DiaCollo: diachronen Kollokationen auf der Spur

— REVISED & CORRECTED DRAFT —

Bryan Jurish[*] **Alexander Geyken**[*] **Thomas Werneke**[†]
jurish@bbaw.de geyken@bbaw.de werneke@zzf-pdm.de

[*]Berlin-Brandenburgische Akademie der Wissenschaften
[†]Zentrum für Zeithistorische Forschung

Abstract

Wir präsentieren DiaCollo, ein Softwarewerkzeug zur effizienten Extraktion, zum Vergleich und zur interaktiven Visualisierung von Kollokationen aus einem diachronen Textkorpus. Im Gegensatz zu konventioneller Kollokationssoftware eignet sich DiaCollo für die Verarbeitung diachroner Kollokationsdaten: Kollokationspaare, deren Assoziationsstärke vom Zeitpunkt ihres Auftretens abhängt. Durch das Aufspüren von Veränderungen in den charakteristischen Kollokaten eines Worts im zeitlichen Verlauf kann DiaCollo dazu beitragen, ein klareres Bild von Veränderungsprozessen der Wortsemantik zu zeichnen.

1 Einführung

In den letzten Jahren sind immer mehr große diachrone Textkorpora zu Forschungszwecken verfügbar gemacht geworden (e.g. Geyken et al., 2011; Davies, 2012). Die durch diese Korpora abgedeckten großen Zeitspannen stellen diverse Herausforderungen an konventionelle Techniken der maschinellen Verarbeitung natürlicher Sprache, die ihrerseits oft auf impliziten Annahmen der Korpushomogenität basieren – insbesondere der zeitliche Achse betreffend. Tatsächlich haben sogar vermeintlich synchrone Zeitungskorpora eine nicht triviale temporale Extension und können bei entsprechender Behandlung zeitabhängige Phänomene zeigen (Scharloth et al., 2013). In dieser Arbeit gehen wir auf das Problem der automatischen Erstellung von Kollokationensprofilen (Church & Hanks, 1990; Evert, 2005) in diachronen Korpora ein, indem wir ein neues, explizit für diesen Zweck entwickeltes Softwarewerkzeug "DiaCollo" vorstellen, das dem Benutzer ermöglicht, für jede Abfrage selber die Granularität der diachronen Achse frei zu wählen. Im Gegensatz zu konventionellen Kollokationswerkzeugen wie dem DWDS Wortprofil[1] (Didakowski & Geyken, 2013) oder dem Sketch Engine[2] (Kilgarriff & Tugwell, 2002) eignet sich DiaCollo zur Extraktion und Analyse diachroner Kollokationsdaten: Kollokationspaare, deren Assoziationsstärke von dem Zeitpunkt ihres Auftretens abhängt. Durch das Aufspüren von Veränderungen in den typischen Kollokaten eines Worts im zeitlichen Verlauf und Anwendung von J. R. Firths berühmtem Prinzip "you shall know a word by the company it keeps", kann DiaCollo helfen, ein klareres Bild diachroner Veränderungen im Wortgebrauch zu liefern.

[1] http://zwei.dwds.de/wp
[2] http://www.sketchengine.co.uk

2 Implementierung

DiaCollo ist als modulare Perl Bibliothek implementiert, einschliesslich wiederverwendbaren Klassen zum Umgang mit nativen Binärindexstrukturen. DiaCollo Indizes sind für Hochlastumgebungen geeignet, da kein persistenter Server-Prozess benötigt wird und jeglicher Laufzeitzugriff auf native Indexstrukturen über direkten Dateisystem I/O stattfindet. Über die programmatische API der Perl-Module hinaus bietet DiaCollo sowohl eine Befehlszeilenschnittstelle als auch einen öffentlich zugänglichen RESTful Webservice mit einer formularbasierten Benutzerschnittstelle zur Auswertung von Datenbankanfragen und einer interaktiven Visualisierung der Anfrageergebnisse. Ein öffentlich zugängliches Web-Frontend für das Korpus des Deutschen Textarchivs ist unter `http://kaskade.dwds.de/dstar/dta/diacollo/` zu finden; der vollständige Quellcode ist via CPAN[3] erhältlich.

2.1 Anfragen & Parameter

DiaCollo ist ein anfrageorientierter Dienst. Er behandelt eine Benutzeranfrage als eine Menge von *(Parameter=Wert)*-Paaren und liefert ein korrespondierendes Profil für den/die angefragten Term(e) zurück. Die Parameter werden wie bei einem üblichen Web-Formular an den Service RESTfully via HTTP GET oder POST Anfrage überreicht. Jede Anfrage muss einen *query* Parameter enthalten, der den oder die zu profilierenden Zielterm(e) spezifiziert. Der *date* Parameter selegiert die gewünschte Zeitspanne, während die Granulariät der zurückgelieferten Profildaten mithilfe des *slice* Parameters durch Angabe der Grösse einer einzelnen Profilepoche festgelegt werden kann. Kollokatkandidaten können über den *groupby* Parameter gefiltert werden, und die Bereinigung ("pruning") der zurückzuliefernden 'besten' Kollokaten wird von den Parametern *score*, *kbest* und *global* gesteuert.

2.2 Profile & Diffs

Das Ergebnis einer einfachen DiaCollo Anfrage wird als tabellarisches Profil der k-best Kollokate für den/die angefragte(n) Term(e) in jedem der angefragten Zeit-Subintervalle ausgegeben ("Epochen" oder "Slices", e.g. Dekaden), die mit den Parameter *date* und *slice* spezifiziert wurden. Als Alternative kann der Benutzer auch ein Vergleichs- bzw. "Diff"-Profil anfordern, um die salientesten Unterschiede zwischen zwei unabhängigen Anfragen hervorzuheben; z.B. zwischen zwei verschiedenen Worten oder zwischen den Vorkommen eines Wortes in verschiedenen Zeitintervallen, Teilkorpora oder lexikalischen Umgebungen.

2.3 Indizes, Attribute & Aggregierung

DiaCollo benutzt eine interne "native" Indexstruktur über alle Inhaltswörter des Eingabekorpus, um Kollokationsprofile zu berechnen. Jedes indizierte Wort wird als n-Tupel linguistisch relevanten Token- oder Dokumentattribute behandelt, zusätzlich zum Dokumentdatum. Die Attribute *Lemma (l)* und *Pos (p)* ("part-of-speech") werden per Default indiziert. Die Anfrageparameter *query* und *groupby* werden als logische Konjunktionen von Suchkriterien bezüglich dieser Attribute interpretiert, um die genauen zu profilierenden Token-Tupel zu selegieren. Um eine feinkörnigere Auswahl von Profilzielen zu ermöglichen, unterstützt DiaCollo den gesamten Umfang der DDC-Abfragesprache[4] (Sokirko, 2003; Jurish et al., 2014), wenn

[3] `http://metacpan.org/release/DiaColloDB`
[4] `http://www.ddc-concordance.org`

die DiaCollo-Instanz mit einem zugrundeliegenden DDC Server assoziiert ist.

2.4 Scoring & Pruning

DiaCollo weist jedem Kollokat w_2 eines unären Profils für einen Zielterm w_1 mittels einer benutzerspezifizierten *Scorefunktion* einen reellwertigen Assoziationswert ("score") zu. Zu den unterstützten Scorefunktionen zählen absolute und logarithmische Frequenzen (*f, lf*), normierte absolute und logarithmische Frequenzen pro Mio. Token (*fm, lfm*), das *pointwise mutual information* × log Frequenz Produkt (*mi*), und der von Rychlý (2008) vorgeschlagene skalierte log-Dice Koeffizient (*ld*). Kollokatkandidaten werden nach Assoziationswert absteigend geordnet und die k-besten Kandidaten jeder Epoche ausgewählt und zurückgegeben. Für "diff" Anfragen werden unabhängige Profile p_a und p_b jeweils für die *query* und *bquery* Parameter berechnet. Nach der Sortierung anhand der selegierten Scorefunktion wird ein Vergleichsprofil $p_{\mathrm{a-b}}$ berechnet als $p_{\mathrm{a-b}} : w_2 \mapsto p_\mathrm{a}(w_2) - p_\mathrm{b}(w_2)$ für jeden der bis zu $2k$ Kollokate $w_2 \in k\text{-best}(p_\mathrm{a}) \cup k\text{-best}(p_\mathrm{b})$, wonach die k-besten von diesen Kandidaten mit den größten absoluten Unterschieden $|p_{\mathrm{a-b}}(w_2)|$ selegiert und zurückgegeben werden.

2.5 Ausgabe & Visualisierung

DiaCollo unterstützt verschiedene Ausgabeformate für die zurückgelieferte Profildaten, darunter TAB-getrennten Text, natives JSON für die weitere automatische Verarbeitung sowie einfaches tabellarisches HTML. Zusätzlich zu den statischen tabellarischen Formaten bietet der Webservice-Plugin auch mehrere interaktive Online-Visualisierungen für diachrone Profildaten, unter anderem zweidimensionale Zeitreihen mithilfe der Highcharts JavaScript Bibliothek, Flash-basierten Motion Charts mithilfe der Google Motion Charts Bibliothek und dynamische Bubble- und Tag-Cloud Visualisierungen mithilfe der D3.js Bibliothek. Das HTML sowie die D3-basierten Formate bieten eine intuitive farbkodierte Repräsentation der Assoziationsscores (bzw. Score-Unterschiede bei "diff"-Profilen) für jedes Kollokationspaar sowie Hyperlinks zu den zugrundeliegenden Korpus-Treffer ("KWIC-links") für jeden abgebildeten Datenpunkt. Beispiele für die Zeitreihen-, Tag-Cloud- und Bubble-Visualisierungen sind in den Abbildungen 1–3 zu finden.

3 Fallstudien

3.1 Zeitgeschichte

Ein Anwendungsgebiet für DiaCollo stellt die (zeit-)historische Forschung dar. Insbesondere politische Prozesse können durch die zeitscheibenbasierte Analyse von DiaCollo auf neue Weise beschrieben werden. Das Potenzial von DiaCollo soll hier kurz am Beispiel des Begriffs "Krise" in der Wochenzeitung *DIE ZEIT* skizziert werden (Abb. 2). Da der Begriff "Krise" eine inhärent instabile Situation bezeichnet, kann davon ausgegangen werden, dass die damit assoziierten Diskursumgebungen im zeitlichen Verlauf stark variieren. Dies sollte sich auch in den Kollokationsprofilen von DiaCollo niederschlagen. Mittels DiaCollo lassen sich Eigennamen (Personen, Orte, Institutionen) extrahieren, die als Kollokationspartner von "Krise" in der *ZEIT* auftreten. Eine Analyse der Zeitreihe dieser Kollokate zeigt, dass mittels DiaCollo korrekt politische Krisen (etwa in den 2000ern "CDU"), ökonomische Krisen (in den

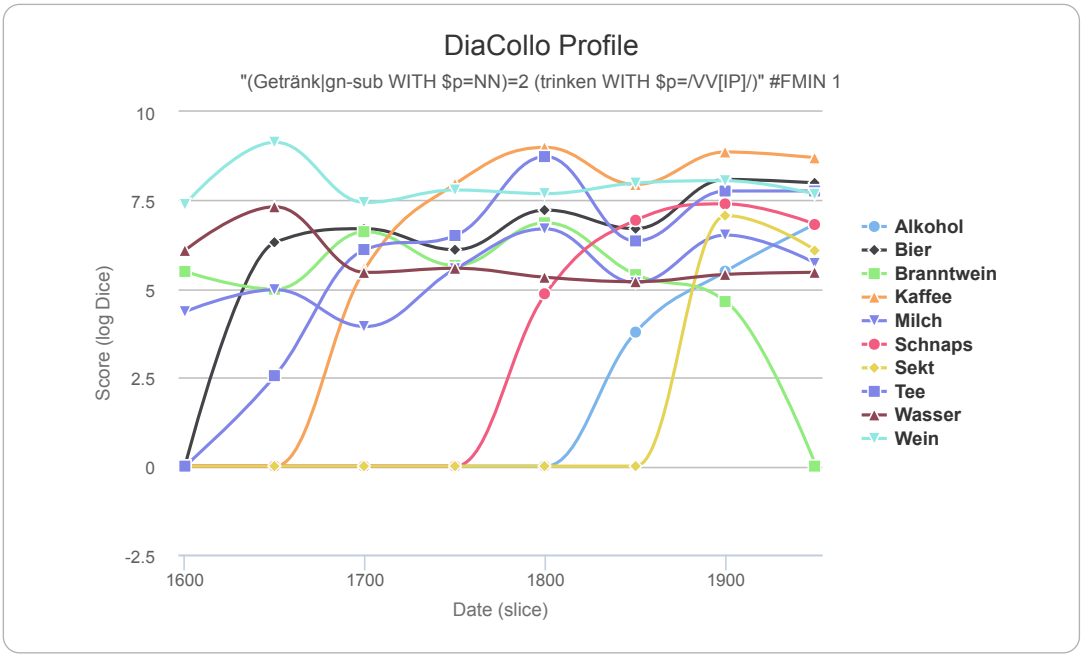

Abbildung 1: DiaCollo Zeitreihe für die zehn global besten nominale Hyponyme des GermaNet (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010) SynSets *Getränk* unmittelbar links vom Verb *trinken* in 50-Jahres Epochen über den Gesamtbestand des Deutschen Textarchivs und DWDS-Kernkorpus.

1980ern "AEG", in den 1990ern die Finanzkrise in Südostasien), aber auch konflikthafte Krisen (z.B. Jugoslawien in den 1990ern) ermittelt werden können. Je nach Granularität der Abfrage werden auch die Ergebnisse komplexer und damit auch für den Experten interessanter. Ihre Interpretation bedarf dann in der Regel weiterer manueller Aufarbeitung, z.B. mithilfe der von DiaCollo bereitgestellten Verknüpfung mit der zugrundeliegenden Textbasis.

3.2 Lexikographie

Ein weiteres Anwendungsgebiet von DiaCollo ist die Lexikographie. Da Kollokationen und die Beschreibung des Bedeutungsspektrums (Lesarten) eines Wortes eng miteinander zusammenhängen, lassen sich aus zeitlichen Verläufen von Kollokationen wichtige lexikographische Befunde ableiten: die Verlagerung der Gewichtung von Lesarten untereinander oder das Verschwinden einer Lesart zugunsten einer anderen können dadurch ebenso nachverfolgt werden wie das Auftauchen von neuen Lesarten (Neosemanteme). Bekannte Beispiele hierfür sind Wörter wie 'Maus' (als Computermaus) oder 'Ampel' (in der politischen Bedeutung), die seit den späten 1980er bzw. den frühen 1990er Jahren im öffentlichen Sprachgebrauch sind. Ein komplexeres Beispiel stellt das Adjektiv 'autofrei' dar. Dieses ist im Duden definiert als "keinen Autoverkehr aufweisend". Eine genauere Sicht auf die Korpusbelege ergibt, dass das Wort zwei Unterbedeutungen aufweist: erstens die "(per Verordnung) auferlegte Autofreiheit", die durch die Ölkrise in den 1970er Jahren in der Kollokation 'autofreier Sonntag' erstmals auftrat und später in Verbindungen wie 'autofreie Innenstädte' ein Bedeutungserweiterung erfuhr. In den 1990er Jahren bildet sich die zweite Bedeutung des Wortes heraus, bei der der Verzicht

Abbildung 2: DiaCollo dynamische Tag-Cloud Visualisierung der zehn besten Eigennamenkollokaten für "Krise" in der Wochenzeitung *DIE ZEIT* für die Epochen 1980–1989 (oben links), 1990–1999 (oben rechts), 2000–2009 (unten links) und 2010–2014 (unten rechts).

auf das Auto auf Selbstverpflichtung beruht[5]. Diese Lesart ist durch Kollokationen wie 'autofreie Wohnanlage' oder 'autofreie Siedlung' gekennzeichnet. Mit DiaCollo lassen sich beide Bedeutungen nicht nur unterscheiden, sondern auch in ihrem zeitlichen Verlauf nachverfolgen (Abb. 3).

4 Zusammenfassung

Wir haben hier DiaCollo vorgestellt, ein neues Softwarewerkzeug für die effiziente Extraktion, den Vergleich und die interaktive Visualisierung von Kollokationen, speziell zugeschnitten auf die besonderen Anforderungen diachroner Textkorpora. Darüber hinaus haben wir anhand von zwei Fallstudien skizziert, wie DiaCollo als modularer Webservice-Plugin Forscher in den Geistes- und Sozialwissenschaften dabei unterstützen kann, ein klareres Bild der diachronen Variation in der Verwendung eines Wortes zu erhalten.

[5] vgl. `http://zwei.dwds.de/wb/autofrei`

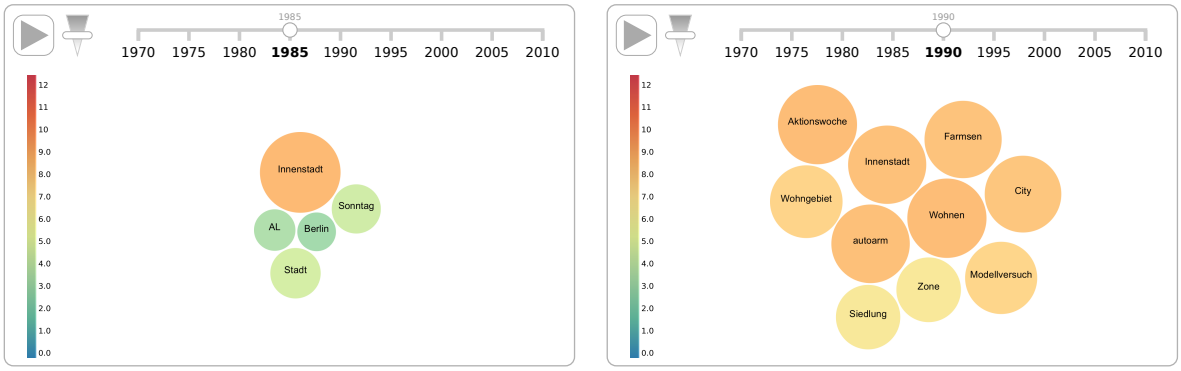

Abbildung 3: DiaCollo dynamische Bubble-Chart Visualisierung der zehn besten Kollokaten des Ajdketivs *autofrei* im aggregierten DWDS Zeitungskorpus für die Epochen 1985–1989 (links) und 1990–1994 (rechts).

Literatur

K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

M. Davies. Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English. *Corpora*, 7(2):121–157, 2012. URL `http://davies-linguistics.byu.edu/ling450/davies_corpora_2011.pdf`.

J. Didakowski and A. Geyken. From DWDS corpora to a German word profile – methodological problems and solutions. In A. Abel and L. Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information*, (OPAL X/2012). IDS, Mannheim, 2013. URL `http://www.dwds.de/static/website/publications/pdf/didakowski_geyken_internetlexikografie_2012_final.pdf`.

S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, Institut für maschinelle Sprachverarbeitung, Universität Stuttgart, 2005. URL `http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/`.

A. Geyken, S. Haaf, B. Jurish, M. Schulz, J. Steinmann, C. Thomas, and F. Wiegand. Das deutsche Textarchiv: Vom historischen Korpus zum aktiven Archiv. In S. Schomburg, C. Leggewie, H. Lobin, and C. Puschmann, editors, *Digitale Wissenschaft. Stand und Entwicklung digital vernetzter Forschung in Deutschland*, pages 157–161, 2011. URL `http://www.hbz-nrw.de/dokumentencenter/veroeffentlichungen/Tagung_Digitale_Wissenschaft.pdf#page=159`.

B. Hamp and H. Feldweg. GermaNet – a lexical-semantic net for German. In *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997.

V. Henrich and E. Hinrichs. GernEdiT – the GermaNet editing tool. In *Proceedings LREC 2010*, pages 2228–2235, 2010. URL `http://www.lrec-conf.org/proceedings/lrec2010/pdf/264_Paper.pdf`.

B. Jurish, C. Thomas, and F. Wiegand. Querying the deutsches Textarchiv. In U. Kruschwitz, F. Hopfgartner, and C. Gurrin, editors, *Proceedings of the Workshop "Beyond Single-Shot Text Queries: Bridging the Gap(s) between Research Communities" (MindTheGap 2014)*, pages 25–30, Berlin, Germany, 4th March 2014. URL http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

A. Kilgarriff and D. Tugwell. Sketching words. In M.-H. Corréard, editor, *Lexicography and Natural Language Processing: A Festschrift in Honour of B. T. S. Atkins*, EURALEX, pages 125–137, 2002. URL http://www.kilgarriff.co.uk/Publications/2002-KilgTugwell-AtkinsFest.pdf.

P. Rychlý. A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, pages 6–9, 2008. URL http://www.fi.muni.cz/usr/sojka/download/raslan2008/13.pdf.

J. Scharloth, D. Eugster, and N. Bubenhofer. Das Wuchern der Rhizome. linguistische Diskursanalyse und Data-driven Turn. In D. Busse and W. Teubert, editors, *Linguistische Diskursanalyse. Neue Perspektiven*, pages 345–380. VS Verlag, Wiesbaden, 2013. URL http://www.scharloth.com/files/Rhizom_Zeit.pdf.

A. Sokirko. A technical overview of DWDS/Dialing Concordance. Talk delivered at the meeting *Computational linguistics and intellectual technologies*, Protvino, Russia, 2003. URL http://www.aot.ru/docs/OverviewOfConcordance.htm.