The CHQL Query Language for Conceptual History Using Google Books Ngrams

Christoph Schmidt-Petri*[1], Martin Schäler , Michael Schefczyk , Klemens Böhm , and Jens Willkomm

[1]Karlsruhe Institute of Technology – Germany

Abstract

The CHQL Query Language for Conceptual History Using Google Books Ngrams
1. Introduction

The digitization of large time-labeled bibliographies has resulted in corpora such as the Google Ngram data set. Such corpora extremely accurately how individual words are used over time. They are expected to reveal novel insights into the evolution of language and society, provided adequate analysis systems are available. Developing a comprehensive query algebra that allows domain experts to formalize complex hypotheses would be a major contribution to successfully unlock this potential and be especially helpful to conceptual historians.

In conceptual history, as exemplified by the work of Reinhart Koselleck (Olsen, 2012), researchers examine the evolution of concepts represented by words such as 'peace' or 'freedom'. In exploring the history of a concept, scholars commonly make use of, but are not restricted to, word-usage frequencies, word contexts, sentiment analysis, how words refer and relate to and contrast with each other, or they look for word pairs or word families whose usage is correlated (Brunner, Conze, Koselleck, 2004; Ritter, Gründer, Gabriel, 2007).

In this paper, we propose a query algebra for empirical analyses of temporal text corpora, the Conceptual History Query Language (CHQL). A *temporal text corpus* in our sense is a set of words and word chains, i.e., ngrams, together with their usage frequency at various points of time. Our query language is meant to be useful for domain experts, i.e., be descriptive and complete (match all actual and potential hypotheses of conceptual history), and bear optimization potential to allow fast query processing on large data sets.

2. The CHQL query language

This section shows in the abstract how the operators of CHQL allow searching for concept types. A formal definition of all of our operators is given in (Willkomm, Schmidt-Petri, Schäler, Schefczyk, & Böhm, 2018).

Conceptual history claims that pragmatic properties of historical, cultural and economic relevance are incorporated in concepts - whether individual users are aware of this or not. It attempts to track changes of particular concepts (such as 'socialism') over time to determine how their pragmatic relevance changes (for instance, it might mostly express generic hopes

*Speaker

at some moment and mostly specific fears at some other). Thus, concepts will be categorised as belonging to a particular concept *type* at a particular moment in time.

Conceptual historians typically read and interpret large masses of texts which provide a variety of information types (e.g. word frequencies, what words appear in the context, how these words function pragmatically (individually as well as in sentences etc.)). Because we want to do the same using *Distant Reading* techniques (Moretti 2013), these information types need to be translated into observable data characeristics for which individual operators in the query language are defined. Converting 'expert knowledge' into computable items is the main challenge of our project.

Data characteristics are quantitative feature either directly present in our data (e.g., the usage frequency of the word 'socialism' in 1848), or a derived piece of information (e.g. the difference between the usage frequency of words 'socialism' and 'communism' from 1848 to 1989). We describe which data characteristics are needed to simulate Koselleck's information needs and explain our realization of all data characteristics and their implementation as operators.

One of Koselleck's implicit assumptions is that each concept type has specific characteristics. In our terminology: any concept type can be described using a specific combination of information types. For example, Koselleck may plausibly be read as claiming that words that form a *parallel concept* (concept type) would have 'similar' *word frequencies* and have a significant number of identical *surrounding words* (information types). By contrast, *counter concepts* would also have similar word frequencies yet their surrounding words would behave differently. For instance, if 'enlightenment' and 'reason' are parallel concepts for a particular period, their relative word frequencies should be similar, and if 'emancipation' occurs near 'enlightenment', it should occur near 'reason' too, and both concepts should be endorsed rather than criticised (in some sense). By contrast, if 'East' and 'West' are counter concepts, their word contexts should contain different words, and there should be some sort of contrast in attitude between them.

If every concept type has its own specific linguistic and pragmatic properties and hence should be representable by a specific *combination* of information types, it should be possible to develop a system that finds these information types in large corpora that are not amenable to conventional close reading. To this end, we need a formal definition of any information type which is observable and quantifiable.

We present an incomplete list of some of the data characteristics with the information type they are intended to represent, such as individual context, topic grouping, sentence structure, frequency data, and sentiment analysis, and how they are implemented in CHQL.

3. Results

Using CHQL, we have tested the toy hypotheses that 1) 'East' and 'West' have acquired a political context after 1945, whereas 'North' and 'South' haven't, and that 2) the former have turned into counter concepts in the political sphere, their contexts expressing diverging attitudes, whereas the latter have remained parallel concepts in the geographical sphere. First results indicate that CHQL is suited to perform such queries.

Bibliography

Abiteboul, S., Hull, R., & Vianu, V. (1995). *Foundations of Databases: The Logical Level.* Addison-Wesley.

Andersen, N. Å. (2003). *Discursive Analytical Strategies: Understanding Foucault, Koselleck, Laclau, Luhmann.* Policy Press.

Brunner, O., Conze, W., & Koselleck, R. (2004). *Geschichtliche Grundbegriffe* (Vols. 1–

8). Klett-Cotta.

Congress, T. L. (2013, 8 30). *The Contextual Query Language.* Retrieved from https://www.loc.gov/standards/sru/

Hai-Jew, S. (2017). *Data Analytics in Digital Humanities.* Springer.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Empirical Methods in Natural Language Processing.*

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016, 5 30). Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change.

Jakubiček, M., Kilgarriff, A., McCarthy, D., & Rychl, P. (2010). Fast Syntactic Searching in Very Large Corpora for Many Languages. *Pacific Asia Conference on Language, Information and Computation (PACLIC).*

Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., & Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. *Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 169-174).

Maier, D. (1983). *Theory of Relational Databases.* Computer Science Press.

Moretti, F. (2013). *Distant Reading.* Verso Books.

Naber, D. (2005). OpenThesaurus: Ein offenes deutsches Wortnetz. *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn*, (pp. 422-433).

O'Connor, M., & Das, A. (2009). SQWRL: a Query Language for OWL. *OWL: Experiences and Directions (OWLED).*

Olsen, N. (2012). *History in the Plural: An Introduction to the Work of Reinhart Koselleck.* Berghahn Books.

Prabhakaran, V., Hamilton, W. L., McFarland, D., & Jurafsky, D. (2016). Predicting the Rise and Fall of Scientific Topics from Trends in their Rhetorical Framing. *Annual Meeting of the Association for Computational Linguistics (ACL)*, (pp. 1170-1180). doi:10.18653/v1/p16-1111

Ritter, J., Gründer, K., & Gabriel, G. (2007). *Historisches Wörterbuch der Philosophie* (Vols. 1–13). Schwabe.

Snodgrass, R. (1987). The temporal query language TQuel. *ACM Transactions on Database Systems, 12*, 247-298. doi:10.1145/22952.22956

Snodgrass, R. T. (Ed.). (1995). *The TSQL2 Temporal Query Language.* Springer.

Warwick, C. (2012). *Digital Humanities in Practice.* Facet Publishing.

Willkomm, J., Schmidt-Petri, C., Schäler, M., Schefczyk, M., & Böhm, K. (2018). A Query Algebra for Temporal Text Corpora. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries - JCDL '18* (pp. 183-192). Fort Worth, Texas, USA: ACM Press.

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative Textanalyse. *Diagnostica, 54*, 85-98. doi:10.1026/0012-1924.54.2.85

Zeldes, A., Lüdeling, A., Ritz, J., & Chiarcos, C. (2009). ANNIS: a search tool for multi-layer annotated corpora. *Proceedings of Corpus Linguistics.* doi:10.18452/13437

Keywords: query language, Google books, ngrams, conceptual history, Begriffsgeschichte, Reinhart Koselleck