Volume 28 — Number 2 — 2013 — ISSN 2190-6858

JLCL
Journal for Language Technology
and Computational Linguistics

Webkorpora in Computerlinguistik und Sprachforschung

Web Corpora for Computational Linguistics and Linguistic Research

Herausgegeben von /*Edited by*
Roman Schneider, Angelika Storrer und
Alexander Mehler

GSCL Gesellschaft für Sprachtechnologie & Computerlinguistik

Contents

Impressum

Herausgeber	Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL)
Aktuelle Ausgabe	Band 28 – 2013 – Heft 2
Herausgeber	Roman Schneider, Angelika Storrer, Alexander Mehler
Anschrift der Redaktion	Lothar Lemnitzer Berlin-Brandenburgische Akademie der Wissenschaften Jägerstr. 22/23 10117 Berlin lemnitzer@bbaw.de
ISSN	2190-6858
Erscheinungsweise	2 Hefte im Jahr, Publikation nur elektronisch
Online-Präsenz	www.jlcl.org

Editorial

"Webkorpora in Computerlinguistik und Sprachforschung" war das Thema eines Workshops, der von den beiden GSCL-Arbeitskreisen „Hypermedia“ und „Korpuslinguistik“ am Institut für Deutsche Sprache (IDS) in Mannheim veranstaltet wurde, und zu dem sich am 27.09. und 28.09.2012 Experten aus universitären und außeruniversitären Forschungseinrichtungen zu Vorträgen und Diskussionen zusammenfanden. Der facettenreiche Workshop thematisierte Fragen der Gewinnung, der Aufbereitung und der Analyse von Webkorpora für computerlinguistische Anwendungen und sprachwissenschaftliche Forschung. Einen Schwerpunkt bildeten dabei die speziellen Anforderungen, die sich gerade im Hinblick auf deutschsprachige Ressourcen ergeben. Im Fokus stand weiterhin die Nutzung von Webkorpora für die empirisch gestützte Sprachforschung, beispielsweise als Basis für sprachstatistische Analysen, für Untersuchungen zur Sprachlichkeit in der internetbasierten Kommunikation oder für die korpusgestützte Lexikographie. Zusätzlich gab es eine Poster-/Demosession, in der wissenschaftliche und kommerzielle Projekte ihre Forschungswerkzeuge und Methoden vorstellen konnten. Eine Übersicht über das Gesamtprogramm sowie Abstracts und Folien der Workshopvorträge sind online unter *http://hypermedia.ids-mannheim.de/gscl-ak/workshop12.html* einsehbar.

Ausgewählte Beiträge des Workshops finden sich nun – zum Teil als Ergebnis von im Anschluss an das Treffen angebahnten wissenschaftlichen Kooperationen – im vorliegenden Themenheft des Journal for Language Technology and Computational Linguistics (JLCL). Dabei wird ein breites Spektrum aktueller Forschungsfragen rund um die Thematik „Webkorpora“ abgedeckt:

- Michael Beißwenger und Lothar Lemnitzer geben einen Überblick über Motivation, Konzeption und laufende Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK), in dem ein deutschsprachiges Korpus zu Genres der internetbasierten Kommunikation aufgebaut wird. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS) konzipiert. Zunächst werden grundsätzliche Probleme der Repräsentation struktureller und linguistischer Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der Text Encoding Initiative (TEI) angesprochen, dann folgt eine Skizze möglicher Anwendungsszenarien.
- Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski und Torsten Zesch zeigen, wie sehr große Korpora für linguistische bzw. computerlinguistische Anwendungen auf Basis von Webinhalten – und deren Charakteristika – kompiliert werden können. Ihr Hauptaugenmerk liegt dabei auf den Bereichen Crawling, Vor- bzw. Weiterverarbeitung sowie Qualitätskontrolle. Weiterhin geben die Autoren einen Einblick in die Nutzung dieser Korpora für NLP-/CL-relevante Forschungsfragen.
- Bryan Jurish und Kay-Michael Würzner präsentieren eine neuartige Methode für die Segmentierung von Text in Sätze bzw. Token. Dabei nutzen sie Hidden Markov Modelle und berechnen Modellparameter unter Heranziehung der Segmentierung in etablierten Korpora bzw. Baumbanken. Die Verfahren werden an verschiedenen Korpora evaluiert, u.a. einem deutschen Korpus zur internetbasierten Kommunikation.
- Sabine Schulte im Walde und Stefan Müller präsentieren zwei Fallstudien, in denen die Eignung von Webkorpora für die automatische Akquisition lexikalisch-

semantischen Wissens untersucht wird. Hierfür vergleichen sie zwei unterschiedlich gut aufbereitete deutsche Webkorpora mit einem deutschen Wikipedia-Korpus und einem Zeitungskorpus. Die Fallstudien zeigen, wie sich verschiedene Parameter (z.B. Korpusgröße, Aufbereitung/Filterung, Domänenspezifik) auf die Qualität der Ergebnisse auswirken.

Wir danken allen Gutachtern und den JLCL-Herausgebern für die tatkräftige Unterstützung.
Wir hoffen, dass die Leser dieses JLCL-Themenhefts die darin enthaltenen Beiträge ebenso interessant und inspirierend finden wie wir.

Dezember 2013

Roman Schneider, Angelika Storrer, Alexander Mehler

Michael Beißwenger, Lothar Lemnitzer

Aufbau eines Referenzkorpus zur deutschsprachigen internetbasierten Kommunikation als Zusatzkomponente für die Korpora im Projekt „Digitales Wörterbuch der deutschen Sprache“ (DWDS)

Abstract

Dieser Beitrag gibt einen Überblick über die laufenden Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK), in dem ein Korpus zur Sprachverwendung in der deutschsprachigen internetbasierten Kommunikation aufgebaut wird. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache“ (*DWDS*, http://www.dwds.de) konzipiert, die die geschriebene deutsche Sprache seit 1900 dokumentieren.

Wir geben einen Überblick über die Motivation und Konzeption des Korpus sowie über die Projektziele (Abschnitte 2 und 3) und berichten über ausgewählte Anforderungen und Vorarbeiten im Zusammenhang mit der Korpuserstellung: a) die Integration des Korpus in die Korpusinfrastruktur des DWDS-Projekts (Abschnitt 4); b) die Entwicklung eines Schemas für die Repräsentation der strukturellen und linguistischen Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der *Text Encoding Initiative* (TEI-P5) (Abschnitt 5). Der Artikel schließt mit einer Skizze der Anwendungsszenarien für das Korpus in der korpusgestützten Sprachanalyse und der gegenwartssprachlichen Lexikographie (Abschnitt 6) sowie mit einem Ausblick (Abschnitt 7).

1. Einleitung

Dieser Beitrag gibt einen Überblick über die laufenden Arbeiten im Projekt „Deutsches Referenzkorpus zur internetbasierten Kommunikation“ (DeRiK), in dem in einer Kooperation der TU Dortmund und der Berlin-Brandenburgischen Akademie der Wissenschaften, Zentrum Sprache, seit 2010 ein Korpus zur Sprachverwendung in der deutschsprachigen internetbasierten Kommunikation aufgebaut wird. Am Projekt beteiligt sind neben den Verfassern dieses Beitrags Angelika Storrer (Dortmund) sowie Alexander Geyken und Maria Ermakova (Berlin). Die Basis des Korpus bilden nicht beliebige Webtexte, sondern die sprachlichen Äußerungen in solchen Webgenres, die in der englischsprachigen Forschung im Forschungsfeld „Computer-Mediated Communication“ (*CMC*) und in der deutschsprachigen Forschung unter dem Oberbegriff „Internetbasierte Kommunikation“ (*IBK*) untersucht werden.[1] Im Fokus stehen dabei Kommunikationstechnologien, die auf der Infrastruktur des Internet und seiner Dienste aufsetzen und die für die Realisierung

1 Überblicke zu den sprachlichen und kommunikativen Besonderheiten internetbasierter Kommunikation bieten z.B. Herring (1996; 2010/2011), Crystal (2001; 2011), Runkehl et al. (1998), Beißwenger (2001), Beißwenger & Storrer (2008) und Storrer (2013).

dialogischer interpersonaler Kommunikation konzipiert sind. Prominente Beispiele für Genres internetbasierter Kommunikation sind Chats und Instant-Messaging-Dialoge, Diskussions-Threads in Online-Foren und in Wikis, Threads mit Nutzerkommentaren in Weblogs, Videoplattformen (z.B. *YouTube*) und auf den Profilseiten sozialer Netzwerke (z.B. *Facebook*), die Kommunikation anhand von *Twitter*-Postings (Tweets) sowie in multimodalen Kommunikationsumgebungen wie *Skype*, MMORPGs („Massively Multi-Player Online Role Playing Games") und in „virtuellen Welten" (*SecondLife* u.a.).

Der Fokus von DeRiK liegt auf der schriftlichen Sprachverwendung in der internetbasierten Kommunikation. Das Korpus ist als eine Zusatzkomponente zu den Korpora im BBAW-Projekt „Digitales Wörterbuch der deutschen Sprache" (*DWDS*, http://www.dwds.de) konzipiert, die die geschriebene deutsche Sprache seit 1900 dokumentieren.

In den folgenden Abschnitten geben wir einen Überblick über die Motivation und Konzeption des Korpus sowie über die Projektziele (Abschnitte 2 und 3) und berichten über ausgewählte Anforderungen und Vorarbeiten im Zusammenhang mit der Korpuserstellung:

- die Integration des Korpus in die Korpusinfrastruktur des DWDS-Projekts (Abschnitt 4);
- die Entwicklung eines Schemas für die Repräsentation der strukturellen und linguistischen Besonderheiten von IBK-Korpora auf der Basis der Repräsentationsformate der *Text Encoding Initiative* (TEI-P5) (Abschnitt 5).

Der Artikel schließt mit einer Skizze der Anwendungsszenarien für das Korpus in der korpusgestützten Sprachanalyse und der gegenwartssprachlichen Lexikographie (Abschnitt 6) sowie mit einem Ausblick (Abschnitt 7).

2. Motivation: Ein Blick in die Korpuslandschaft

Gegenwärtig gibt es erst wenige Korpora zu Genres internetbasierter Kommunikation ('IBK-Korpora'). Bei den meisten existierenden Korpora mit IBK-Daten handelt es sich um Ressourcen, die für die interne Nutzung in einzelnen Forschungsprojekten aufgebaut wurden und die für die Fachcommunitys nicht ohne Weiteres zugänglich sind. Die unbefriedigende Abdeckung des wichtigen Kommunikationsbereichs „internetbasierte Kommunikation" in der Korpuslandschaft zur deutschen Gegenwartssprache (wie auch zu vielen anderen Sprachen) ist darauf zurückzuführen, dass in Bezug auf die Erhebung, Dokumentation, linguistische Annotation und Beschreibung von IBK-Daten wie auch auf Aspekte ihrer Bereitstellung für Forschungs- und Lehrzwecke derzeit noch viele offene Fragen und Herausforderungen bestehen. Forschungs- und Klärungsbedarf besteht u.a. hinsichtlich der folgenden Punkte:

- **Rechtliche und ethische Fragen:** Unter welchen Bedingungen können Sprachdaten aus Genres internetbasierter Kommunikation für die wissenschaftliche Nutzung archiviert, für linguistische Analysezwecke aufbereitet und annotiert und im Rahmen von Korpora als Forschungsressource bereitgestellt werden?
- **Fragen der Strukturrepräsentation und der Interoperabilität:** Welche Formate für die Repräsentation von Textgenres lassen sich sinnvoll für die Strukturbe-

schreibung von IBK-Korpora adaptieren? Wie müssen existierende Repräsentationsschemata angepasst werden, um die Struktur von Threads und Logfiles darzustellen? Welche Modellierungseinheiten können aus existierenden Schemata übernommen werden und für welche Einheiten bedarf es neuer, IBK-spezifisch angepasster Modelle?

- **Fragen der linguistischen Annotation:** Wie können Werkzeuge für die automatische Sprachverarbeitung (für die Tokenisierung, Normalisierung, Lemmatisierung, das Part-of-speech-Tagging und die syntaktische Annotation) sowie die von ihnen verwendeten Tagsets für den Umgang mit orthographischen Normabweichungen und IBK-spezifischen Stilelementen (z.B. Emoticons, Aktionswörter, Hashtags, Adressierungen) sowie für die Behandlung von Phänomenen der konzeptionellen Mündlichkeit angepasst werden?

- **Fragen der Integration von IBK-Ressourcen in bestehende Korpusinfrastrukturen:** Wie können IBK-Daten in existierende Infrastrukturen für die Verwaltung, Bereitstellung und Abfrage von Sprachkorpora integriert werden? Wie lassen sie sich vergleichend mit anderen Typen von Korpora (Textkorpora, Korpora gesprochener Sprache) nutzen und analysieren? Wie sind IBK-Daten an der Nutzerschnittstelle von Korpusabfragesystemen zu präsentieren? Wie können IBK-Korpora anhand von Metadaten beschrieben werden?

Diese und weitere Herausforderungen beim Aufbau von IBK-Korpora sind in der Forschung bekannt (vgl. z.B. Beißwenger & Storrer 2008, King 2009, Storrer 2013) und aktuell Thema verschiedener Netzwerke und Arbeitsgruppen – u.a. des DFG-Netzwerks „Empirische Erforschung internetbasierter Kommunikation“ (*Empirikom*, http://www.empirikom.net) und des GSCL-Arbeitskreises *Social Media /Internetbasierte Kommunikation* (http://gscl.org/ak-ibk.html) –, für ihre Bearbeitung gibt es bislang aber erst wenige „Best Practice“-Beispiele. Einen Überblick über IBK-Korpora auf dem Stand von 2008 geben Beißwenger & Storrer (2008). Beispiele für das Englische sind das *NPS Chat Corpus* (Forsyth et al. 2007) sowie das *Queer Chat-Room Corpus* (King 2009), für das Flämische das *Netlog Corpus* (Kestemont et al. 2012). Für das Deutsche existiert mit dem von 2002 bis 2008 an der TU Dortmund aufgebauten *Dortmunder Chat-Korpus* eine Ressource zur Sprachverwendung und sprachlichen Variation in der Chat-Kommunikation, die seit 2005 auch online zur Verfügung steht (http://www.chatkorpus.tu-dortmund.de, Beißwenger 2013). Für das Französische gibt es mit den *Learning and Teaching Corpora* (*LETEC*) eine Sammlung mit Daten aus multimodalen Lehr-/Lernarrangements (Reffay et al. 2012); mit *CoMeRe* (*Corpus de communication médiée par les réseaux*, http://comere.org) befindet sich darüber hinaus gegenwärtig ein genreheterogenes Korpus zur französischsprachigen internetbasierten Kommunikation im Aufbau. Im Projekt *Web2Corpus_it* (*Corpus Italiano di Communicazione mediata dal Computer,* http://www.glottoweb.org/web2corpus/) soll ein ausgewogenes IBK-Korpus für das Italienische entstehen. Beispiele für Referenzkorpora zur Gegenwartssprache, die Teilkorpora zu IBK-Genres integrieren, sind das niederländische *SoNaR*-Korpus (*Stevin Nederlandstalig Referentiecorpus*, Reynaert et al. 2010, Oostdijk et al. 2013)

sowie für das Estnische das *National Corpus of Estonian Language* (http://www.cl.ut.ee/korpused/segakorpus/).

Ein wichtiges Desiderat für das Deutsche ist ein Referenzkorpus zur internetbasierten Kommunikation, das für linguistische Analysezwecke aufbereitet ist, in einem anerkannten Repräsentationsformat zur Verfügung gestellt wird, vergleichende Analysen mit der Sprachverwendung in redigierten Texten (Textkorpora) ermöglicht und das in Forschung und Lehre als Basis für empirische Untersuchungen zur sprachlichen Variation in der internetbasierten Kommunikation sowie für die datengestützte Vermittlung ihrer sprachlichen und kommunikativen Besonderheiten genutzt werden kann. Das DeRiK-Projekt strebt an, einen Beitrag zur Schließung dieser Lücke in der Korpuslandschaft zur deutschen Gegenwartssprache zu leisten.

3. Zur Konzeption des Korpus

Ziel des DeRiK-Projekts ist der Aufbau einer Zusatzkomponente zu den DWDS-Korpora, die die deutschsprachige internetbasierte Kommunikation dokumentiert und durch deren Integration in die DWDS-Korpusinfrastruktur es u.a. möglich werden soll, die schriftliche Sprachverwendung im Internet vergleichend mit der schriftlichen Sprachverwendung in redigierten Texten zu untersuchen. Redigierte Texte sind in den Korpora des DWDS-Projekts bereits umfangreich dokumentiert: Das Kernkorpus umfasst Texte aus den Textsortenbereichen Belletristik, Wissenschaft, Zeitung und Gebrauchstexte für alle Dekaden seit 1900 (vgl. Geyken 2007). Die schriftliche Sprachverwendung im Netz ist in den Korpora hingegen bislang nicht berücksichtigt. Die DeRiK-Komponente soll diesen wichtigen Kommunikationsbereich nicht nur für die empirische Forschung zu IBK-Phänomenen, sondern auch für den Bereich der lexikographischen Bearbeitung der deutschen Gegenwartssprache korpuslinguistisch erschließen.

Unter der letztgenannten Perspektive erweitert DeRiK die Ressourcen für die Erarbeitung des DWDS-Wörterbuchs in zweierlei Weise: Zum einen wird das Korpus zahlreiche Beispiele für schriftliche Sprachverwendung im Duktus der konzeptionellen Mündlichkeit (i.S.v. Koch & Oesterreicher 1994) umfassen. Dies ermöglicht es, in die Wörterbuchartikel auch Belegbeispiele aus Kontexten informeller Schriftlichkeit aufzunehmen, was insbesondere für die Beschreibung umgangssprachlich markierter Lexik oder von sprachlichen Einheiten, die wichtige Aufgaben bei der Handlungskoordination in Gesprächen übernehmen (z.B. Interjektionen), bedeutsam ist. Zum anderen ist zu erwarten, dass in Daten zur Sprachverwendung in IBK-Genres zahlreiche Wörter und Wortbedeutungen dokumentiert sind, die mit traditionellen Quellen arbeitenden Lexikographen vermutlich entgehen (vgl. die Beispiele in Abschnitt 4). Ein regelmäßig aktualisiertes, hinsichtlich der Genres breit gestreutes Referenzkorpus zur internetbasierten Kommunikation stellt somit für die korpusbasierte Lexikographie eine wichtige Ergänzung zu den bisher typischerweise verwendeten Ressourcen dar. Durch ihre Verfügbarkeit kommt man dem Versprechen der gegenwartssprachlichen Lexikographie, den Wortschatz einer Sprache in seiner Gänze zu beschreiben, näher.

DeRiK ist konzipiert als

- ein **Referenzkorpus** zur internetbasierten Kommunikation, das der Fachcommunity als Basis für korpusgestützte Untersuchungen und für die Vermittlung der

sprachlichen Besonderheiten von IBK-Genres in der Lehre zur Verfügung gestellt wird;

- ein **ausgewogenes Korpus**, das – soweit rechtlich möglich – Daten aus den meistgenutzten IBK-Genres umfasst und die einzelnen Genres nach Popularität gegeneinander gewichtet;
- ein **zeitlich gestaffeltes Korpus**, bei dem – analog zur Datenerhebung für das DWDS-Kernkorpus – die Datenerhebung nicht nur einmalig, sondern mehrfach in regelmäßigen Abständen erfolgen soll, wodurch es möglich wird, auch sprachlichen Wandel *innerhalb* der internetbasierten Kommunikation darzustellen;
- ein **annotiertes Korpus**, das neben einer linguistischen Basisannotation auch Annotationen zu charakteristischen sprachlichen und strukturellen Besonderheiten bei der Sprachverwendung im Netz umfasst.

Die anvisierte Größe des Korpus sind 10 Millionen Token je Dekade, beginnend mit dem Jahr 2010 (dem Jahr des Projektbeginns). Für die Zusammensetzung des Korpus wurde ein Idealschlüssel entwickelt, der von den Ergebnissen der jährlich durchgeführten *ARD/ZDF-Onlinestudie* (http://www.ard-zdf-onlinestudie.de/, vgl. van Eimeren & Frees 2013) ausgeht und aus den in der Studie beschriebenen Präferenzen deutscher Internetnutzer für internetbasierte Kommunikationstechnologien und der Online-Affinität verschiedener Altersgruppen einen Faktor für die Gewichtung unterschiedlicher IBK-Genres bei der Festlegung der Datensets für einen Erhebungszeitraum ableitet. Der Schlüssel nutzt verschiedene Teilergebnisse der Studie als Grundlage: zum einen die Verbreitung der Internetnutzung nach Altersgruppen (vgl. Tab. 1), zum anderen die Präferenzen der Nutzer für bestimmte Typen von Online-Anwendungen (vgl. Tab. 2).

	1997	2000	2003	2006	2009	2010	2011	2012	2013
Gesamt	6,5	28,6	53,5	59,5	67,1	69,4	73,3	75,9	77,2
14-19 J.	6,3	48,5	92,1	97,3	97.5	100,0	100,0	100,0	100,0
20-29 J.	13,0	54,6	81,9	87,3	95,2	98,4	98,2	98,6	97,5
30-39 J.	12,4	41,1	73,1	80,6	89,4	89,9	94,4	97,6	95,5
40-49 J.	7,7	32,2	67,4	72,0	80,2	81,9	90,7	89,4	88,9
50-59 J.	3,0	22,1	48,8	60,0	67,4	68,9	69,1	76,8	82,7
ab 60 J.	0,2	4,4	13,3	20,3	27,1	28,2	34,5	39,2	42,9

Tab. 1: Internetnutzer in Deutschland 1997 bis 2013 nach Alter: mindestens gelegentliche Nutzung, in %. Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=421

	Ge-samt	14-29	30-49	50-69	ab 70
Suchmaschinen nutzen	83	90	87	76	61
senden/empfangen von E-Mails	79	80	85	73	64
zielgerichtet bestimmte Angebote/informationen suchen	72	80	77	64	50
einfach so im Internet surfen	44	57	45	35	22
Onlinecommunitys nutzen	39	76	38	13	7
sog. "Apps" auf Mobilgeräten nutzen, um ins Internet zu gehen	35	60	35	17	8
Homebanking	34	33	39	31	31
Videoportale nutzen	32	65	28	11	7
Chatten	26	59	20	9	3
Herunterladen von Dateien	23	35	22	15	6
Kartenfunktionen nutzen	20	27	20	15	10
Onlinespiele	16	23	17	9	7
Audios im Internet herunterladen/anhören	14	31	12	5	0
Musikdateien aus dem Internet	14	33	9	4	0
Video/TV zeitversetzt	13	24	11	11	4
live im Internet Radio hören	13	22	11	8	2
RSS-feeds/Newsfeeds	10	18	10	4	4
Gesprächsforen	10	15	12	4	2
Ortungsdienste für ortsbezogene Informationen nutzen	10	14	8	9	5

Tab. 2: Onlineanwendungen 2013 nach Alter: mindestens einmal wöchentlich genutzt, in % (Ausschnitt). Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=423. Anwendungstypen mit Relevanz für DeRiK (= internetbasierte Kommunikationstechnologien oder Anwendungen mit entsprechenden Funktionen) sind in der Tabelle hervorgehoben.

Die Nutzung von Onlinecommunitys wurde in der ARD/ZDF-Onlinestudie zudem auch in einer Teilstudie zu den Präferenzen für ausgewählte „Social Media"- bzw. „Web 2.0"-Anwendungen abgefragt. In dieser Teilstudie wird weiter differenziert nach privaten und

beruflichen Communitys; daneben wird die Nutzung der Wikipedia, von Weblogs und von Twitter sowie von Videoportalen und Fotocommunitys dargestellt (vgl. Tab. 3 sowie im Detail Busemann 2013). Der ideale Schlüssel für DeRiK sieht vor, die in den Tab. 2 und 3 dargestellten Nutzungspräferenzen für die verschiedenen Altergruppen mit einem Faktor zu multiplizieren, der der Online-Affinitität der jeweiligen Altersgruppe entspricht und der sich aus den in Tab. 1 dargestellten Zahlen ableitet. So würde beispielsweise die Präferenz der 14-29-Jährigen für die Nutzung von Chats in der Studie aus 2013 (59%, s. Tab. 2) mit dem Faktor 0,9875 multipliziert (= Durchschnitt aus 100% und 97,5% Online-Nutzung bei der Altersgruppe der 14-19- und der 20-29-Jährigen, Tab. 1), die Präferenz der 30-49-Jährigen (20%) hingegen mit dem Faktor 0,922 (= Durchschnitt aus 95,5% und 88,9% Online-Nutzung bei der Altersgruppe der 30-39- und der 40-49-Jährigen) usf. Der exakte Anteil von Daten aus einem bestimmten Anwendungstyp (z.B. Chats) in der Gesamtdatenmenge für einen Erhebungszeitraum ergäbe sich schließlich aus dem Durchschnitt der gewichteten Präferenzen aller Altersgruppen im Verhältnis zu den ermittelten Werten für alle anderen relevanten Anwendungstypen. Für Anwendungstypen, deren Präferenzen in beiden Teilstudien der ARD/ZDF-Onlinestudie abgefragt wurden (Onlinecommunitys, Videoportale) sollen dabei jeweils die Werte aus der Teilstudie „Onlineanwendungen“ zugrunde gelegt werden.

In Bezug auf die in Tab. 3 dargestellten Werte zur Nutzung von Wikipedia, Weblogs und Twitter ist zu bedenken, dass bei allen drei (Typen von) Anwendungen die rezeptive Nutzung höher ist als die produktive Nutzung: Viele Onliner nutzen die Wikipedia als Nachschlagewerk, nur ein Teil der Nachschlagenden tritt aber selbst als Autor von Wikipedia-Artikeln und als Diskutant auf Diskussionsseiten in Erscheinung; dennoch können die Diskussionsseiten auch von nicht aktiv zum Ausbau der Anwendung beitragenden Nutzern eingesehen werden. Tweets, die in Twitter gepostet werden, können von einer Vielzahl von Nutzern (auch solchen, die nicht selbst aktiv „twittern“) gelesen werden; Gleiches gilt für Kommentare zu Weblog-Einträgen. Da die Frage der Gewichtung der rein rezeptiven gegenüber der auch produktiven Nutzung in Bezug auf Wikipedia, Twitter und Weblogs schwierig zu beantworten ist, ist bislang vorgesehen, die bei der Berechnung des Idealschlüssels unberücksichtigt zu lassen: Da in der Wikipedia geführte schriftliche Diskussionen sowie Tweets und Weblog-Kommentare auch für nur rezeptiv Zugreifende jederzeit einsehbar sind und ein Wechsel von der ausschließlich rezeptiven zur auch produktiven Nutzung der benannten Anwendungstypen jederzeit möglich ist, wird hier vorerst nicht weiter differenziert.

Bei der Entscheidung, aus welchen Online-Anwendungen konkret Daten für die einzelnen Anwendungstypen erhoben werden, soll Wert darauf gelegt werden, Vielfalt – und damit sprachliche Variation bei der Nutzung ein- und desselben Anwendungstyps – abzubilden. So wird beispielsweise für Chats angestrebt, die Daten nicht sämtlich aus demselben Chat-Angebot zu erheben und zudem die Nutzung von Chats in unterschiedlichen Handlungsbereichen abzubilden (Freizeitkommunikation, berufliche Nutzung, Nutzung in Lehr/Lernkontexten). Analog soll auch bei allen anderen Anwendungstypen verfahren werden.

Da die Onlinestudie jährlich aktualisiert wird, wird es möglich sein, den Schlüssel je Erhebungszeitraum an die jeweils aktuellen Zahlen anzupassen und dabei ggf. auch neu aufkommende IBK-Genres zu berücksichtigen.

	Gesamt	14-19	20-29	30-39	40-49	50-59	60-69	ab 70
Wikipedia	74	95	93	81	77	61	47	32
Videoportale (z.B. YouTube)	60	91	87	71	62	43	25	13
private Netzwerke u. Communitys	46	87	80	55	38	21	16	6
Fotosammlungen, Communitys	27	28	38	37	26	16	17	13
berufliche Netzwerke u. Communitys	10	5	14	19	13	4	2	0
Weblogs	16	18	31	19	17	7	3	5
Twitter	7	22	10	7	5	3	4	0

Tab. 3: Nutzung von Web 2.0-Anwendungen nach Alter 2013: gelegentliche Nutzung, in %. Quelle: http://www.ard-zdf-onlinestudie.de/index.php?id=397

Der Idealschlüssel wird bei der Datenerhebung allerdings nur mit Einschränkungen umgesetzt werden können: Aufgrund der unklaren Rechtslage in Bezug auf die Erhebung, Aufbereitung, Bereitstellung und Nutzung von IBK-Daten in Korpora werden bis auf Weiteres für das Korpus nur Daten aus solchen Kommunikationsumgebungen im Netz erhoben werden können, bei denen die Nutzung unproblematisch bzw. durch explizit deklarierte Lizenzen geregelt ist – u.a. Daten aus Anwendungen, deren Inhalte unter CC-BY-SA („Creative Commons: Namensnennung – Weitergabe unter gleichen Bedingungen“[2]) oder vergleichbaren Modellen für eine Bearbeitung und Weitergabe lizenziert sind.

4. Integration des Korpus als Zusatzkomponente in das DWDS-Korpusframework

Das Digitale Wörterbuch der deutschen Sprache ist ein digitales lexikalisches System, das an der Berlin-Brandenburgischen Akademie der Wissenschaften entwickelt wurde und weiterentwickelt wird. Das System eröffnet den Nutzern einen integrierten Zugang zu drei verschiedenen Typen von Ressourcen (Geyken 2007, Klein & Geyken 2010):

2 http://creativecommons.org/licenses/by-sa/2.0/de/legalcode (der Lizenztext) und http:// creativecommons.org/licenses/by-sa/2.0/de/ (eine verständliche Zusammenfassung des Inhalts).

a) **Lexikalische Ressourcen:** ein gegenwartssprachliches Wörterbuch, basierend auf dem retrodigitalisierten *Wörterbuch der deutschen Gegenwartssprache* (WDG, Klappenbach & Steinitz 1964-1977), das *Etymologische Wörterbuch des Deutschen* (Pfeifer 1993), die Erstbearbeitung des *Deutschen Wörterbuchs* von Jacob Grimm und Wilhelm Grimm (Jacob Grimm & Wilhelm Grimm, 1852-1971) sowie ein Thesaurus *(Openthesaurus)*.

b) **Korpora:** Das DWDS bietet ein ausgewogenes Korpus des 20. und des frühen 21. Jahrhunderts und darüber hinaus Zeitungskorpora und Spezialkorpora. Die jüngsten Texte stammen momentan aus dem Jahr 2010.

c) **Statistische Ressourcen für Wörter und Wortkombinationen:** Angeboten werden Wortprofile und Wortverlaufskurven für den gegenwartssprachlichen Wortschatz; die Auswertungen basieren auf dem Kernkorpus und den Zeitungskorpora.

Ergebnisse zu Suchanfragen auf diesen Ressourcen werden in einer panelbasierten, nutzerkonfigurierbaren Sicht (s. Abb. 1) präsentiert. Jede Ressource wird in einem eigenen Fenster angezeigt (*Panel*), eine Sicht (*View*) besteht aus einem oder mehreren Fenstern. Das System stellt dem Benutzer eine Reihe vorkonfigurierter Sichten zu Verfügung (Standardsicht, Korpussicht usw.). Darüber hinaus kann jeder registrierte Benutzer sich aus den bestehenden Ressourcen/Panels eine oder mehrere private Sichten erstellen und nutzen (mehr dazu in Klein & Geyken 2010: Abschnitt 6.4).

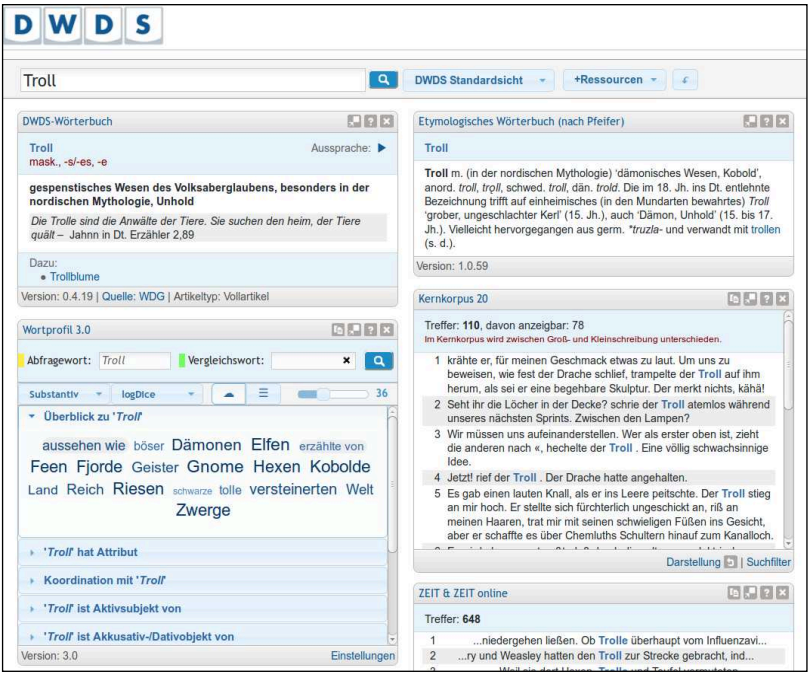


Abb. 1: Panelbasierte Sicht des DWDS, Stichwort ‚Troll'.

Nutzer können sich auf diese Weise schnell ein Bild über ein Wort (oder eine Wortkombination), seine Bedeutung und seinen Gebrauch in der geschriebenen Gegenwartssprache machen.

Der modulare Aufbau des Digitalen Lexikalischen Systems – Informationen aus heterogenen Ressourcen werden „unter dem Dach" einer Suchanfrage angeboten – erleichtert die Integration weiterer Ressourcen, sofern diese hinsichtlich Format und Annotation zu den vorhandenen Ressourcen kompatibel sind. Die vorhandenen Ressourcen wurden wie folgt aufbereitet:

- **Linguistische Annotation:** Die Korpora wurden mit den an der BBAW entwickelten sprachtechnologischen Werkzeugen linguistisch annotiert. Die Segmentierung der Texte, die Tokenisierung und die Wortartenannotation erfolgen mit dem Part-of-Speech-Tagger *moot* (Jurish, 2003). Für die Lemmatisierung der Textwörter wird die *TAGH*-Morphologie verwendet (Geyken & Hanneforth 2006). Dies ermöglicht die Formulierung komplexerer linguistischer Abfragen, z.B. Suchmuster als Kombinationen von Wortformen, Lemmata und Wortarten. Die für das DWDS entwickelte Suchmaschine *DDC* kann solche komplexen Abfragen bearbeiten. Auch komplexe statistische Auswertungen, z.B. zu typischen Wortkombi-

nationen in bestimmten syntaktischen Beziehungen, bauen auf diese Annotationen auf.

- **Korpusrepräsentation:** Korpora und lexikalische Ressourcen sind durchgängig in XML nach den Kodierungsregeln der Text Encoding Initiative (TEI-P5) ausgezeichnet. Dies umfasst sowohl die Auszeichnung der Primärdaten als auch die Bereitstellung der Metadaten.

Das DeRiK-Korpus soll konform zu den genannten Anforderungen aufbereitet werden. Hinsichtlich der linguistischen Annotation müssen die an der BBAW entwickelten sprachtechnologischen Verfahren an die sprachlichen Besonderheiten schriftlicher internetbasierter Kommunikation angepasst werden, um ein qualitativ akzeptables Analyseergebnis zu erzielen. Vor allem müssen diejenigen Phänomene berücksichtigt und angemessen behandelt werden, die charakteristisch für das interaktionsorientierte Schreiben in sozialen Medien sind, z.B. Schnellschreibphänomene und Phänomene geschriebener Umgangssprache sowie Elemente der „Netzsprache", die in redigierten Texten nur in Ausnahmefällen auftreten (z.B. Emoticons, Aktionswörter, Adressierungen). Für die Bearbeitung dieser Aufgabe wird derzeit in Kooperation mit dem DFG-Netzwerk *Empirikom* eine Community-Shared-Task zur automatischen linguistischen Analyse und Annotation deutscher IBK-Daten vorbereitet (vgl. http://empirikom.net/bin/view/Themen/SharedTask). Vorschläge für die Erweiterung des STTS-Tagsets um Einheiten für das POS-Tagging von „Netzsprache"-Phänomenen sind in Bartz et al. (2013) beschrieben und werden in der Arbeitsgruppe zur Erweiterung von STTS diskutiert (s. JLCL 2013, Heft 1).

Hinsichtlich der Repräsentation der Primärdaten müssen die DeRiK-Daten in einem TEI-kompatiblen Format repräsentiert werden. Für diesen Zweck wurde von 2010 bis 2012 ein auf den TEI-P5-Formaten basierendes – und deshalb TEI-konformes – und auf die Besonderheiten der schriftlichen internetbasierten Kommunikation angepasstes Repräsentationsschema entwickelt. Die Eckpunkte dieses Schemas sowie seiner geplanten Weiterentwicklung im Zusammenhang mit der Special Interest Group „Computer-Mediated Communication" der TEI werden im folgenden Abschnitt beschrieben.

5. TEI-Repräsentation der Korpusdaten: Stand der Arbeiten und Perspektiven

Da IBK-Korpora einen vergleichsweise neuen Typus von Korpora mit spezifischen strukturellen und linguistischen Besonderheiten darstellen (vgl. Storrer 2013a: Abschnitt 4), existieren in den 'Digital Humanities' bislang keine Standards oder Quasi-Standards für die Repräsentation der in ihnen erfassten Datentypen und Genres. Wer bislang annotierte Korpora mit Sprachdaten aus Genres internetbasierter Kommunikation aufbaut, muss dafür i.d.R. eigene Annotationsschemata entwickeln. So wurde beispielsweise für das Dortmunder Chat-Korpus ein Schema definiert, das speziell auf die Strukturmodellierung von Chat-Mitschnitten (Logfiles), die Annotation unterschiedlicher Typen von Nutzerbeiträgen sowie die Auszeichnung ausgewählter „Netzsprache"-Phänomene zugeschnitten ist (vgl. Beißwenger 2013).

Auch für DeRiK stellt sich das Problem fehlender Standards im Bereich der Repräsentation von Korpusdokumenten mit Sprachdaten aus IBK-Genres. Da die bereits in der DWDS-

Korpusinfrastruktur vorhandenen Ressourcen in TEI-P5 repräsentiert sind, sollen auch die DeRiK-Daten auf der Basis von TEI-P5 annotiert werden. Zum gegenwärtigen Zeitpunkt finden sich in den in TEI-P5 enthaltenen Formaten allerdings noch keine Modelle, die auf die Repräsentation der strukturellen und sprachlichen Besonderheiten von IBK-Genres zugeschnitten sind oder die sich ohne Weiteres für die Repräsentation von IBK-Daten übernehmen lassen. Allerdings bietet das Encoding Framework der TEI die Möglichkeit, vorhandene Modelle auf die Erfordernisse neuer, bislang noch nicht im Standard berücksichtigter Genres anzupassen (in der TEI-Terminologie als *customization* bezeichnet):
Because the TEI Guidelines must cover such a broad domain and user community, it is essential that they be customizable: both to permit the creation of manageable subsets that serve particular purposes, and also to permit usage in areas that the TEI has not yet envisioned. (http://www.tei-c.org/Guidelines/Customization/)
Bei der *customization* erlaubt ist u.a. das Hinzufügen neuer, die Entfernung nicht benötigter und die Umbenennung vorhandener TEI-Elemente, die Anpassung von Inhaltsmodellen, das Hinzufügen und Entfernen von Attributen sowie die Modifikation von Attributwertlisten zu Elementen. Solange dabei bestimmte Regeln für die Spezifikation neuer Elemente und Attribute und für die Dokumentation individuell vorgenommener Modifikationen eingehalten werden, bleiben auf diese Weise erzeugte Repräsentationsschemata – obwohl sie Modelle umfassen, die selbst nicht Teil des Standards sind – kompatibel zu TEI-P5.

Das für DeRiK entwickelte Schema macht von dieser Möglichkeit der Anpassung Gebrauch. Die Basis für das Schema bildet das Modul „Basic text structure“ aus TEI-P5; dieses Modul wird für die Zwecke der Annotation von IBK-Genres spezifisch erweitert und modifiziert. Zentrale Komponenten des Schemas sind

- eine Komponente für die Beschreibung der *Makrostruktur* von Genres schriftlicher internetbasierter Kommunikation;
- eine Komponente für die Beschreibung ausgewählter „netztypischer“ Stilelemente innerhalb einzelner Nutzerbeiträge (= *Mikrostruktur* von IBK).

Im Folgenden geben wir einen Überblick über einige wichtige Eckpunkte des Schemas; eine ausführliche Beschreibung findet sich in Beißwenger et al. (2012), das zugehörige ODD-Dokument sowie Annotationsbeispiele können unter http://empirikom.net/bin/view/Themen/CmcTEI heruntergeladen werden.

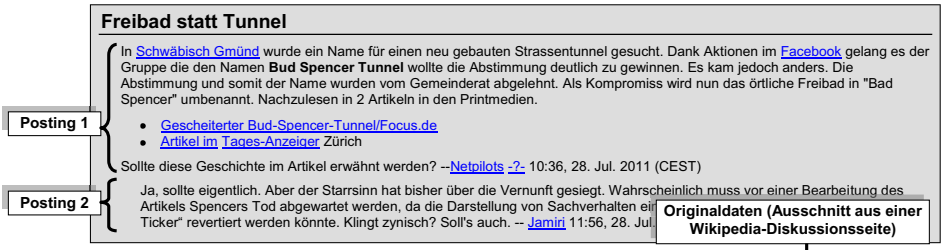
Die grundlegende Modellierungseinheit des Schemas auf der Ebene der *Makrostruktur* von IBK-Dokumenten bildet das *Posting*, das spezifiziert ist als eine Zeichensequenz, die zu einem bestimmten Zeitpunkt von einem Nutzer – etwa durch Betätigung der Eingabetaste – *en bloc* an den Server übermittelt und anschließend als neuer Nutzerbeitrag am Bildschirm angezeigt wird.

Postings werden in unterschiedlichen IBK-Genres auf unterschiedliche Arten zu größeren Einheiten zusammengeordnet. Das Schema unterscheidet zwischen zwei Typen von IBK-Makrostrukturen:

- dem Strukturtyp ‚Logfile’, bei dem Postings, auf- oder absteigend, linear chronologisch nach ihren Eintreffenszeitpunkten beim Server dargestellt werden; die Abfolge der Postings wird dabei vom Server festgelegt;

- dem Strukturtyp ‚Thread', bei dem Postings entlang einer *oben/unten*- und einer *links/rechts*-Dimension auf der Bildschirmseite platziert werden. *Oben/unten* symbolisiert dabei prototypischerweise eine *vorher/nachher*-Relation, unterschiedliche Grade der Rechtseinrückung auf der *links/rechts*-Dimension symbolisieren thematische Bezüge auf Vorgängerpostings. In manchen Systemen (z.B. klassischen Foren mit Baumstrukturdarstellung) wird die Platzierung auf der *oben/unten*- und auf der *links/rechts*-Dimension automatisch erzeugt; spezifiziert der Nutzer ein bestimmtes Vorgängerposting als Bezugsbeitrag für sein eigenes Posting, so wird das eigene Posting nach der Verschickung relativ zum Bezugsbeitrag um eine Ebene eingerückt. Auf Wikipedia-Diskussionsseiten können Nutzer die Platzierung ihrer Postings – sowohl in der Horizontalen wie auch in der Vertikalen – hingegen selbst frei festlegen.

Zu jedem Posting wird ein Nutzer als Autor spezifiziert, der Name des Nutzer wird dabei durch eine ID ersetzt. Die Modellierung der als Posting-Autoren im Dokument dokumentierten Nutzer inklusive der zu ihnen im Dokument gegebenen Informationen (z.B. Inhalt der individuellen Nutzersignatur) wird in Personenprofilen im Dokument-Header gespeichert. Über eine ID-Referenz im <posting>-Element kann die ID auf den tatsächlichen Nutzernamen bezogen werden. Für die Bereitstellung des Korpus lassen sich die Korpusdokumente durch die Trennung der Nutzerinformationen von den Postings einfach anonymisieren. Abb. 2 zeigt am Beispiel eines Ausschnitts aus einer Wikipedia-Diskussionsseite die Annotation zweier Postings. Die Nutzer sind im Element <listPerson> modelliert, die Postings sind über das Attribut *@who* den Einträgen in der Liste zugeordnet. Die Veröffentlichungszeitpunkte der Postings sind im Element <timeline> modelliert; die Postings sind über das *@synch*-Attribut den Einträgen in der Timeline zugeordnet.




```
<listPerson>
        <person xml:id="A01">
                <persName>Netpilots</persName>
                <signatureContent><ref target="http://de.wikipedia.org/wiki/Benutzer:Netpilots">Netpilots</ref><ref
                target="http://de.wikipedia.org/wiki/Benutzer_Diskussion:Netpilots">-Ì¡-</ref></signatureContent>
        </person>
        <person xml:id="A02">
                <persName>Jamiri</persName>
                <signatureContent><ref
                target="http://de.wikipedia.org/wiki/Benutzer:Jamiri">Jamiri</ref></signatureContent>
        </person>
        …
</listPerson>
<front>
        <timeline>
                <when xml:id="t01" absolute="2011-07-27T16:46:00"/>
                <when xml:id="t02" absolute="2011-07-28T10:36:00"/>
                …
        </timeline>
</front>
<body>
        <div type="thread">
                <head>Freibad statt Tunnel</head>
                <posting synch="#t01" who="#A07">
                        <p>In<ref target="http://de.wikipedia.org/wiki/Schw%C3%A4bisch_Gm%C3%BCnd">Schwäbisch
                        Gmünd</ref> wurde ein Name für einen neu gebauten Strassentunnel gesucht. Dank Aktionen im <ref
                        target="http://de.wikipedia.org/wiki/Facebook">Facebook</ref> gelang es der Gruppe die den Namen
                        Bud Spencer Tunnel wollte die Abstimmung deutlich zu gewinnen. Es kam jedoch anders. Die
                        Abstimmung und somit der Name wurden vom Gemeinderat abgelehnt. Als Kompromiss wird nun das
                        örtliche Freibad in "Bad Spencer" umbenannt. Nachzulesen in 2 Artikeln in den Printmedien.</p>
                        <list>
                                <item><ref target="http://www.focus.de/panorama/welt/stuermische-ratssitzung-kein-bud-
                                spencer-tunnel-in-schwaebisch-gmuend_aid_649932.html,">Gescheiterter Bud-Spencer-
                                Tunnel/Focus.de</ref></item>
                                <item><ref target="http://www.tagesanzeiger.ch/leben/gesellschaft/Grosse-Hysterie-um-einen-
                                alten-Mann-/story/17754241">Artikel im</ref> <ref target="http://de.wikipedia.org/wiki/Tages-
                                Anzeiger">Tages-Anzeiger</ref> Zürich</item>
                        </list>
                        <p>Sollte diese Geschichte im Artikel erwähnt werden? -- <autoSignature/></p>
                </posting>
                <posting synch="#t02" who="#A06" indentLevel="1">
                        <p>Ja, sollte eigentlich. Aber der Starrsinn hat bisher über die Vernunft gesiegt. Wahrscheinlich muss
                        vor einer Bearbeitung des Artikels Spencers Tod abgewartet werden, da die Darstellung von
                        Sachverhalten einer noch lebenden Person sonst als „Live-Ticker“ revertiert werden könnte. Klingt
                        zynisch? Soll's auch. -- <autoSignature/></p>
                </posting>
                …
        </div>
</body>
```

Abb. 2: Encoding-Beispiel: Wikipedia-Diskussionsseite.

Während auf der *Makroebene* einzelne Postings annotiert sowie Strukturen oberhalb der Postingebene modelliert werden, beschreiben wir auf der *Mikroebene* von IBK-Dokumenten

sprachliche Besonderheiten *innerhalb* von Postings. Die Annotation bezieht sich hier also auf die schriftlichen Beiträge der Verfasser zum Kommunikationsgeschehen, die als Postings an den Server geschickt werden. Von besonderem Interesse für die korpusgestützte Analyse und die lexikographische Bearbeitung internetbasierter Kommunikation sind dabei solche Einheiten, die als „typisch netzsprachlich" gelten:

- *Emoticons*, die durch die Kombination von Interpunktions-, Buchstaben- und Sonderzeichen gebildet werden, ikonisch fundiert sind und daher übereinzelsprachlich verwendet werden können. Emoticons dienen typischerweise der emotionalen Kommentierung, der Markierung von Ironie oder der Bewertung von Partneräußerungen. In unterschiedlichen Kulturkreisen haben sich unterschiedliche Stile herausgebildet (z.B. westlicher, japanischer, koreanischer Stil), deren Verwendung aber nicht auf die jeweiligen Ursprungskulturen beschränkt geblieben ist. So sind in vielen deutschsprachigen Online-Communities neben den „klassischen" Emoticons westlichen Stils inzwischen u.a. auch japanische Emoticons gebräuchlich.

- *Aktionswörter* (*interaction words*), die zur sprachlichen Beschreibung von Gesten, mentalen Zuständen oder Handlungen verwendet werden und als Emotions- oder Illokutionsmarker, Ironiemarker oder für die spielerische Nachbildung fiktiver Handlungen verwendet werden. Sie sind einzelsprachlich gebunden und basieren auf einem Wort – typischerweise einem unflektierten Verbstamm –, das entweder alleine steht (*lach*, *freu*, *grübel*) oder um weitere Einheiten erweitert ist (*malnachdenk*, *stirnrunzel*).

- *Adressierungen*, mit denen Kommunikationsbeiträge an einen bestimmten anderen Kommunikationsbeteiligten oder eine Gruppe von Kommunikationsbeteiligten adressiert werden und die häufig zur Sicherstellung von thematischen Bezügen auf Vorbeiträge (des/der benannten Adressaten) verwendet werden (*@tinchen*, *@alle*).

- *„Interaction templates"*, die ähnliche Funktionen übernehmen wie Emoticons und Aktionswörter, bei denen es sich aber weder um tastaturschriftlich erzeugte noch um frei formulierte sprachliche Einheiten handelt, sondern um Einheiten, die – als statische oder animierte Grafiken oder als Kombinationen aus Grafiken und vordefinierten Textbausteinen – von den Nutzern durch Eingabe bestimmter Codes oder per Auswahl aus einem Grafikmenü in ihre Beiträge eingefügt werden. Beispiele sind die sog. „Grafik-Smileys" sowie Vorlagenbausteine, die in der Wikipedia durch Eingabe bestimmter Codes in Beiträge auf Diskussionsseiten integriert werden können (z.B. im Rahmen von Abstimmungen, vgl. die beiden Beispiele in Abb. 3).

Die vier Typen von Einheiten haben gemeinsam, dass sie typischerweise nicht syntaktisch integriert sind und sowohl im linken oder rechten Außenfeld von Sätzen auftreten wie auch an nahezu beliebiger Position in Form von Parenthesen eingeschoben werden können. Funktional sind sie spezialisiert auf Aufgaben im Bereich der Handlungskoordination im Dialog, der emotionalen Kommentierung und der Respondierung vorangegangener Partneräußerungen. Sie ähneln damit dabei den Interjektionen beziehungsweise den Einheiten, die die

„Grammatik der deutschen Sprache“ (GDS, Zifonun et al. 1997, online: GRAMMIS, s. http://hypermedia.ids-mannheim.de/) als *Interaktive Einheiten* beschreibt und zu denen neben den Interjektionen auch die Responsive (*ja*, *nein*) gehören.

Das DeRiK-Schema führt für die oben beschriebenen Einheiten jeweils eigene Annotationskategorien ein, die in Anlehnung an die Kategorie der GDS einer gemeinsamen Oberkategorie „interaction sign“ zugeordnet sind und damit als IBK-spezifische Erweiterungen der Kategorie der interaktiven Einheiten dargestellt werden. Jeder Einheitentyp ist durch ein eigenes XML-Element beschrieben und kann durch eine Reihe von Attributen subklassifiziert werden. Die Schemakomponente für die Beschreibung der „interaction signs“ ist im Detail in Beißwenger et al. (2012: Abschnitt 3.5.1) beschrieben und begründet. Eine Übersicht über die Kategorie der „interaction signs“ gibt Abb. 3.

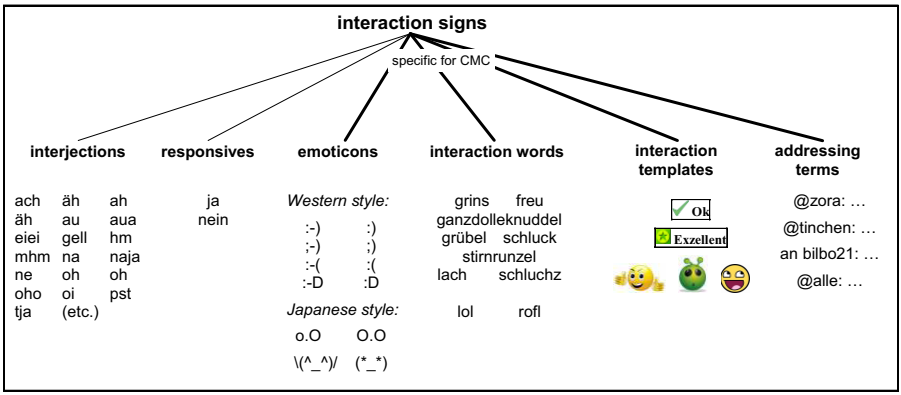


Abb. 3: „Interaction signs“.

Das für DeRiK entwickelte TEI-Schema bildet eine der Grundlagen für die Arbeit in der 2013 neu eingerichteten Special Interest Group (SIG) „Computer-Mediated Communication“ der *Text Encoding Initiative* (http://www.tei-c.org/Activities/SIG/CMC/). In der Gruppe arbeiten Vertreterinnen und Vertreter von Korpusprojekten zu verschiedenen europäischen Sprachen – darunter auch die DeRiK-Beteiligten – gemeinsam an der Erarbeitung von Vorschlägen für die Integration von Modellen für die Annotation von IBK-Genres in das Encoding Framework der TEI. Dabei wird u.a. angestrebt, eine Basisstruktur für die Repräsentation von IBK-Korpora zu entwickeln, die auch die Darstellung der Kommunikation in multimodalen Online-Umgebungen erlaubt, sowie ein Schema für die Repräsentation von Metadaten zu Sprachdaten aus Online-Umgebungen zu entwickeln. Das DeRiK-TEI-Schema soll, wo erforderlich, an die von der SIG entwickelten Vorschläge angepasst werden, um die DeRiK-Daten interoperabel zu den Korpusdaten der übrigen in der SIG beteiligten Projekte zu repräsentieren.

6. Anwendungsszenarien: Korpusgestützte Sprachanalyse und gegenwartssprachliche Lexikographie

Für den Bereich der korpusgestützten linguistischen Sprachanalyse wird das DeRiK-Korpus neue Möglichkeiten eröffnen, die Sprachverwendung in sozialen, internetbasierten Medien sowie deren Auswirkungen auf die geschriebene deutsche Gegenwartssprache empirisch zu untersuchen. Durch seine Integration in die Korpusinfrastruktur des DWDS wird es u.a. möglich,

- auf Basis eines Referenzkorpus typische Stilmerkmale des interaktionsorientierten Schreibens im Netz (z.B. Schnellschreibphänomene, Phänomene der konzeptionellen Mündlichkeit, Emoticons und Aktionswörter) für unterschiedliche Genres und Kontexte internetbasierter Kommunikation qualitativ und quantitativ zu untersuchen;
- die Variationsbreite bei der Verwendung dieser Stilmerkmale innerhalb der internetbasierten Kommunikation empirisch darzustellen und für diesen Bereich Faktoren sprachlicher Variation unter den Bedingungen technischer Vermittlung herauszuarbeiten;
- das Auftreten dieser Phänomene in IBK-Genres und in der redigierten Schriftlichkeit außerhalb des Netzes vergleichend zu untersuchen;
- den Wandel der Sprachverwendung innerhalb der internetbasierten Kommunikation zu untersuchen;
- durch vergleichende Untersuchungen zu sprachlichen Merkmalen in IBK-Daten und in den DWDS-Textkorpora charakteristische Unterschiede des *text-* und des *interaktionsorientierten Schreibens* herauszuarbeiten und für didaktische Zwecke (z.B. die Vermittlung im Deutschunterricht) fruchtbar zu machen.[3]

Darüber hinaus kann das Korpus auch als Ressource für die Sprachdidaktik genutzt werden – beispielsweise bei der Entwicklung didaktischer Materialien für den Bereich „Reflexion über Sprache“ des sprachbezogenen Deutschunterrichts oder dazu, im Modus des „Forschenden Lernens“ mit Schülerinnen und Schülern Besonderheiten der Sprachverwendung und der sprachlichen Variation in sozialen Medien anhand der DWDS-Korpora (inkl. DeRiK) selbst zu erarbeiten.

Da die Ressourcen des DWDS-Projekts an der BBAW in erster Linie als Basis für die Aktualisierung eines gegenwartssprachlichen Wörterbuches verwendet werden (vgl. Abschnitt 4), wird DeRiK darüber hinaus als Ressource für die gegenwartssprachliche Lexikographie Verwendung finden. Durch die Miteinbeziehung des IBK-Korpus wird es möglich, Phänomene des lexikalischen Wandels im Gegenwartsdeutschen, die auf die Sprachverwendung im Netz zurückzuführen sind, korpusgestützt lexikographisch zu bearbeiten. Im Folgenden präsentieren wir zwei Beispiele für sprachliche Zeichen, die, zumindest in einer Bedeutung, der Domäne der internetbasierten Kommunikation entstammen und für die eine angemessene lexikographische Beschreibung daher ohne die Konsultation von Sprachdaten

3 Zur Unterscheidung zwischen text- und interaktionsorientiertem Schreiben und ihrer didaktischen Relevanz für die Bewertung der schriftlichen Sprachverwendung in der internetbasierten Kommunikation vgl. Storrer (2012, 2013).

aus IBK-Genres nicht gelingen kann. Die Beispiele veranschaulichen die Notwendigkeit der Einbeziehung von IBK-Ressourcen für eine umfassende gegenwartssprachliche Lexikographie.

‚Troll'

Zum Stichwort *Troll* verzeichnet das auf dem „Wörterbuch der deutschen Gegenwartssprache" basierende DWDS-Wörterbuch eine Bedeutung (s. auch Abb. 1):

„gespenstisches Wesen des Volksaberglaubens, besonders in der nordischen Mythologie, Unhold"

Der Grund dafür, dass dieses Wort als Bezeichnung für eine 'Person, die in internetbasierten Kommunikationsumgebungen eine störende Rolle einnimmt, meist in provokativer Absicht', nicht in diesem Wörterbuch verzeichnet ist, ist der Zustand des WDG, das den Sprachgebrauch bis etwa Mitte der siebziger Jahre verzeichnet und zurzeit aktualisiert wird. Aber auch die Korpora des DWDS (die jüngsten Texte dort sind aus dem Jahr 2010) erzählen keine andere Geschichte. Der ‚Troll' (und das ‚Trollen') als – von den meisten als lästig empfundener – Teil der Netzkultur kommt in diesen Korpora (noch nicht) vor. Umso wichtiger ist es, die Möglichkeit zur Recherche in einem aktuellen Korpus der internetbasierten Kommunikation zu haben, spätestens dann, wenn der Artikel „Troll" im DWDS-Wörterbuch überarbeitet werden soll.

‚lol'

Im Jahr 2011 wurde das Akronym 'lol' für 'laughing out loud', ein in der internetbasierten Kommunikation häufig verwendetes Aktionswort, durch Aufnahme in das „Oxford English Dictionary" geadelt. Das OED erkennt damit die lexikalische Produktivität der netzbasierten Kommunikation an.

Im DWDS-Wörterbuch findet sich hierzu, aus den oben bereits genannten Gründen, nichts. Die Korpora des DWDS sind da etwas ergiebiger. Allein 46 Treffer findet man in der ZEIT. In den meisten dieser Belege sehen es die Autoren allerdings als notwendig an, das Akronym aufzulösen – vertrauen also nicht auf sofortiges Verstehen des Kürzels bei ihrer Leserschaft.

In der Jugendsprache geht diese Abkürzung neuerdings eigene Wege. Immer öfter vernimmt man die Adjektivbildung 'lollig', die aus 'lol' gebildet wurde und laut Auskunft einiger Sprecher eindeutig positiv-anerkennend gemeint ist (etwa: „echt witzig").

Wortneuschöpfungen im Bereich gruppenspezifischer Sprachen, wie etwa der Jugendsprache, sind oft Gelegenheitsbildungen. Ihre Verwendung wird deshalb von Lexikographen erst einmal längere Zeit beobachtet, ehe über eine (Nicht-)Aufnahme ins Wörterbuch entschieden wird. Stellt man dann fest, dass eine Neuschöpfung sich in der Netzkultur oder auch darüber hinaus durchsetzen konnte, wird für die Aufnahme dieses Wortes entschieden. Meist sind dann auch die Verwendungsregeln und -formen so stabil, dass eine verlässliche lexikographische Beschreibung möglich ist. Die Bedeutung(en) und die unterschiedlichen kommunikativen Funktionen kann man aber nur nachvollziehen und angemessen beschreiben, wenn man ein Korpus internetbasierter Kommunikation mit einer gewissen historischen Tiefe zur Verfügung hat. Deshalb sind die Nachhaltigkeit des Projektes, die eine dauerhafte Weiterentwicklung des Korpus ermöglicht, und die Interoperabilität der Daten, die ihre

Verknüpfung mit anderen Korpora gewährleistet, von zentraler Bedeutung für die lexikographische Anwendung.

7. Ausblick

IBK-Korpora stellen einen Korpustyp neuer Art dar, zu dem im Bereich der 'Digital Humanities' noch in verschiedener Hinsicht Forschungsbedarf besteht: Diverse grundlegende Fragen hinsichtlich des Aufbaus, der Annotation und der Integration in Korpusinfrastrukturen sind für diesen Korpustyp noch nicht abschließend geklärt (vgl. Storrer 2013a: Abschnitt 4).

DeRiK versteht sich deshalb als ein Projekt, dessen primäres Ziel zwar der Aufbau und die Bereitstellung eines Korpus ist, das aber zugleich auch den Rahmen dafür bildet, methodische Fragen in Bezug auf die Integration und Beschreibung von IBK-Daten in linguistische Korpora zu sondieren und Lösungsvorschläge dafür zu erarbeiten. Zentrale Fragen sind u.a. die angemessene und an Standards (z.B. TEI) orientierte Modellierung der Primärdaten, die Repräsentation von Metadaten zu IBK-Genres sowie Fragen der linguistischen Annotation und der Anpassung von Werkzeugen für die automatische Sprachverarbeitung. Entscheidend für den Erfolg der Unternehmung ist die enge Zusammenarbeit mit Initiativen und Projekten, die sich im nationalen und internationalen Rahmen mit diesen Fragen befassen – als Beispiele seien hier das schon erwähnte DFG-Netzwerk „Empirische Erforschung internetbasierter Kommunikation", die Special Interest Group „Computer-Mediated Communication" im Rahmen der TEI sowie die in Vorbereitung befindliche Shared Task zur linguistischen Annotation deutschsprachiger IBK-Daten genannt. Von der engen Zusammenarbeit mit Infrastrukturprojekten im Bereich der Digital Humanities wie CLARIN (http://www.clarin.eu) und, national, CLARIN-D (http://de.clarin.eu) erwarten wir Fortschritte in Richtung auf eine umfassende Interoperabilität der entstehenden Ressourcen. Im Rahmen der Special Interest Group sollte es gelingen, ein Referenzschema zu entwickeln, das als Grundlage für die Annotation im Aufbau befindlicher Korpora zu verschiedenen europäischen Sprachen und unterschiedlichen IBK-Genres dienen kann und das damit die Interoperabilität und Verknüpfbarkeit dieser Ressourcen ermöglicht – die damit für die korpusgestützte Bearbeitung auch von sprachübergreifenden Forschungsfragen von Nutzen sein können. CLARIN ist auch ein geeigneter Rahmen, um das entstehende Korpus a) für die interessierten Fachcommunitys sichtbar zu machen – hierfür bietet das „Virtual Language Observatory"[4] eine geeignete Plattform – und b) die Daten zu distribuieren. Dadurch, dass nur Texte akquiriert werden, deren Lizenzstatus eine Redistribuierung der Daten (und Annotationen) erlauben, können interessierte Nutzer nicht nur über die DWDS-Webseite in den Daten recherchieren, sondern sich das gesamte Korpus oder Teile davon herunterladen und in ihrer individuellen Arbeitsumgebung nutzen.

Aus lexikographischer Perspektive ist die kontinuierliche Weiterentwicklung von IBK-Korpora von besonderer Bedeutung, damit Prozesse der Lexikalisierung und des Bedeutungswandels auch in diesem Bereich erfasst und beschrieben werden können. Umso wichtiger ist es, dass für DeRiK grundlegende Fragen der Modellierung und Annotation von

4 S. http://www.clarin.eu/vlo.

IBK-Daten bereits in einer frühen Projektphase geklärt werden. Aktuelle Arbeitsschwerpunkte liegen daher zum einen auf der Entwicklung von Standardisierungsvorschlägen im Kontext der TEI und zum anderen auf der Anpassung von Sprachverarbeitungswerkzeugen. Daneben ist für 2014 die Erhebung einer ersten Tranche von Sprachdaten aus verschiedenen, rechtlich unproblematischen Online-Umgebungen geplant.

8. Literatur

Bartz, T.; Beißwenger, M. & Storrer, A. (2013, im Erscheinen). Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: *Journal for Language Technology and Computational Linguistics (Themenheft „Das STTS-Tagset für Wortartentagging – Stand und Perspektiven").*

Beißwenger, M. (Hrsg., 2001). *Chat-Kommunikation. Sprache, Interaktion, Sozialität & Identität in synchroner computervermittelter Kommunikation. Perspektiven auf ein interdisziplinäres Forschungsfeld.* Stuttgart: ibidem.

Beißwenger, M. (2013). Das Dortmunder Chat-Korpus. In: *Zeitschrift für germanistische Linguistik* 41/1, pp. 161-164.

Beißwenger, M. & Storrer, A. (2008). Corpora of Computer-Mediated Communication. In Lüdeling, A. & Kytö, M. (eds), *Corpus Linguistics*. *An International Handbook,* vol. 1. Berlin, de Gruyter, pp. 292-308.

Beißwenger, M., Ermakova, M., Geyken, A., Lemnitzer, L. & Storrer, A. (2012). A TEI schema for the Representation of the Computer-mediated Communication. In: *Journal of the Text Encoding Initiative* 3. http://jtei.revues.org/476 (DOI: 10.4000/jtei.476).

Busemann, K. (2013): Wer nutzt was im Social Web? Ergebnisse der ARD/ZDF-Onlinestudie 2013. Media Perspektiven 7-8/2013, 373–385. http://www.ard-zdf-onlinestudie.de/fileadmin/Onlinestudie/PDF/Busemann.pdf

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Crystal, D. (2011). *Internet Linguistics. A Student Guide.* New York: Routledge.

Forsyth, E. N. & Martell, C. H. (2007). Lexical and Discourse Analysis of Online Chat Dialog. In *Proceedings of the First IEEE International Conference on Semantic Computing (ICSC 2007)*, pp. 19-26.

Geyken, A. (2007). The DWDS corpus: A reference corpus for the German language of the 20th century. In: Fellbaum, Ch. (ed.), *Collocations and Idioms*. London: continuum, pp. 23-40.

Geyken A. & Hanneforth T. (2006). TAGH: A Complete Morphology for German based on Weighted Finite State Automata. In: Yli-Jyrä A, Karttunen L, Karhumäki J, (eds.), *Finite State Methods and Natural Language Processing.* Berlin/Heidelberg: Springer, 55-66.

Grimm, J. & Grimm W. (1852-1971). *Deutsches Wörterbuch. Erstbearbeitung*, 33 Bände, Leipzig:Hirzel Verlag

Herring, S. C. (ed., 1996). *Computer-Mediated Communication. Linguistic, Social and Cross-Cultural Perspectives*. Amsterdam/Philadelphia: John Benjamins (Pragmatics and Beyond New Series 39).

Herring, S. C. (ed., 2010/2011). Computer-Mediated Conversation. *Special Issue of Language@Internet,* vol. 7/8. http://www.languageatinternet.org/articles/2010, http://www.languageatinternet.org /articles/2011

Jurish, B. (2003). *A Hybrid Approach to Part-of-Speech Tagging. Final report, project 'Kollokationen im Wörterbuch'.* Berlin: BBAW; 2003. http://www.dwds.de/dokumentation/tagger/.

Kestemont, M., Peersman, C., De Decker, B., De Pauw, G., Luyckx, K., Morante, R. Vaassen, F., van de Loo, J., Daelemans, W. (2012). The Netlog Corpus. A Resource for the Study of Flemish Dutch Internet Language. In: *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Paris, pp. 1569-1572.

King, B. W. (2009). Building and Analysing Corpora of Computer-Mediated Communication. In: Baker, P. (ed.), *Contemporary corpus linguistics*. London: Continuum, pp. 301-320.

Klappenbach, R. & Steinitz, W. (eds., 1964-1977). *Wörterbuch der deutschen Gegenwartssprache (WDG).* 6 Bände. Berlin: Akademie-Verlag.

Klein, W., Geyken, A. (2010). Das Digitale Wörterbuch der Deutschen Sprache (DWDS). In Heid, U., Schierholz, S., Schweickard, W., Wiegand, H. E., Gouws, R. H. & Wolski, W. (eds), *Lexicographica*. pp. 79-96.

Koch, P. & Oesterreicher, W. (1994). Schriftlichkeit und Sprache. In: Günther, Hartmut/ Ludwig, Otto (eds.), *Schrift und Schriftlichkeit. Ein interdisziplinäres Handbuch internationaler Forschung.* Bd. 1. Berlin u.a., 587-604.

Oostdijk, N., M. Reynaert, V. Hoste & I. Schuurman (2013). The construction of a 500-million-word reference corpus of contemporary written Dutch. In J. Odijk & P. Spyns (eds.), *Essential Speech and Language Technology for Dutch.* Springer.

Openthesaurus. http://www.opentesaurus.org.

Pfeifer, W. (1993). *Etymologisches Wörterbuch des Deutschen*. Berlin:Akademie-Verlag, 2. Aufl.

Reffay, C., Betbeder, M.-L. & Chanier, T. (2012). Multimodal Learning and Teaching Corpora Exchange: Lessons learned in 5 years by the Mulce project. In: *International Journal of Technology Enhanced Learning (IJTEL)* 4, vol. 1/2. DOI: 10.1504/IJTEL.2012.048310

Reynaert, N., Oostdijk, O., De Clercq, O., van den Heuvel, H. & de Jong, F. (2010). Balancing SoNaR: IPR versus Processing Issues in a 500-Million-Word Written Dutch Reference Corpus. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*. Paris, pp. 2693-2698.

Runkehl, K., Siever, T. & Schlobinski, P. (1998). *Sprache und Kommunikation im Internet. Überblick und Analysen.* Opladen: Westdeutscher Verlag.

Storrer, A. (2012). Neue Text- und Schreibformen im Internet: Das Beispiel Wikipedia. In: Juliane Köster & Helmuth Feilke (Hrsg.): *Textkompetenzen für die Sekundarstufe II.* Freiburg: Fillibach, 277-304.

Storrer, A. (2013). Sprachstil und Sprachvariation in sozialen Netzwerken. In Frank-Job, B., Mehler, A., Sutter, T. (eds), *Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW.* Wiesbaden: VS Verlag für Sozialwissenschaften, pp. 329-364.

Storrer, A. (2013a, im Druck). Sprachverfall durch internetbasierte Kommunikation? Linguistische Erklärungsansätze – empirische Befunde. In: *Sprachverfall? Dynamik – Wandel – Variation. Jahrbuch des Instituts für Deutsche Sprache*.

[TEI-P5] TEI Consortium (ed., 2007). *TEI P5: Guidelines for Electronic Text Encoding and Interchange.* http://www.tei-c.org/Guidelines/P5/ (accessed 30 April 2013).

van Eimeren, B. & Frees, B. (2013): Rasanter Anstieg des Internetkonsums – Onliner fast drei Stunden täglich im Netz. Ergebnisse der ARD/ZDF-Onlinestudie 2013. In: Media Perspektiven 7-8/2013, 358–372. http://www.ard-zdf-onlinestudie.de/fileadmin/Onlinestudie/PDF/Eimeren_Frees.pdf

Zifonun, G., Hoffmann, L. & Strecker, B. (1997): *Grammatik der deutschen Sprache.* 3 Bände. Berlin/ New York: de Gruyter.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, Torsten Zesch

Scalable Construction of High-Quality Web Corpora

Abstract

In this article, we give an overview about the necessary steps to construct high-quality corpora from web texts. We first focus on web crawling and the pros and cons of the existing crawling strategies. Then, we describe how the crawled data can be linguistically pre-processed in a parallelized way that allows the processing of web-scale input data. As we are working with web data, controlling the quality of the resulting corpus is an important issue, which we address by showing how corpus statistics and a linguistic evaluation can be used to assess the quality of corpora. Finally, we show how the availability of extremely large, high-quality corpora opens up new directions for research in various fields of linguistics, computational linguistics, and natural language processing.

1 Introduction

The availability of large corpora is a prerequisite for both empirical linguistic research in many fields such as phraseology, graphemics, morphology, syntax, etc. and for many applications in language technology such as spell checking, language models, statistical machine translation, collocation extraction, or measuring semantic textual similarity. The web is certainly an abundant source of text documents and many projects have are already constructed corpora from web texts. Most of them either provide a final corpus in the form of a collection of sentences, paragraphs, or documents, e. g. COW [85], Gigaword [76], LCC [52], UMBC WebBase [57], and WaCky [11]. Some projects only provide aggregated information like word-level (Google Web 1T 5-Grams [24]) or syntactic n-gram counts [50]. As the documents found on the web can be quite noisy from a linguistic point of view, all the resulting corpora have been criticized (among other reasons) for being biased or for not properly reflecting the research question at hand. In this article, we argue that we can address these challenges by carefully controlling all steps of the corpus creation process, including (i) crawling, (ii) preprocessing, and (iii) an assessment of corpus quality.

It will become clear that there are diverse notions of "corpus quality" (and, consequently, "noise"), which depend on the intended use of the corpus. In empirically oriented theoretical linguistics, carefully selected sampling procedures and non-destructive cleaning is important, while for many tasks in computational linguistics and language technology, aggressive cleaning is fundamental to achieve good results. Technically, however, these differences merely result in different software configurations, but not

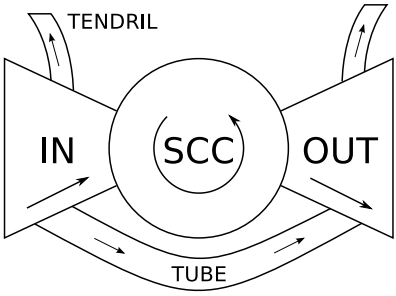


Figure 1: Schematic structure of the web according to [26], as depicted in [86, 9]

in substantially different software architectures. Also, all projects for web corpus construction considered in this paper put an emphasis on improving processing speed and collecting substantial amounts of data.

One major focus of this article is on web crawling, as it might be relatively easy to crawl a large English corpus, but in order to collect a big corpus for "smaller" languages more sophisticated crawling strategies are required. We further describe how the crawled data can be linguistically pre-processed, including ways of parallelization that allow the processing of web-scale corpora. When working with web data, controlling the quality of the resulting corpus is an important issue, which we address by showing how corpus statistics and a linguistic evaluation can be used to assess the quality of corpora. Finally, we show how the availability of extremely large, high-quality corpora opens up new directions for research in various fields of linguistics, computational linguistics, and natural language processing.

2 Web Crawling

Web crawling is the process of fetching web documents by recursively following hyperlinks. Web documents link to unique addresses (URLs) of other documents, thus forming a directed and cyclic graph with the documents as nodes and the links as edges. Each node has an in-degree (the number of nodes linking to it) and an out-degree (number of nodes linked to by it). It is usually reported that the in-degrees in the web graph are distributed according to a power law [7, 72].

Macroscopically, the web graph has been described as having primarily a tripartite structure [26, 87], as shown in Figure 1: Pages with an in-degree of 0 form the IN component, pages with an out-degree of 0 form the OUT component, pages with both in- and out-degree larger than 0 form the strongly connected component (SCC). Within SCC, there is a link path from each page to each other page. Since web crawling relies on pages being either known beforehand or having at least one in-link, it reaches only pages in SCC and OUT and pages in IN which are known at the start.

Besides pages with an in-degree of 0, certain other pages also cannot be reached, such as pages on servers which only serve certain IP addresses, pages requiring a login, form input, etc. There are approaches to access this so-called deep web (e. g. [70]), but we are unaware that any such techniques were used in web corpus construction so far. Besides the problems of accessing the deep web, crawling is complicated by the fact that the web is not static. Pages appear and disappear at very high rates. The decay of web sites is known as *link rot* [10]. Even worse, many web sites are dynamic and generate or remix content for each request based on input parameters, the client's IP, and geolocation.

A major issue is the extremely large size of the web. Even large search engines index only fractions of the web (web pages they classify as relevant), but already in 2008 it was announced that Google had indexed a trillion pages.[1] Thus, building corpora from the web starts by sampling from a population and will usually result in a corpus that is orders of magnitude smaller than the web itself.

2.1 Sampling

The population from which the sample is taken is the set of web documents. There are several options for sampling: We can take random (preferably uniform) samples, cluster samples, stratified samples, or non-random samples like systematic/theoretical samples (an overview of sampling techniques can be found in Chapter 3 of [23]). In most variants, crawling is some form of random sampling, whereas classical balanced corpus construction is a variant of stratified and/or theoretical sampling, cf. Section 5.2. It is rather complicated to do the same stratified sampling with web data, because (i) the relative sizes of the strata in the population are not known, and (ii) it would be required to start with a crawled data set from which the corpus strata are sampled, as web documents are not archived and pre-classified like many traditional sources of text. Web documents have to be *discovered* through the crawling process, and cannot be taken from the shelves. If a crawling procedure favors the inclusion of, for example, highly popular web pages (e. g. pages which are linked to by many other pages), it might be argued that a theoretical or systematic sample is drawn. Still, the basics of the initial discovery procedure would remain the same.

In this section, we have discussed the nature of the web to the extent that it is relevant for crawling. A more detailed overview is given in [86]. Additional information can also be found in [72, 75], although these sources are more concerned with implementation details than with empirically sound sampling and its relation to crawling strategies, which we cover in the next section.

2.2 Crawling Strategies

Crawling strategies can be divided into graph traversal techniques (each node is only visited once) and random walks (nodes might be revisited), cf. [49]. Available off-the-

[1] `http://googleblog.blogspot.de/2008/07/we-knew-web-was-big.html`

shelf crawler software like Heritrix[2] [74] or Nutch[3] commonly applies some kind of graph traversal, mostly a (modified) Breadth-First Search (BFS). In a BFS, the crawler starts with a set of known URLs (the *seeds*), downloads them, extracts the contained URLs and adds them to the queue, then crawls again etc. This strategy leads to the exhaustive collection of a local connected sub-component of the web graph. BFS can thus introduce a bias toward the local neighborhood of the seed URLs, possibly not discovering relevant/interesting material in other parts of the graph. For example, a URL/host analysis for WaCky [11] and COW corpora showed that these corpora contain many documents from only a few hosts, most likely due to BFS crawling [85]. Although there was only one URL from each host in the crawling seed set for deWaC, these hosts account for 84% of the documents in the resulting corpus. In a crawl of the top level domain `.se`, we found that 95% of the corpus documents came from the hosts already represented in the seed set, and one single host alone (`blogg.se`) accounted for over 70% of the documents [85]. Furthermore, in a series of theoretical and experimental papers [3, 65, 64, 71], it was found that BFS is biased toward pages with high in-degrees, and that this bias is impossible to correct for. Whether such biased (and hence not uniform) random sampling is acceptable is an important design decision. It is especially problematic from the perspective of empirical linguistics, where the question of sampling should receive close attention (cf. Sections 2.1 and 5.2). However, an important advantage of BFS is that it allows crawling a huge amount of web pages in a short time.

Alternative strategies, which deliver considerably less yield in the same time compared to BFS, are random walks (RW). A random walk is characterized by recursively selecting a single out-link from the current page and following this link. Pages might be revisited in the process (something which implementations of BFS crawlers go to great lengths to avoid). Sampling based on random walks also delivers samples which are biased, but in contrast to BFS, these biases can be corrected for, for example by rejection sampling [58, 81]. Such approaches have not yet been used in web corpus construction, although, for example, [31] mention unbiased sampling as possible future work. The COW project (see Section 3.2) is currently working on their own crawler (CLARA) which implements diverse bias-free crawling algorithms based on random walks. A long-term goal is linguistic characterization for single languages based on uniform random samples, i. e. the characterization of the distribution of linguistic features in the web graph.

Refinements There is a number of refinements that can be applied to modify the basic crawling strategy. We briefly discuss three such refinements: scoped crawling, focused crawling, and optimized crawling.

Scoped crawling imposes constraints on the kind of crawled documents. The idea is to avoid downloading content which would not be included in the final corpus anyway. A scoped crawl is restricted by accepting only documents from certain URLs, IP ranges, etc. Restricting the scope of a crawl to a national top level domain is

[2] `http://webarchive.jira.com/wiki/display/Heritrix/`
[3] `http://nutch.apache.org/`

standard procedure in monolingual web corpus construction. Since documents are never exclusively written in one language under any top level domain, additional language detection/filtering is required. Also, with the recent proliferation of top level domains, it is no longer guaranteed that the relevant documents in one language can be found under its associated top level domain. A non-scoped but focused crawler (as discussed below) might be the better solution in the long run.

Focused crawling imposes even stricter constraints and tries to efficiently discover specific types of information (languages, genres, topics, etc.) which cannot simply be inferred from address ranges, domains, or hosts. A focused crawler guesses for each harvested link the kind of document it points to. Normally, the link URL and the surrounding text (in all documents where the link appears) are analyzed in order to predict the contents of the linked document. Various heuristics and machine learning methods are used for the prediction [4, 28, 29, 30, 54, 73, 82, 92]. For language detection based purely on URLs, see [15]. For topic and genre detection by URL analysis, see [2, 14].

What we call *optimized crawling* is similar to focused crawling, but it is merely intended to make the crawl more effective, so that more usable text documents can be downloaded in a shorter time, wasting less bandwidth. Contrary to focused crawling, there is not necessarily a specific restriction on the contents of the crawled documents. The crawl is biased towards documents which, according to some metric, have a high relevance. The bias can be implemented, e. g. by giving promising URLs a better chance of being queued compared to less promising URLs. This is an important aspect in crawling for search engine applications, where usually a bias toward pages with a high PageRank is desirable. In [1], Online Page Importance Computation (OPIC) is suggested, which is basically a method of guessing the relevance of the pages while the crawl is going on.

An interesting experiment concerning linguistically motivated crawl optimization was reported in [93].[4] The authors integrate their post-processing into their own crawler and collect statistics about the final yield from each encountered host after the removal of material which is not good corpus material. Hosts receive penalty for low yield, up to a point where they get effectively blacklisted. Although the optimization effect is reported to be moderate, the method is in principle suitable for collecting more corpus data in a shorter time.

2.3 Strategies Used by Existing Web Corpus Projects

We now give an overview of the crawling and post-processing strategies used by previous web corpus projects.

The WaCky project (*Web as Corpus kool ynitiative*) has released web corpora in a number of European languages [11]. A similar project is COW (*Corpora from the Web*) [85]. Both used the Heritrix crawler, which means that a variant of BFS was applied, restricted to national top level domains. However, the Heritrix strategy is not

[4] The crawler is available as SpiderLing: `http://nlp.fi.muni.cz/trac/spiderling/`

pure BFS, because a system of queues in a multi-threaded implementation is used to optimize the download process. This system prefers downloading all documents from one host exhaustively in one batch. Taking care of one host en bloc makes a number of tasks more efficient, e. g. keeping track of robots exclusion information for the host, obeying delay times between requests to the host, and caching DNS information. This leads to an overall BFS with a preference for host-internal breadth.

The UMBC WebBase corpus [57] consists of 3.3 billion tokens of "good quality English" extracted from the February 2007 crawl of the Stanford WebBase project. The aim was to compile a "large and balanced text corpus" for word co-occurrence statistics and distributional semantic models. Suitable text was selected based on various heuristics and de-duplicated at paragraph level.

The *Leipzig Corpora Collection* (LCC) uses mainly two approaches in parallel for collecting textual data in various languages [52]. On the one hand, the distributed web crawler FindLinks[5] is used, which implements a BFS strategy. There is no restriction to specific top level domains, and it therefore discovers a large number of domains while following the existing link structure of the web. The crawl is split into so-called rounds, where the URLs found in the documents of one round are crawled in the next round. To prevent the size of the rounds from exponential increase, the number of pages per domain is limited to between five and twenty, depending on the top-level domain. On the other hand, news pages are crawled exhaustively using httrack.[6] The directory of AbyZNewsLinks provides a very comprehensive list of URLs for news in about 120 languages which are used as seeds.[7]

The largest publicly available resource derived from a web corpus was compiled by researchers at Google Inc. as a basis for the *Google Web 1T 5-gram* database, or Web1T5 for short [24]. According to the authors, this corpus consists of about 1 trillion words of English text. However, it is not distributed in full-text form, but only as a database of frequency counts for n-grams of up to five words. Throughout, n-grams with fewer than 40 occurrences were omitted from the database. Due to these restrictions, this resource is of limited use for linguistic purposes, but it can be – and has been – applied to certain types of analyses such as collocation identification (cf. also Section 4.2).

So far, we have focused on projects which have created publicly available corpora. Another (not publicly available) project was described in [77], which resulted in a 70 billion token corpus for English based on the ClueWeb09 dataset.[8] According to an informal and otherwise unpublished description on the ClueWeb09 web page, it was crawled with Nutch and "best-first search, using the OPIC metric", which is biased toward documents which are relevant to search engine applications.[9]

Finally, there are completely different kinds of web corpus projects, which use stratified non-random sampling as mentioned in Section 2.1, such as the German DeRiK corpus

[5] `http://wortschatz.uni-leipzig.de/findlinks/`
[6] `http://www.httrack.com/`
[7] `http://www.abyznewslinks.com/`
[8] `http://lemurproject.org/clueweb09/`
[9] `http://boston.lti.cs.cmu.edu/Data/web08-bst/planning.html` on 2012-04-22.

project of computer-mediated communication [16]. Such corpora are designed with a specific purpose in mind, and they are orders of magnitude smaller than the large crawled web corpora we discuss in this paper.

3 Processing Crawled Data

In this section, we describe approaches to web data processing, i. e. the steps taken to transform crawled data into a corpus which can be used for linguistic research or applications in language technology. The purpose of corpus construction is to provide a sufficient amount of data for a given purpose. For some research questions, a small or a medium sized corpus contains enough data, but in many cases the objects of interest are very rare, and a large corpus is needed. Such questions include research on:

- phraseological units (especially those falling out of use)
- neologisms (might also require certain registers not available in traditional corpora)
- statistical data for very infrequent words (as used, for instance, in distributional semantics)
- word co-occurrences for infrequent words
- rare (morpho-)syntactic phenomena of high theoretical importance
- (structured) n-gram counts for language modeling

We will take a detailed look at two projects that perform processing of crawled data for uses in (computational) linguistics: WebCorpus and COW. Both projects share the following high-level architecture: Data from crawls is filtered to retain only textual material in a desired target language, and unified regarding the encoding. Depending on the unit of interest, which could be either a document or a sentence, a de-duplication step removes redundant material. The resulting large raw corpus is subsequently annotated with automatic linguistic processing (such as POS-tagging). Finally, statistics and indices are generated. The projects were designed towards slightly different goals, differ in their implementation, address scaling differently, and target overlapping but somewhat different usage scenarios for the resulting corpora. Regarding technological aspects, there are no fundamental differences between COW and WebCorpus, as they both implement the same high-level design pattern: a pipeline that performs various pre-processing and filtering steps, and results in a corpus.

3.1 WebCorpus Project

This section describes WebCorpus[10] that implements corpus data processing based on the Hadoop MapReduce framework.[11] After motivating this choice of parallelization technology, we flesh out the generic pipeline steps with concise descriptions of techniques and software used in their implementation. The overall framework is entirely written in Java, and is available under the open-source Apache Software License 2.0.[12]

[10] http://sourceforge.net/projects/webcorpus/
[11] http://hadoop.apache.org/
[12] http://code.google.com/p/dkpro-bigdata

As the size of web corpora usually ranges in the order of gigabytes to terabytes, processing requires considerable amounts of time. Sole optimization of code execution speed, e. g. through the use of advanced compilers, is insufficient as runtime will stay proportional to the input size. A better suited approach to reduce runtime is the (massive) use of parallelization through distributing independently processable parts of the data, which in our case can be done by, e. g. dividing data archives into single records or texts into sentences.

One method of distributing such "embarrassingly parallel" [48] problems, which has recently gained popularity, is MapReduce [33]. Its core idea is to perform processing in two consecutive steps: the map phase and the reduce phase. In the map phase, input is read and split in a defined manner appropriate for a specific *input format*, e. g. one that reads text documents and creates one split per line. These are then processed in parallel by multiple *mappers*, which can themselves produce an arbitrary number of intermediate results, e. g. extracted tokens. In a second step, *reducers* will combine these intermediate results, for example by counting all occurrences of identical tokens and producing a sorted list of tokens with frequencies.

A notable advantage of this method is that it scales arbitrarily with increasing amounts of input data: Additional mappers and reducers can run on different cores or on different machines. Since all intermediate data is processed independently in each of the two phases, runtime can be reduced almost linearly by adding more computing hardware. Reducing processing time is therefore merely a matter of available resources.

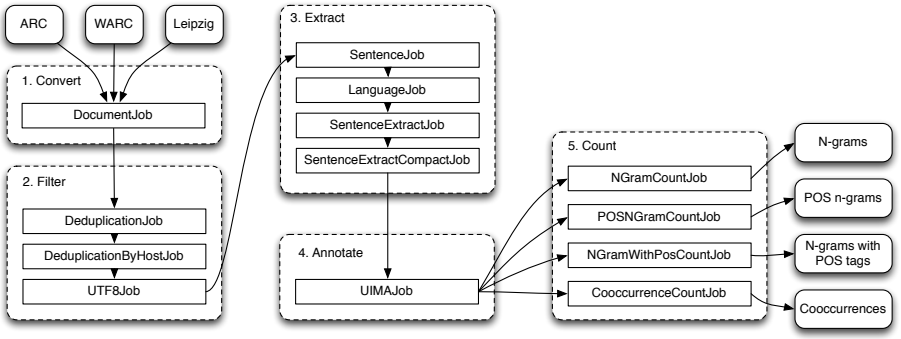


Figure 2: WebCorpus processes data in a pipeline fashion to allow for error recovery and reuse of intermediate data. Lines represent the data flow between jobs.

Figure 2 illustrates data flow in the WebCorpus project. The data is processed in a pipeline fashion, where results from each job are written to disk before being passed onto the next job, which in turn has to read the data back from disk. Especially for large-scale computations that involve long running jobs and many machines, the scheme of saving intermediary steps makes it possible to use partial results even if an outage

occurs. Also, this makes it possible to reuse output data from any job in other contexts, e. g. to use de-duplicated sentences and their counts to determine frequent boilerplate text. In a production environment, intermediate outputs can be deleted automatically once they are not needed by any further steps in the pipeline.

For the WebCorpus project, the Hadoop framework was chosen, which has become the de-facto standard implementation of MapReduce. Native Hadoop code is written in Java, as is most of the WebCorpus code, although other programming environments exist for Hadoop. One such example is the data-flow oriented scripting language *Pig Latin*.[13] Its capability of performing joins across several MapReduce tables allows the computation of significance measures on word co-occurrences.

Convert Since the input can be given in various formats, such as HTML, ARC[14], or WARC[15] the pipeline starts with a conversion step that extracts documents from the input and stores them in a unified format: One document per record is stored along with its metadata from the crawl, i. e. URL, time of download, etc. In the first pipeline step, those different input formats are read and split into parts that can be processed in parallel. For ARC/WARC, this is done by reading input archives using the open-source library JWAT[16] and generating a split for each archive record. Although web archives can contain text documents in virtually any format such as PDF, word processor, presentations, or even images, only HTML documents are used for processing. In order to extract content from HTML and remove boilerplate text, we use the *html2text* package of the Findlinks project, which itself makes use of the Jericho HTML parser.[17] Encoding of all HTML documents was normalized by utilizing the *encodingdetector* package of the Leipzig ASV toolbox [20].[18] In this step, removal of unwanted (i. e. non-textual) material, DOM-tree based cleaning [12] as well as normalization of white space are also performed. Also, parameters that are irrelevant to the document content can be removed, e. g. session IDs and tracking parameters added by Google WebAnalytics. Once plain text documents have been extracted from the input, we perform additional normalization of white spaces and mark paragraphs in the text with XML-like tags. Additionally, metadata such as URL or time of download are kept and also enclosed in XML tags in the resulting text documents. Normalization also extends to metadata: different URL strings may denote the same location, due to possible permutations of query parameters. For example, `?lang=de&page=2` and `?page=2&lang=de` are equivalent – a simple and effective way to resolve such cases is to sort query parameters alphabetically.

[13] `http://pig.apache.org/`
[14] `https://archive.org/web/researcher/ArcFileFormat.php`
[15] `http://archive-access.sourceforge.net/warc/`
[16] `https://sbforge.org/display/JWAT/`
[17] `http://jericho.htmlparser.net/docs/`
[18] `http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/`

Filter In order to reduce the amount of data that has to be processed in later steps, undesirable documents are filtered as early as possible in the pipeline. Such documents may be duplicate documents, documents with a broken encoding, documents that are too long or too short, or documents not written in the target language. In WebCorpus, we tackle de-duplication using an approach that incorporates a Bloom filter [22] for probabilistic detection of duplicates residing on the same host (as defined by the URL), which produced satisfactory results while using very few resources. See Section 3.2 for a more sophisticated method for finding document duplicates.

Encoding detection can be performed with different types of heuristic approaches. Depending on the input format, encoding has either already been normalized (usually to UTF-8) or encoding information is contained in the metadata. We thus only have to catch cases of wrongly converted or non-convertible documents. For this, we do not need to incorporate a full-fledged system for encoding detection such as ICU.[19] It suffices to look for malformed UTF-8 bytes and malformed UTF-8 byte sequences.

While all filtering steps described so far operate on documents, language detection is performed on sentences. Language filtering is performed with the *LanI* language identification system from the ASV Toolbox. Based on word frequencies and the size of the underlying corpus, LanI returns the most probable languages of a given sentence. If LanI does not return the expected language as the most likely one, then the sentence is either filtered from the data, or it is considered for further processing if one of following conditions are met: the sentence is not at the beginning or at the end of a paragraph, and its length is shorter than a configurable value, e. g. 200 characters. In this way, content from mixed-language documents can be used for all target languages. After language detection has been performed, subsequent steps only operate on a predefined target language and ignore material of other languages.

Extract While all subsequent WebCorpus steps can also be run on documents, we describe a sentence-centric pipeline from this point on. It is advantageous for many applications – in particular those based on a statistical analysis of the corpus data – to remove duplicated sentences, i. e. to keep only one copy of every distinct sentence. This removes artifacts that are valid sentences but are vastly overrepresented because of their appearance in boilerplate text, e. g.: "You are not allowed to edit this post."[20] These artifacts lead to unwanted bias in the frequency distribution of contained terms, which hurt linguistic analysis.

In the extraction step, we apply a language-independent, extensible rule-based system for splitting all the documents that passed the filtering step into sentences. For specific languages, this could easily be replaced by a sentence splitter tuned to the target language. Filtering based on sentence length or other metrics can also be performed, cf. [56]. To extract the sentences in a basic format, two Hadoop jobs are employed.

[19] International Components for Unicode: `http://site.icu-project.org/`

[20] Boilerplate is text which is usually not written by humans, but inserted automatically from a template, etc. Typical boilerplate material includes navigational elements, copyright notices, "read more" snippets, date and user name strings.

The first job outputs all sentences along with their source URLs. The second job de-duplicates the sentences in its reducer, computes the sentence frequency and outputs the frequency as well as up to ten URLs per sentence.

Annotate In this step, we use the Unstructured Information Management Architecture (UIMA) [44] in order to allow for arbitrary annotations and to flexibly support future annotation requirements. Since UIMA has recently gained popularity, there is a variety of annotators available that can be reused easily, e. g. from OpenNLP[21] or the DKPro framework.[22] Currently, the OpenNLP tokenizer and POS taggers for German and English are included. The sentences are processed in parallel by Hadoop mappers that pass the sentences through the UIMA pipeline. Each component in the pipeline writes annotations to the Common Analysis Structure (CAS), an efficient in-memory database used by UIMA to store annotations. In the final step of the pipeline, all annotation stored in the CAS are written to disk as XML files. While the produced XML documents tend to be verbose, it was found to be sufficient to compress them with *gzip* to keep required storage space and disk I/O overhead in reasonable ranges.

Count Finally, based on the annotated corpus, statistical analyses are performed. First, patterns are generated from the corpus, e. g. all token unigrams, all verb-object dependencies, or any other pattern that can be defined on top of the annotations. Afterward, the generated patterns are collected, counted, and written in a suitable exchange format for further processing or indexing. For a scalable implementation of co-occurrence significance measures and second order similarities, the reader is referred to the JoBimText project [21].[23]

Figure 3 illustrates the process of annotating and pattern counting using the Hadoop framework: Sentences are tokenized and annotated with POS tags, which are stored in the UIMA CAS. Then, a map step extracts patterns of interest, in this case POS-tagged bigrams. In the reduce step, these are counted across the whole corpus.

With WebCorpus, it was possible to process 600 GB of German web-crawled text (already stripped from non-textual material, as provided by the Findlinks[24] project [59]) on a (comparatively small) Hadoop cluster with 8 nodes providing 64 cores in total. The preprocessing from the raw input to a total of 130 Million de-duplicates sentences (over 2 Gigatokens) took about 8 hours. The annotation with POS tags and the counting of 1-5-grams with and without POS tags required another 11 hours. Since most steps scale linearly with the amount of input data, the speed can be increased easily by adding more machines with sufficient disk space, without ever having to care about RAM limits.

[21] http://opennlp.apache.org/
[22] http://code.google.com/p/dkpro-core-asl/
[23] http://sourceforge.net/p/jobimtext/
[24] http://wortschatz.uni-leipzig.de/findlinks/

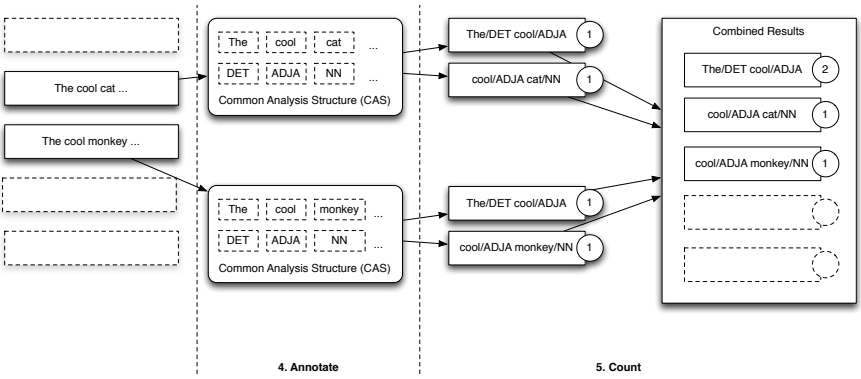


Figure 3: Annotation, pattern gathering and count example

3.2 COW Project

We describe here the most recent (COW2013) workflow, which was first tested on the UKCOW2012 corpus that is already available.[25] In the COW project, everything up to but not including tokenization is implemented in the `texrex` software suite.[26] The software is available under (modified) GNU (Lesser) General Public License(s) and is written in the object-oriented dialect of FreePascal.[27] The latter choice is motivated by the speed and exceptionally good platform support of the compiler. Even integrating the tools into Java-based environments is unproblematic, because the compiler can compile directly (i. e. without code translations) to JVM byte code.

First, there is a *conversion* step, which includes encoding normalization to UTF-8 using ICU, as well as HTML stripping. To allow for later analysis of link structures, link relations between the corpus documents are extracted in the HTML stripping process. Then, documents undergo certain quality assessment steps, most importantly boilerplate detection using a Multi-Layer Perceptron (MLP) trained on document-internal features [85] and text quality assessment [84].[28] Instead of simply removing potential boilerplate material, it is preserved and marked as potential boilerplate by storing the output of the MLP as paragraph metadata. Also, the overall text quality, which is measured as the lack of otherwise highly frequent words, only leads to the removal of the document if it is extremely low. Otherwise, the text quality metric is added as document metadata, and the document is preserved. This allows corpus users to perform queries restricted

[25] `http://hpsg.fu-berlin.de/cow/?action=corpora`
[26] `http://sourceforge.net/projects/texrex/`
[27] `http://www.freepascal.org/`
[28] Notice that recent versions of the COW software uses 39 features as input for the boilerplate classifier, not just the 9 features described in the original paper. A comprehensive list of features for this task is described in [91].

to potentially non-noisy regions of the corpus, while still being able to see the noisier regions (cf. Section 5.2 for why this is desirable).

Removal of near-duplicate documents is performed by a conservative implementation of w-Shingling without clustering as described in [25] and using an efficient implementation of Rabin hashes [78]. All important parameters (w, fingerprint size, and sensitivity) are configurable. Duplicated documents are removed, but for future versions of the software, a method of recording the degree to which each document, which was not removed, had duplicates in the original crawl data is being developed.

In the *annotation* step, standard tools for POS tagging, chunking, and named entity recognition are used, but the tools vary from language to language to ensure best possible results. For example, named entity recognition for German turned out to work best using the Stanford NER tool with the deWaC-generalized classifier described in [41], whereas for Swedish, the best option is the Stockholm Tagger, which also performs NER [83].[29,30]

In addition to finding the optimal annotation tools and models, the noisy data in web documents usually requires hand-crafted pre- and post-processing scripts (or even modified rules and models for the software used) in each annotation step. For example, special language-specific rule sets for the Ucto tokenizer were developed to handle documents which contain emoticons, non-standard use of punctuation, etc.[31] An introduction to the problems of linguistic processing of noisy data can be found in Chapter 4 of [86].

Finally, because whole documents cannot be redistributed under German copyright legislation, shuffled versions are produced for public release in the *shuffle* step. Sentence-wise shuffled corpora contain single sentences in random order (unique, with the frequency of the sentence in the original corpus as metadata). However, for tasks like distributional semantics, there will be in-document shuffles (sentences within the documents are shuffled) and windowed shuffles (words are shuffled within windows of n words). For such applications, it is a commendable option to use only those documents classified as being of high quality and the paragraphs which are most likely not boilerplate.

4 Assessing Corpus Quality

Large-scale web corpora as discussed in this article are often designed to replace and extend a traditional general-language reference corpus such as the British National Corpus (BNC) [5]. In contrast to more specialized samples used to study text types unique to computer-mediated communication (e. g. [16]), web corpora are expected to be similar to "traditional" corpora compiled from newspapers, magazines, books, essays, letters, speeches, etc. Their advantages lie in (i) better accessibility (e. g. there is no German reference corpus whose full text can freely be accessed by researchers),

[29] http://www.nlpado.de/~sebastian/software/ner_german.shtml
[30] www.ling.su.se/english/nlp/tools/stagger/
[31] http://ilk.uvt.nl/ucto/

(ii) much larger size, allowing for a more sophisticated and reliable statistical analysis (web corpora are now typically 10 to 100 times larger than the BNC), (iii) inclusion of up-to-date material so that recent trends and developments can be tracked (most of the texts included in the BNC were produced between 1985 and 1993), and (iv) a broader range of authors and genres (such as fan fiction, semi-personal diaries, blogs, and forum discussions) than can be found in traditional corpora.

For these reasons, it is of great importance to control the quality of the material included in a web corpus and to assess its usefulness for linguistic purposes. Section 4.1 describes how general corpus statistics can be used to select high-quality text and detect problematic material in a web corpus. Section 4.2 presents a comparative evaluation of a number of existing web corpora, which are tested on the linguistic task of collocation identification.

4.1 Corpus Statistics

In this section, we briefly describe several straightforward methods for comparing corpora based on quality metrics, and show how these metrics can be used to identify problems with corpus quality.

Corpora can be characterized – and thus compared – by a wide range of statistical parameters. Three types of conclusions can be drawn from such corpus statistics:

- When comparing similar corpora (same language, genre, and size): If they differ in some key parameters, this may indicate quality problems.
- When comparing corpora of the same language, but with different genre or subject area: Parameters that correlate with genre or subject area can be identified.
- When comparing corpora of different languages: Systematic correlations between parameters can be identified. Features that are relevant for typological classifications or the classification into language families can be identified.

Furthermore, such corpus statistics can also be used to assess corpus quality. Quality assurance is a major challenge, especially when dealing with hundreds of corpora, possibly containing millions of sentences each, which can hardly be evaluated by hand [35]. For example, extreme values for certain statistics are possible indicators of problematic/noisy objects which require further inspection. Such words, sentences, or even whole documents are often of dubious quality for most types of linguistic research, and removing such objects can therefore improve corpus quality. In other cases, the distribution of some parameter is expected to follow a smooth curve, and any sharp peak can be a good indicator that more in-depth checking is required. Typical statistics used for assessing corpus quality are:

- distribution of word, sentence, or document lengths;
- distributions of characters or n-grams; and
- agreement with certain empirical laws of language such as Zipf's Law [96].

If anomalies are found by examining such statistics, further cleaning can be applied to rectify the problems which cause the anomalies. In particular, such metrics can be indicative of necessary modifications of the processing chain.

For the Leipzig Corpora Collection, there is a *Technical Report Series on Corpus Creation* that suggests various corpus statistics as indicators of corpus quality.[32] The following features turned out to be good indicators for potential flaws in corpora:

- C1: largest domains represented in the corpus and their size
- C2: number of sources per time period (in the case of continuous crawling)
- W1: graph showing the length distribution of words
- W2: list of n most frequent words in the corpus
- W3: list of longest words among the n most frequent & longest words overall
- W4: words ending in a capitalized stop word
- A1: list of characters (alphabet) and their frequencies
- A2: possibly missing abbreviations (left neighbors of the full stop with additional internal full stops)
- S1: shortest and longest sentences
- S2: length distribution of sentences (measured both in words and in characters)

The inspection of such statistics can reveal different kinds of preprocessing problems:

- problems with crawling (C1, C2),
- wrong language (W3, A1),
- wrong character set (A1, W3, S1),
- near-duplicate sentences (S2),
- problems with sentence segmentation (W4, S1),
- predominance of a certain subject area (C1, W2),
- many malformed sentences (W4, A2, S1, S2).

Figure 4 shows an example for feature S2. We computed the sentence length distribution (in characters) for two corpora. The Hindi news corpus shows the expected shape of the distribution, while the distribution in the Sundanese Wikipedia is highly atypical. On closer examination, the peaks turned out to be the result of boilerplate material and near duplicates, which should have been removed.

4.2 Linguistic Evaluation: The Collocation Identification Task

The linguistic usefulness of web corpora as representative samples of general language (rather than web-specific genres) can be evaluated by comparison with a traditional reference corpus, using frequent general-language words and constructions as test items. The underlying assumption is that an ideal web corpus should agree with the reference corpus on the syntactic and lexical core of the language, while offering better coverage of

[32] `http://asvdoku.informatik.uni-leipzig.de/corpora/index.php?id=references`

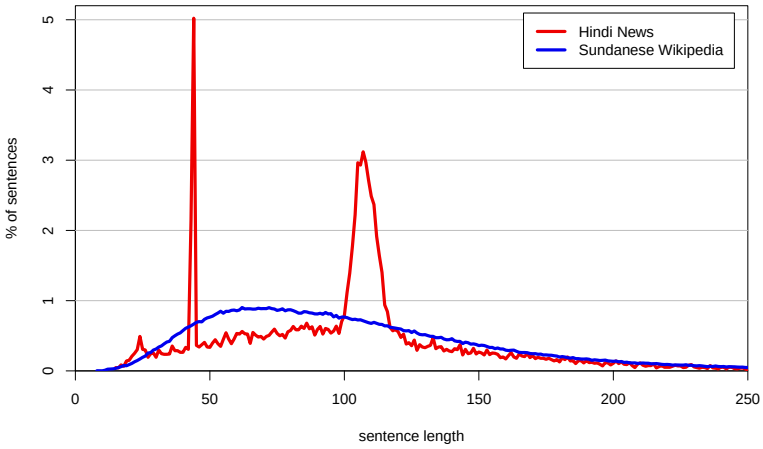


Figure 4: Distribution of sentence lengths (measured in characters) in a corpus of Hindi newspapers vs. a corpus made from the Sundanese Wikipedia.

less frequent words and construction, highly specialized expressions, and recently coined words. Evaluation criteria range from direct correlation of frequency counts [61] to assessing the benefit of web-derived data for a natural language processing application such as measuring semantic textual similarity [9, 57].

For the experiments reported here, we selected a task that plays a central role in the fields of corpus linguistics and computational lexicography: the automatic identification of collocations and other lexicalized multiword expressions (MWE) based on statistical association scores that are computed from the co-occurrence frequency of a word combination within a specified span and the marginal frequencies of the individual words. This task has two important advantages: (i) unlike frequency correlation, it allows us to assess the linguistic quality of the web data, not just its surface similarity to a reference corpus; (ii) unlike complex NLP applications, the evaluation results directly reflect corpus frequency data and do not depend on a large number of system components and parameters.

We evaluated a number of English-language web corpora that differ in their size, composition and annotation (cf. Sec. 2.3): 0.9 billion tokens from the English corpus of the *Leipzig Corpora Collection* (LCC) [52]; the ukWaC corpus of British English web pages compiled by the WaCky initiative [11];[33] the aggressively filtered UMBC WebBase corpus [57]; a preliminary version of the public release of UKCOW2012 [85]; and the Google Web 1T 5-gram database [24]. As a point of reference, we used the British

[33] Due to technical limitations of the corpus indexing software, only the first 2.1 billion tokens of the corpus could be used, omitting approx. 5% of the data.

name	size (tokens)	corpus type	POS tagged	lemmatized	basic unit
BNC	0.1 G	reference corpus	+	+	text
WP500	0.2 G	Wikipedia	+	+	fragment
Wackypedia	1.0 G	Wikipedia	+	+	article
ukWaC	2.1 G	web corpus	+	+	web page
WebBase	3.3 G	web corpus	+	+	paragraph
UKCOW	4.0 G	web corpus	+	+	sentence
LCC	0.9 G	web corpus	+	−	sentence
LCC ($f \geq k$)	0.9 G	web n-grams	+	−	n-gram
Web1T5 ($f \geq 40$)	1000.0 G	web n-grams	−	−	n-gram

Table 1: List of corpora included in the linguistic evaluation, with size in billion tokens, type of corpus, linguistic annotation and unit of sampling.

National Corpus (BNC), a balanced sample of written and spoken British English from the early 1990s, comprising a total of 110 million tokens of text [5]. We also included Wackypedia, a corpus derived from a 2009 snapshot of the English Wikipedia.[34] This represents an alternative approach to building large and up-to-date general-language corpora from online sources. While similar in size to the corpora obtained from web crawls, it covers a much narrower range of genres and should consist mostly of standard-conformant edited written English. Finally, WP500 is a subset of the Wackypedia corpus containing the first 500 words from each article, resulting in a much smaller (230 million instead of 1 billion tokens) and more balanced sample.[35] Table 1 summarizes characteristics of the corpora included in the evaluation.

The comparison of five different web corpora allows us to study the influence of different aspects of web corpus compilation with respect to linguistic usefulness. Corpus size ranges from less than 1 billion words (LCC) to 4 billion words (UKCOW), while Web1T5 offers two orders of magnitude more data. All corpora except for Web1T5 were automatically tagged with part-of-speech annotation; ukWaC, WebBase and UKCOW were also lemmatized. LCC, WebBase and Web1T5 were compiled from broad, unrestricted crawls of the web, while ukWaC and UKCOW were restricted to the `.uk` domain. The corpora also represent different strategies for selecting and cleaning web pages, as well as different levels of de-duplication (entire web pages for ukWaC, paragraphs for WebBase, and individual sentences for UKCOW and LCC).

N-gram databases such as Web1T5 are an increasingly popular format of distribution: they circumvent copyright issues, protect the intellectual property of the resource owner, and are usually much smaller in size than the underlying corpus. Such databases record the frequencies of all word forms (or lemmas) in the corpus, as well as those

[34] `http://wacky.sslmit.unibo.it/doku.php?id=corpora`

[35] In addition to truncating each article, disambiguation pages, listings and "messy" pages were removed with heuristic filters.

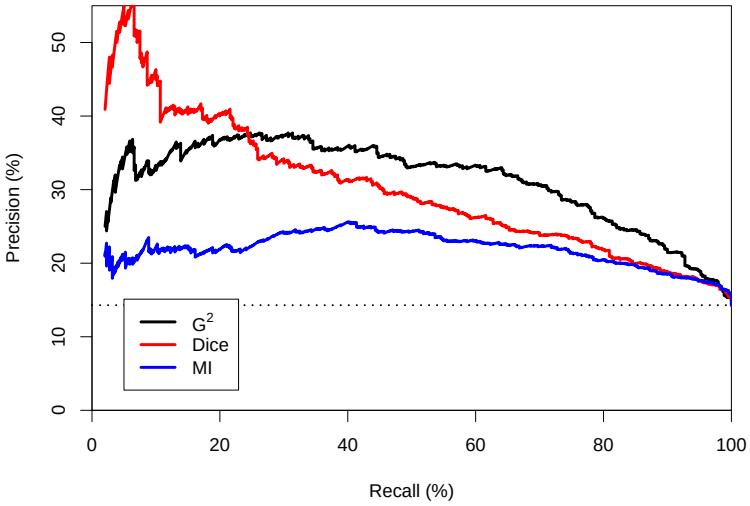


Figure 5: Precision-recall graphs for identification of lexicalized verb-particle combinations (VPC) based on co-occurrence data from British National Corpus.

of bigrams, trigrams, etc. of consecutive words. Co-occurrence frequencies for word pairs can be computed by summing over suitable n-grams of different sizes [38]. With a 5-gram database such as Web1T5, the largest possible span size is 4 words. However, most available databases omit n-grams below a certain frequency threshold in order to keep the amount of data manageable (e.g. $f < 40$ in the case of Web1T5). As a result, co-occurrence frequencies for spans of more than a single adjacent word are systematically underestimated; [38] refers to such data as quasi-collocations. In order to assess the usefulness of quasi-collocations obtained from n-gram databases, we collected all n-grams with $f \geq 5$ and $f \geq 10$ from LCC as described in Sec. 3.1. They are contrasted with the full-text LCC corpus in the evaluation below.

Our evaluation follows the methodology proposed by [39, 40]. It is based on a gold standard of candidate MWE, among which all true positives have been identified by human raters. The candidates are then ranked according to various statistical association measures computed from their observed and expected co-occurrence frequencies in a given corpus. Ideally, all true positives should be ranked at the top of the list (high association scores) and non-lexicalized word combinations at the bottom of the list (low association scores). In order to quantify this intuition, n-best lists of the highest-ranked candidates are evaluated in terms of precision (= percentage of true positives in the n-best lists) and recall (= percentage of true positives in the gold standard that are also found in the n-best list).

Results for many different values of n can be collected in precision-recall graphs as

shown in Fig. 5. Each point on the red curve represents some n-best list according to the Dice association measure. Its position on the x-axis specifies the recall achieved by this n-best list; its position on the y-axis specifies the precision achieved by the n-best list. Different association measures can thus be compared at a glance. For a meaningful interpretation, the precision-recall graphs should always be compared to the baseline precision (i. e. the overall percentage of true positives in the gold standard) indicated by the dashed horizontal line.

Fig. 5 illustrates a fairly typical situation in which no single best association measure can be identified. The red curve achieves highest precision for low recall values, whereas the black curve is considerably better for high recall values. In order to ensure a unique ranking of association measures, we use average precision (AP) as a global evaluation criterion. This measure averages precision values across all recall points from 0% to 100%; it corresponds to the area under the precision-recall graph. High AP indicates a good ranking of the candidates and a well-balanced trade-off between precision and recall. An ideal association measure would achieve an AP of 100% by collecting all true positives at the top of the ranking. A random shuffling of the candidates leads to an AP close to the baseline precision. In Fig. 5, the red curve has an AP of 30.13% and the black curve has an AP of 31.06%; the baseline AP is 14.29%. We would thus conclude that the ranking corresponding to the black curve is slightly better on the whole than the ranking corresponding to the red curve.

We evaluate the web corpora on two collocation identification tasks that focus on different aspects of multiword expressions and different types of data. The first task is concerned with the distinction between compositional and non-compositional verb-particle combinations (VPC). It shows whether the corpus-based measures help to separate semantically opaque word combinations from semi-compositional or habitual collocations, and whether they are suitable for combinations involving highly frequent grammatical words (the particles). The second task is concerned with a more intuitive notion of collocations as habitual word combinations [45]. This task shows how well corpus-based measures correspond to the intuitions of lexicographers collected in a collocations dictionary [18], and whether they are suitable for low-frequency combinations of content words. It is also relevant for applications in language education, where such collocations help to improve the language production skills of advanced learners [17].

English verb-particle combinations Our first evaluation study is based on a gold standard of 3,078 English verb-particle combinations that were manually classified as non-compositional (true positive, e. g. *carry on*, *knock out*) or compositional (false positive, e. g. *bring together*, *peer out*) [8]. This data set has previously been used for evaluation purposes by the multiword expressions community and thus allows a direct comparison. For example, [79] report a best result of 26.41% AP based on a subset of the BNC, which is clearly outperformed by the web corpora in our study (cf. Table 2).

We extracted co-occurrence counts for the word pairs from each corpus, using a 3-word span to the right of the verb (L0/R3 in the notation of [37]), which gave consistently best results in preliminary experiments. POS tags were used to filter the corpus data if

corpus	POS filter	size (tokens)	average precision G^2	t	MI	Dice	MI^2	X^2
BNC	+	0.1 G	31.06	29.15	22.58	30.13	30.97	**32.12**
WP500	+	0.2 G	28.01	25.73	27.81	30.29	29.98	**31.56**
Wackypedia	+	1.0 G	28.03	25.70	27.39	30.35	30.10	**31.58**
ukWaC	+	2.2 G	30.01	27.82	25.76	30.54	30.98	**32.66**
WebBase	+	3.3 G	30.34	27.80	27.95	31.74	32.02	**33.95**
UKCOW	+	4.0 G	32.31	30.00	26.43	32.00	32.96	**34.71**
LCC	+	0.9 G	25.61	24.83	22.14	**26.82**	25.09	26.38
LCC ($f \geq 5$)	+	0.9 G	26.95	26.45	25.54	**27.78**	25.96	27.66
LCC ($f \geq 10$)	+	0.9 G	27.34	26.81	27.13	27.85	25.95	**28.09**
LCC	−	0.9 G	24.67	23.63	21.41	25.36	23.88	**25.63**
LCC ($f \geq 5$)	−	0.9 G	25.45	24.79	23.54	**26.30**	24.55	26.21
LCC ($f \geq 10$)	−	0.9 G	25.84	25.16	25.28	26.49	24.71	**26.63**
Web1T5 ($f \geq 40$)	−	1000.0 G	26.61	26.12	21.67	**27.82**	25.72	27.14

Table 2: Evaluation results (in terms of average precision) for identification of lexicalized verb-particle combinations (VPC). The best result for each corpus is highlighted in bold font. The baseline precision for this task is 14.29%.

available (first word tagged as lexical verb, second word tagged as preposition, particle or adverb). The verbs were lemmatized if possible. For corpora without lemma annotation, morphological expansion was applied, i.e. frequency counts for all inflected forms of each verb were aggregated. Since the performance of an association measure can vary in unpredictable ways for different corpora and evaluation tasks, we present results for a range of standard, widely-used association measures: the log-likelihood ratio (G^2), t-score (t), Mutual Information (MI), the Dice coefficient (Dice), a heuristic variant of Mutual Information (MI^2) and Pearson's chi-squared test with continuity correction (X^2). See [37] for details and references.

Table 2 summarizes the evaluation results for the VPC task. The average precision of the best association measure for each corpus is highlighted in bold font. The overall best result is obtained from the UKCOW web corpus, with an AP of 34.71%. Comparing Wackypedia, ukWaC, WebBase, and UKCOW, we find that AP increases with corpus size for the web-derived corpora. Scaling up does indeed pay off: the largest corpora clearly outperform the reference corpus BNC. However, they need more than an order of magnitude more data to achieve the same AP of 32.12%.

Surprisingly, Web1T5 achieves only low precision despite its huge size. There are several possible explanations for this observation: distribution as an n-gram database with frequency threshold, poor quality of the corpus (i. e. little filtering and cleaning of web pages), as well as lack of POS tagging and lemmatization. The first hypothesis is clearly ruled out by the LCC results, which show no detrimental effect for n-gram frequency thresholds of $f \geq 5$ and $f \geq 10$.[36] In fact, AP improves by more than 1%

[36] Taken in relation to corpus size, these thresholds are much more aggressive than the $f \geq 40$ threshold of Web1T5.

when the thresholds are applied. Evaluating LCC with and without a POS filter, we found a small benefit from using the annotation. Reliable lemmatization appears to be even more important and might account for the considerably lower AP of LCC compared to the similar-sized Wackypedia.

We can summarize the conclusions from this first evaluation study as follows. Web corpora indeed seem to be a valid replacement for a traditional reference corpus such as the BNC, provided that suitable strategies are used to select and clean up web pages and the resulting corpus is enriched with linguistic annotation. The precise implementation of this procedure and the level of de-duplication do not appear to play an important role: AP grows with corpus size regardless of other parameters. Nonetheless, diversity may be a relevant factor. Web corpora of more than a billion words are needed to achieve the same performance as the 100-million-word BNC, and WP500 discards almost 80% of the Wackypedia text without a noticeable loss of precision. Web1T5 shows that sheer size cannot make up for "messy" content and lack of annotation. The distribution format of an n-gram database is not detrimental *per se* for MWE identification and similar tasks, though.

BBI Collocations The second evaluation study is based on a gold standard of 36,328 two-word lexical collocations from the BBI Combinatory Dictionary [18], which are close to the notion of collocation put forward by J. R. Firth (cf. [13]) and have important applications in language education [17]. In contrast to the first task, there is no fixed list of candidates annotated as true and false positives: a different candidate set is extracted from each corpus and evaluated by comparison with the gold standard. Candidate sets were obtained by collecting all co-occurrences of nouns, verbs, adjectives and adverbs within a span of three words to the left and right (L3/R3), allowing only words that appear in one or more entries of the BBI dictionary.[37] POS tags were used as a filter if available. Since the gold standard does not distinguish collocations between lemmas from collocations that are restricted to particular word forms, all candidate pairs were lemmatized. For corpora without lemma annotation, morphological expansion was used to aggregate frequency counts for all possible inflected forms of each candidate pair, based on word form–lemma correspondences obtained from automatically tagged and lemmatized corpora.

In order to ensure a fair comparison, all candidate lists were truncated to the most frequent 1 million word pairs. While this reduces coverage of the gold standard for the larger Web corpora, our previous experience suggests that high-frequency candidates can be ranked more reliably than lower-frequency data, resulting in better AP scores if frequency thresholds are applied. Preliminary experiments with the WebBase corpus showed quite stable results for candidate lists ranging from 1 million to 10 million word

[37] [13] evaluated different span sizes (3, 5 and 10 words) on a subset of the BBI data, obtaining better results (i.e. higher AP) for smaller spans. For the present experiment, we decided to use a three-word span as a good compromise between precision and coverage. This choice was corroborated by additional preliminary tests on the full BBI data with span sizes ranging from 1 to 10 words.

corpus	size (tokens)	average precision (up to 50% recall) G^2	t	MI	Dice	MI^2	X^2
BNC	0.1 G	22.76	18.48	16.68	20.26	**25.16**	24.89
WP500	0.2 G	21.68	17.08	16.31	19.71	**24.41**	24.33
Wackypedia	1.0 G	21.45	16.90	17.41	19.98	**24.64**	24.56
ukWaC	2.2 G	17.22	13.58	15.63	16.30	**20.39**	20.33
WebBase	3.3 G	19.30	15.06	17.77	18.21	22.77	**22.88**
UKCOW	4.0 G	20.65	16.35	18.28	19.29	**24.15**	24.04
LCC	0.9 G	15.02	12.11	13.58	14.22	**17.78**	17.67
LCC ($f \geq 5$)	0.9 G	15.80	13.35	13.99	15.22	**18.59**	18.50
LCC ($f \geq 10$)	0.9 G	15.72	13.44	14.46	15.30	**18.51**	18.39
Web1T5	1000.0 G	11.95	10.73	11.08	11.38	**13.50**	13.18

Table 3: Evaluation results for identification of BBI collocations. The best result for each corpus is highlighted in bold font.

pairs. Since coverage (and thus the highest recall percentage that can be achieved by a ranking) differs between corpora, it is not meaningful to specify a baseline precision. As a global evaluation criterion, we use average precision for recall points up to 50% (AP50).

The gold standard used here is not without problems. Recent collocations found in the web corpora may not be attested due to the age of the BBI dictionary (published in 1986). Moreover, true positives may be missing due to introspective bias or the space constraints of a paper dictionary. Manual inspection of small samples from the ranked candidate lists revealed hardly any candidates that were marked as false positives despite being clearly collocational in present-day English. We believe therefore that the AP scores of web corpora are underestimated by less than 1%, a margin of error that has no substantial influence on the conclusions drawn from the evaluation. However, a more comprehensive manual validation and quantitative analysis is certainly desirable and is planned for future work.

The results of the second study are shown in Table 3. In this case, corpus composition plays a much more important role than size. Comparing the web corpora LCC, ukWaC, WebBase, and UKCOW, we find a linear increase of AP50 with corpus size that does not seem to depend strongly on other aspects of corpus compilation. However, the 4-billion word UKCOW corpus is still outperformed by the reference corpus BNC. Even a 230-million-word subset of Wikipedia (WP500) is better suited for the identification of lexical collocations than web corpora. The results for LCC confirm our previous observation that n-gram databases with frequency thresholds need not be inferior to full-text corpora, provided that the analysis can be carried out within a window of 5 words. Finally, the extremely low precision of Web1T5 can be explained by its "messy" content and the lack of POS tagging.

In summary, our evaluation study has shown that web corpora can be a valid

replacement for a traditional reference corpus such as the BNC, with much better coverage and up-to-date information. This is not simply a matter of scaling up, though. Unless web pages are carefully selected, cleaned, de-duplicated and enriched with linguistic annotation, even a gigantic text collection such as Web1T5 offers little value for linguistic research. In some tasks, web corpora are still inferior to traditional corpora (and even to other online material such as Wikipedia articles), but this is likely to change with a further increase in corpus size.

5 Working with Web Corpora / Linguistic Applications

5.1 Accessing Very Large Corpora

For most linguistic applications of web corpora, an essential step is to scan the text and annotation with sophisticated search patterns, often based on linguistic annotation such as part-of-speech tags. Ideally, it should be possible to execute simple queries interactively and obtain the search results within a matter of seconds. Several dedicated corpus query engines can be used for this purpose, such as the IMS Open Corpus Workbench, the NoSketch Engine, and Poliqarp.[38,39,40] Most of these query engines will comfortably handle corpora of several billion words even on state-of-the-art commodity hardware, which is sufficient for all publicly available web corpora with linguistic annotation (including, in particular, LCC, UKCOW, ukWaC, and UMBC WebBase). They can easily be scaled up to much larger amounts of data by sharding the text across a grid of servers (with simple wrapper programs to collect and merge result sets). Specialized, highly scalable web/text search engines such as Apache Lucene[41] do not offer the expressiveness and complexity required by most linguistic corpus queries.

For certain applications, it can be convenient to provide web corpora in the form of pre-compiled n-gram databases. Because of their tabular form, such data sets can easily be stored in a relational database, enabling indexed retrieval and flexible aggregation of frequency counts. See [38, 66] for suggestions and a suitable database design. Compact storage and interactive access for full n-gram counts from trillion-word web corpora can only be achieved with specially designed software packages [80, 47].

5.2 Corpus Linguistics and Theoretical Linguistics

Compiling static web corpora in conjunction with an efficient retrieval system and a powerful query language have been argued to be the best way of exploiting web data for linguistic purposes [69]. While large web corpora have unique features to offer to the linguist and are thus particularly attractive in many respects (see Section 4), there are also a number of potential disadvantages (or, more neutrally speaking, characteristics), which have consequences for their use in linguistic research.

[38] http://cwb.sourceforge.net/
[39] http://nlp.fi.muni.cz/trac/noske/
[40] http://poliqarp.sourceforge.net/
[41] http://lucene.apache.org/core/

Composition A long-standing and much-discussed problem is the question of representativeness: In order to make valid inferences about a particular language on the basis of corpus data (in terms of fundamental research), corpus linguists often require a general purpose corpus to be representative of that language. For other, more task-oriented linguistic applications (such as extracting collocations or creating thesauri), representativeness might not strictly be required, as long as reasonably good results are obtained from an application perspective. In traditional corpus creation, representativeness is usually approached by following an elaborated stratified sampling scheme (cf. also Section 2.1), resulting in a corpus that has the desired composition in terms of genres, topic domains, and socio-linguistic factors. The difficulty lies in determining the "correct" composition in the first place, and different proposals exist as to how the importance of individual text types should be assessed. Seminal papers include [6, 19]. [68] discusses representativeness in web corpora. For a critical stance on representativeness and balance, see [60]. A recently suggested new approach [63] is characterized by the creation of a very large *primordial sample* with a lot of meta data which is in principle relevant for stratification. From this sample, users can then create specialized *virtual corpora* according to their desired sampling frame. This is an interesting idea for web corpora as well, because they are usually very large. However, the lack of suitable meta data stands in the way of the creation of virtual corpora. Also, if the primordial web document sample itself is heavily biased (cf. 2.2), then it is unclear whether the virtual corpora can adequately model the user's sampling frame.

That said, in the real-world construction of web corpora, there are at two major design procedures with respect to corpus composition:

Random The corpus should be a "random sample" from the population of web texts in a particular language. It should thus reflect the distribution of text types, topics etc. of (a particular segment of) the web. What can we expect then, in comparison to other corpora? An impressionistic estimate is given in [46], according to which legal, journalistic, commercial, and academic texts make up for the major part of prose texts on the web. If true, it would not be unreasonable to expect a web corpus built from a random sample to be comparable in terms of composition to traditional, "general purpose" corpora, since these kinds of texts dominate the composition of traditional corpora as well. On the other hand, [88] finds that, in a given web corpus, texts about the arts, humanities, and social sciences are under-represented compared to the BNC, while texts from technical fields are over-represented. However, this estimate is based on a web corpus made from a breadth-first crawl that was seeded with URLs from search engines. It is thus not likely to be representative of the population of texts on the web, as has been argued in Section 2.2.

Balanced The corpus should be balanced in such a way that no genre or topic is heavily over-represented. Such an approach is taken, e. g. by [31], who argue that the distribution of genres and topics on the web is probably different from what should be contained in a general purpose corpus, and thus a random sample from the population of web pages is not actually desirable. The aim of creating a corpus that is balanced in this sense conflicts with the idea of statistical representativeness.

Building a corpus by random crawling still allows for statements about the composition, however, but they can only be made after the corpus has been constructed, and they have to be based either on estimates from small hand-annotated samples or on automatic classification. See Table 4, reproduced from [85] for an assessment of the genre/text type composition of DECOW2012 and ESCOW2012 (Spanish) based on a small manually annotated sample. The classification scheme is based on the scheme suggested in [88] with only a few added categories. Notice the high number of documents for which authorship cannot be determined as well as the (surprisingly) low proportion of truly fictional texts. The *Audience* classification is a simple evaluation of a document's readability. The *Quasi-Spontaneous* classification for *Mode* refers to forum discussions and the like, whereas *Blogmix* refers to web pages containing a mix of written text plus a forum-like discussion. The amount of documents containing potentially non-standard language is thus estimated at 27% for DECOW2012. Interestingly, the substantial difference to the Spanish ESCOW2012 in this regard (only 11.5%) could be an interesting result related to the linguistic and socio-linguistic characterization of national top-level domains. However, given that biased crawling algorithms and potentially biased search engine-derived seed URLs were used, this must be taken with a grain of salt.

Metadata One of the most serious drawbacks of web corpora is their almost complete lack of document metadata, which is essential for a thorough linguistic interpretation. Such metadata includes information about date of publication/utterance, age, sex, etc. of the author/speaker, text type, and topic domain. Linguists often adjust their hypotheses according to the distribution of such categories in a given corpus. However, no such data is encoded in web documents in a reliable and standardized way. As an additional challenge, some web documents resist a straightforward classification along such traditional categories because of their complex editing history (e. g., multiple authors producing subsequent versions of a text over a period of time). In any event, the only feasible way to gather such metadata for each document in a crawled web corpus is automatic document classification. This can be done with high accuracy for some dimensions of classification, but it is never free of errors. Although some corpus linguists might find this problematic, the size of web corpora usually allows users to extract concordances large enough to make any hypotheses testable with high reliability despite such error rates, and automatic generation of metadata is not just the only option, but in fact a valid one. Furthermore, it is worth pointing out that the lack of metadata is not unique to web corpora. Consider, for instance, corpora containing mostly newspaper articles (like the German DeReKo [63]), where authorship cannot always be attributed to specific individuals.

Document structure The suitability of a web corpus for a specific area of linguistic research depends to a considerable extent on the methods of non-linguistic post-processing applied to the raw data, cf. Section 3. For instance, if we remove boilerplate or treat sentences containing emoticons as noise and delete them, then we run the risk

	DECOW2012		**ESCOW2012**	
Type	%	CI ±%	%	CI ±%
Authorship				
Single, female	**6.0**	2.8	**5.0**	2.5
Single, male	**11.5**	3.7	**16.5**	4.3
Multiple	**36.0**	5.6	**16.5**	4.3
Corporate	**21.0**	4.7	**20.5**	4.7
Unknown	**25.5**	5.0	**41.5**	5.7
Mode				
Written	**71.0**	5.0	**86.0**	4.0
Spoken	**1.0**	3.0	**2.5**	1.8
Quasi-Spontaneous	**22.5**	4.9	**3.5**	2.1
Blogmix	**4.5**	2.4	**8.0**	3.2
Audience				
General	**75.5**	5.0	**94.0**	2.8
Informed	**17.0**	4.4	**2.5**	1.8
Professional	**7.5**	3.0	**3.5**	2.1
Aim				
Recommendation	**12.5**	3.8	**7.0**	3.0
Instruction	**4.5**	2.4	**6.0**	2.8
Information	**36.0**	5.5	**41.5**	5.7
Discussion	**47.0**	5.8	**44.5**	5.8
Fiction	**0.0**	0.0	**1.0**	1.2
Domain				
Science	**2.5**	1.8	**5.0**	2.5
Technology	**14.0**	4.0	**6.5**	2.9
Medical	**4.5**	2.4	**4.0**	2.3
Pol., Soc., Hist.	**21.5**	4.8	**21.0**	4.7
Business, Law	**10.0**	3.5	**12.5**	3.8
Arts	**8.5**	3.2	**8.5**	3.2
Beliefs	**5.0**	2.5	**3.0**	2.0
Life, Leisure	**34.0**	5.5	**39.5**	5.7

Table 4: Text category/genre distribution in DECOW2012 and ESCOW2012 with 90% confidence interval ($n = 200$)

of removing material (e. g. headings or even entire paragraphs) that would have been crucial in the linguistic analysis of discourse-related phenomena, such as co-reference, information structure, and rhetorical structure. Similarly, research on text types and genres is likely to be affected by the removal of such material. Even studies on syntax may require the context of individual sentences to be available, for instance when examining the syntax of sentence connectors. The same is true for some computational linguistics tasks such as distributional semantics based on larger units than sentences (like LSA [67]). Corpora of single sentences are inadequate for such research. Unfortunately, even corpora containing full documents like the COW corpora have often to be distributed in a shuffled form (randomly sorted single sentences) to avoid legal problems with copyright claims. At least for tasks in computational linguistics, corpora containing full documents within which the order of the sentences has been randomized might be a viable solution, as mentioned in Section 3.2.

Duplication Web corpora consisting of complete documents usually contain a certain amount of duplicated material, even if some form of de-duplication was applied in post-processing. Again, this kind of problem concerns non-web corpora as well, albeit to a much lesser extent. For example, newspaper corpora sometimes contain multiple instances of news agency material that was reprinted by several newspapers (see, e.g. 62 on near-duplicate detection in large, non-web corpora). As described in Section 3, in a web corpus (containing only individual sentences) duplication can be dealt with by removing all but one instance of each sentence. The publicly available versions of the COW corpora are distributed in a similar way. The frequency of each sentence in the full corpus is also recorded as metadata, allowing users to reconstruct statistical information for frequent sentences that are not boilerplate, like *Hello!* or *Thank you.*

5.3 Quantitative Typology and Language comparison

The availability of corpora for a large variety of languages is a necessary requirement for performing typological studies. Creating high-quality web corpora, as described in this article, opens up further possibilities for this field.

Linguistic typology is concerned with the classification of the world's languages into types based on phonological, morphological, syntactic and other features [55]. This allows researchers to gain insights into the structure of language by studying possible or preferred types [32]. A main objective of quantitative typology is the search for systematic patterns of variation of language and the linguistic features they are based on. Thus it aims at finding, for example, absolute or statistical language universals [90, 36]. Absolute universals are rules which necessarily hold for all languages. Therefore, they are rare and typically of a very general nature. Statistical universals, on the other hand, describe preferences on a global or local scale. Among them are conditional universals (or implicational universals), which make assumptions about combinations of types or features that are typically preferred or avoided. Many studies in quantitative typology are concerned with parameters of language and their relations. Some works

are concerned with systematic interlingual correlations between the different parameters measured on the text resources [42, 43]. Correlations between measured parameters and typological parameters are also analyzed [43, 53].

In most cases, the starting point for research in quantitative typology is collections of textual resources, which are usually manually created. By using automatically collected web corpora instead, the necessary manual work can be greatly reduced. Based on these corpora simple, features like sentence, word, syllable, morpheme, etc. can be determined. While for many properties nearly exact values can be computed, some can only be approximated - especially when independence of language is aspired [51]. Possible features are:

- average word length in characters,
- average sentences length in words or characters,
- text coverage of the most frequent words,
- slope of Zipf's Law,
- different measurements of vocabulary richness,
- entropy on word and character level,
- average number of syllables per word or sentence,
- average syllable length,
- amount of affixes, or
- ratio of prefixes and suffixes.

In order to reliably compute these features, high quality web corpora are required [34], as properties such as size, text type, or subject area can have strong influence on typological investigations [51]. By keeping the main properties of corpora constant, statistical significance of typological analyses can be increased. Table 5 illustrates this influence. For each feature, it shows the ratio of cross-language standard deviation to standard deviation across other properties of corpora. Let us consider the example of average sentence length in words. First, we calculate the standard deviation of this feature across languages. In addition, we determine the standard deviation when varying corpus size (from 10,000 to 10,000,000 sentences), text type (random web text, Wikipedia, ...), or subject area (culture, science, economy, politics, ...). Since for most features a high variation between languages is expected, the ratio of the standard deviations should typically be larger than one. In the case of average sentence length in words, we measure a cross-language standard deviation which is 13 times larger than when modifying the subject area of the texts analyzed. Compared to the standard deviation dependent on corpus size it is even 107 times larger.

Utilizing methods like tests of significance, relationships with classical typological parameters - some kind of language universals - can be discovered. The following examples were identified by analyzing corpora in 700 languages and comparing measured features to available classical typological parameters taken from the World Atlas of Language Structures.[42] We found significant relations between:

[42] http://wals.info

measurement	language / corpus size	language / text type	language / subject area
average sentence length in words	107.41	8.65	13.20
average sentence length in characters	77.03	6.23	7.67
ratio of suffixes an prefixes	18.78	17.69	25.84
syllables per sentence	30.25	8.22	7.33
Type-Token Ratio	1.16	8.21	6.13
Turing's Repeat Rate	238.95	6.37	8.69
slope of Zipf's Law	3.27	11.35	11.25
text coverage of the top 100 words	530.85	7.93	8.75

Table 5: Comparison of standard deviations of corpus-based measurements. Quotient of cross-language standard deviation and other properties of corpora such as corpus size. Values larger than 1 imply a higher cross-language standard deviation.

- ratio of suffixes and prefixes and position of case marking (end of word vs. beginning of word): $p<0.001\%$, mean values of 10.48 and 0.7 and sample sizes of 57 and 11,
- average length of words of a language and its morphological type (concatenative vs. isolating): $p<1\%$, mean values of 8.43 and 6.95 and sample sizes of 68 and 8,
- measured amount of affixation of a language and its morphological type (concatenative vs. isolating): $p<0.5\%$, mean values of 21.20 and 10.06 and sample sizes of 68 and 8,
- average number of syllables per sentence and word order (SOV vs. SVO): $p<0.001\%$, mean values of 56.95 and 45.27 and sample sizes of 111 and 137,
- average number of syllables per word and morphological type (concatenative vs. isolating): $p<5\%$, mean values of 2.06 and 1.64 and sample sizes of 68 and 8 and
- average number of syllables per sentence and morphological type (concatenative vs. isolating): $p<5\%$, mean values of 21.76 and 27.47 and sample sizes of 68 and 8.

Typological parameters can also be predicted based on measurements on corpora. Using methods such as supervised machine learning, knowledge about different features of corpora can be combined to determine typological properties of a language. Table 6 shows results of the prediction of morphological type.

In addition, the branch of quantitative typological work concerned with vocabulary-based language comparison uses orthographic or phonetic similarity of words or letter n-grams to measure the similarity or relatedness of language pairs. Word lists for such analyses are usually compiled manually [94, 95, 27], but can also be extracted from corpora [89]. However, this comes with high demands for quality and comparability of the corpora used. One the one hand contamination of corpora with text in other languages can lead to unusually high similarity values. For web corpora especially

features used	accuracy
Baseline	50.0%
Words per sentence	74.4%
Number of word forms	87.8%
Words per sentence, number of word forms, syllables per word	91.8%

Table 6: Probability of correct prediction of morphological type of a language (concatenative vs. isolating) using different features and equal sample sizes for both classes.

English remnants are a common problem. On the other hand properties of corpora such as subject area can have influence on the analysis. For very similar languages such as the North Germanic genus, usage of varying subject areas across languages can lead to changes in resulting genus internal ranking of similarities.

6 Summary and Outlook

As discussed in this article, the availability of large corpora is a prerequisite for research in empirical linguistics as well as language technology. We showed how very large, high quality corpora can be constructed from the web despite of all the noise that makes the construction process a challenging enterprise. We focused on three main problems: crawling, processing, and quality control. Crawling is especially important if web corpora for a variety of languages, not just for English, should be created. We then introduced two projects, WebCorpus and COW, that both allow creating large-scale, linguistically annotated corpora from crawled data. We argued that when working with web data, controlling the quality of the resulting corpus is an important issue that we addressed with corpus statistics and a linguistic evaluation based on two collocation identification tasks. Finally, we showed how the availability of extremely large, high-quality web corpora fits in with research in various fields of linguistics, computational linguistics, and natural language processing.

Taking it all together, several desiderata for web corpora creation clearly emerge. Crawling procedures can be optimized in diverse ways, depending on corpus design goals. Focused and optimized crawling should be implemented to discover more interesting data using less bandwidth, storage, and time resources. This is especially true for smaller languages, for which documents are hard to find using non-selective crawling. Bias-free crawling should be used to derive truly representative samples from the web for fundamental research and linguistic web characterization.

The selection of optimal tools and models for linguistic post-processing and their integration into automatable tool chains (for example the UIMA architecture) is not yet completed for many languages. Also, the evaluation of the accuracy of such tools on web data is important under any corpus design goal, and researchers in this area can greatly benefit from joining efforts. For example, to make the COW processing

chain use the advantages of the WebCorpus technology, the COW annotators would just have to be added to the UIMA framework. Further linguistic annotations (like chunking and parsing) as well as automatic text classification and metadata generation need to be added in a similar fashion to improve the quality of web corpora. This is a prerequisite for their acceptance within (theoretical, empirical, and computational) linguistic research communities, as well as for more detailed quantitative analysis of linguistic phenomena.

Continued work on massive parallelization using architectures like Hadoop is the only way of meeting the computational requirements for corpora in the region of 100 billion or even trillions of tokens. As we have learned from the evaluation studies described in Section 4.2, quality increases with corpus size: a high-quality web corpus of 100 billion tokens should outperform traditional reference corpora in most application tasks.

Acknowledgments

The second evaluation study reported in Section 4.2 is based on joint work with Sabine Bartsch.

References

[1] Serge Abiteboul, Mihai Preda, and Gregory Cobena. Adaptive on-line page importance computation. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 280–290, New York, NY, USA, 2003. ACM.

[2] Myriam Abramson and DavidW. Aha. What's in a URL? Genre Classification from URLs. Technical report, AAAI Technical Report WS-12-09, 2009.

[3] Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore. On the Bias of Traceroute Sampling: or, Power-law Degree Distributions in Regular Graphs. In *Proceedings of the Thirty-seventh Annual ACM Symposium on Theory of Computing*, STOC '05, pages 694–703, New York, NY, USA, 2005. ACM.

[4] George Almpanidis, Constantine Kotropoulos, and Ioannis Pitas. Combining Text and Link Analysis for Focused Crawling – An Application for Vertical Search Engines. *Inf. Syst.*, 32(6):886–908, 2007.

[5] Guy Aston and Lou Burnard. *The BNC Handbook*. Edinburgh University Press, Edinburgh, 1998.

[6] Sue Atkins, Jeremy Clear, and Nicholas Ostler. Corpus design criteria. *Literary and Linguistic Computing*, 7(1):1–16, 1992.

[7] Ricardo Baeza-Yates, Carlos Castillo, and Efthimis N. Efthimiadis. Characterization of national Web domains. *ACM Trans. Internet Technol.*, 7(2), 2007.

[8] Timothy Baldwin. A resource for evaluating the deep lexical acquisition of English verb-particle constructions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 1–2, Marrakech, Morocco, 2008.

[9] Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. In *Proceedings of First Joint Conference on Lexical and Computational Semantics (*SEM), Montreal, Canada*, pages 435–440, Montreal, Canada, 2012.

[10] Ziv Bar-Yossef, Andrei Z Broder, Ravi Kumar, and Andrew Tomkins. Sic transit gloria telae: towards an understanding of the web's decay. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 328–337, New York, NY, USA, 2004. ACM.

[11] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009.

[12] Marco Baroni, Francis Chantree, Adam Kilgarriff, and Serge Sharoff. Cleaneval: a Competition for Cleaning Web Pages. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco, 2008.

[13] Sabine Bartsch and Stefan Evert. Towards a Firthian notion of collocation. In Andrea Abel and Lothar Lemnitzer, editors, *Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network Internet Lexicography*, OPAL – Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim, to appear.

[14] Eda Baykan, Monika Henzinger, Ludmila Marian, and Ingmar Weber. Purely URL-based Topic Classification. In *Proceedings of the 18th international conference on {World Wide Web}*, pages 1109–1110, 2009.

[15] Eda Baykan, Monika Henzinger, and Ingmar Weber. Web Page Language Identification Based on URLs. In *Proceedings of the VLDB Endowment*, pages 176–187, 2008.

[16] Michael Beiß wenger, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer, and Angelika Storrer. DeRiK: A German Reference Corpus of Computer-Mediated Communication. In *Proceedings of Digital Humanities 2012*, 2012.

[17] Morton Benson. The structure of the collocational dictionary. *International Journal of Lexicography*, 2:1–14, 1989.

[18] Morton Benson, Evelyn Benson, and Robert Ilson. *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. John Benjamins, Amsterdam, New York, 1986.

[19] Douglas Biber. Representativeness in Corpus Design. *Literary and Linguistic Computing*, 8(4):243–257, 1993.

[20] Chris Biemann, Uwe Quasthoff, Gerhard Heyer, and Florian Holz. ASV Toolbox – A Modular Collection of Language Exploration Tools. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC) 2008*, 2008.

[21] Chris Biemann and Martin Riedl. Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity. *Journal of Language Modelling*, 1(1), 2013.

[22] Burton H Bloom. Space/Time Trade-offs in Hash Coding with Allowable Errors. *Commun. ACM*, 13(7):422–426, July 1970.

[23] Jürgen Bortz. *Statistik für Human- und Sozialwissenschaftler*. Springer, Berlin etc., 7 edition, 2010.

[24] Thorsten Brants and Alex Franz. *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia, PA, 2006. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T13`.

[25] Andrei Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig. Syntactic Clustering of the Web. *Comput. Netw. ISDN Syst.*, 29(8-13):1157–1166, September 1997.

[26] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Raymie Stata, Andrew Tomkins, and Janet L Wiener. Graph Structure in the Web. In *In Proceedings of the 9th International World Wide Web conference on Computer Networks: The International Journal of Computer and Telecommunications Networking*, pages 309–320. North-Holland Publishing Co, 2000.

[27] Cecil H Brown, Eric W Holman, Søren Wichmann, and Viveka Velupillai. Automated classification of the worlds languages: A description of the method and preliminary results. *STUF-Language Typology and Universals*, 61(4):285–308, 2008.

[28] Soumen Chakrabarti, Byron Dom, and Piotr Indyk. Enhanced Hypertext Categorization Using Hyperlinks. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, SIGMOD '98, pages 307–318. ACM, 1998.

[29] Soumen Chakrabarti, Martin van den Berg, and Byron Dom. Focused Crawling: A New Approach to Topic-specific Web Resource Discovery. *Computer Networks*, 31:1623–1640, 1999.

[30] Junghoo Cho, Hector García-Molina, and Lawrence Page. Efficient Crawling through URL ordering. In *Proceedings of the 8th International World Wide Web Conference*, 1998.

[31] Massimiliano Ciamarita and Marco Baroni. Measuring Web-Corpus Randomness: A Progress Report. pages 127–158.

[32] Michael Cysouw. Quantitative methods in typology. In G. Altmann, R. Köhler, and R. Piotrowski, editors, *Quantitative linguistics: an international handbook*, pages 554–578. Mouton de Gruyter, 2005.

[33] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *OSDI'04: Proceedings of the 6th Symposium on Operating Systems Design and Implementation*. USENIX Association, 2004.

[34] Thomas Eckart, Uwe Quasthoff, and Dirk Goldhahn. The influence of corpus quality on statistical measurements on language resources. In *LREC*, pages 2318–2321, 2012.

[35] Thomas Eckart, Uwe Quasthoff, and Dirk Goldhahn. Language Statistics-Based Quality Assurance for Large Corpora. In *Proceedings of Asia Pacific Corpus Linguistics Conference 2012, Auckland, New Zealand*, 2012.

[36] Halvor Eifring and Rolf Theil. *Linguistics for students of Asian and African languages*. Institutt for østeuropeiske og orientalske studier, 2004.

[37] Stefan Evert. Corpora and collocations. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook*, chapter 58. Mouton de Gruyter, Berlin, New York, 2008.

[38] Stefan Evert. Google Web 1T5 n-grams made easy (but not for the computer). In *Proceedings of the 6th Web as Corpus Workshop (WAC-6)*, pages 32–40, Los Angeles, CA, 2010.

[39] Stefan Evert and Brigitte Krenn. Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 188–195, Toulouse, France, 2001.

[40] Stefan Evert and Brigitte Krenn. Using small random samples for the manual evaluation of statistical association measures. *Computer Speech and Language*, 19(4):450–466, 2005.

[41] Manaal Faruqui and Sebastian Padó. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany, 2010.

[42] Gertraud Fenk-Oczlon and August Fenk. The mean length of propositions is 7 plus minus 2 syllables – but the position of languages within this range is not accidental. In G. D'Ydevalle, editor, *Cognition, Information Processing, and Motivation*, pages 355–359. North Holland: Elsevier Science Publisher, 1985.

[43] Gertraud Fenk-Oczlon and August Fenk. Crosslinguistic correlations between size of syllables, number of cases, and adposition order. In G. Fenk-Oczlon and Ch. Winkler, editors, *Sprache und Natürlichkeit. Gedenkband für Willi Mayerthaler*, pages 75–86. Gunther Narr, Tübingen, 2005.

[44] David Ferrucci and Adam Lally. UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment. *Natural Language Engineering*, 10(3-4):327–348, 2004.

[45] J. R. Firth. A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pages 1–32. The Philological Society, Oxford, 1957.

[46] William Fletcher. Facilitating the compilation and Dissemination of Ad-hoc web corpora. In Guy Aston, Silvia Bernardini, and Dominic Stewart und Silvia Bernadini, editors, *Corpora and language learners*, pages 273–300. Benjamins, Amsterdam, 2004.

[47] Michael Flor. A fast and flexible architecture for very large word n-gram datasets. *Natural Language Engineering*, 19(1):61–93, 2013.

[48] Ian Foster. *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[49] Minas Gjoka, Maciej Kurant, Carter T Butts, and Athina Markopoulou. A Walk in Facebook: a Case Study of Unbiased Sampling of Facebook. In *Proceedings of IEEE INFOCOM 2010*, San Diego, 2011. IEEE.

[50] Yoav Goldberg and Jon Orwant. A dataset of syntactic-ngrams over time from a very large corpus of English books. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM 2013)*, Atlanta, GA, 2013. Data sets available from `http://commondatastorage.googleapis.com/books/syntactic-ngrams/index.html`.

[51] D. Goldhahn. *Quantitative Methoden in der Sprachtypologie: Nutzung korpusbasierter Statistiken (Doctoral dissertation)*. University of Leipzig, Leipzig, Germany, 2013.

[52] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[53] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Finding Language Universals: Multivariate Analysis of Language Statistics using the Leipzig Corpora Collection. In *Leuven Statistics Days 2012*, KU Leuven, 2012.

[54] Daniel Gomes and Mário J Silva. Characterizing a national community web. *ACM Trans. Internet Technol.*, 5(3):508–531, 2005.

[55] Joseph H Greenberg. Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113, 1963.

[56] Erla Hallsteinsdóttir, Thomas Eckart, Chris Biemann, Uwe Quasthoff, and Matthias Richter. Íslenskur orðasjóður - Building a Large Icelandic Corpus. In *Proceedings of NODALIDA-07*, Tartu, Estonia, 2007.

[57] Lushan Han, Abhay L Kashyap, Tim Finin, James Mayfield, and Johnathan Weese. UMBC_EBIQUITY-CORE: Semantic Textual Similarity Systems. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, 2013.

[58] Monika R Henzinger, Allan Heydon, Michael Mitzenmacher, and Marc Najork. On near-uniform URL sampling. In *Proceedings of the 9th International World Wide Web conference on Computer Networks: The International Journal of Computer and Telecommunications Networking*, pages 295–308. North-Holland Publishing Co, 2000.

[59] Gerhard Heyer and Uwe Quasthoff. Calculating communities by link analysis of URLs. In *Proceedings of the 4th international conference on Innovative Internet Community Systems*, IICS'04, pages 151–156, Berlin, Heidelberg, 2006. Springer-Verlag.

[60] Susan Hunston. Collection Strategies and Design Decisions. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. {A}n International Handbook*, pages 154–168. Walter de Gruyter, Berlin, 2008.

[61] Frank Keller and Mirella Lapata. Using the Web to Obtain Frequencies for Unseen Bigrams. *Computational Linguistics*, 29(3):459–484, 2003.

[62] Marc Kupietz. Near-duplicate detection in the ids corpora of written german. Technical Report kt-2006-01, Institut für Deutsche Sprache, Mannheim, 2005.

[63] Marc Kupietz, Cyril Belica, Holger Keibel, and Andreas Witt. The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, pages 1848–1854, Valletta, Malta, 2010. European Language Resources Association (ELRA).

[64] Maciej Kurant, Minas Gjoka, Carter T Butts, and Athina Markopoulou. Walking on a graph with a magnifying glass: stratified sampling via weighted random walks. In *Proceedings of the ACM SIGMETRICS joint international conference on Measurement and modeling of computer systems*, SIGMETRICS '11, pages 281–292, New York, NY, USA, 2011. ACM.

[65] Maciej Kurant, Athina Markopoulou, and Patrick Thiran. On the bias of BFS (Breadth First Search). In *International Teletraffic Congress (ITC 22)*, 2010.

[66] Yan Chi Lam. Managing the Google Web 1T 5-gram with relational database. *Journal of Education, Informatics, and Cybernetics*, 2(2), 2010.

[67] Thomas K Landauer, Danielle S McNamara, Simon Dennis, and Walter Kintsch, editors. *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, Mahwah, 2007.

[68] Geoffrey Leech. New resources or just better old ones? the holy grail of representativeness. pages 133–149. 2007.

[69] Anke Lüdeling, Stefan Evert, and Marco Baroni. Using the web for linguistic purposes. pages 7–24.

[70] Jayant Madhavan, Loredana Afanasiev, Lyublena Antova, and Alon Halevy. Harnessing the Deep Web: Present and Future. In *4th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2009.

[71] Arun S Maiya and Tanya Y Berger-Wolf. Benefits of bias: towards better characterization of network sampling. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 105–113, New York, NY, USA, 2011. ACM.

[72] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *An Introduction to Information Retrieval*. CUP, Cambridge, 2009.

[73] Filippo Menczer, Gautam Pant, and Padmini Srinivasan. Topical web crawlers: Evaluating adaptive algorithms. *ACM Trans. Internet Technol.*, 4(4):378–419, 2004.

[74] Gordon Mohr, Michael Stack, Igor Ranitovic, Dan Avery, and Michele Kimpton. Introduction to Heritrix, an Archival Quality Web Crawler. In *Proceedings of the 4th International Web Archiving Workshop (IWAW'04)*, 2004.

[75] Christopher Olston and Marc Najork. *Web Crawling*, volume 4(3) of *Foundations and Trends in Information Retrieval*. now Publishers, Hanover, MA, 2010.

[76] Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. *English Gigaword Fifth Edition*. Linguistic Data Consortium, Philadelphia, USA, 2011.

[77] Jan Pomikálek, Miloš Jakubíček, and Pavel Rychlý. Building a 70 Billion Word Corpus of English from ClueWeb. In *Proceedings of LREC 08*, pages 502–506, Istanbul, 2012.

[78] Michael O. Rabin. Fingerprinting by random polynomials. Technical Report TR-CSE-03-01, Center for Research in Computing Technology, Harvard University, Harvard, 1981.

[79] Carlos Ramisch, Paulo Schreiner, Marco Idiart, and Aline Villavicencio. An evaluation of methods for the extraction of multiword expressions. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE 2008)*, pages 50–53, Marrakech, Morocco, 2008.

[80] Patrick Riehmann, Henning Gruendl, Bernd Froehlich, Martin Potthast, Martin Trenkmann, and Benno Stein. The Netspeak WordGraph: Visualizing keywords in context. In *Proceedings of the 4th IEEE Pacific Visualization Symposium (PacificVis '11)*, pages 123–130. IEEE, 2011.

[81] Paat Rusmevichientong, David M Pennock, Steve Lawrence, and C Lee Giles. Methods for sampling pages uniformly from the World Wide Web. In *In AAAI Fall Symposium on Using Uncertainty Within Computation*, pages 121–128, 2001.

[82] Mejdl S. Safran, Abdullah Althagafi, and Dunren Che. Improving Relevance Prediction for Focused Web Crawlers. In *IEEE/ACIS 11th International Conference on Computer and Information Science (ICIS), 2012*, pages 161–166, 2012.

[83] Andreas Salomonsson, Svetoslav Marinov, and Pierre Nugues. Identification of entities in swedish. In *Proceedings ot The Fourth Swedish Language Technology Conference*, pages 62–63, Lund, 2012.

[84] Roland Schäfer, Adrien Barbaresi, and Felix Bildhauer. The Good, the Bad, and the Hazy: Design Decisions in Web Corpus Construction. In *Proceedings of WAC8*, 2013.

[85] Roland Schäfer and Felix Bildhauer. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, 2012. ELRA.

[86] Roland Schäfer and Felix Bildhauer. *Web Corpus Construction*. Synthesis Lectures on Human Language Technologies. Morgan and Claypool, San Francisco etc., 2013.

[87] M Ángeles Serrano, Ana Maguitman, Marián Boguñá, Santo Fortunato, and Alessandro Vespignani. Decoding the structure of the WWW: A comparative analysis of Web crawls. *ACM Trans. Web*, 1(2), 2007.

[88] Serge Sharoff. Creating general-purpose corpora using automated search engine queries. In *WaCky! Working papers on the Web as Corpus. Gedit*, 2006.

[89] Anil Kumar Singh and Harshit Surana. Can corpus based measures be used for comparative study of languages? In *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, pages 40–47. Association for Computational Linguistics, 2007.

[90] Jae Jung Song. *Linguistic Typology: Morphology and Syntax*. Harlow, Longman, 2001.

[91] Miroslav Spousta, Michal Marek, and Pavel Pecina. Victor: The web-page cleaning tool. pages 12–17.

[92] Padmini Srinivasan, Filippo Menczer, and Gautam Pant. A General Evaluation Framework for Topical Crawlers. *Inf. Retr.*, 8(3):417–447, 2005.

[93] Vít Suchomel and Jan Pomikálek. Effcient Web Crawling for Large Text Corpora. In Adam Kilgarriff and Serge Sharoff, editors, *Proceedings of the seventh Web as Corpus Workshop*, pages 40–44, 2012.

[94] Morris Swadesh. Lexico-statistic dating of prehistoric ethnic contacts: with special reference to north american indians and eskimos. *Proceedings of the American philosophical society*, 96(4):452–463, 1952.

[95] Morris Swadesh. Towards greater accuracy in lexicostatistic dating. *International journal of American linguistics*, 21(2):121–137, 1955.

[96] George Kingsley Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA, 1949.

Bryan Jurish, Kay-Michael Würzner

Word and Sentence Tokenization with Hidden Markov Models

We present a novel method ("WASTE") for the segmentation of text into tokens and sentences. Our approach makes use of a Hidden Markov Model for the detection of segment boundaries. Model parameters can be estimated from pre-segmented text which is widely available in the form of treebanks or aligned multi-lingual corpora. We formally define the WASTE boundary detection model and evaluate the system's performance on corpora from various languages as well as a small corpus of computer-mediated communication.

1 Introduction

Detecting token and sentence boundaries is an important preprocessing step in natural language processing applications since most of these operate either on the level of words (e.g. syllabification, morphological analysis) or sentences (e.g. part-of-speech tagging, parsing, machine translation). The primary challenges of the tokenization task stem from the ambiguity of certain characters in alphabetic and from the absence of explicitly marked word boundaries in symbolic writing systems. The following German example illustrates different uses of the dot character[1] to terminate an abbreviation, an ordinal number, or an entire sentence.

(1) Am 24.1.1806 feierte E. T. A. Hoffmann seinen 30. Geburtstag.
On 24/1/1806 celebrated E. T. A. Hoffmann his 30th birthday.
'On 24/1/1806, E. T. A. Hoffmann celebrated his 30$^{\text{th}}$ birthday.'

Recently, the advent of instant written communications over the internet and its increasing share in people's daily communication behavior has posed new challenges for existing approaches to language processing: computer-mediated communication (CMC) is characterized by a creative use of language and often substantial deviations from orthographic standards. For the task of text segmentation, this means dealing with unconventional uses of punctuation and letter-case, as well as genre-specific elements such as emoticons and inflective forms (e.g. "*grins*"). CMC sub-genres may differ significantly in their degree of deviation from orthographic norms. Moderated discussions from the university context are almost standard-compliant, while some passages of casual chat consist exclusively of metalinguistic items.

(2) schade, dass wien so weit weg ist d.h. ich hätt´´s sogar überlegt
shame, that Vienna so far away is i.e. I have$_{\text{SUBJ}}$-it even considered
'It's a shame that Vienna is so far away; I would even have considered it.'

[1] ".", ASCII/Unicode codepoint 0x2E, also known as "full stop" or "period".

In addition, CMC exhibits many structural similarities to spoken language. It is in dialogue form, contains anacolutha and self-corrections, and is discontinuous in the sense that utterances may be interrupted and continued at some later point in the conversation. Altogether, these phenomena complicate automatic text segmentation considerably.

In this paper, we present a novel method for the segmentation of text into tokens and sentences. Our system uses a Hidden Markov Model (HMM) to estimate the placement of segment boundaries at runtime, and we refer to it in the sequel as "WASTE".[2] The remainder of this work is organized as follows: first, we describe the tasks of tokenization and EOS detection and summarize some relevant previous work on these topics. Section 2 contains a description of our approach, including a formal definition of the underlying HMM. In Section 3, we present an empirical evaluation of the WASTE system with respect to conventional corpora from five different European languages as well as a small corpus of CMC text, comparing results to those achieved by a state-of-the-art tokenizer.

1.1 Task Description

Tokenization and EOS detection are often treated as separate text processing stages. First, the input is segmented into atomic units or word-like tokens. Often, this segmentation occurs on whitespace, but punctuation must be considered as well, which is often not introduced by whitespace, as for example in the case of the commas in Examples (1) and (2). Moreover, there are tokens which may contain internal whitespace, such as cardinal numbers in German, in which a single space character may be used as thousands separator. The concept of a token is vague and may even depend on the client application: *New York* might be considered a single token for purposes of named entity recognition, but two tokens for purposes of syntactic parsing.

In the second stage, sentence boundaries are marked within the sequence of word-like tokens. There is a set of punctuation characters which typically introduce sentence boundaries: the "usual suspects" for sentence-final punctuation characters include the question mark ("*?*"), exclamation point ("*!*"), ellipsis ("…"), colon ("*:*"), semicolon ("*;*"), and of course the full stop ("."). Unfortunately, any of these items can mislead a simple rule-based sentence splitting procedure. Apart from the different uses of the dot character illustrated in Ex. (1), all of these items can occur sentence-internally (e.g. in direct quotations like "*'Stop!' he shouted.*"), or even token-internally in the case of complex tokens such as URLs. Another major difficulty for EOS detection arises from sentence boundaries which are not explicitly marked by punctuation, as e.g. for newspaper headlines.

[2] An acronym for "Word and Sentence Token Estimator", and occasionally for "Weird And Strange Tokenization Errors" as well.

1.2 Existing Approaches

Many different approaches to tokenization and EOS detection have been proposed in the literature. He and Kayaalp (2006) give an interesting overview of the characteristics and performance of 13 tokenizers on biomedical text containing challenging tokens like DNA sequences, arithmetical expressions, URLs, and abbreviations. Their evaluation focuses on the treatment of particular phenomena rather than general scalar quantities such as precision and recall, so a clear "winner" cannot be determined. While He and Kayaalp (2006) focus on existing and freely available tokenizer implementations, we briefly present here the theoretical characteristics of some related approaches. All of these works have in common that they focus on the disambiguation of the dot character as the most likely source of difficulties for the text segmentation task.

Modes of evaluation differ for the various approaches, which makes direct comparisons difficult. Results are usually reported in terms of error rate or accuracy, often focusing on the performance of the disambiguation of the period. In this context, Palmer and Hearst (1997) define a lower bound for EOS detection as "the percentage of possible sentence-ending punctuation marks [...] that indeed denote sentence boundaries." The Brown Corpus (Francis and Kucera, 1982) and the Wall Street Journal (WSJ) subset of the Penn Treebank (Marcus et al., 1993) are the most commonly used test corpora, relying on the assumption that the manual assignment of a part-of-speech (PoS) tag to a token requires prior manual segmentation of the text.

Riley (1989) trains a decision tree with features including word length, letter-case and probability-at-EOS on pre-segmented text. He uses the 25 million word AP News text database for training and reports 99.8% accuracy for the task of identifying sentence boundaries introduced by a full stop in the Brown corpus. Grefenstette and Tapanainen (1994) use a set of regular rules and some lexica to detect occurrences of the period which are not EOS markers. They exhibit rules for the treatment of numbers and abbreviations and report a rate of 99.07% correctly recognized sentence boundaries for their rule-based system on the Brown corpus. Palmer and Hearst (1997) present a system which makes use of the possible PoS-tags of the words surrounding potential EOS markers to assist in the disambiguation task. Two different kinds of statistical models (neural networks and decision trees) are trained from manually PoS-tagged text and evaluated on the WSJ corpus. The lowest reported error rates are 1.5% for a neural network and 1.0% for a decision tree. Similar results are achieved for French and German.

Mikheev (2000) extends the approach of Palmer and Hearst (1997) by incorporating the task of EOS detection into the process of PoS tagging, and thereby allowing the *disambiguated* PoS tags of words in the immediate vicinity of a potential sentence boundary to influence decisions about boundary placement. He reports error rates of 0.2% and 0.31% for EOS detection on the Brown and the WSJ corpus, respectively. Mikheev's treatment also gives the related task of abbreviation detection much more attention than previous work had. Making use of the internal structure of abbreviation candidates, together with the surroundings of clear abbreviations and a list of frequent

abbreviations, error rates of 1.2% and 0.8% are reported for Brown and WSJ corpus, respectively.

While the aforementioned techniques use pre-segmented or even pre-tagged text for training model parameters, Schmid (2000) proposes an approach which can use raw, unsegmented text for training. He uses heuristically identified "unambiguous" instances of abbreviations and ordinal numbers to estimate probabilities for the disambiguation of the dot character, reporting an EOS detection accuracy of 99.79%. More recently, Kiss and Strunk (2006) presented another unsupervised approach to the tokenization problem: the Punkt system. Its underlying assumption is that abbreviations may be regarded as collocations between the abbreviated material and the following dot character. Significant collocations are detected within a training stage using log-likelihood ratios. While the detection of abbreviations through the collocation assumption involves type-wise decisions, a number of heuristics involving its immediate surroundings may cause an abbreviation candidate to be reclassified on the token level. Similar techniques are applied to possible ellipses and ordinal numbers, and evaluation is carried out for a number of different languages. Results are reported for both EOS- and abbreviation-detection in terms of precision, recall, error rate, and unweighted F score. Results for EOS detection range from $\mathrm{F} = 98.83\%$ for Estonian to $\mathrm{F} = 99.81\%$ for German, with a mean of $\mathrm{F} = 99.38\%$ over all tested languages; and for abbreviation detection from $\mathrm{F} = 77.80\%$ for Swedish to $\mathrm{F} = 98.68\%$ for English with a mean of $\mathrm{F} = 90.93\%$ over all languages.

A sentence- and token-splitting framework closely related to the current approach is presented by Tomanek et al. (2007), tailored to the domain of biomedical text. Such text contains many complex tokens such as chemical terms, protein names, or chromosome locations which make it difficult to tokenize. Tomanek et al. (2007) propose a supervised approach using a pair of conditional random field classifiers to disambiguate sentence- and token-boundaries in whitespace-separated text. In contrast to the standard approach, EOS detection takes place first, followed by token-boundary detection. The classifiers are trained on pre-segmented data, and employ both lexical and contextual features such as item text, item length, letter-case, and whitespace adjacency. Accuracies of 99.8% and 96.7% are reported for the tasks of sentence- and token-splitting, respectively.

2 The WASTE Tokenization System

In this section, we present our approach to token- and sentence-boundary detection using a Hidden Markov Model to simultaneously detect both word and sentence boundaries in a stream of candidate word-like segments returned by a low-level scanner. Section 2.1 briefly describes some requirements on the low-level scanner, while Section 2.2 is dedicated to the formal definition of the HMM itself.

2.1 Scanner

The scanner we employed in the current experiments uses Unicode[3] character classes in a simple rule-based framework to split raw corpus text on whitespace and punctuation. The resulting pre-tokenization is "prolix" in the sense that many scan-segment boundaries do not in fact correspond to actual word or sentence boundaries. In the current framework, only scan-segment boundaries can be promoted to full-fledged token or sentence boundaries, so the scanner output must contain at least these.[4] In particular, unlike most other tokenization frameworks, the scanner also returns whitespace-only pseudo-tokens, since the presence or absence of whitespace can constitute useful information regarding the proper placement of token and sentence boundaries. In Ex. (3) for instance, whitespace is crucial for the correct classification of the apostrophes.

(3) Consider Bridges' poem's "And peace was 'twixt them."

2.2 HMM Boundary Detector

Given a prolix segmentation as returned by the scanner, the task of tokenization can be reduced to one of classification: we must determine for each scanner segment whether or not it is a word-initial segment, and if so, whether or not it is also a sentence-initial segment. To accomplish this, we make use of a Hidden Markov Model which encodes the boundary classes as hidden state components, in a manner similar to that employed by HMM-based chunkers (Church, 1988; Skut and Brants, 1998). In order to minimize the number of model parameters and thus ameliorate sparse data problems, our framework maps each incoming scanner segment to a small set of salient properties such as word length and typographical class in terms of which the underlying language model is then defined.

2.2.1 Segment Features

Formally, our model is defined in terms of a finite set of *segment features*. In the experiments described here, we use the observable features CLASS, CASE, LENGTH, STOP, ABBR, and BLANKS together with the hidden features BOW, BOS, and EOS to specify the language model. We treat each feature f as a function from candidate tokens (scanner segments) to a characteristic finite set of possible values $\mathrm{rng}(f)$. The individual features and their possible values are described in more detail below, and summarized in Table 1.

- [CLASS] represents the typographical class of the segment. Possible values are given in Table 2.

[3] Unicode Consortium (2012), `http://www.unicode.org`

[4] In terms of the evaluation measures described in Section 3.2, the pre-tokenization returned by the scanner places a strict upper bound on the recall of the WASTE system as a whole, while its precision can only be improved by the subsequent procedure.

- [CASE] represents the letter-case of the segment. Possible values are '`cap`' for segments in all-capitals, '`up`' for segments with an initial capital letter, or '`lo`' for all other segments.
- [LENGTH] represents the length of the segment. Possible values are '1' for single-character segments, '≤3' for segments of length 2 or 3, '≤5' for segments of length 4 or 5, or '>5' for longer segments.
- [STOP] contains the lower-cased text of the segment just in case the segment is a known *stopword*; i.e. only in conjunction with [CLASS: `stop`]. We used the appropriate language-specific stopwords distributed with the Python NLTK package whenever available, and otherwise an empty stopword list.
- [BLANKS] is a binary feature indicating whether or not the segment is separated from its predecessor by whitespace.
- [ABBR] is a binary feature indicating whether or not the segment represents a known abbreviation, as determined by membership in a user-specified language-specific abbreviation lexicon. Since no abbreviation lexica were used for the current experiments, this feature was vacuous and will be omitted henceforth.[5]
- [BOW] is a hidden binary feature indicating whether or not the segment is to be considered token-initial.
- [BOS] is a hidden binary feature indicating whether or not the segment is to be considered sentence-initial.
- [EOS] is a hidden binary feature indicating whether or not the segment is to be considered sentence-final. Sentence boundaries are only predicted by the final system if a [+EOS] segment is immediately followed by a [+BOS] segment.

Among these, the feature STOP is *context-independent* in the sense that we do not allow it to contribute to the boundary detection HMM's transition probabilities. We call all other features *context-dependent* or *contextual*. An example of how the features described above can be used to define sentence- and token-level segmentations is given in Figure 1.

2.2.2 Language Model

Formally, let $F_{\text{surf}} = \{\text{CLASS}, \text{CASE}, \text{LENGTH}, \text{STOP}, \text{BLANKS}\}$ represent the set of *surface features*, let $F_{\text{noctx}} = \{\text{STOP}\}$ represent the set of *context-independent features*, and let $F_{\text{hide}} = \{\text{BOW}, \text{BOS}, \text{EOS}\}$ represent the set of *hidden features*, and for any finite set of features $F = \{f_1, f_2, \ldots, f_n\}$ over objects from a set S, let $\bigwedge F$ be a composite feature

[5] Preliminary experiments showed no significant advantage for models using non-empty abbreviation lexica on any of the corpora we tested. Nonetheless, the results reported by Grefenstette and Tapanainen (1994); Mikheev (2000); and Tomanek et al. (2007) suggest that in some cases at least, reference to such a lexicon can be useful for the tokenization task.

f	rng(f)	**Hidden?**	**Description**
CLASS	{`stop`, `alpha`, `num`, ...}	no	typographical class
CASE	{`lo`, `up`, `cap`}	no	letter-case
LENGTH	$\{1, \leq 3, \leq 5, > 5\}$	no	segment length
STOP	finite set $\mathcal{S}_{\mathcal{L}}$	no	stopword text
BLANKS	$\{+, -\}$	no	leading whitespace?
ABBR	$\{+, -\}$	no	known abbreviation?
BOW	$\{+, -\}$	yes	beginning-of-word?
BOS	$\{+, -\}$	yes	beginning-of-sentence?
EOS	$\{+, -\}$	yes	end-of-sentence?

Table 1: Features used by the WASTE tokenizer model.

Class	**Description**
`stop`	language-specific stopwords
`roman`	segments which may represent roman numerals
`alpha`	segments containing only alphabetic characters
`num`	segments containing only numeric characters
`$.`	period
`$,`	comma
`$:`	colon
`$;`	semicolon
`$?`	sentence-final punctuation (‘?’, ‘!’, or ‘...’)
`$(`	left bracket (‘(’, ‘[’, or ‘{’)
`$)`	right bracket (‘)’, ‘]’, or ‘}’)
`$-`	minus, hyphen, en- or em-dash
`$+`	plus
`$/`	backward or forward slash
`$"`	double quotation marks
`$'`	single quotation mark or apostrophe
`$~`	other punctuation marks, superscripts, vulgar fractions, *etc.*
`other`	all remaining segments

Table 2: Typographical classes used by the WASTE tokenizer model.

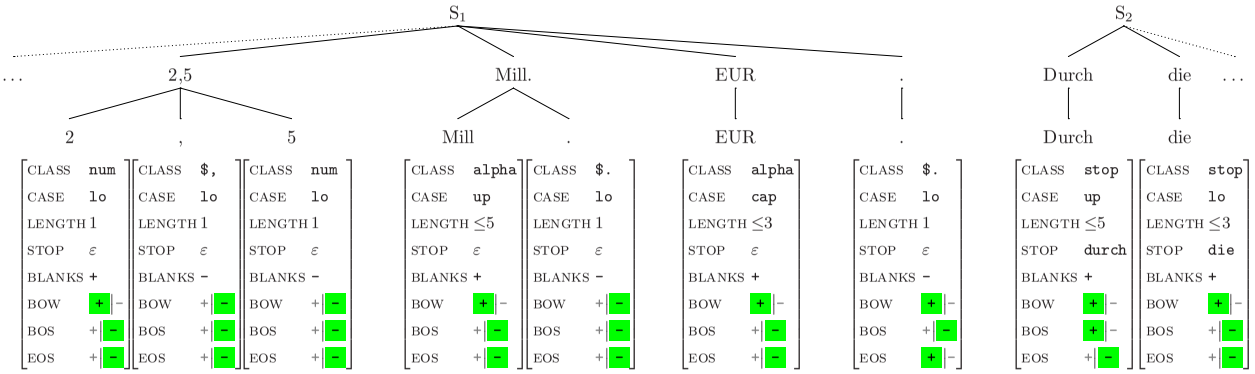


Figure 1: Different levels of the WASTE representation for the German text fragment "... 2,5 Mill. Eur. Durch die ..." depicted as a tree. Nodes of depth 0, 1, and 2 correspond to the levels of sentences, tokens, and segments, respectively. The feature values used by the HMM boundary detector are given below the corresponding segments. Boldface values of "hidden" features on a green background indicate the correct assignment, while gray values on a white background indicate possible but incorrect assignments.

function representing the conjunction over all individual features in F as an n-tuple:

$$\begin{aligned} \bigwedge F &: S \to \mathrm{rng}(f_1) \times \mathrm{rng}(f_2) \times \ldots \times \mathrm{rng}(f_n) \\ &: x \mapsto \langle f_1(x), f_2(x), \ldots, f_n(x) \rangle \end{aligned} \tag{1}$$

Then, the boundary detection HMM can be defined in the usual way (Rabiner, 1989; Manning and Schütze, 1999) as the 5-tuple $D = \langle \mathcal{Q}, \mathcal{O}, \Pi, A, B \rangle$, where:

1. $\mathcal{Q} = \mathrm{rng}\left(\bigwedge (F_{\mathrm{hide}} \cup F_{\mathrm{surf}} \backslash F_{\mathrm{noctx}})\right)$ is a finite set of model *states*, where each state $q \in \mathcal{Q}$ is represented by a 7-tuple of values for the contextual features CLASS, CASE, LENGTH, BLANKS, BOW, BOS, and EOS;
2. $\mathcal{O} = \mathrm{rng}\left(\bigwedge F_{\mathrm{surf}}\right)$ is a finite set of possible *observations*, where each observation is represented by a 5-tuple of values for the surface features CLASS, CASE, LENGTH, BLANKS, and STOP;
3. $\Pi : \mathcal{Q} \to [0,1] : q \mapsto p(Q_1 = q)$ is a probability distribution over $\mathcal{Q}$ representing the model's *initial state probabilities*;
4. $A : \mathcal{Q}^k \to [0,1] : \langle q_1, \ldots, q_k \rangle \mapsto p(Q_i = q_k | Q_{i-k+1} = q_1, \ldots, Q_{i-1} = q_{k-1})$ is a conditional probability distribution over state k-grams representing the model's *state transition probabilities*; and
5. $B : \mathcal{Q} \times \mathcal{O} \to [0,1] : \langle q, o \rangle \mapsto p(O = o | Q = q)$ is a probability distribution over observations conditioned on states representing the model's *emission probabilities*.

Using the shorthand notation w_i^{i+j} for the string $w_i w_{i+1} \ldots w_{i+j}$, and writing $f_{\mathcal{O}}(w)$ for the observable features $\left[\bigwedge F_{\mathrm{surf}}\right](w)$ of a given segment w, the model D computes the probability of a segment sequence w_1^n as the sum of path probabilities over all possible generating state sequences:

$$p(W = w_1^n) \quad = \quad \sum_{q_1^n \in \mathcal{Q}^n} p(W = w_1^n, Q = q_1^n) \tag{2}$$

Assuming suitable boundary handling for negative indices, joint path probabilities themselves are computed as:

$$p(W = w_1^n, Q = q_1^n) \quad = \quad \prod_{i=1}^{n} p(q_i | q_{i-k+1}^{i-1}) p(w_i | q_i) \tag{3}$$

Underlying these equations are the following assumptions:

$$p(q_i | q_1^{i-1}, w_1^{i-1}) \quad = \quad p(q_i | q_{i-k+1}^{i-1}) \tag{4}$$

$$p(w_i | q_1^{i}, w_1^{i-1}) \quad = \quad p(w_i | q_i) \;=\; p(O_i = f_{\mathcal{O}}(w_i) | Q_i = q_i) \tag{5}$$

Equation (4) asserts that state transition probabilities depend on at most the preceding $k-1$ states and thus on the contextual features of at most the preceding $k-1$ segments.

Equation (5) asserts the independence of a segment's surface features from all but the model's current state, formally expressing the context-independence of F_{noctx}. In the experiments described below, we used scan-segment trigrams ($k = 3$) extracted from a training corpus to define language-specific boundary detection models in a supervised manner. To account for unseen trigrams, the empirical distributions were smoothed by linear interpolation of uni-, bi-, and trigrams (Jelinek and Mercer, 1980), using the method described by Brants (2000) to estimate the interpolation coefficients.

2.2.3 Runtime Boundary Placement

Having defined the disambiguator model D, it can be used to predict the "best" possible boundary placement for an input sequence of scanner segments W by application of the well-known *Viterbi algorithm* (Viterbi, 1967). Formally, the Viterbi algorithm computes the state path with maximal probability for the observed input sequence:

$$\textsc{Viterbi}(W, D) = \underset{\langle q_1,\ldots,q_n\rangle \in \mathcal{Q}^n}{\arg\max}\ p(q_1, \ldots, q_n, W|D) \tag{6}$$

If $\langle q_1, \ldots, q_n\rangle = \textsc{Viterbi}(W, D)$ is the optimal state sequence returned by the Viterbi algorithm for the input sequence W, the final segmentation into word-like tokens is defined by placing a word boundary immediately preceding all and only those segments w_i with $i = 1$ or $q_i[\textsc{bow}] = +$. Similarly, sentence boundaries are placed before all and only those segments w_i with $i = 1$ or $q_i[\textsc{bow}] = q_i[\textsc{bos}] = q_{i-1}[\textsc{eos}] = +$. Informally, this means that every input sequence will begin a new word and a new sentence, every sentence boundary must also be a word boundary, and a high-level agreement heuristic is enforced between adjacent EOS and BOS features.[6] Since all surface feature values are uniquely determined by observed segment and only the hidden segment features BOW, BOS, and EOS are ambiguous, only those states q_i need to be considered for a segment w_i which agree with respect to surface features, which represents a considerable efficiency gain, since the Viterbi algorithm's running time grows exponentially with the number of states considered per observation.

3 Experiments

In this section, we present four experiments designed to test the efficacy of the WASTE boundary detection framework described above. After describing the corpora and software used for the experiments in Section 3.1 and formally defining our evaluation criteria in Section 3.2, we first compare the performance of our approach to that of the Punkt system introduced by Kiss and Strunk (2006) on corpora from five different European languages in Section 3.3. In Section 3.4 we investigate the effect of training corpus size on HMM-based boundary detection, while Section 3.5 deals with the effect

[6] Although either of the EOS or BOS features on its own is sufficient to define a boundary placement model, preliminary experiments showed substantially improved precision for the model presented above using both EOS and BOS features together with an externally enforced agreement heuristic.

Corpus	Sentences	Words	Segments
cz	21,656	487,767	495,832
de	50,468	887,369	936,785
en	49,208	1,173,766	1,294,344
fr	21,562	629,810	679,375
it	2,860	75,329	77,687
chat	11,416	95,102	105,297

Table 3: Corpora used for tokenizer training and evaluation.

of some common typographical conventions. Finally, Section 3.6 describes some variants of the basic WASTE model and their respective performance with respect to a small corpus of computer-mediated communication.

3.1 Materials

Corpora We used several freely available corpora from different languages for training and testing. Since they were used to provide ground-truth boundary placements for evaluation purposes, we required that all corpora provide both word- and sentence-level segmentation. For English (en), we used the Wall Street Journal texts from the Penn Treebank (Marcus et al., 1993) as distributed with the Prague Czech-English Dependency Treebank (Cuřín et al., 2004), while the corresponding Czech translations served as the test corpus for Czech (cz). The TIGER treebank (de; Brants et al., 2002) was used for our experiments on German.[7] French data were taken from the 'French Treebank' (fr) described by Abeillé et al. (2003), which also contains annotations for multi-word expressions which we split into their components.[8] For Italian, we chose the Turin University Treebank (it; Bosco et al., 2000). To evaluate the performance of our approach on non-standard orthography, we used a subset of the Dortmund Chat Corpus (chat; Beißwenger and Storrer, 2008). Since pre-segmented data are not available for this corpus, we extracted a sample of the corpus containing chat logs from different scenarios (media, university context, casual chats) and manually inserted token and sentence boundaries. To support the detection of (self-)interrupted sentences, we grouped each user's posts and ordered them according to their respective timestamps. Table 3 summarizes some basic properties of the corpora used for training and evaluation.

[7] Unfortunately, corresponding (non-tokenized) raw text is not included in the TIGER distribution. We therefore semi-automatically de-tokenized the sentences in this corpus: token boundaries were replaced by whitespace except for punctuation-specific heuristics, e.g. no space was inserted before commas or dots, or after opening brackets. Contentious cases such as date expressions, truncations or hyphenated compounds were manually checked and corrected if necessary.

[8] Multi-word tokens consisting only of numeric and punctuation subsegments (e.g. "*16/9*" or "*3,2*") and hyphenated compounds ("*Dassault-électronique*") were not split into their component segments, but rather treated as single tokens.

Software The WASTE text segmentation system described in Sec. 2 was implemented in C++ and Perl. The initial prolix segmentation of the input stream into candidate segments was performed by a traditional `lex`-like scanner generated from a set of 49 hand-written regular expressions by the scanner-generator `RE2C` (Bumbulis and Cowan, 1993).[9] HMM training, smoothing, and runtime Viterbi decoding were performed by the `moot` part-of-speech tagging suite (Jurish, 2003). Viterbi decoding was executed using the default beam pruning coefficient of one thousand in `moot`'s "streaming mode," flushing the accumulated hypothesis space whenever an unambiguous token was encountered in order to minimize memory requirements without unduly endangering the algorithm's correctness (Lowerre, 1976; Kempe, 1997). To provide a direct comparison with the Punkt system beyond that given by Kiss and Strunk (2006), we used the `nltk.tokenize.punkt` module distributed with the Python NLTK package. Boundary placements were evaluated with the help of GNU `diff` (Hunt and McIlroy, 1976; MacKenzie et al., 2002) operating on one-word-per-line "vertical" files.

Cross-Validation Except where otherwise noted, WASTE HMM tokenizers were tested by 10-fold cross-validation to protect against model over-fitting: each test corpus C was partitioned on true sentence boundaries into 10 strictly disjoint subcorpora $\{c_i\}_{1\leq i\leq 10}$ of approximately equal size, and for each evaluation subcorpus c_i, an HMM trained on the remaining subcorpora $\bigcup_{j\neq i} c_j$ was used to predict boundary placements in c_i. Finally, the automatically annotated evaluation subcorpora were concatenated and evaluated with respect to the original test corpus C. Since the Punkt system was designed to be trained in an unsupervised fashion from raw untokenized text, no cross-validation was used in the evaluation of Punkt tokenizers.

3.2 Evaluation Measures

The tokenization method described above was evaluated with respect to the ground-truth test corpora in terms of *precision*, *recall*, and the harmonic precision-recall average F, as well as an intuitive scalar error rate. Formally, for a given corpus and a set B_{relevant} of boundaries (e.g. token- or sentence-boundaries) within that corpus, let $B_{\text{retrieved}}$ be the set of boundaries of the same type predicted by the tokenization procedure to be evaluated. Tokenizer precision (pr) and recall (rc) can then be defined as:

$$\text{pr} = \frac{\text{tp}}{\text{tp}+\text{fp}} = \frac{|B_{\text{relevant}} \cap B_{\text{retrieved}}|}{|B_{\text{retrieved}}|} \tag{7}$$

$$\text{rc} = \frac{\text{tp}}{\text{tp}+\text{fn}} = \frac{|B_{\text{relevant}} \cap B_{\text{retrieved}}|}{|B_{\text{relevant}}|} \tag{8}$$

where following the usual conventions $\text{tp} = |B_{\text{relevant}} \cap B_{\text{retrieved}}|$ represents the number of *true positive* boundaries predicted by the tokenizer, $\text{fp} = |B_{\text{retrieved}} \setminus B_{\text{relevant}}|$ represents

[9] Of these, 31 were dedicated to the recognition of special complex token types such as URLs and e-mail addresses.

the number of *false positives*, and $\text{fn} = |B_{\text{relevant}} \backslash B_{\text{retrieved}}|$ represents the number of *false negatives*.

Precision thus reflects the likelihood of a true boundary given its prediction by the tokenizer, while recall reflects the likelihood that that a boundary will in fact be predicted given its presence in the corpus. In addition to these measures, it is often useful to refer to a single scalar value on the basis of which to compare tokenization quality. The unweighted harmonic precision-recall average F (van Rijsbergen, 1979) is often used for this purpose:

$$\text{F} = \frac{2 \times \text{pr} \times \text{rc}}{\text{pr} + \text{rc}} \tag{9}$$

In the sequel, we will also report tokenization *error rates* (Err) as the ratio of errors to all predicted or true boundaries:[10]

$$\text{Err} = \frac{\text{fp} + \text{fn}}{\text{tp} + \text{fp} + \text{fn}} = \frac{|B_{\text{relevant}} \triangle B_{\text{retrieved}}|}{|B_{\text{relevant}} \cup B_{\text{retrieved}}|} \tag{10}$$

To allow direct comparison with the results reported for the Punkt system by Kiss and Strunk (2006), we will also employ the scalar measure used there, which we refer to here as the "Kiss-Strunk error rate" (Err_{KS}):

$$\text{Err}_{\text{KS}} = \frac{\text{fp} + \text{fn}}{\text{number of all candidates}} \tag{11}$$

Since the Kiss-Strunk error rate only applies to sentence boundaries indicated by a preceding full stop, we assume that the "number of candidates" referred to in the denominator is simply the number of dot-final tokens in the corpus.

3.3 Experiment 1: WASTE *versus* Punkt

We compared the performance of the HMM-based WASTE tokenization architecture described in Sec. 2 to that of the Punkt tokenizer described by Kiss and Strunk (2006) on each of the five conventional corpora from Table 3, evaluating the tokenizers with respect to both sentence- and word-boundary prediction.

3.3.1 Sentence Boundaries

Since Punkt is first and foremost a disambiguator for sentence boundaries indicated by a preceding full stop, we will first consider the models' performance on these, as given in Table 4. For all tested languages, the Punkt system achieved a higher recall on dot-terminated sentence boundaries, representing an average relative recall error reduction rate of 54.3% with respect to the WASTE tokenizer. WASTE exhibited greater precision however, providing an average relative precision error reduction rate of 73.9% with respect to Punkt. The HMM-based WASTE technique incurred the fewest errors

[10] $A \triangle B$ represents the *symmetric difference* between sets A and B: $A \triangle B = (A \backslash B) \cup (B \backslash A)$.

Corpus	Method	tp	fp	fn	pr%	rc%	F%	$\mathrm{Err}_{\mathrm{KS}}$%
cz	WASTE	20,808	158	143	**99.25**	99.32	**99.28**	**1.20**
	PUNKT	19,892	1,019	46	95.13	**99.77**	97.39	4.52
de	WASTE	40,887	128	420	**99.69**	98.98	**99.33**	**1.23**
	PUNKT	41,068	399	292	99.04	**99.29**	99.17	1.56
en	WASTE	47,309	151	109	**99.68**	99.77	**99.73**	**0.40**
	PUNKT	46,942	632	77	98.67	**99.84**	99.25	1.09
fr	WASTE	19,944	66	203	**99.67**	98.99	99.33	1.24
	PUNKT	19,984	128	112	99.36	**99.44**	**99.40**	**1.10**
it	WASTE	2,437	15	169	**99.39**	93.51	**96.36**	**6.02**
	PUNKT	2,595	233	0	91.76	**100.00**	95.70	7.64

Table 4: Performance on dot-terminated sentences, evaluation following Kiss and Strunk (2006).

Corpus	Method	tp	fp	fn	pr%	rc%	F%	Err%
cz	WASTE	21,230	242	425	**98.87**	**98.04**	**98.45**	**3.05**
	PUNKT	20,126	1,100	1,529	94.82	92.94	93.87	11.55
de	WASTE	47,453	965	3,014	98.01	**94.03**	**95.98**	**7.74**
	PUNKT	41,907	497	8,560	**98.83**	83.04	90.25	17.77
en	WASTE	48,162	452	1,045	**99.07**	**97.88**	**98.47**	**3.01**
	PUNKT	47,431	724	1,776	98.50	96.39	97.43	5.01
fr	WASTE	20,965	143	596	**99.32**	**97.24**	**98.27**	**3.40**
	PUNKT	20,744	230	817	98.90	96.21	97.54	4.80
it	WASTE	2,658	16	201	**99.40**	**92.97**	**96.08**	**7.55**
	PUNKT	2,635	240	223	91.65	92.20	91.92	14.95

Table 5: Overall performance on sentence boundary detection.

overall, and those errors which it did make were more uniformly distributed between false positives and false negatives, leading to higher F values and lower Kiss-Strunk error rates for all tested corpora except French.

It is worth noting that the error rates we observed for the Punkt system as reported in Table 4 differ from those reported in Kiss and Strunk (2006). In most cases, these differences can be attributed to the use of different corpora. The most directly comparable values are assumedly those for English, which in both cases were computed based on samples from the Wall Street Journal corpus: here, we observed a similar error rate (1.10%) to that reported by Kiss and Strunk (1.65%), although Kiss and Strunk observed fewer false positives than we did. These differences may stem in part from incompatible criteria regarding precisely which dots can legitimately be regarded as sentence-terminal, since Kiss and Strunk provide no formal definition of what exactly constitutes a "candidate" for the computation of Eq. (11). In particular, it is unclear how sentence-terminating full stops which are not themselves in sentence-final position – as often occurring in direct quotations (e.g. "*He said 'stop.'*") – are to be treated.

Despite its excellent recall for dot-terminated sentences, the Punkt system's performance dropped dramatically when considering all sentence boundaries (Table 5), including those terminated e.g. by question marks, exclamation points, colons, semi-

Corpus	Method	tp	fp	fn	pr%	rc%	F%	Err%
cz	WASTE	487,560	94	206	**99.98**	**99.96**	**99.97**	**0.06**
	PUNKT	463,774	10,445	23,993	97.80	95.08	96.42	6.91
de	WASTE	886,937	533	431	**99.94**	**99.95**	**99.95**	**0.11**
	PUNKT	882,161	9,082	5,208	98.98	99.41	99.20	1.59
en	WASTE	1,164,020	9,228	9,745	**99.21**	**99.17**	**99.19**	**1.60**
	PUNKT	1,154,485	22,311	19,281	98.10	98.36	98.23	3.48
fr	WASTE	625,554	2,587	4,255	**99.59**	**99.32**	**99.46**	**1.08**
	PUNKT	589,988	61,236	39,822	90.60	93.68	92.11	14.62
it	WASTE	74,532	132	796	**99.82**	**98.94**	**99.38**	**1.23**
	PUNKT	71,028	3,514	4,302	95.29	94.29	94.78	9.91

Table 6: Overall performance on word-boundary detection.

colons, or non-punctuation characters. Our approach outperformed Punkt on global sentence boundary detection for all languages and evaluation modes except precision on the German TIGER corpus (98.01% for the WASTE tokenizer vs. 98.83% for Punkt). Overall, WASTE incurred only about half as many sentence boundary detection errors as Punkt ($\mu = 49.3\%, \sigma = 15.3\%$). This is relatively unsurprising, since Punkt's rule-based scanner stage is responsible for detecting any sentence boundary not introduced by a full stop, while WASTE can make use of token context even in the absence of a dot character.

3.3.2 Word Boundaries

The differences between our approach and that of Kiss and Strunk become even more apparent for word boundaries. As the data from Table 6 show, WASTE substantially outperformed Punkt on word boundary detection for all languages and all evaluation modes, reducing the number of word-boundary errors by over 85% on average ($\mu = 85.6\%, \sigma = 16.0\%$). Once again, this behavior can be explained by Punkt's reliance on strict rule-based heuristics to predict all token boundaries except those involving a dot on the one hand, and WASTE's deferral of all final decisions to the model-dependent runtime decoding stage on the other. In this manner, our approach is able to adequately account for both "prolix" target tokenizations such as that given by the Czech corpus – which represents e.g. adjacent single quote characters (") as separate tokens – as well as "terse" tokenizations such as that of the English corpus, which conflates e.g. genitive apostrophe-s markers (*'s*) into single tokens. While it is almost certainly true that better results for Punkt than those presented in Table 6 could be attained by using additional language-specific heuristics for tokenization, we consider it to be a major advantage of our approach that it does not require such fine-tuning, but rather is able to learn the "correct" word-level tokenization from appropriate training data.

Although the Punkt system was not intended to be an all-purpose word-boundary detector, it was specifically designed to make reliable decisions regarding the status of word boundaries involving the dot character, in particular abbreviations (e.g. "*etc.*",

Corpus	Method	tp	fp	fn	pr%	rc%	F%	Err%
cz	WASTE	0	0	0	–	–	–	–
	PUNKT	0	2,658	0	0.00	–	–	100.00
de	WASTE	3,048	58	101	98.13	**96.79**	**97.46**	**4.96**
	PUNKT	2,737	23	412	**99.17**	86.92	92.64	13.71
en	WASTE	16,552	521	145	**96.95**	**99.13**	**98.03**	**3.87**
	PUNKT	15,819	1,142	878	93.27	94.74	94.00	11.32
fr	WASTE	1,344	30	68	**97.82**	**95.18**	**96.48**	**6.80**
	PUNKT	1,315	60	97	95.64	93.13	94.37	10.67
it	WASTE	182	11	12	**94.30**	**93.81**	**94.06**	**11.22**
	PUNKT	153	76	41	66.81	78.87	72.34	43.33

Table 7: Performance on word boundary detection for dot-final words.

"*Inc.*") and ordinals ("*24.*"). Restricting the evaluation to dot-terminated words containing at least one non-punctuation character produces the data in Table 7. Here again, WASTE substantially outperformed Punkt for all languages[11] and all evaluation modes except for precision on the German corpus (98.13% for WASTE vs. 99.17% for Punkt), incurring on average 62.1% fewer errors than Punkt ($\sigma = 15.5\%$).

3.4 Experiment 2: Training Corpus Size

It was mentioned above that our approach relies on supervised training from a pre-segmented corpus to estimate the model parameters used for runtime boundary placement prediction. Especially in light of the relatively high error-rates observed for the smallest test corpus (Italian), this requirement raises the question of how much training material is in fact necessary to ensure adequate runtime performance of our model. To address such concerns, we varied the amount of training data used to estimate the HMM's parameters between 10,000 and 100,000 tokens,[12] using cross-validation to compute averages for each training-size condition. Results for this experiment are given in Figure 2.

All tested languages showed a typical logarithmic learning curve for both sentence- and word-boundary detection, and word-boundaries were learned more quickly than sentence boundaries in all cases. This should come as no surprise, since any non-trivial corpus will contain more word boundaries than sentence boundaries, and thus provide more training data for detection of the former. Sentence boundaries were hardest to detect in the German corpus, which is assumedly due to the relatively high frequency of punctuation-free sentence boundaries in the TIGER corpus, in which over 10% of the sentence boundaries were not immediately preceded by a punctuation

[11] We ignore the Czech data for the analysis of dot-final word boundaries, since the source corpus included token boundaries before every word-terminating dot character, even for "obvious" abbreviations like "*Ms.*" and "*Inc.*" or initials such as in "[John] *D.* [Rockefeller]".

[12] Training sizes for Italian varied between 7,000 and 70,000 due to the limited number of tokens available in that corpus as a whole.

Case	Punct	tp	fp	fn	pr%	rc%	F%	Err%
+	+	47,453	965	3,014	**98.01**	**94.03**	**95.98**	**7.74**
+	−	33,597	1,749	16,862	95.05	66.58	78.31	35.65
−	+	44,205	1,185	6,262	97.39	87.59	92.23	14.42
−	−	4,814	3,277	45,645	59.50	9.54	16.44	91.04

Table 8: Effect of typographical conventions on sentence detection for the TIGER corpus (de).

character,[13] *vs.* only 1% on average for the other corpora ($\sigma = 0.78\%$). English and French were the most difficult corpora in terms of word boundary detection, most likely due to apostrophe-related phenomena including the English genitive marker *'s* and the contracted French article *l'*.

3.5 Experiment 3: Typographical Conventions

Despite the lack of typographical clues, the WASTE tokenizer was able to successfully detect over 3300 of the unpunctuated sentence boundaries in the German TIGER corpus ($\text{pr} = 94.1\%, \text{rc} = 60.8\%$). While there is certainly room for improvement, the fact that such a simple model can perform so well in the absence of explicit sentence boundary markers is encouraging, especially in light of our intent to detect sentence boundaries in non-standard computer-mediated communication text, in which typographical markers are also frequently omitted. In order to get a clearer idea of the effect of typographical conventions on sentence boundary detection, we compared the WASTE tokenizer's performance on the German TIGER corpus with and without both punctuation (±Punct) and letter-case (±Case), using cross-validation to train and test on appropriate data, with the results given in Table 8.

As hypothesized, both letter-case and punctuation provide useful information for sentence boundary detection: the model performed best for the original corpus retaining all punctuation and letter-case distinctions. Also unsurprisingly, punctuation was a more useful feature than letter-case for German sentence boundary detection,[14] the [−Case, +Punct] variant achieving a harmonic precision-recall average of $\text{F} = 92.23\%$. Even letter-case distinctions with no punctuation at all sufficed to identify about two thirds of the sentence boundaries with over 95% precision, however: this modest success is attributable primarily to the observations' STOP features, since upper-cased sentence-initial stopwords are quite frequent and almost always indicate a preceding sentence boundary.

[13] Most of these boundaries appear to have been placed between article headlines and body text for the underlying newspaper data.

[14] In German, not only sentence-initial words and proper names, but also all common nouns are capitalized, so letter-case is not as reliable a clue to sentence boundary placement as it might be for English, which does not capitalize common nouns.

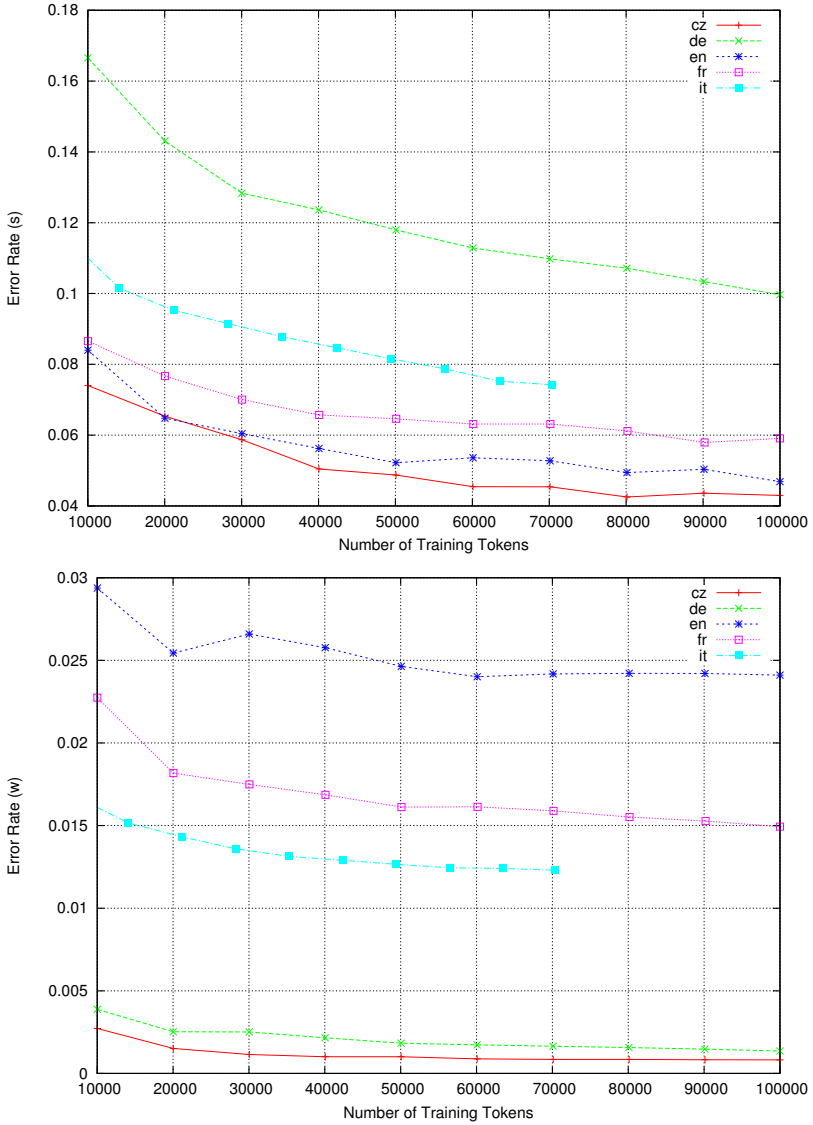


Figure 2: Effect of training corpus size on sentence boundary detection (top) and word boundary detection (bottom).

Model	tp	fp	fn	pr%	rc%	F%	Err%
CHAT	7,052	1,174	4,363	85.73	61.78	71.81	43.98
CHAT[+force]	11,088	1,992	327	84.77	**97.14**	90.53	17.30
CHAT[+feat]	10,524	784	891	**93.07**	92.19	**92.63**	**13.73**
TIGER	2,537	312	8,878	89.05	22.23	35.57	78.37
TIGER[+force]	10,396	1,229	1,019	89.43	91.07	90.24	17.78

Table 9: Effect of training source on sentence boundary detection for the chat corpus.

3.6 Experiment 4: Chat Tokenization

We now turn our attention to the task of segmenting a corpus of computer-mediated communication, namely the chat corpus subset described in Sec. 3.1. Unlike the newspaper corpora used in the previous experiments, chat data is characterized by non-standard use of letter-case and punctuation: almost 37% of the sentences in the chat corpus were not terminated by a punctuation character, and almost 70% were not introduced by an upper-case letter. The chat data were subdivided into 10,289 distinct observable utterance-like units we refer to as *posts*; of these, 9479 (92.1%) coincided with sentence boundaries, accounting for 83% of the sentence boundaries in the whole chat corpus. We measured the performance of the following five distinct WASTE tokenizer models on sentence- and word-boundary detection for the chat corpus:

- CHAT: the standard model as described in Section 2, trained on a disjoint subset of the chat corpus and evaluated by cross-validation;
- CHAT[+force]: the standard model with a supplemental heuristic forcing insertion of a sentence boundary at every post boundary;
- CHAT[+feat]: an extended model using all features described in Section 2.2.1 together with additional binary contextual surface features BOU and EOU encoding whether or not the corresponding segment occurs at the beginning or end of an individual post, respectively;
- TIGER: the standard model trained on the entire TIGER newspaper corpus; and
- TIGER[+force]: the standard TIGER model with the supplemental [+force] heuristic for sentence boundary insertion at every post boundary.

Results for chat corpus sentence boundary detection are given in Table 9 and for word boundaries in Table 10. From these data, it is immediately clear that the standard model trained on conventional newspaper text (TIGER) does not provide a satisfactory segmentation of the chat data on its own, incurring almost twice as many errors as the standard model trained by cross-validation on chat data (CHAT). This supports our claim that chat data represent unconventional and non-standard uses of model-relevant features, in particular punctuation and capitalization. Otherwise, differences between the various cross-validation conditions CHAT, CHAT[+force], and CHAT[+feat] with respect to word-boundary placement were minimal.

Model	tp	fp	fn	pr%	rc%	F%	Err%
CHAT	93,167	1,656	1,934	98.25	97.97	98.11	3.71
CHAT[+force]	93,216	1,612	1,885	98.30	98.02	98.16	3.62
CHAT[+feat]	93,235	1,595	1,866	**98.32**	**98.04**	**98.18**	**3.58**
TIGER	91,603	5,889	3,499	93.96	96.32	95.13	9.30
TIGER[+force]	91,611	5,971	3,491	93.88	96.33	95.09	9.36

Table 10: Effect of training source on word boundary detection for the chat corpus.

Sentence-boundary detection performance for the standard model (CHAT) was similar to that observed in Section 3.5 for newspaper text with letter-case but without punctuation, EOS recall in particular remaining unsatisfactory at under 62%. Use of the supplemental [+force] heuristic to predict sentence boundaries at all post boundaries raised recall for the newspaper model (TIGER[+force]) to over 91%, and for the cross-validation model (CHAT[+force]) to over 97%. The most balanced performance however was displayed by the extended model CHAT[+feat] using surface features to represent the presence of post boundaries: although its error rate was still quite high at almost 14%, the small size of the training subset compared to those used for the newspaper corpora in Section 3.3 leaves some hope for improvement as more training data become available, given the typical learning curves from Figure 2.

4 Conclusion

We have presented a new method for estimating sentence and word token boundaries in running text by coupling a prolix rule-based scanner stage with a Hidden Markov Model over scan-segment feature bundles using hidden binary features BOW, BOS, and EOS to represent the presence or absence of the corresponding boundaries. Language-specific features were limited to an optional set of user-specified stopwords, while the remaining observable surface features were used to represent basic typographical class, letter-case, word length, and leading whitespace.

We compared our "WASTE" approach to the high-quality sentence boundary detector Punkt described by Kiss and Strunk (2006) on newspaper corpora from five different European languages, and found that the WASTE system not only substantially outperformed Punkt for all languages in detection of both sentence- and word-boundaries, but even outdid Punkt on its "home ground" of dot-terminated words and sentences, providing average relative error reduction rates of 62% and 33%, respectively. Our technique exhibited a typical logarithmic learning curve, and was shown to adapt fairly well to varying typographical conventions given appropriate training data.

A small corpus of computer-mediated communication extracted from the Dortmund Chat Corpus (Beißwenger and Storrer, 2008) and manually segmented was introduced and shown to violate some typographical conventions commonly used for sentence boundary detection. Although the unmodified WASTE boundary detector did not perform as well as hoped on these data, the inclusion of additional surface features

sensitive to observable *post boundaries* sufficed to achieve a harmonic precision-recall average F of over 92%, representing a relative error reduction rate of over 82% with respect to the standard model trained on newspaper text, and a relative error reduction rate of over 38% with respect to a naïve domain-specific splitting strategy.

Acknowledgements

Research was supported by the *Deutsche Forschungsgemeinschaft* grants KL 337/12-2 and KL 955/19-1. We are grateful to Sophie Arana, Maria Ermakova, and Gabriella Pein for their help in manual preparation of the chat corpus used in section 3.6. The WASTE system described above is included with the open-source `moot` package, and can be tested online at `http://www.dwds.de/waste/`. Finally, we would like to thank this article's anonymous reviewers for their helpful comments.

References

Abeillé, A., Clément, L., and Toussenel, F. (2003). Building a treebank for French. In Abeillé, A., editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 165–187. Springer.

Beißwenger, M. and Storrer, A. (2008). Corpora of Computer-Mediated Communication. In Lüdeling, A. and Kytö, M., editors, *Corpus Linguistics. An International Handbook*, pages 292–308. De Gruyter.

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly, Sebastopol, CA.

Bosco, C., Lombardo, V., Vassallo, D., and Lesmo, L. (2000). Building a treebank for Italian: a data-driven annotation schema. In *Proceedings LREC '00*, Athens, Greece.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Brants, T. (2000). TnT – a statistical part-of-speech tagger. In *Proceedings of the Sixth Applied Natural Language Processing Conference ANLP-2000*, Seattle, WA.

Bumbulis, P. and Cowan, D. D. (1993). RE2C: a more versatile scanner generator. *ACM Letters on Programming Languages and Systems*, 2(1-4):70–84.

Church, K. W. (1988). A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the 2nd Conference on Applied Natural Language Processing (ANLP)*, pages 136–143.

Cuřín, J., Čmejrek, M., Havelka, J., Hajič, J., Kuboň, V., and Žabokrtskỳ, Z. (2004). Prague Czech-English dependency treebank version 1.0. *Linguistic Data Consortium, Catalog No.: LDC2004T25*.

Francis, W. N. and Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and Grammar*. Houghton Mifflin.

Grefenstette, G. and Tapanainen, P. (1994). What is a word, What is a sentence? Problems of Tokenization. In *Proceedings of the 3rd Conference on Computational Lexicography and Text Research*, COMPLEX '94, pages 79–87.

He, Y. and Kayaalp, M. (2006). A Comparison of 13 Tokenizers on MEDLINE. Technical Report LHNCBC-TR-2006-03, U.S. National Library of Medicine.

Hunt, J. W. and McIlroy, M. D. (1976). An algorithm for differential file comparison. Computing Science Technical Report 41, Bell Laboratories.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In Gelsema, E. S. and Kanal, L. N., editors, *Pattern Recognition in Practice*, pages 381–397. North-Holland Publishing Company, Amsterdam.

Jurish, B. (2003). A hybrid approach to part-of-speech tagging. Technical report, Project "Kollokationen im Wörterbuch", Berlin-Brandenburg Academy of Sciences, Berlin.

Kempe, A. (1997). Finite state transducers approximating hidden markov models. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 460–467.

Kiss, T. and Strunk, J. (2006). Unsupervised multilingual sentence boundary detection. *Computational Linguistics*, 32(4):485–525.

Lowerre, B. T. (1976). *The HARPY speech recognition system*. PhD thesis, Carnegie Mellon University.

MacKenzie, D., Eggert, P., and Stallman, R. (2002). *Comparing and Merging Files with GNU diff and patch*. Network Theory Ltd., Bristol, UK.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

Marcus, M. P., Santorini, B., Marcinkiewicz, M. A., and Taylor, A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Mikheev, A. (2000). Tagging Sentence Boundaries. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 264–271.

Palmer, D. D. and Hearst, M. A. (1997). Adaptive multilingual sentence boundary disambiguation. *Computational Linguistics*, 23(2):241–267.

Rabiner, L. R. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Riley, M. D. (1989). Some applications of tree-based modelling to speech and language. In *Proceedings of the DARPA Speech and Natural Language Workshop*, Human Language Technology Workshops '89, pages 339–352.

Schmid, H. (2000). Unsupervised Learning of Period Disambiguation for Tokenisation. Internal report, *Institut für Maschinelle Sprachverarbeitung*, Universität Stuttgart.

Skut, W. and Brants, T. (1998). Chunk tagger - statistical recognition of noun phrases. *CoRR*, cmp-lg/9807007.

Tomanek, K., Wermter, J., and Hahn, U. (2007). Sentence and token splitting based on conditional random fields. In *Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics*, pages 49–57.

Unicode Consortium (2012). *The Unicode Standard, Version 6.2.0*. The Unicode Consortium, Mountain View, CA.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth-Heinemann, Newton, MA.

Viterbi, A. J. (1967). Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transications on Information Theory*, pages 260–269.

Sabine Schulte im Walde, Stefan Müller

Using Web Corpora for the Automatic Acquisition of Lexical-Semantic Knowledge

This article presents two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare three German web corpora and a traditional newspaper corpus on modelling two types of semantic relatedness: (1) Assuming that free word associations are semantically related to their stimuli, we explore to which extent stimulus–associate pairs from various associations norms are available in the corpus data. (2) Assuming that the distributional similarity between a noun–noun compound and its nominal constituents corresponds to the compound's degree of compositionality, we rely on simple corpus co-occurrence features to predict compositionality. The case studies demonstrate that the corpora can indeed be used to model semantic relatedness, (1) covering up to 73/77% of verb/noun–association types within a 5-word window of the corpora, and (2) predicting compositionality with a correlation of $\rho = 0.65$ against human ratings. Furthermore, our studies illustrate that the corpus parameters *domain, size* and *cleanness* all have an effect on the semantic tasks.

1 Motivation

Distributional models assume that the contexts of a linguistic unit (such as a word, a multi-word expression, a phrase, a sentence, etc.) provide information about the meaning of the linguistic unit (Firth, 1957; Harris, 1968). They have been widely applied in data-intensive lexical semantics (among other areas), and proven successful in diverse research issues, such as the representation and disambiguation of word senses (Schütze, 1998; McCarthy et al., 2004; Springorum et al., 2013), selectional preference modelling (Herdagdelen and Baroni, 2009; Erk et al., 2010; Schulte im Walde, 2010), compositionality of compounds and phrases (McCarthy et al., 2003; Reddy et al., 2011; Boleda et al., 2013; Schulte im Walde et al., 2013), or as a general framework across semantic tasks (Baroni and Lenci, 2010; Padó and Utt, 2012), to name just a few examples.

While distributional models are well-established in computational linguistics, from a cognitive point of view the relationship between meaning and distributional models has been more controversial (Marconi, 1997; Glenberg and Mehta, 2008), because distributional models are expected to cover the linguistic ability of how to use language (*inferential abilities*), but they do not incorporate a knowledge of the world (*referential abilities*). Since distributional models are very attractive – the underlying parameters

being accessible from even low-level annotated corpus data – we are interested in maximising the benefit of distributional information for lexical semantics, and in exploring the potential and the limits with regard to individual semantic tasks. More specifically, our work addresses distributional approaches with respect to semantic relatedness. As resources for the distributional knowledge we rely on web corpora, because (a) these corpora are the largest language corpora currently available, and size matters (Banko and Brill, 2001; Curran and Moens, 2002; Pantel et al., 2004; Hickl et al., 2006), and (b) web corpora are domain-independent, in comparison to e.g. large newspaper collections, and thus cover the potentially widest breadth of vocabulary.

In this article, we present two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare three German web corpora (differing in domain, size and cleanness) on modelling two types of semantic relatedness:

(1) semantic associations, and

(2) the compositionality of noun-noun compounds.

Concerning task (1), we in addition compare the usage of the web corpora against applying a traditional newspaper corpus.

Semantic Relatedness Task (1) assumes that free word associations (i.e., words that are spontaneously called to mind by a stimulus word) represent a valuable resource for cognitive and computational linguistics research on semantic relatedness. This assumption is based on a long-standing hypothesis, that *associations reflect meaning components of words* (Nelson et al., 1997, 2000; McNamara, 2005), and are thus semantically related to the stimuli. Our first semantic relatedness task will explore to which extent lexical-semantic knowledge – as represented by various collections of associations – is available in the corpus data.

Semantic Relatedness Task (2) is the prediction of compositionality for a set of German noun-noun compounds, i.e., the degree of semantic relatedness between a compound and its constituents. We rely on simple corpus co-occurrence features to instantiate a distributional model of the compound nouns and their nominal constituents, and use the cosine measure to rate the distributional similarity between the compounds and the constituents. Assuming that the distributional similarity between a compound and a constituent corresponds to the compound–constituent degree of compositionality, we compare our predictions against human compositionality ratings.

The remainder of the article is organised as follow. Section 2 first introduces the (web) corpora that are used for our semantic tasks. Section 3 then describes the usage of these corpora with regard to our two tasks, (1) the availability of association knowledge and (2) predicting the compositionality of noun compounds. Section 4 summarises and discusses the results and insights.

2 Corpora

The main goal of our work is to explore the potential and the limits of distributional information for modelling semantic relatedness. We are thus interested in (a) exploiting the largest available corpus data but at the same time (b) comparing the usage of different types of corpora. Accordingly, corpus criteria that are central to our work are

- corpus domain(s) and
- corpus size.

Our case studies in Section 3 make use of the following four corpora:

1. *WebKo*

   The *WebKo* corpus is a slightly cleaned version of *deWaC*, a standard cross-domain German web corpus created by the *WaCky* group (Baroni et al., 2009). The cleaning was performed through simple heuristics to identify and disregard implausible domains and to repair dates and incorrectly separated tokens.

2. *SdeWaC*

   The *SdeWaC* corpus (Faaß et al., 2010; Faaß and Eckart, 2013) is a more severely cleaned version of *WebKo*. The corpus cleaning had focused on removing duplicates from the *deWaC*, and on disregarding sentences that were syntactically ill-formed (following Quasthoff et al. (2006) regarding heuristics such as the number of commas per sentence, the number of spaces in proportion to sentence length, etc.; and relying on a parsability index provided by the standard dependency parser by Schiehlen (2003)). The SdeWaC is freely available and can be downloaded from `http://wacky.sslmit.unibo.it/`.

   A special feature of the SdeWaC is that the sentences in the corpus have been sorted alphabetically, so going beyond the sentence border is likely to entering a sentence that did not originally precede or follow the sentence of interest. This will have an effect on window co-occurrence in Section 3.

3. *German Wikipedia (Wiki-de)*

   The *Wiki-de* corpus is also a German web corpus, based on the official Wikipedia dump[1] `dewiki-20110410` from April 10, 2011. The dump has been downloaded and processed by André Blessing, *Institut für Maschinelle Sprachverarbeitung (IMS)*. The processing was performed using the *Java Wikipedia Library (JWPL)*[2], an open-source Java-based application programming interface to access and parse the information in the Wikipedia articles (Zesch et al., 2008). The encyclopaedic knowledge is expected to complement the knowledge induced from the other two web corpora, cf. Roth and Schulte im Walde (2008).

[1] `http://dumps.wikimedia.org/dewiki/`
[2] `http://www.ukp.tu-darmstadt.de/software/jwpl/`

4. *Huge German Corpus (HGC)*

   The *HGC* corpus is a large collection of German newspaper corpora, containing data from *Frankfurter Rundschau*, *Stuttgarter Zeitung*, *VDI-Nachrichten*, *die tageszeitung (taz)*, *Gesetzestexte (German Law Corpus)*, *Donaukurier*, and *Computerzeitung* from the 1990s. We used the HGC in contrast to the above three web corpora, to explore the influence of the more restricted corpus domain.

Table 1 provides an overview of the corpus sizes. All of the corpora have been tokenised and part-of-speech tagged with the *Tree Tagger* (Schmid, 1994).

	WebKo	SdeWaC	Wiki-de	HGC
sentences	71,585,693	45,400,446	23,205,536	9,255,630
words (tokens)	1,520,405,616	884,356,312	432,131,454	204,813,118
words (types)	14,908,301	9,220,665	7,792,043	3,193,939

Table 1: Overview of corpora.

3 Web Corpora and Semantic Relatedness: Two Case Studies

This section as the main part of the article explores whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compare the German corpora introduced in Section 2 on modelling two types of semantic relatedness, the availability of semantic associates (Section 3.1) and the prediction of compositionality for German noun-noun compounds (Section 3.2).

3.1 Availability of Semantic Associates in Web Corpora

Our first semantic relatedness task will explore to which extent lexical-semantic knowledge – as represented by three collections of association norms – is available in the (web) corpus data. Section 3.1.1 first introduces the relevant background on association norms, before Section 3.1.2 provides an explicit description of our hypotheses. Section 3.1.3 then describes the actual distributional explorations.

3.1.1 Background and Earlier Work on Association Norms

Associations are commonly obtained by presenting *target stimuli* to the participants in an experiment, who then provide *associate responses*, i.e., words that are spontaneously called to mind by the stimulus words. The quantification of the resulting stimulus–association pairs (i.e., how often a certain association is provided for a certain stimulus) is called *association norm*. Table 2 (as taken from Schulte im Walde et al. (2008)) provides the 10 most frequent associate responses for the ambiguous verb *klagen* and the ambiguous noun *Schloss* as examples.

klagen 'complain, moan, sue'		
Gericht	'court'	19
jammern	'moan'	18
weinen	'cry'	13
Anwalt	'lawyer'	11
Richter	'judge'	9
Klage	'complaint'	7
Leid	'suffering'	6
Trauer	'mourning'	6
Klagemauer	'Wailing Wall'	5
laut	'noisy'	5

Schloss 'castle, lock'		
Schlüssel	'key'	51
Tür	'door'	15
Prinzessin	'princess'	8
Burg	'castle'	8
sicher	'safe'	7
Fahrrad	'bike'	7
schließen	'close'	7
Keller	'cellar'	7
König	'king'	7
Turm	'tower'	6

Table 2: Associate responses and associate frequencies for example stimuli.

Association norms have a long tradition in psycholinguistic research, where the implicit notion that associations reflect meaning components of words has been used for more than 30 years to investigate semantic memory. One of the first collections of word association norms was done by Palermo and Jenkins (1964), comprising associations for 200 English words. The *Edinburgh Association Thesaurus* (Kiss et al., 1973) was a first attempt to collect association norms on a larger scale, and also to create a network of stimuli and associates, starting from a small set of stimuli derived from the Palermo and Jenkins norms. A similar motivation underlies the association norms from the University of South Florida (Nelson et al., 1998),[3] who grew a stimulus-associate network over more than 20 years, from 1973. More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 stimulus words. In another long-term project, Simon de Deyne and Gert Storms are collecting associations to Dutch words, cf. `www.smallworldofwords.com`. Previously, they performed a three-year collection of associations to 1,424 Dutch words (de Deyne and Storms, 2008b). Smaller sets of association norms have also been collected for example for German (Russell and Meseck, 1959; Russell, 1970), Dutch (Lauteslager et al., 1986), French (Ferrand and Alario, 1998) and Spanish (Fernández et al., 2004) as well as for different populations of speakers, such as adults vs. children (Hirsh and Tree, 2001).

In parallel to the interest in collecting association norms, researchers have analysed association data in order to get insight into semantic memory and – more specifically – issues concerning semantic relatedness. For example, Clark (1971) classified stimulus–association relations into sub-categories of paradigmatic and syntagmatic relations, such as synonymy and antonymy, selectional preferences, etc. Heringer (1986) concentrated on syntagmatic associations to a small selection of 20 German verbs. He asked his subjects to provide question words as associations (e.g., *wer* 'who', *warum* 'why'), and used the responses to investigate the valency behaviour of the verbs. Spence and Owens (1990) showed that associative strength and word co-occurrence are correlated. Their

[3] `http://www.usf.edu/FreeAssociation/`

investigation was based on 47 pairs of semantically related concrete nouns, as taken from the Palermo and Jenkins norms, and their co-occurrence counts in a window of 250 characters in the 1-million-word Brown corpus. Church and Hanks (1989) were the first to apply information-theoretic measures to corpus data in order to predict word association norms for lexicographic purposes. Our own work analysed German noun and verb associations at the syntax-semantics interface (Schulte im Walde et al., 2008). Schulte im Walde and Melinger (2008) performed a more in-depth analysis of window co-occurrence distributions of stimulus–response pairs. Roth and Schulte im Walde (2008) explored whether dictionary and encyclopaedic information provides more world knowledge about associations than corpus co-occurrence, and found that the information in the three resource types complements each other.

In experimental psychology, association norms have been used extensively to conduct studies with variations of the semantic priming technique to investigate (among other things) word recognition, knowledge representation and semantic processes (see McNamara (2005) for a review of methods, issues, and findings). In the last decade, association norms have also found their way into lexical-semantic research in computational linguistics. For example, Rapp (2002) developed corpus-based approaches to predict paradigmatic and syntagmatic associations; de Deyne and Storms (2008a) created semantic networks from Dutch associations; and Schulte im Walde (2008) used associations to German verbs to select features for automatic semantic classification.

3.1.2 Associations, Semantic Relatedness, and Corpus Co-Occurrence

Our analyses to follow rely on three well-known hypotheses: (i) the psycholinguistic notion that *associations reflect meaning components of words* (Nelson et al., 1997, 2000; McNamara, 2005); (ii) the *co-occurrence hypothesis* that associations are related to the textual co-occurrence of the stimulus–association pairs (Miller, 1969; Spence and Owens, 1990; McKoon and Ratcliff, 1992; Plaut, 1995); and (iii) the *distributional hypothesis* that contexts of a word provide information about its meaning (Firth, 1957; Harris, 1968). Figure 1 illustrates the triangle that combines the three hypotheses and at the same time bridges the gap between long-standing assumptions in psycholinguistic and computational linguistics research, regarding associations, semantic memory, and corpus co-occurrence.

According to these three hypotheses, association norms thus represent a valuable resource for cognitive and computational linguistics research on semantic relatedness. In the current study, we exploit three collections of association norms to explore the availability of lexical-semantic knowledge in the corpora introduced in Section 2, assuming that associations reflect semantic knowledge that can be captured by distributional information. Accordingly, for each of the norms we check the availability of the semantic associates in corpus co-occurrence.

3.1.3 *Study (1):* Association Norms and Corpus Co-Occurrence

Our analyses make use of the following three association norms:

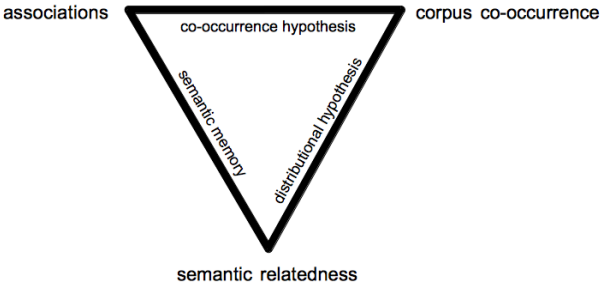


Figure 1: Association–meaning triangle.

1. Associations to German verbs, collected in 2004 (Schulte im Walde et al., 2008):
   - 330 verbs including 36 particle verbs
   - 44–54 participants per stimulus
   - 38,769/79,480 stimulus–association types/tokens

2. Associations to German nouns, collected in 2003/4 (Melinger and Weber, 2006):
   - 409 nouns referring to picturable objects
   - 100 participants per stimulus
   - 30,845/116,714 stimulus–association types/tokens

3. Associations to German noun compounds, collected in 2010–2012:

   a) based on a web experiment with German noun compounds and constituents (Schulte im Walde et al., 2012):
      - 996 compounds+constituents for 442 concrete, depictable compounds
      - 10–36 participants per stimulus
      - 28,238/47,249 stimulus–association types/tokens

   b) based on an Amazon Mechanical Turk (AMT) experiment with a subset of the above compounds+constituents (Borgwaldt and Schulte im Walde, 2013):
      - 571 compounds+constituents for 246 noun-noun compounds
      - 2–120 (mostly: 30) participants per stimulus
      - 26,415/59,444 stimulus–association types/tokens

Relying on these three norms enables us to check on the availability of semantic associate knowledge, with regard to two major word classes of the stimuli (verbs and nouns), and a morphologically restricted subset of one of these classes, compound nouns, where we only take the noun compound stimuli from the norms 3a) and 3b) into account.

Table 3 presents the proportions of the stimulus–associate types that were found in a 5-word window from each other in the various corpora, i.e., the association appeared at least once in a maximum of five words to the left or to the right of the respective stimulus. Since the SdeWaC corpus is sorted alphabetically and thus going beyond the sentence border was likely to entering a sentence that did not originally precede or follow the sentence of interest (cf. Section 2), we distinguish two modes: *sentence-internal (int)*, where window co-occurrence refers to five words to the left and right BUT within the same sentence, and *sentence-external (ext)*, the standard window co-occurrence that goes beyond sentence borders. We applied the *int* mode also to the other corpora, to make the co-occurrence numbers more comparable.

Table 3 shows that the association type coverage is monotonous with regard to the size of the corpora and the co-occurrence mode, with one exception: the larger the corpus, the stronger the coverage of the association data; but for the verb stimuli, the 430-million-word corpus Wiki-de has a lower coverage than the 200-million-word newspaper corpus HGC. Unsurprisingly, including sentence-external co-occurrence material increases the coverage, in comparison to only taking sentence-internal co-occurrence into account.

Norms	Size	Corpora						
		HGC		Wiki-de		WebKo		SdeWaC
		200		*430*		*1,500*		*880*
		ext	int	ext	int	ext	int	int
verbs	38,769	54	51	50	47	73	70	68
nouns	30,845	53	51	56	54	77	76	72
compound nouns	14,326	23	22	25	23	49	47	42

Table 3: Coverage of association types across corpora in a 5-word window.

Table 4 compares the same corpora on association coverage as Table 3, but with regard to sub-corpora of similar sizes, thus looking at the effects of the corpus domains and corpus cleaning. Since the HGC as the smallest corpus contains approx. 200 million words, and accidentally the sizes of the other corpora are rough multiples of 200 million words, we created sub-corpora of approx. 200 million words for each corpus. Table 4 compares the corpus coverage of the verb–association and the noun-association types with regard to the whole HGC, both parts of Wiki-de, two parts (out of eight) of WebKo and all four parts of SdeWaC.

Table 4 shows that the exact coverage of the association types varies from (sub-)corpus to (sub-)corpus but is within the range $[49\%, 57\%]$ for HGC, WebKo and SdeWaC. For Wiki-de, however, the coverage is clearly lower, within the range $[41\%, 49\%]$. Two facts are surprising: (i) We would have expected the association coverage of HGC to be

below that of the web corpora, because the domain is more restricted. The coverage is however compatible with the WebKo and SdeWaC coverage scores. (ii) We would have expected the association coverage of Wiki-de to be compatible with that of the other two web corpora, because the encyclopaedic knowledge was expected to provide relevant semantic knowledge. The coverage is however below that of the other corpora.

Norms	Size	Corpora								
		HGC	Wiki-de		WebKo		SdeWaC			
		ext					int			
verbs	38,769	54	44	41	53	55	56	55	53	49
nouns	30,845	53	49	44	54	56	57	57	55	52

Table 4: Coverage of association types across 200-million-word corpora in a 5-word window.

Altogether, the above two tables demonstrate that the corpora are indeed capturing lexical-semantic knowledge with regard to the various stimulus–association pairs, finding up to 73/77/49% of the types in a small 5-word window of the stimuli in the largest corpus (WebKo). Table 5 (focusing on the SdeWaC) shows that the coverage is even stronger when looking at the stimulus–association pair *tokens* (in comparison to the *types*), reaching 78/87/58% coverage (in comparison to 68/72/42%). Looking at a larger 20-word window, these numbers go up to 84/91/67%. Thus, we induce that we can indeed find lexical-semantic knowledge in our corpora, and stronger related pairs are even more likely to be covered than weaker pairs (which can be induced from the fact that the token coverage is larger than the type coverage).

Norms	Size	SdeWaC			
		window: 5		window: 20	
		types	tokens	types	tokens
verbs	38,769/79,480	68	78	76	84
nouns	30,845/116,714	72	87	80	91
compound nouns	14,326/33,065	42	58	51	67

Table 5: Coverage of association tokens vs. types in *SdeWaC*.

In addition, the larger the corpora are, the stronger is the availability of the semantic knowledge. Since web corpora are the largest corpora available, they therefore represent the best choice for lexical-semantic knowledge, at least with regard to the specific instance of semantic association data. Corpus size outperforms extensive corpus cleaning (when comparing WebKo with SdeWaC in Table 3), but the difference is small (70/76/47% vs. 68/72/42%), taking into consideration that the size difference is large (1,500 vs. 880 million words), so the cleaning obviously had a positive effect on the corpus quality. Finally, general web corpora such as the deWaC (of which WebKo and SdeWaC are

both subsets) seem overall more suitable for general lexical-semantic knowledge than the more structured and entry-based Wikipedia corpus.

Focusing on stimulus–association pairs with compound stimuli in Tables 3 and 5 shows that their coverage is significantly below the coverage of non-compound stimuli (note that most of the stimuli in the verb and noun association norms are not morphologically complex). We assume that this is due to two facts: (i) German noun compounds are notoriously productive, so their corpus coverage is expected to be lower than for non-complex nouns. (ii) Noun compounds are more difficult to process than non-complex nouns because of their morphological complexity, even for standard tools. In sum, we expected a lower co-occurrence corpus coverage for compound–associate pairs in comparison to pairs with morphologically non-complex stimuli, because the stimuli are more sparse. Considering this restriction, the coverage rates are actually quite impressive, reaching 58/67% token coverage in a 5/20-word window of the SdeWaC. Comparing across corpora, the two deWaC corpora capture clearly larger proportions of compound–associate pairs than the HGC and Wiki-de, confirming the usefulness of the two web corpora for the availability of the semantic associate knowledge.

Concerning the stimulus–association pairs that are *not* found in corpus co-occurrence, the reader is referred to Schulte im Walde et al. (2008) and Roth and Schulte im Walde (2008) for detailed explorations. As expected, a serious proportion of these associations reflects world knowledge and is therefore not expected to be found in the immediate context of the stimuli at all, for example *mampfen–lecker* 'munch–yummy', *auftauen–Wasser* 'defrost–water', *Ananas–gelb* 'pineapple–yellow' and *Geschenk–Überraschung* 'present–surprise'. These cases pose a challenge to empirical models of word meaning.

To complete our first analysis on the availability of lexical-semantic knowledge in (web) corpora, Table 6 presents a selection of stimulus–association pairs from different domains, with different morphological complexity of the stimuli, and with difference strengths, accompanied by their 20-word window co-occurrence coverage in our corpora. Comparing WebKo *ext* with WebKo *int* demonstrates that reducing the 20-word windows to sentence-internal windows has a severe effect on the association coverage: all co-occurrence counts for the former condition are larger than for the latter condition, in some cases even twice as much. Furthermore, the table illustrates that typically the larger the corpora the more instances of the stimulus–associate pairs were found. However, the domains of the corpora also play a role: The HGC outperforms Wiki-de in more cases than vice versa, and in four cases (*Affe–Urwald*, *Blockflöte–Musik*, *Polizist–grün*, *Telefonzelle–gelb*), the HGC even outperforms SdeWaC, which is approximately four times as large (however restricted to sentence-internal co-occurrence). Regarding our examples, it is also obvious that the total co-occurrence counts involving compounds are lower than those involving simplex stimuli. Even though the sample is too small to generalise this intuition, it confirms our insight from Table 5 that associations to compounds are covered less than associations to simplex words.

Stimulus–Associate Pairs			Corpus				
			HGC	Wiki-de	WebKo		SdeWaC
				ext	ext	int	int
			200	*430*	*1,500*		*880*
Affe 'monkey'	*Urwald* 'jungle'	15	10	1	89	48	3
analysieren 'analyse'	*untersuchen* 'analyse'	8	56	172	1,613	685	451
bedauern 'regret'	*Mitleid* 'pity'	11	3	0	29	12	10
Blockflöte 'flute'	*Musik* 'music'	23	26	73	90	43	22
Fliegenpilz 'toadstool'	*giftig* 'poisonous'	34	0	9	40	14	9
Kuh 'cow'	*melken* 'milk'	12	71	28	880	683	200
Obstkuchen 'fruit cake'	*backen* 'bake'	7	1	1	17	12	5
Polizist 'police-man'	*grün* 'green'	45	65	9	120	61	52
rollen 'roll'	*Kugel* 'bowl'	15	96	10	654	483	277
schleichen 'crawl'	*leise* 'quiet'	36	10	4	569	428	88
Schlittenhund 'sledge dog'	*Winter* 'winter'	10	1	5	6	3	3
Telefonzelle 'phone box'	*gelb* 'yellow'	25	16	5	17	14	6
verbrennen 'burn'	*heiß* 'hot'	15	42	55	534	348	194

Table 6: Examples of co-occurring stimulus–association pairs across corpora.

3.2 Predicting the Degree of Compositionality of German Noun-Noun Compounds

Our second semantic relatedness task will predict the compositionality of German noun-noun compounds, i.e., the semantic relatedness between a compound and its constituents. The distributional model for this task describes the compound nouns and their nominal constituents by simple corpus co-occurrence features, and relies on the distributional similarity between the compounds and the constituents to rate their semantic relatedness. Section 3.2.1 first introduces the relevant background on German noun compounds and describes our compound data, before Section 3.2.2 presents human ratings on the compositionality of the compounds. Section 3.2.3 then describes the actual distributional explorations.

3.2.1 German Noun-Noun Compounds

Compounds are combinations of two or more simplex words. Traditionally, a number of criteria (such as compounds being syntactically inseparable, and that compounds have a specific stress pattern) have been proposed, in order to establish a border between compounds and non-compounds. However, Lieber and Stekauer (2009a) demonstrated

that none of these tests are universally reliable to distinguish compounds from other types of derived words. Compounds have thus been a recurrent focus of attention within theoretical, cognitive, and in the last decade also within computational linguistics. Recent evidence of this strong interest are the *Handbook of Compounding* on theoretical perspectives (Lieber and Stekauer, 2009b), and a series of workshops and special journal issues on computational perspectives (Journal of Computer Speech and Language, 2005; Language Resources and Evaluation, 2010; ACM Transactions on Speech and Language Processing, 2013).[4]

Our focus of interest is on German noun-noun compounds (see Fleischer and Barz (2012) for a detailed overview and Klos (2011) for a recent detailed exploration), such as *Ahornblatt* 'maple leaf' and *Feuerwerk* 'fireworks', where both the grammatical head (in German, this is the rightmost constituent) and the modifier are nouns. More specifically, we are interested in the degrees of compositionality of German noun-noun compounds, i.e., the semantic relatedness between the meaning of a compound (e.g., *Feuerwerk*) and the meanings of its constituents (e.g., *Feuer* 'fire' and *Werk* 'opus').

Our work is based on a selection of noun compounds by von der Heide and Borgwaldt (2009), who created a set of 450 concrete, depictable German noun compounds according to four compositionality classes: compounds that are transparent with regard to both constituents (e.g., *Ahornblatt* 'maple leaf'); compounds that are opaque with regard to both constituents (e.g., *Löwenzahn* 'lion+tooth → dandelion'); compounds that are transparent with regard to the modifier but opaque with regard to the head (e.g., *Feuerzeug* 'fire+stuff → lighter'); and compounds that are opaque with regard to the modifier but transparent with regard to the head (e.g., *Fliegenpilz* 'fly+mushroom → toadstool'). From the compound set by von der Heide and Borgwaldt, we disregarded noun compounds with more than two constituents (in some cases, the modifier or the head was complex itself) as well as compounds where the modifiers were not nouns. Our final set comprises a subset of their compounds: 244 two-part noun-noun compounds.

3.2.2 Compositionality Ratings

von der Heide and Borgwaldt (2009) collected human ratings on compositionality for all their 450 compounds. The compounds were distributed over 5 lists, and 270 participants judged the degree of compositionality of the compounds with respect to their first as well as their second constituent, on a scale between 1 (definitely opaque) and 7 (definitely transparent). For each compound–constituent pair, they collected judgements from 30 participants, and calculated the rating mean and the standard deviation.

Table 7 presents example mean ratings for the compound–constituent ratings, accompanied by the standard deviations. We selected two examples each from our set of 244 noun-noun compounds, according to five categories of mean ratings: the compound–constituent ratings were (1) high or (2) mid or (3) low with regard to both constituents; the compound–constituent ratings were (4) low with regard to the modifier but high with regard to the head; (5) vice versa.

[4] www.multiword.sourceforge.net

Compounds			Mean Ratings	
whole	literal meanings of constituents		modifier	head
Ahornblatt 'maple leaf'	maple	leaf	**5.64** ± 1.63	**5.71** ± 1.70
Postbote 'post man'	mail	messenger	**5.87** ± 1.55	**5.10** ± 1.99
Seezunge 'sole'	sea	tongue	**3.57** ± 2.42	**3.27** ± 2.32
Windlicht 'storm lamp'	wind	light	**3.07** ± 2.12	**4.27** ± 2.36
Löwenzahn 'dandelion'	lion	tooth	**2.10** ± 1.84	**2.23** ± 1.92
Maulwurf 'mole'	mouth	throw	**2.21** ± 1.68	**2.76** ± 2.10
Fliegenpilz 'toadstool'	fly/bow tie	mushroom	**1.93** ± 1.28	**6.55** ± 0.63
Flohmarkt 'flea market'	flea	market	**1.50** ± 1.22	**6.03** ± 1.50
Feuerzeug 'lighter'	fire	stuff	**5.87** ± 1.01	**1.90** ± 1.03
Fleischwolf 'meat chopper'	meat	wolf	**6.00** ± 1.44	**1.90** ± 1.42

Table 7: Examples of compound ratings.

3.2.3 *Study (2):* Predicting Compositionality by Similarity of Corpus Co-Occurrence

In this study, the goal of our experiments is to predict the degree of compositionality of our set of noun-noun compounds as presented in the previous section, by relying on the similarities between the compound and constituent distributional properties. The distributional properties of our noun targets (i.e., the compounds as well as the nominal constituents) are instantiated by a standard vector space model (Turney and Pantel, 2010; Erk, 2012) using window co-occurrence with varying window sizes. We restrict window co-occurrence to nouns, i.e., the dimensions in the vector spaces are all nouns co-occurring with our target nouns within the specified window sizes.[5] For example, for a window size of 5, we count how often our target nouns appeared with any nouns in a window of five words to the left and to the right. As in our first case study, we distinguish the two modes *sentence-internal (int)* and *sentence-external (ext)*.

In all our vector space experiments, we first induce co-occurrence frequency counts from our corpora, and then calculate *local mutual information (LMI)* values (Evert, 2005), to instantiate the empirical properties of our target nouns. LMI is a measure from information theory that compares the observed frequencies O with expected frequencies E, taking marginal frequencies into account: $LMI = O \times log \frac{O}{E}$, with E representing the product of the marginal frequencies over the sample size.[6] In comparison to (pointwise) mutual information (Church and Hanks, 1990), LMI improves the problem of propagating low-frequent events.

Relying on the LMI vector space models, the *cosine* determines the distributional similarity between the compounds and their constituents, which is in turn used to predict the compositionality between the compound and the constituents, assuming that the stronger the distributional similarity (i.e., the *cosine* values), the larger the degree of compositionality. The vector space predictions are evaluated against the human ratings on the degree of compositionality (cf. Section 3.2.2), using the Spearman

[5] See Schulte im Walde et al. (2013) for variations of this noun vector space.
[6] See `http://www.collocations.de/AM/` for a detailed illustration of association measures.

Rank-Order Correlation Coefficient ρ (Siegel and Castellan, 1988). The ρ correlation is a non-parametric statistical test that measures the association between two variables that are ranked in two ordered series.

Figure 2 presents the correlation coefficient ρ (i.e., the quality of the predictions) for our three web corpora WebKo, SdeWaC and Wiki-de across the window sizes $1, 2, 5, 10, 20$. As in our first study, WebKo once more outperforms the SdeWaC corpus, for both modes *ext* and *int*, reaching an optimal prediction of $\rho = 0.6497$ (WebKo (ext)) when relying on a 20-word noun window. For larger windows, the difference between WebKo (ext) and SdeWaC is even significant, according to the Fisher r-to-z transformation. The difference between WebKo (int) and SdeWaC is marginal, however, especially taking into account that WebKo is twice as big as SdeWaC. As in our previous study, the task performance relying on Wiki-de is significantly worse than when relying on WebKo or SdeWaC.

As none of the window lines in Figure 2 has reached an optimal correlation with a window size of 20 yet (i.e., the correlation values are still increasing), we enlarged the window size up to 100 words, in order to check on the most successful window size. For SdeWaC and Wiki-de, the correlations slighly increase, but for WebKo they do not increase. The optimal prediction is still performed using WebKo (ext) and a window size of 20 (see above).

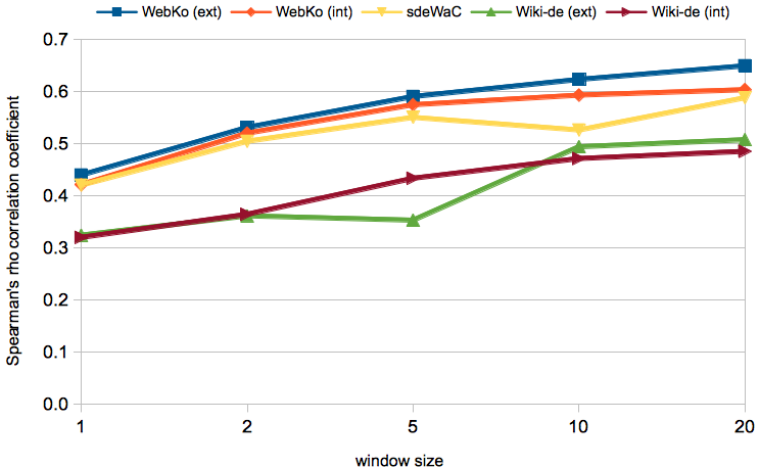


Figure 2: Window-based ρ correlations across corpora.

Figure 3 compares the three corpora on the same task, prediction of compositionality, but with regard to sub-corpora of similar sizes, thus looking at the effects of the corpus domains and corpus cleaning. The left-hand part of the plot breakes WebKo down into two sub-corpora of approx. 800 million words, and compares the prediction quality with

SdeWaC. This part of the plot once more confirms that "real" windows going beyond the sentence border clearly outperform window information restricted to the sentence level, reaching ρ correlations of 0.6494/0.6265 (WebKo (ext)) in comparison to 0.6048/0.5820 (WebKo (int)). The difference between the directly comparable WebKo (int) and SdeWaC vanishes, in contrast, as the WebKo (int) sub-corpora do not consistently outperform the SdeWaC ρ correlation of 0.5883. So in this study the corpus cleaning does not have an obvious effect on the semantic task.

The right-hand part of the plot breakes down SdeWaC into two sub-corpora of approx. 440 million words, and compares the prediction quality with Wiki-de (in both window modes). In this case, the difference in prediction quality persists: SdeWaC does not only outperform the directly comparable Wiki-de (int), $\rho = 0.4857$, but also Wiki-de (ext), $\rho = 0.5080$, which is in an advantegeous position, as Wiki-de (ext) is roughly of the same size as SdeWaC but window-based beyond the sentence border. So again, the web corpus Wiki-de performs worse than both other web corpora.

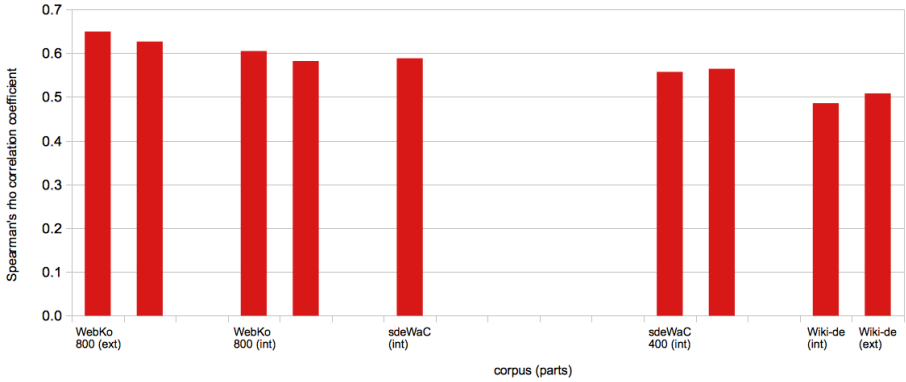


Figure 3: Window-based ρ correlations across corpus parts.

4 Summary and Discussion

The previous section presented two case studies to explore whether and how web corpora can be used to automatically acquire lexical-semantic knowledge from distributional information. For this purpose, we compared three German web corpora and one newspaper corpus, differing in domain, size and cleanness. Both case studies demonstrated that the corpora can indeed be used to model semantic relatedness: In the first study, we found up to 73/77% of verb/noun–association types within a 5-word window.[7]

[7] Schulte im Walde et al. (2008), Schulte im Walde and Melinger (2008) and Roth and Schulte im Walde (2008) present related work on co-occurrence analyses of the association norms.

Adhering to the standard assumption that *associates reflect meaning components of words*, we thus presented strong evidence for the *co-occurrence hypothesis* as well as for the *distributional hypothesis*, that semantic associations can be captured by corpus co-occurrence. In the second study, we could predict the compositionality of German noun-noun compounds with regard to their constituents with a ρ correlation of up to 0.6497.[8] We were therefore successful in inducing semantic compositionality, based on the distributional information in the web corpora.

Comparing the web corpora with a standard newspaper corpus (only done in the first study), we found that the availability of association knowledge in the newspaper corpus was actually compatible with the availability in the web corpora. So there was no effect concerning the more restricted domain, even though the association knowledge is completely open-domain.

In contrast, we found an effect of web corpus size in both studies: the larger the corpus, (1) the stronger the coverage of the association data, and (2) the better the prediction of compositionality. The differences between the larger (and noisier) WebKo data and the smaller (and cleaner) SdeWaC data were small (in study (1)) and negligible (in study (2)), indicating that the corpus cleaning had a positive effect on the corpus quality. Comparing the general web corpora WebKo and SdeWaC with the more structured and entry-based Wikipedia corpus, we demonstrated that the former are more suitable both (1) for general semantic knowledge and also (2) for predicting noun compound compositionality. We would have expected the encyclopeadic knowledge in Wiki-de to be compatible with the association knowledge in task (1) but the coverage is below that of the other two web corpora, even if we take the size into account.

Summarising, this article confirmed the suitability of web corpora for the automatic acquisition of lexical-semantic knowledge with regard to two very diverse semantic case studies. At the same time, we showed that newspaper corpora – if equal in size – provide suffient semantic relatedness knowledge, too, and are thus less restricted in their domains than commonly assumed.

Acknowledgements

The research presented in this article was funded by the DFG Heisenberg Fellowship SCHU-2580/1-1 (Sabine Schulte im Walde) and the DFG Sachbeihilfe SCHU-2580/2-1 (Stefan Müller). In addition, we thank the anonymous reviewer for valuable comments.

[8] Schulte im Walde et al. (2013) present related work on predicting the degree of compositionality of the noun-noun compounds.

References

Banko, M. and Brill, E. (2001). Scaling to Very Very Large Corpora for Natural Language Disambiguation. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and Lenci, A. (2010). Distributional Memory: A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36(4):673–721.

Boleda, G., Baroni, M., The Pham, N., and McNally, L. (2013). Intensionality was only alleged: On Adjective-Noun Composition in Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 35–46, Potsdam, Germany.

Borgwaldt, S. and Schulte im Walde, S. (2013). A Collection of Compound–Constituent and Compound Whole Ratings for German Noun Compounds. Manuscript.

Church, K. W. and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pages 76–83, Vancouver, Canada.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Clark, H. H. (1971). Word Associations and Linguistic Theory. In Lyons, J., editor, *New Horizon in Linguistics*, chapter 15, pages 271–286. Penguin.

Curran, J. and Moens, M. (2002). Improvements in Automatic Thesaurus Extraction. In *Proceedings of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 59–66, Philadelphia, PA.

de Deyne, S. and Storms, G. (2008a). Word Associations: Network and Semantic Properties. *Behavior Research Methods*, 40(1):213–231.

de Deyne, S. and Storms, G. (2008b). Word Associations: Norms for 1,424 Dutch Words in a Continuous Task. *Behavior Research Methods*, 40(1):198–205.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning: A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–763.

Evert, S. (2005). *The Statistics of Word Co-Occurrences: Word Pairs and Collocations*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Faaß, G. and Eckart, K. (2013). SdeWaC – a Corpus of Parsable Sentences from the Web. In *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pages 61–68, Darmstadt, Germany.

Faaß, G., Heid, U., and Schmid, H. (2010). Design and Application of a Gold Standard for Morphological Analysis: SMOR in Validation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 803–810, Valletta, Malta.

Fernández, A., Diez, E., Alonso, M. A., and Beato, M. S. (2004). Free-Association Norms for the Spanish Names of the Snodgrass and Vanderwart Pictures. *Behavior Research Methods, Instruments and Computers*, 36(3):577–583.

Ferrand, L. and Alario, F.-X. (1998). French Word Association Norms for 366 Names of Objects. *L'Ann'ee Psychologique*, 98(4):659–709.

Firth, J. R. (1957). *Papers in Linguistics 1934-51*. Longmans, London, UK.

Fleischer, W. and Barz, I. (2012). *Wortbildung der deutschen Gegenwartssprache*. de Gruyter.

Glenberg, A. M. and Mehta, S. (2008). Constraint on Covariation: It's not Meaning. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):241–264.

Harris, Z. (1968). Distributional Structure. In Katz, J. J., editor, *The Philosophy of Linguistics*, Oxford Readings in Philosophy, pages 26–47. Oxford University Press.

Herdagdelen, A. and Baroni, M. (2009). BagPack: A General Framework to Represent Semantic Relations. In *Proceedings of the EACL Workshop on Geometrical Models for Natural Language Semantics*, pages 33–40, Athens, Greece.

Heringer, H. J. (1986). The Verb and its Semantic Power: Association as the Basis for Valence. *Journal of Semantics*, 4:79–99.

Hickl, A., Williams, J., Bensley, J., Roberts, K., Rink, B., and Shi, Y. (2006). Recognizing Textual Entailment with LCC's GROUNDHOG System. In *Proceedings of the Second PASCAL Challanges Workshop on Recognising Textual Entailment*, pages 80–85, Venice, Italy.

Hirsh, K. W. and Tree, J. (2001). Word Association Norms for two Cohorts of British Adults. *Journal of Neurolinguistics*, 14(1):1–44.

Kiss, G. R., Armstrong, C., Milroy, R., and Piper, J. (1973). An Associative Thesaurus of English and its Computer Analysis. In *The Computer and Literary Studies*. Edinburgh University Press.

Klos, V. (2011). *Komposition und Kompositionalität*. Number 292 in Germanistische Linguistik. Walter de Gruyter, Berlin.

Lauteslager, M., Schaap, T., and Schievels, D. (1986). *Schriftelijke Woordassociatienormen voor 549 Nederlandse Zelfstandige Naamworden*. Swets and Zeitlinger.

Lieber, R. and Stekauer, P. (2009a). Introduction: Status and Definition of Compounding. In Lieber and Stekauer (2009b), chapter 1, pages 3–18.

Lieber, R. and Stekauer, P., editors (2009b). *The Oxford Handbook of Compounding*. Oxford University Press.

Marconi, D. (1997). *Lexical Competence*. MIT Press, Cambridge, MA.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a Continuum of Compositionality in Phrasal Verbs. In *Proceedings of the ACL-SIGLEX Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 73–80, Sapporo, Japan.

McCarthy, D., Koeling, R., Weeds, J., and Carroll, J. (2004). Finding Predominant Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

McKoon, G. and Ratcliff, R. (1992). Spreading Activation versus Compound Cue Accounts of Priming: Mediated Priming Revisited. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:1155–1172.

McNamara, T. P. (2005). *Semantic Priming: Perspectives from Memory and Word Recognition*. Psychology Press, New York.

Melinger, A. and Weber, A. (2006). Database of Noun Associations for German. URL: www.coli.uni-saarland.de/projects/nag/.

Miller, G. (1969). The Organization of Lexical Memory: Are Word Associations sufficient? In Talland, G. A. and Waugh, N. C., editors, *The Pathology of Memory*, pages 223–237. Academic Press, New York.

Nelson, D. L., Bennett, D., and Leibert, T. (1997). One Step is not Enough: Making Better Use of Association Norms to Predict Cued Recall. *Memory and Cognition*, 25:785–796.

Nelson, D. L., McEvoy, C. L., and Dennis, S. (2000). What is Free Association and What does it Measure? *Memory and Cognition*, 28:887–899.

Nelson, D. L., McEvoy, C. L., and Schreiber, T. A. (1998). The University of South Florida Word Association, Rhyme, and Word Fragment Norms. http://www.usf.edu/FreeAssociation/.

Padó, S. and Utt, J. (2012). A Distributional Memory for German. In *Proceedings of the 11th Conference on Natural Language Processing*, pages 462–470, Vienna, Austria.

Palermo, D. and Jenkins, J. J. (1964). *Word Association Norms: Grade School through College*. University of Minnesota Press, Minneapolis.

Pantel, P., Ravichandran, D., and Hovy, E. (2004). Towards Terascale Knowledge Acquisition. In *Proceedings of the 20th International Conference of Computational Linguistics*, pages 771–777, Geneva, Switzerland.

Plaut, D. C. (1995). Semantic and Associative Priming in a Distributed Attractor Network. In *Proceedings of the 17th Annual Conference of the Cognitive Science Society*, volume 17, pages 37–42.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus Portal for Search in Monolingual Corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 1799–1802, Genoa, Italy.

Rapp, R. (2002). The Computation of Word Associations: Comparing Syntagmatic and Paradigmatic Approaches. In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, Taiwan.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Roth, M. and Schulte im Walde, S. (2008). Corpus Co-Occurrence, Dictionary and Wikipedia Entries as Resources for Semantic Relatedness Information. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1852–1859, Marrakech, Morocco.

Russell, W. A. (1970). The complete German Language Norms for Responses to 100 Words from the Kent-Rosanoff Word Association Test. In Postman, L. and Keppel, G., editors, *Norms of Word Association*, pages 53–94. Academic Press, New York.

Russell, W. A. and Meseck, O. (1959). Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 6:191–211.

Schiehlen, M. (2003). A Cascaded Finite-State Parser for German. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, pages 163–166, Budapest, Hungary.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging using Decision Trees. In *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Schulte im Walde, S. (2008). Human Associations and the Choice of Features for Semantic Verb Classification. *Research on Language and Computation*, 6(1):79–111.

Schulte im Walde, S. (2010). Comparing Computational Approaches to Selectional Preferences: Second-Order Co-Occurrence vs. Latent Semantic Clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 1381–1388, Valletta, Malta.

Schulte im Walde, S., Borgwaldt, S., and Jauch, R. (2012). Association Norms of German Noun Compounds. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 632–639, Istanbul, Turkey.

Schulte im Walde, S. and Melinger, A. (2008). An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):89–128.

Schulte im Walde, S., Melinger, A., Roth, M., and Weber, A. (2008). An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation*, 6(2):205–238.

Schulte im Walde, S., Müller, S., and Roller, S. (2013). Exploring Vector Space Models to Predict the Compositionality of German Noun-Noun Compounds. In *Proceedings of the 2nd Joint Conference on Lexical and Computational Semantics*, pages 255–265, Atlanta, GA.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–123. Special Issue on Word Sense Disambiguation.

Siegel, S. and Castellan, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA.

Spence, D. P. and Owens, K. C. (1990). Lexical Co-Occurrence and Association Strength. *Journal of Psycholinguistic Research*, 19:317–330.

Springorum, S., Schulte im Walde, S., and Utt, J. (2013). Detecting Polysemy in Hard and Soft Cluster Analyses of German Preposition Vector Spaces. In *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pages 632–640, Nagoya, Japan.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning: Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

von der Heide, C. and Borgwaldt, S. (2009). Assoziationen zu Unter-, Basis- und Oberbegriffen. Eine explorative Studie. In *Proceedings of the 9th Norddeutsches Linguistisches Kolloquium*, pages 51–74.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1646–1652, Marrakech, Morocco.

Author Index

Michael Beißwenger
Technische Universität Dortmund
beisswenger@tu-dortmund.de

Chris Biemann
Technische Universität Darmstadt
biem@cs.tu-darmstadt.de

Felix Bildhauer
Freie Universität Berlin
felix.bildhauer@fu-berlin.de

Stefan Evert
Friedrich-Alexander-Universität Erlangen-Nürnberg
stefan.evert@fau.de

Dirk Goldhahn
Universität Leipzig
dgoldhahn@informatik.uni-leipzig.de

Bryan Jurish
Berlin-Brandenburgische Akademie der Wissenschaften

jurish@bbaw.de

Lothar Lemnitzer
Berlin-Brandenburgische Akademie der Wissenschaften
lemnitzer@bbaw.de

Stefan Müller
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
muellesn@ims.uni-stuttgart.de

Uwe Quasthoff
Universität Leipzig
quasthoff@informatik.uni-leipzig.de

Roland Schäfer
Freie Universität Berlin
roland.schaefer@fu-berlin.de

Sabine Schulte im Walde
Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart
schulte@ims.uni-stuttgart.de

Johannes Simon
Technische Universität Darmstadt
johannes.simon@gmail.com

Leonard Swiezinski
Technische Universität Darmstadt
leonard@swiezinski.de

Kay-Michael Würzner
Berlin-Brandenburgische Akademie der Wissenschaften
wuerzner@bbaw.de

Torsten Zesch
Universität Duisburg-Essen
torsten.zesch@uni-due.de