Lexikographie: Explizite und implizite Verortung in den Digital Humanities

Lindemann, David
david.lindemann@uni-hildesheim.de
Universität Hildesheim, Deutschland

Kliche, Fritz
fritz.kliche@uni-hildesheim.de
Universität Hildesheim, Deutschland

Kutzner, Kristin
kutzner@uni-hildesheim.de
Universität Hildesheim, Deutschland

Zusammenfassung

Ziel der hier vorgestellten Studie ist eine Beschreibung der Schnittmenge von Diskursräumen in der Lexikographie bzw. Metalexikographie und den Digital Humanities (DH). Dabei geht es um die Bestimmung von explizit bzw. implizit als Teil der DH aufzufassenden Beiträgen zu lexikographischen Themen und, andersherum, von lexikographierelevanten Themen, die in den DH diskutiert werden. Zur Bestimmung der Diskursräume, von Schnitt- und disjunktiven Mengen, werden Volltexte und Metadaten analysiert, bibliometrische Netzwerke (Autoren- bzw. Zitationsnetzwerke) verglichen und Topic Modelings vorgenommen.

Einleitung

Der Einzug digitaler Methoden und Werkzeuge in die Geistes- und Sozialwissenschaften, genauer: der als „computational turn“ (Berry 2011) bezeichnete methodisch-epistemologische Quantensprung, lässt sich in allen Disziplinen der Humanities beobachten. In der Sprachwissenschaft hat sich dieser Wandel bekanntermaßen besonders deutlich in der Etablierung der Computerlinguistik als eigene Disziplin niedergeschlagen. Neben computerlinguistischen Verfahren der Textanalyse sind eine maschinenlesbare Wissensrepräsentation und -organisation, sind Formate für digitale Editionen und komputationell erstellte Visualisierungen heute in allen textbasierten Disziplinen in Gebrauch.

Der angesprochene Wandel lässt sich ebenfalls in der Lexikographie feststellen. Die Lexikographie bzw. Metalexikographie, als solche bereits seit geraumer Zeit als Disziplin emanzipiert (Tarp 2008; Wiegand 2013), haben den Übergang zum digitalen Medium inzwischen vollzogen (cf. zum frühen Stand der Dinge De Schryver 2003) und sind beständig dabei, ihr komputationell informiertes methodisches Instrumentarium weiterzuentwickeln (Heid 2013). Als zentrale Aspekte gelten hier der Einzug korpuslinguistischer Verfahren in die Lexikographie (Hanks 2008; Heid 2008), komputationelle Methoden zur Datenrepräsentation (Spohr 2012), speziell auch für die digitale Edition historischer Wörterbücher (Lemnitzer u. a. 2013) und zur Implementierung funktionsgerichteter Benutzerschnittstellen (Heid 2014) sowie zur Wörterbuchbenutzungsforschung (Müller-Spitzer 2014).

In der hier vorgestellten Studie gehen wir der Frage nach, wie sich der gemeinsame Diskursraum als Schnittmenge von Lexikographie und Digital Humanities mit quantitativen Methoden definieren lässt. Nach einer nicht exhaustiven und von Hand durchgeführten Voruntersuchung folgen wir der Ausgangshypothese, der gemeinsame Diskursraum sei um ein Vielfaches größer als man annehmen könnte, folgte man allein denjenigen Themen, die als lexikographierelevant gelten können und die in Publikationen diskutiert werden, die explizit zum Bereich der DH gehören (vgl. Abb. 1).

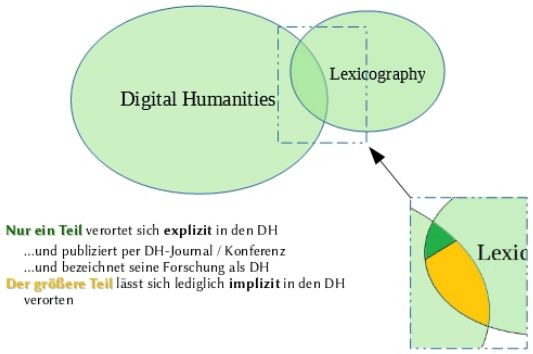

Abb. 1: Ausgangshypothese: Explizite und implizite Schnittmengen.

Diejenigen Arbeiten, die im DH-Kontext veröffentlicht werden und explizit einem Thema der Lexikographie zugeordnet werden, sind recht leicht über relevante Schlüsselwörter bestimmbar. Dazu tritt die Gruppe jener Publikationen, die zur eingangs skizzierten Schnittmenge zu zählen sind, ohne dass sie sich selbst ausdrücklich den Digital Humanities zuordnen. Es ist das Ziel dieser Untersuchung, zu bestimmen, welche in der Lexikographie diskutierten Themen und welche Autoren zu dieser Gruppe gerechnet werden und also eine Zurechnung zu den Digital Humanities *implizieren* können.

Voruntersuchung und Zwischenergebnis: Explizite Verortung in den DH

Als Voruntersuchung zum benannten Gegenstand haben wir eine Recherche in den Archiven bedeutender englischsprachiger Zeitschriften der Digital Humanities [1] sowie in den Proceedings der ADHO-Jahreskonferenzen [2] durchgeführt. Über die Suchbegriffe „Lexicography“ und „Dictionary“ finden sich in den genannten Archiven 31 englischsprachige Beiträge, die sich mit lexikographischen Themen befassen, und die sich qua Erscheinen in DH-Medien zu denjenigen Publikationen zählen lassen, die sich explizit in den Digital Humanities verorten.

Eine manuelle Zuordnung lexikographierelevanter Schlüsselwörter zu den genannten 31 Beiträgen ergibt das in Tabelle 1 wiedergegebene Bild; dabei sind mehrfache Zuordnungen möglich. Zunächst lässt sich ohne Verwunderung feststellen, dass in allen Beiträgen die digitale Repräsentation lexikalischer Daten eine Rolle spielt, allerdings mit unterschiedlichen Fragestellungen, Herangehensweisen und Zielsetzungen. Die drei größten Themencluster haben wir hier, in dieser Reihenfolge, mit den Schlagwörtern „e-Wörterbücher / Visualisierung lexikalischer Daten“, „Historische Lexikographie“ und „Korpuslinguistik“ bezeichnet. Ersteres benennt Fragen der Produktion digitaler Wörterbücher einschließlich neuer Methoden der Visualisierung, letzteres die Erstellung und Nutzung elektronischer Textkorpora zu einer Reihe lexikographischer Zwecke. Beide Bereiche sind durch die Heraufkunft digitaler Methoden überhaupt erst möglich geworden und haben die Lexikographie revolutioniert. Die Historische Lexikographie kann als philologische Disziplin gelten, die sich mit der Edition historischer lexikalischer Datensammlungen befasst; die diesem Schlüsselwort zugeordneten Beiträge befassen sich grundsätzlich mit Methoden digitaler Edition, einem Kernbereich der DH.

Schlüsselwort (Topic)	Zählung
Digitale Wissensrepräsentation / Formate	31
e-Wörterbücher / Visualisierung lexikalischer Daten	14
Historische Lexikographie	10
Korpuslinguistik	9
Wörterbuchnetz	4
NLP-Lexicon	2
Bilingual Dictionary Drafting	2
Autorenwörterbuch	2
Dialektologie	1

Tabelle 1: Schlüsselwörter, manuelles Clustering, manuelle Zählung

Unter den weniger häufig gewählten Schlüsselwörtern sticht das „Wörterbuchnetz“ hervor, das Strategien zur Vernetzung lexikalischer Ressourcen bezeichnet. Hinzu kommen noch lexikalische Datensammlungen zur Anwendung in der maschinellen Sprachverarbeitung („NLP-Lexicon“), Methoden zum Entwurf zweisprachiger Wörterbuchinhalte („Bilingual Dictionary Drafting“), das „Autorenwörterbuch“, also Extraktionen aus Korpora, die aus dem Schaffen jeweils einer Literatin oder eines Literaten bestehen, sowie in einem Fall dialektologische Forschung mit digitalen Methoden.

Methode für die Bestimmung implizit in den DH verorteter Arbeiten

Wir haben ein Textkorpus erstellt, das im Zeitraum 2000 bis zur Gegenwart (2018) erschienene englischsprachige Beiträge aus Zeitschriften, Kongressakten und Handbüchern zu den Digital Humanities (Subkorpus DH) und der Lexikographie (Subkorpus Lexicog) enthält; Tabelle 2 gibt die Titel der verarbeiteten Quellen wieder. Dabei wurden die ausgewählten Zeitschriften bzw. Sammelbände jeweils vollständig berücksichtigt; die Beiträge wurden zusammen mit Metadaten, u. a. Verfasser (Name und Affiliation), Datum, Textsorte, Umfang und Identifier (ISBN, DOI), im Tool Zotero [3] verwaltet. Die Volltexte wurden semiautomatisch bereinigt und zusammen mit Metadatensätzen in das Korpus aufgenommen. Darüber hinaus wurden die in den Volltexten enthaltenen bibliographischen Referenzen extrahiert (GROBID, Lopez 2009).

DH / 1.422 (41%)	Digital Humanities Quarterly: http://www.digitalhumanities.org/dhq / 284
	DSH (ex LLC): https://academic.oup.com/dsh / 886
	TEI Journal of the Text Encoding Initiative: http://jtei.revues.org / 63
	Digital Studies/Le champ numerique: https://www.digitalstudies.org/ / 152
	Blackwell Companion to DH: Schreibman et al. (ed.) 2004 / 37
Lexikog / 2.056 (59%)	IJL: http://ijl.oxfordjournals.org/ 282
	Lexikos: http://lexikos.journals.ac.za/pub / 376
	Dictionaries (Journal of the DSNA): https://muse.jhu.edu/journal/540 / 257
	Euralex: https://euralex.org/publications/ / 782
	eLex: https://elex.link/ / 202
	HSK 5/4: Gouws et al. (ed.) 2013 / 110
	The Routledge Handbook of Lexicography: Fuertes-Olivera (ed.) 2018 / 47

Tabelle 2: Quellen für das DH/Lexikog Textkorpus / Zahl der Volltexte

Topic Modeling

Unüberwachtes Topic Modeling (LDA, eingesetztes Tool: MALLET (McCallum 2002)) soll es uns ermöglichen, die in Abb. 1 grob skizzierten Mengen als sich überschneidende Diskursräume zu bestimmen und zu visualisieren. Unsere Ergebnisse zeigen die relative Relevanz in beiden Subkorpora von 50 durch den LDA-Algorithmus bestimmten, jeweils mit einer Reihe von Schlüssel-Tokens repräsentierten Topics. Eine Reihe von Anhaltspunkten spricht für das zuverlässige Funktionieren der Methode: Die Liste der Topics, die besonders DH-relevant seien, wird von den Tokens „digital humanities computing tools“ angeführt, die Liste der Lexikographie-Topics von „dictionary dictionaries english words word learners language“.

Bei der Ansicht der im Mittelfeld befindlichen Topics, also Themen, die in beiden Subcorpora als relevant bezeichnet sind, stellt sich heraus, dass sich hier nicht nur der digitale Wandel als Thema widerspiegelt, sondern dass darüber hinaus weitere Topics den gemeinsamen Diskursraum von Lexikographie und DH ausmachen. Inmitten von Zeilen, die, quasi erwartungsgemäß, Tokens wie „information model data structure process analysis“ oder „corpus words frequency texts word corpora table“ sowie eine ganze Reihe von Namen natürlicher Sprachen enthalten, ist hier etwa das mit den Tokens „women male female gender woman man people [...] black [...] girl feminist [...]“ repräsentierte Topic auffällig (40 der 100 für dieses Topic relevantesten Beiträge stammen aus dem DH-, 60 aus dem Lexikog-Subkorpus).

Abbildung 2 zeigt für 50 von MALLET bestimmte Topics (Spalten) die Verteilung der 100 jeweils relevantesten Texte über die Subcorpora (Ausgabe von MALLET; DH-Beiträge sind grün, Lexicog-Beiträge violett unterlegt). Es wird deutlich, dass ein Teil der Topics eindeutig einem der Subkorpora zuzurechnen ist, andere Topics dagegen eine starke Durchmischung aufweisen.

Abb. 2: Visualisierung des Topic Modeling

Citation Network

Für alle Artikel des Korpus (siehe Tabelle 2) untersuchten wir die Anzahl der auf Items innerhalb des Netzwerks gerichteten Zitationen. Es zeigten sich 2.431 Zitationen (31% DH, 69% Lexicog). Die Zitationen aus DH sind nur zu 2% auf items aus Lexicog gerichtet; die Zitationen aus Lexicog zu 1% auf items aus DH.

Ergebnisse und Schlussfolgerungen

Die vorgestellten korpuslinguistischen und bibliometrischen Untersuchungen bieten wie beschrieben Aufschluss über Schnitt- und disjunkte Mengen von Themen- und Autorenclustern der Lexikographie und der Digital Humanities. Visualisierungen dieser Cluster und Listen der relevanten Keywords und Autoren werden bereitgestellt. Topic Modeling und Zitationsnetzwerke bilden unterschiedlich große Schnittmengen zwischen beiden Disziplinen ab: Während einerseits deutlich wird, dass eine ganze Reihe von Themen in beiden Disziplinen relevant ist, zitiert man sich vergleichsweise selten gegenseitig.

Die gezeigten Ergebnisse können zunächst zu einer verbesserten gegenseitigen Wahrnehmung in

Lexikographie und Digital Humanities beitragen sowie in der lexikographischen Community das Bewusstsein dafür stärken, ein Gutteil der Disziplin gehöre durch die inhaltliche Überschneidung de facto zum Einflussbereich der Digital Humanities. Dies wiederum kann in der Zukunft zu einer stärkeren expliziten Verortung relevanter lexikographischer Beiträge in den Digital Humanities führen.

Weiterhin haben wir mit den für diese Studie durchgeführten Arbeiten eine annotierte bibliographische Datensammlung angelegt und die dazugehörigen Volltexte mit korpuslinguistischen Methoden annotiert und analysiert. Wir beabsichtigen, diese Sammlung auch weiterhin zu pflegen und öffentlich zugänglich zu machen.

Fußnoten

1. Digital Humanities Quarterly, DSH, TEI Journal
2. http://adho.org
3. http://www.zotero.org

Bibliographie

Berry, David M. (2011): „The Computational Turn: Thinking About the Digital Humanities“. (The Computational Turn). In *Culture Machine* 12 (0).

De Schryver, Gilles-Maurice (2003): „Lexicographers' Dreams in the Electronic#Dictionary Age“. In *International Journal of Lexicography* 16 (2): 143–199.

Fuertes-Olivera, Pedro (ed.) (2018): *The Routledge Handbook of Lexicography*. London: Routledge.

Gouws, Rufus / Heid, Ulrich / Schweickard, Wolfgang / Wiegand, Herbert E. (eds.). (2013): *Dictionaries. An International Encyclopedia of Lexicography*. HSK 5/4. Berlin / Boston: De Gruyter Mouton.

Hanks, Patrick (2008). „The Lexicographical Legacy of John Sinclair“. In *International Journal of Lexicography* 21 (3): 219–229.

Heid, Ulrich (2008). „Corpus linguistics and lexicography“. In Anke Lüdeling / Merja Kytö (ed.) *Corpus Linguistics. An international Handbook*: 131–153. Berlin: Mouton de Gruyter.

Heid, Ulrich (2013): „The impact of computational lexicography“. In Gouws et al. 2013: 24–30.

Heid, Ulrich (2014): „Natural Language Processing Techniques for Improved User-friendliness of Electronic Dictionaries“. In *Proceedings of EURALEX 2012*: 47–61.

Lemnitzer, Lothar / Romary, Laurent / Witt, Andreas (2013): „Representing human and machine dictionaries in markup languages (SGML, XML)“. In Gouws et al. 2013: 1195–1209.

Lopez, Patrice (2009): „GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications“. In *Research and Advanced Technology for Digital Libraries*: 473–474. Lecture Notes in Computer Science. Berlin / Heidelberg: Springer

McCallum, Andrew K. (2002): *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu/.

Müller-Spitzer, Carolin (2014): „Methoden der Wörterbuchbenutzungsforschung“. In *Lexicographica* 30 (1): 112–151.

Schreibman, Susan / Siemens, Ray / Unsworth, John (2004): *A Companion to Digital Humanities*. Oxford: Blackwell

Spohr, Dennis (2012): *Towards a Multifunctional Lexical Resource, Design and Implementation of a Graph-Based Lexicon Model*. Lexicographica Series Maior, 141. Berlin / Boston: De Gruyter.

Tarp, Sven (2008): *Lexicography in the borderland between knowledge and non-knowledge: general lexicographical theory with particular focus on learner's lexicography*. Lexicographica, Series Mayor 134. Tübingen: Niemeyer.

Wiegand, Herbert E. (2013): „Lexikographie und Angewandte Linguistik“. In *Zeitschrift für angewandte Linguistik* 58 (1): 13–39.