquigarrow topic retrieval

Semantic Relations

- **Synonymy** (\sim “is-equivalent”, ...)
- **Hyponymy** (\sim “is-subordinate”, “species-of”, “is-subsumed-by”, ...)
- **Hypernymy** (\sim “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.openththesaurus.de>
- crowd-sourced community project
- free open-source license (LGPL)
- **Example:** <http://kaskade.dwds.de/openththesaurus/?q=Schloss>



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/dtaseм>

- 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

The screenshot shows the DTAQ (Deutsches Textarchiv) interface. On the left, the OCR page 'frege_sinn_1892.0030' is displayed, showing the text of 'Über Sinn und Bedeutung'. On the right, a list of recommended items is shown, with 'hegel_logik02_1816.0092' circled in red. A red arrow points from the circled item to the right, towards the 'hegel_logik02_1816.0092' box.

hegel_logik02_1816.0092

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



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



Use Case: Similarity-based Recommendations: by Book

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

DTAQ zuletzt gelesen · Hilfe · Zufallsseite

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

Informationen

Quelle: OCR
Publikationstyp: Artikel einer Zeitschrift/Zeitung
Umfang: 48 Scans
ca. 72317 Zeichen
ca. 10226 Tokens
ca. 2550 Oberflächentypen

Schriftart: Fraktur
Genre: Wissenschaft :: Philosophie
im DTA seit: 2008-01-17 16:21:23
zuletzt geändert: 2014-08-25 13:37:36
Verfügbarkeit: Text (TEI-XML, HTML, TCF, E-Book-Fassung); CC-BY-NC 3.0
Weitere Informationen: Nutzungsbedingungen.

Metadaten

URN: urn:nbn:de:hbz:5:1-200905191458
Titel: Über Sinn und Bedeutung
Aut.-Daten: Aut. 1, Gottlob
Vorn. 1, Frege
Nachn. 1, Frege
Aut. 1, PND: 118535161 (DNB) · ADB/NDB

Ersch.-Jahr: 1892
Verlag: Pfeffer
Ort: Leipzig
Umfang: 25 S.
erschienen in: Zeitschrift für Philosophie und philosophische Kritik, Neue Folge, Band 100/1, Jg. 1892, S. 25-50
Auflage: Erstausgabe
Bibliothek: Staatsbibliothek zu Berlin – Preussischer Kulturbesitz
Signatur: SBB-PK, in: Nf 11296-100.1892

ähnliche Werke

Steinthal, Heymann: Grammatik, Logik und Psychologie. Ihre Principien und ihr Verhältniss zu einander. Berlin, 1855.

www.deustextarchiv.de/dtaq/book/show/steinthal_grammatik_1855

ähnliche Werke

Steinthal, Heymann: Grammatik, Logik und Psychologie. Ihre Principien und ihr Verhältniss zu einander. Berlin, 1855.

- 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)

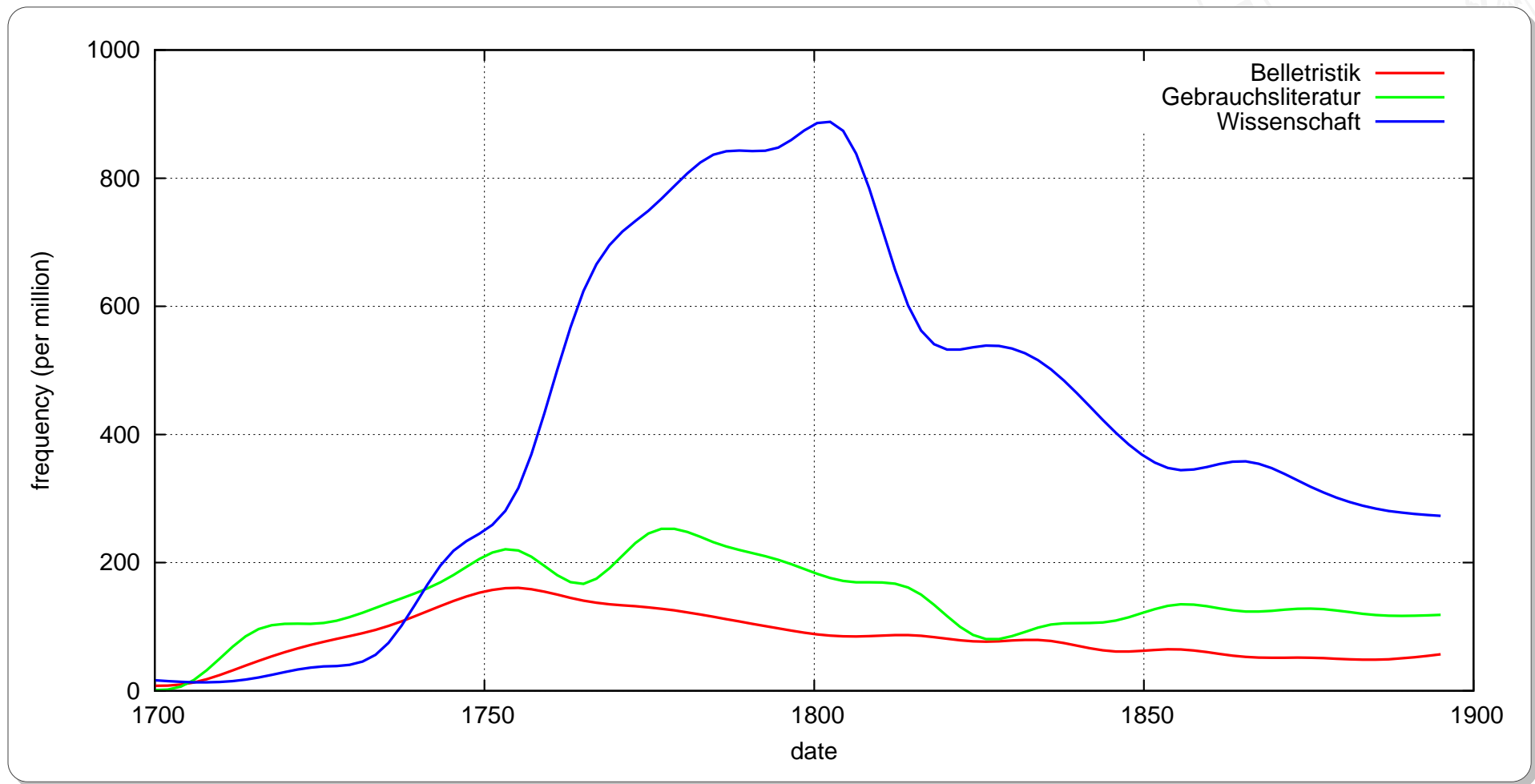
Example: 25 characteristically Kantian terms

book=[^]kant_@25



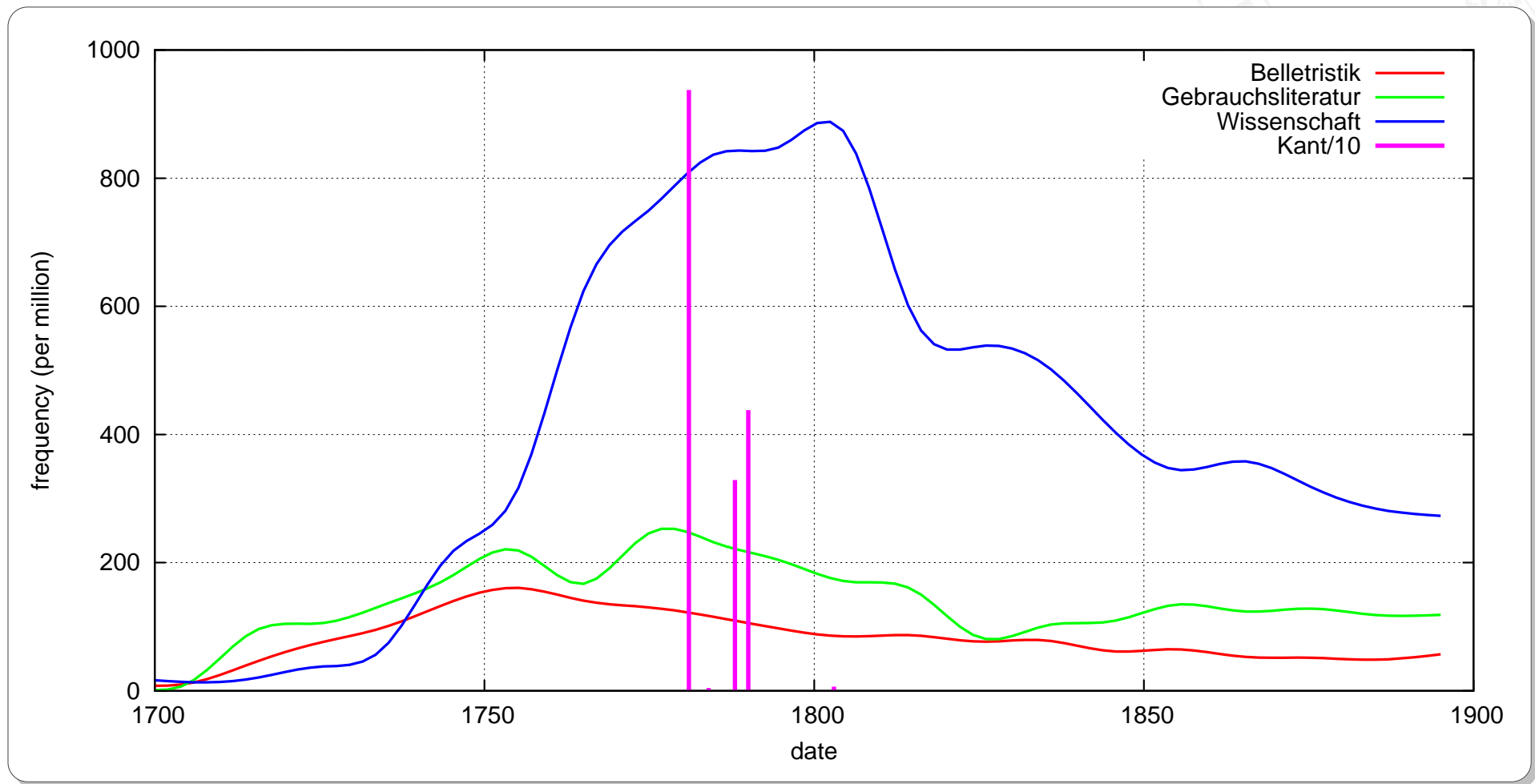
Example: Kantian terms

'book=^kant_@25' | sem



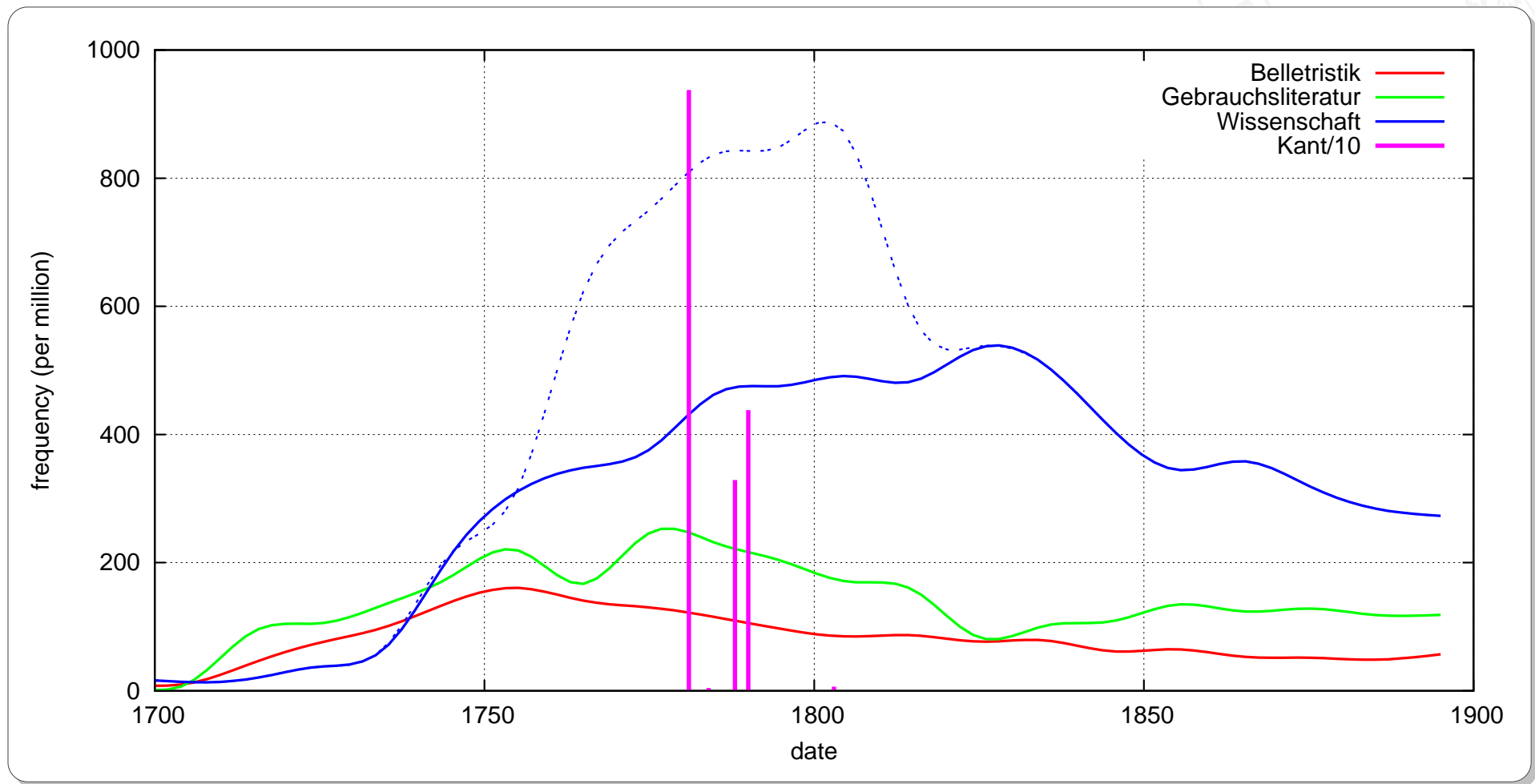
Example: Kantian terms ... in Kant

```
'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
- ontological relations
- corpus research

~> *distributional tag clouds*

~> *manual thesauri*

~> *term expansion*

Use Cases

- similarity-based recommendations
- lexical ambiguity resolution
- distant reading

~> *lateral navigation*

~> *semantic filtering*

~> *topic histograms*

Future Work

- document clustering & genre classification
- improve distributional models, scalability
- improve accessibility for typical tasks
- ...?

(e.g. Blei et al., 2003)

~> *which ones?*

~> *suggestions?*



