Originalveröffentlichung in: Engelberg, Stefan/Lobin, Henning/Steyer, Kathrin/Wolfer, Sascha (Hrsg.): Wortschätze. Dynamik, Muster, Komplexität. - Berlin [u.a.]: de Gruyter, 2018. S. 245-266
(Jahrbuch / Institut für Deutsche Sprache 2017)

Carolin Müller-Spitzer/Sascha Wolfer/Alexander Koplenig (Mannheim)

Quantitative Analyse lexikalischer Daten

Methodenreflexion am Beispiel von Wandel und Sequenzialität

Abstract: Quantitativ ausgerichtete empirische Linguistik hat in der Regel das Ziel, große Mengen sprachlichen Materials auf einmal in den Blick zu nehmen und durch geeignete Analysemethoden sowohl neue Phänomene zu entdecken als auch bekannte Phänomene systematischer zu erforschen. Das Ziel unseres Beitrags ist es, anhand zweier exemplarischer Forschungsfragen methodisch zu reflektieren, wo der quantitativ-empirische Ansatz für die Analyse lexikalischer Daten wirklich so funktioniert wie erhofft und wo vielleicht sogar systembedingte Grenzen liegen. Wir greifen zu diesem Zweck zwei sehr unterschiedliche Forschungsfragen heraus: zum einen die zeitnahe Analyse von produktiven Wortschatzwandelprozessen und zum anderen die Ausgleichsbeziehung von Wortstellungs- vs. Wortstrukturregularität in den Sprachen der Welt. Diese beiden Forschungsfragen liegen auf sehr unterschiedlichen Abstraktionsebenen. Wir hoffen aber, dass wir mit ihnen in großer Bandbreite zeigen können, auf welchen Ebenen die quantitative Analyse lexikalischer Daten stattfinden kann. Darüber hinaus möchten wir anhand dieser sehr unterschiedlichen Analysen die Möglichkeiten und Grenzen des quantitativen Ansatzes reflektieren und damit die Interpretationskraft der Verfahren verdeutlichen.

1 Einleitung

„The language looks different when you look at a lot of it at once." (Sinclair 1991, S. 100). Dies ist, denken wir, der Grundgedanke der quantitativen Linguistik: Man möchte große Mengen von Sprachdaten in den Blick nehmen und aus dieser Vogelperspektive neue Erkenntnisse über Sprache gewinnen. Die empirische Ausrichtung beinhaltet dabei eine Hinwendung zum Sprachgebrauch als zentralem Forschungsgegenstand (vgl. u.a. Gries 2015, S. 725; Allan/Robinson 2012; Bybee 2006, 2015 u.v.m.). Die linguistische Forschung verfügt heute über eine Datenmenge, die für die Linguistinnen und Linguisten bis in die 1990er Jahre noch der Phantasiewelt angehörte. Dies gilt auch für das Deutsche: Allein das Deutsche Referenzkorpus (DeReKo) hat eine Größe von aktuell fast 32 Milliarden lau-

DOI 10.1515/9783110579963-014

fenden Wörtern.[1] Allerdings sind es nur bestimmte Arten von Sprachdaten, die in großer Masse verfügbar sind. So sind beispielsweise in DeReKo v.a. Zeitungstexte enthalten und nur wenige fiktionale Texte (vgl. Kupietz/Lüngen 2014). Diese prozentuale Verteilung hängt nicht damit zusammen, dass unsere Sprachproduktion sich ähnlich verteilt, sondern ist auf eine pragmatische Entscheidung zurückzuführen: Um mit überschaubaren Kosten gegenwartsnah eine große Menge an Sprachdaten zur Verfügung stellen zu können, sind Zeitungstexte das Mittel der Wahl. Dagegen haben beispielsweise die mündlichen Korpora des IDS eine Größe von weniger als zehn Millionen Token. Eigentlich müsste aber die mündliche Sprache für (viele) Aspekte des Sprachgebrauchs der zentrale Forschungsgegenstand sein. Daher wäre es theoretisch wünschenswert, eine sehr große Anzahl von Feldforscherinnen und -forschern in die Welt hinauszuschicken, um Daten in allen denkbaren Arten von Sprachsituationen zu erheben und diese der Forschungsgemeinschaft zur Verfügung zu stellen (vgl. u.a. Schmid 2015; Arppe/Järviki 2007, S. 10; Szmrecsanyi 2016, S. 157). Da dies praktisch aber nicht möglich ist, hat es sich für die empirische quantitative Analyse lexikalischer Daten als Usus etabliert, große schriftsprachliche Korpora als Stellvertreter für Sprachgebrauch heranzuziehen (vgl. u.a. Koplenig 2016; S. 9–39; Evert 2006; Gilquin/Gries 2009).

Wir konzentrieren uns in diesem Beitrag auf die quantitative Analyse lexikalischer Daten. Das Ziel dieses Beitrags ist es nicht, dieses gesamte Feld zu beschreiben, sondern anhand zweier exemplarischer Forschungsfragen methodische Aspekte zu reflektieren. Wir greifen zu diesem Ziel die zwei Themenbereiche Sprachwandel und Sequenzialität heraus. Am Beispiel der Analyse der neuen verbalen Wortbildungsmittel *gegen-* und *fremd-* können wir illustrieren, wie attraktiv auf der einen Seite der quantitativ-empirische Ansatz zur Erforschung dieses Phänomens ist, wie schwierig seine Umsetzung allerdings in der Praxis sein kann (Kap. 2). Am Beispiel der Sequenzialität lässt sich auf der anderen Seite zeigen, wie gewinnbringend es sein kann, mit quantitativen Methoden große Sprachmengen in den Blick zu nehmen. Wir weisen hier (Kap. 3) anhand eines großen Bibelkorpus nach, dass es in allen untersuchten Sprachen einen *Trade-off* zu beobachten gibt zwischen den Informationen, die über die Wortstellungs- und die Wortstrukturregularität vermittelt werden: Jene Sprachen, die viel Information über die Wortstellung vermitteln, übertragen umso weniger über die Wortstruktur und andersherum. Eine solche Analyse v.a. in der Breite von Sprachen ist jedoch wiederum nur mit bestimmten Korpora überhaupt möglich. Unser Beitrag endet mit einer Schlussbemerkung.

1 www1.ids-mannheim.de/kl/projekte/korpora/ (Stand: 30.5.2017). Die Angabe der Größe von DeReKo bezieht sich auf den Stand vom 8.3.2017.

2 Die quantitative Untersuchung von Sprachwandelprozessen am Beispiel der verbalen Wortbildungsmittel *gegen-* und *fremd-*

Sprachwandelprozesse scheinen prädestiniert dazu, mit quantitativen Analysemethoden erforscht zu werden. Gerade der Aufbau großer Textkorpora ermöglicht es, sozusagen mit einem Fangnetz alle lexikalischen Innovationen im Zeitverlauf zu detektieren: Es werden einfach jene neuen Wortoberflächen herausgefiltert, die zeitlich vorher noch nicht vorkamen (vgl. u.a. Keibel/Hennig/Perkuhn 2011; Kerremans et al. 2012; Rohrdantz et al. 2012; Würschinger et al. 2016).[2] Auch Änderungen im Bedeutungsspektrum können über die Veränderung der Kontexte der einzelnen Wörter aufgespürt werden, auch wenn solche Verfahren ungleich datenintensiver sind (vgl. Gulordava/Baroni 2011). Zwei Herausforderungen sind dabei allerdings zu beachten: i) Die Zipf-Verteilung sprachlicher Daten, und ii) die Datenlage.

i) Zipfsche Verteilungen zeichnen sich grob gesagt dadurch aus,[3] dass einige wenige Einheiten sehr häufig auftreten und extrem viele Einheiten äußerst selten (vgl. z.B. Engelberg 2015, S. 206–220). Die Häufigkeiten von Wörtern in einem Korpus sind typischerweise als Zipf-nahe Verteilung organisiert, d.h. sehr wenige lexikalische Einheiten kommen sehr häufig vor, sehr viele dagegen sehr selten. Der Anteil an Hapax legomena, d.h. an lexikalischen Einheiten, die in einem (Sub-)Korpus nur einmal vorkommen, liegt dabei meist etwa bei der Hälfte der Lexeme. In diesem Bereich der sehr selten vorkommenden Lexeme herrscht naturgemäß viel Varianz. Analysiert man beispielsweise alle Ausgaben der Wochenzeitung „Die Zeit" und filtert nur die neuen Wortoberflächen heraus, sind es jedes Jahr etwa 20% aller Lexeme, die davor noch nie vorkamen. Dies sind zum größten Teil keine lexikalischen Innovationen, sondern Bestandteile der sprachlichen Varianz, die in dem Bereich der seltenen Wörter sehr hoch ist. Eine erste Aufgabe bei der Analyse lexikalischer Innovationen ist es daher, die normale sprachliche Varianz von sprachlicher Innovation zu trennen. Weinreich drückt dies so aus:

> Not all variability and heterogeneity in language structure involves change; but all change involves variability and heterogeneity. (Weinreich/Labov/Herzog 1968: 52, vgl. auch Szmrecsanyi 2016)

2 Genau diesen Vorgang – neue Wortoberflächen für bestimmte Jahre nach bestimmten Vorgaben herauszufiltern – kann man für die Wochenzeitung „Die Zeit" über das Tool „Wortschatzwandel in der Zeit" in OWID^{plus} explorieren: www.owid.de/plus/wwzeit2016/ (Stand: 30.5.2017).
3 Eine genaue statistische Beschreibung findet sich in Clauset/Shalizi/Newman (2009).

Das bedeutet in der Praxis, dass lexikalische Innovationen mit häufigkeitsbasierten Analysemethoden schwer von anderen Phänomenen sprachlicher Varianz zu trennen sind und dass in der Regel eine nicht unbeträchtliche Menge manueller Arbeit investiert werden muss, um aus der Fülle neuer Wortoberflächen Kandidaten für Neologismen zu ermitteln.

ii) Zur Datenlage: Viele Innovationsprozesse spielen sich nicht in der geschriebenen Sprache ab, wie sie in den großen verfügbaren Textkorpora erfasst sind. Ein schönes Beispiel ist das „Jugendwort des Jahres", welches Jahr für Jahr bei den Jugendlichen eher für Kopfschütteln sorgt. Ein Kommentar zur entsprechenden Auszeichnung 2015 in Österreich:

> Seit einigen Tagen steht das österreichische Jugendwort des Jahres 2015 fest. Falls ihr es noch nicht mitbekommen habt und euch das Kopfschütteln bis jetzt erspart geblieben ist: Der neue Titelträger ist ‚zach'. Das ist tragisch und an Lameness kaum zu überbieten— nicht weil ‚zach' ein beschissenes Wort ist, sondern viel eher, weil es genauso gut das Jugendwort des Jahres 2005 hätte sein können.[4]

Auch wenn man die Erhebung solcher angeblich neuen Wörter methodisch verbessern würde, bleibt es ein Grundproblem, dass sich solche sprachlichen Innovationsprozesse in der mündlichen Sprache entwickeln und dass es sehr lange dauern kann, bis sie sich in überregionalen Zeitungen finden (falls dies überhaupt eintritt). Doch selbst wenn man sich allgemein gesagt auf die „Sprache der Öffentlichkeit" konzentriert, ist die Datenlage nicht immer so reichhaltig wie erhofft. Zunächst einmal benötigt man für die Detektion lexikalischer Innovationen diachrone Korpora. Für das Deutsche sieht die Situation in dieser Hinsicht nicht so gut aus wie im Englischen, wo z.B. mit dem „Corpus of Historical American English"[5] und vielen anderen historisch ausgerichteten Korpusprojekten (siehe z.B. die Studie in Lansdall-Welfare et al. 2017) interessante Datenquellen zur Verfügung stehen. Für das Deutsche stehen uns erst ab den 1990er Jahren große Datenmengen zur Verfügung. Zum zweiten braucht man für die quantitative Forschung Zugriff auf die Rohdaten. Um das Bild vom Anfang des Abschnitts wieder aufzugreifen: Wenn man neue Wortoberflächen oder neue Bedeutungen herausfiltern will, muss man das Fangnetz auch auswerfen können.[6] Viele Kor-

4 http://www.vice.com/alps/read/goennung-ist-das-wahre-jugendwort-des-jahres-2015-839 (Stand: 31.5.2017).

5 http://corpus.byu.edu/coha/ (Stand: 31.5.2017).

6 Ehrlicherweise hieße es eher *dürfen*: Man könnte beispielsweise das komplette Zeit-Korpus herunterladen, man darf es aber nicht. Und als öffentlich-finanzierte wissenschaftliche Institution halten wir uns an solche Vorgaben des Urheberrechts.

pus-Lizenzvereinbarungen sind allerdings so konstruiert, dass das nicht möglich ist. Und selbst wenn alle diese Bedingungen gegeben sind, werden wir im folgenden Abschnitt sehen, dass die Verteilung sprachlicher Daten ein nicht zu unterschätzender Faktor ist.

Wir werden uns im Folgenden mit einem potenziellen lexikalischen Innovationsprozess beschäftigen, der methodisch etwas interessanter ist als schlicht neue Wortoberflächen zu ermitteln, und zwar die Produktivitätsentwicklung von Morphemen. Genauer wollen wir *gegen-* und *fremd-* in den Blick nehmen und analysieren, ob sich diese Präfixe in den letzten 25 Jahren zu produktiven verbalen Wortbildungsmitteln entwickelt haben oder nicht. Ausgangspunkt unserer Studie ist ein Aufsatz von Klosa (2003), in der sie nachweisen konnte, dass mit *gegen*-Verben ein neues Wortbildungsmuster im Deutschen entstanden ist, d.h. dass Verben wie *gegenlenken*, *gegenfinanzieren* oder *gegenchecken* ein seit den 1990er Jahren nachweisbares neues Muster bilden. Aus unserer eigenen Sprachintuition heraus hatten wir außerdem die Hypothese, dass sich *fremd-* ebenfalls zu einem produktiven Element entwickelt hat, wie es sich in Verben wie *fremdschämen*, *fremderziehen* oder *fremdtwittern* zeigt. Unsere Fragestellungen waren dabei:

- Kann man (mit den heute gegenüber 2003 ausgereifteren Analysemethoden) quantitativ gut nachzeichnen, dass *gegen-* in den letzten 25 Jahren ein produktives verbales Wortbildungselement geworden ist? Zeigt sich das Gleiche für *fremd-*?
- Zeigen sich unterschiedliche Wortbildungsbedeutungen bei den *fremd*-Verben (als Indikator für die Breite des Wortbildungsparadigmas)?
- Ist die korpuslinguistische Evidenz aussagekräftig genug, um zur Beantwortung der Fragen allein auf diese Daten zurückzugreifen?

Die Datengrundlage für diese Studie ist DeReKo 2014-II.[7] Eine erste Übersicht über die Entwicklung der *gegen-* und *fremd*-Verben zeigt, dass die Anzahl der Infinitive, die mit den Präfixen verbunden werden, deutlich steigt (vgl. Abb. 1). Um welche Infinitive es sich dabei handelt, kann interaktiv exploriert werden.[8] Es finden sich neben geläufigeren Verben wie *gegenfinanzieren* oder *gegenargumentieren* auch (vielleicht) weniger geläufige wie *gegenbalancieren* oder *gegenregieren*. So scheint es berechtigt, dass *gegen-* mittlerweile als neues verbales Wortbildungs-

7 Wir danken unserem Kollegen Roman Schneider für die Ermittlung der Daten über die Korpusgrammatik-Datenbank. Für nähere Infos zu den DeReKo-Releases siehe www1.ids-mannheim.de/direktion/kl/projekte/korpora/releases.html?L=1 (Stand: 31.5.2017).

8 https://owid.shinyapps.io/Inf-f3/, https://owid.shinyapps.io/Inf-g3/ (Stand: 31.5.2017).

mittel in der „Wortbildung der deutschen Gegenwartssprache“ (Fleischer/Barz 2012, S. 412) aufgeführt ist. „64%“ der *gegen*-Verben seien „erst seit 1990 belegt; bei den früher belegten nimmt der Gebrauch bis heute stetig zu“ (ebd., S. 412).

Auch bei den *fremd*-Verben zeigt sich eine große Vielfalt. Neben den im Duden online[9] verzeichneten *fremdbeziehen*, *fremdgehen*, *fremdficken*, *fremdschämen*, *fremdsteuern*, *fremdvergeben* finden sich noch viele weitere wie *fremdkomponieren*, *fremdhören*, *fremdknutschen*, *fremdtwittern* oder *fremdwidmen*.

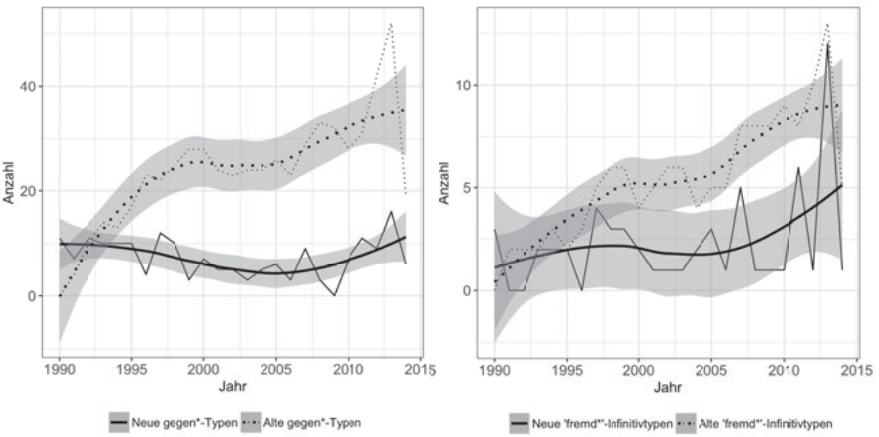


Abb. 1: Anzahl der Kompositionsinfinitive mit *gegen-* (links) bzw. *fremd-* (rechts) in DeReKo von 1990 bis 2015. Die untere Linie zeigt die neu hinzukommenden Infinitivtypen, die obere Linie die Anzahl der Kompositionsinfinitive insgesamt

Die bisherigen Analysen scheinen also die These zu stützen, dass sich mit *fremd-* und *gegen*-Verben zwei neue Wortbildungsmuster seit Ende des 20. Jahrhunderts herausgebildet haben. Nach Klosa (2003) ist es außerdem ein Zeichen für die Komplexität eines Wortbildungsmusters, wenn es über mehrere Wortbildungsbedeutungen verfügt. Für die *gegen*-Verben konnte sie diese Komplexität bereits nachweisen (Klosa 2003, S. 484–492). Auch für *fremd-* zeigen sich unterschiedliche Bedeutungsaspekte:

9 www.duden.de (Stand: 31.5.2017). Die hier genannten *fremd*-Verben beziehen sich auf den Stand von Februar 2017.

1. *fremd-* in der Bedeutungsgruppe „sich außerhalb eines bestehenden Systems etc. bewegen", „eine Bindung verlassen":
   - Vertreter: *fremdgehen*, *fremdknutschen*, *fremdficken*, *fremdhören*
   - Exemplarischer Beleg: „Die echte Reichweite hängt tagtäglich allein von jenen Radiokonsumenten ab, die per Knopfdruck automatisch ihren Sender wählen oder des Öfteren auch *fremdhören*." (Oberösterreichische Nachrichten, 20.4.2000; Eine harte Nuss)
2. *fremd-* in der Bedeutungsgruppe „eine Tätigkeit, die man normalerweise selber übernimmt, jm. anderem übergeben":
   - *fremdbetreuen*, *fremderziehen*, *fremdschreiben*
   - Exemplarischer Beleg: „Dumas hat ein Millionenpublikum – weltweit und in Frankreich. Aber unumstritten ist seine Umbettung nicht. Zu Lebzeiten war Dumas als Vielschreiber verschrien, als ‚Romanfabrikant', der viel *fremdschreiben* ließ." (die tageszeitung, 30.11.2002, S. 10, Ressort: Ausland; In der Gruft mit Zola und Hugo)
3. *fremd-* in der Bedeutungsgruppe „anstelle von jemand anderem selber etwas tun":
   - *fremdschämen*, *fremdtwittern*
   - Exemplarischer Beleg: „Da sich jeder Nutzer kostenlos und ohne Überprüfung ein Konto anlegen und dabei jeden freien Namen registrieren kann, gibt es auch zahlreiche so genannte Fakes, also Nutzer, die unter einem bekannten Namen *fremdtwittern*. Prominente Fake-Beispiele: [...] ZDF-Korrespondent Claus Kleber. Da Kleber selbst kein Konto bei dem Dienst hatte, sicherte sich Birte Oldenburg (‚Zuckerhund') das [sic] Namen ‚claus_kleber'. Seitdem schreibt sie als falscher Kleber unterhaltsame Tweets wie ‚Sitzt meine Frisur eigentlich noch, Gundula Gause?' oder ‚Ich habe jetzt die Handynummer von Obama'." (Hannoversche Allgemeine, 13.11.2008, S. 15; Was zwitscherst Du gerade?)
4. *fremd-* in der Bedeutungsgruppe „von jemand anders gemacht, z.T. von außen aufgezwungen, fremdbestimmt":
   - *fremddiktieren*, *fremdevaluieren*, *fremdentscheiden*
   - Exemplarischer Beleg: „Fazit: Lieber regelmässig mit Events wie WEF, WM- und EM-Anlässen präsent sein als einmal eine hochmütige, *fremddiktierte* Riesenshow und dann in Vergessenheit zu geraten." (Die Südostschweiz, 26.2.2013, S. 21; Die olympischen Ringe)

Kann damit schon als nachgewiesen gelten, dass die *gegen-* und die *fremd*-Verben ein Beispiel für ein quantitativ gut nachweisbares Phänomen des Wortschatzwandels sind? Schaut man sich die Häufigkeit der Verben normiert auf die Gesamtgröße der Korpora an, verändert sich das Bild deutlich (vgl. Abb. 2). Zwar gibt es

absolut immer mehr Verben, die mit *gegen-* oder *fremd-* gebildet werden, die Korpora werden allerdings auch wesentlich größer und relativ zur Korpusgröße werden es weniger Vorkommen. Damit sinkt die normierte Häufigkeit. Es lässt sich also nicht eindeutig feststellen, ob nur mehr Verben mit den beiden Wortbildungselementen gefunden wurden, weil immer mehr Texte für die Analyse zur Verfügung standen, oder ob diese beiden Muster tatsächlich in den 1990er Jahren produktiv geworden sind.[10] Die Datenlage zeigt also nicht eindeutig, dass seit den 1990er Jahren ein „Boom" (Klosa 2003, S. 478) der *gegen*-Verben stattgefunden hat.

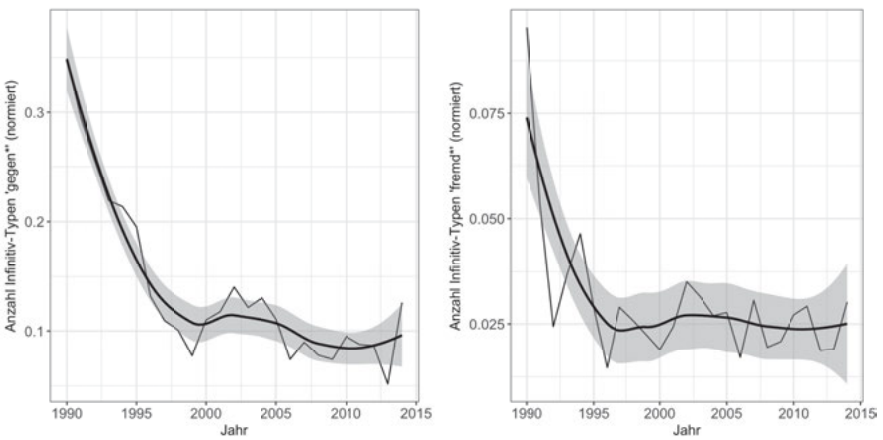


Abb. 2: Anzahl der Kompositionsinfinitive mit *gegen-* (links) bzw. *fremd-* (rechts) in DeReKo von 1990 bis 2015 normiert zur Korpusgröße

Damit kann zwar nach wie vor konstatiert werden, dass es sich bei beiden Wortbildungselementen um ein produktives Wortbildungsmuster handelt, da es viele Verben gibt, die mit ihnen kombiniert werden können, und auch ein Spektrum verschiedener Wortbildungsbedeutungen zu erkennen ist. Was aber mit den Analysen in DeReKo nicht nachgewiesen werden kann, ist, ob es sich dabei um ein Phänomen lexikalischer Innovation handelt. Auch in den Belegen findet man

10 Als Einschränkung zu diesen Mengenangaben muss man ansehen, dass Partikelverben wie die *gegen-* und *fremd-* Verben quantitativ immer schwer zu analysieren sind, da leicht Fälle in der Analyse übersehen werden können, in denen das trennbare Präfix nicht richtig erkannt wurde. Allerdings werden diese beiden Verbgruppen noch nicht oft mit abgetrenntem Präfix verwendet, sodass es dadurch nicht zu deutlichen Verzerrungen gekommen sein sollte.

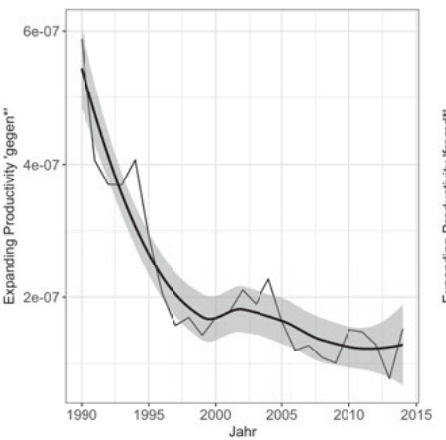
keine Indizien dafür, dass den Verben ein Innovationsstatus zugeschrieben wird, indem sie beispielsweise erklärt oder in Anführungszeichen verwendet werden.

Eine Idee zur Beantwortung dieser Frage könnte sein, andere Verfahren heranzuziehen, mit denen man die Produktivität eines Morphems quantitativ nachweisen kann. Solche Verfahren (grundlegend dazu Baayen 2009) sind zwar v.a. zum Vergleich der Produktivität unterschiedlicher Morpheme entwickelt worden, können aber auch zum Vergleich der Produktivität eines Morphems in unterschiedlichen Zeitschnitten verwendet werden. Zwei Maße und daraus abgeleitete Analysen wollen wir hier kurz demonstrieren: Zum einen die „expanding productivity" nach Baayen (2009, S. 905 f.) und zum anderen die Type-Token-Ratio der einzelnen Wortbildungsmuster über die Zeit. Die Grundidee der Produktivitätsmaße nach Baayen ist, dass die Hapax legomena in einem Korpus als Indikator für Wortschatzwachstum und für das lexikalische Innovationspotenzial angesehen werden können. Möchte man nun die Produktivität eines einzelnen Phänomens untersuchen, wird die Anzahl der Hapax legomena des zu untersuchenden Phänomens (also hier der *gegen-* oder *fremd*-Verben, die in einem Zeitabschnitt nur einmal vorkommen) durch die Anzahl der Hapax legomena in diesem Korpusausschnitt insgesamt geteilt. Damit wird sozusagen der Anteil dieses Phänomens am Innovationspotenzial insgesamt untersucht, d.h. der Beitrag des Morphems zum Wachstum des Wortschatzes. Ein Morphem würde über die Zeit als wachsend produktiv gelten, wenn es einen größeren Anteil an den Hapax legomena einnimmt.

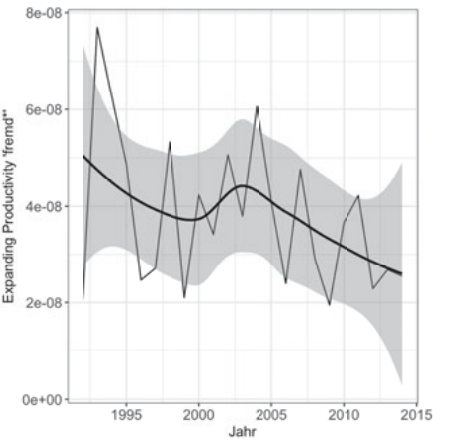
Ein anderes Hilfsmittel könnte sein, die Type-Token-Ratio (TTR) des Phänomens über die Zeit zu untersuchen. Die TTR gilt als ein Maß, um lexikalische Vielfalt zu messen (siehe für eine frühe Untersuchung Templin 1957; zur diachronen Anwendung von TTR-Analysen vgl. Koplenig 2015). Eine TTR-Analyse der *gegen-* und *fremd*-Verben könnte zeigen, ob es z.B. am Anfang der 1990er Jahre nur wenige Typen von *gegen-* und *fremd*-Verben gibt, die dafür aber häufiger vorkommen, und ob deren Vielfalt sich in den folgenden 25 Jahren erhöht. Die Analysen sind in Abbildung 3 zu sehen. Sollte die Produktivität und die TTR über die Jahre steigen, müssten die Kurven auch einen steigenden Verlauf zeigen. Dies ist nicht der Fall. Im Grunde zeigen die Plots mehr Rauschen als eindeutige Befunde, was unseres Erachtens nach vor allem daran liegt, dass die Phänomene insgesamt zu selten sind (vgl. die Ausführungen oben zur Zipf-nahen Verteilung sprachlicher Daten). Welche Auswirkungen die Seltenheit eines Phänomens in der quantitativen Analyse hat, lässt sich auch an den unteren beiden Plots in Abbildung 3 sehen: Hier haben wir die Unsicherheit des Samplings explizit gezeigt, d.h. jeder Samplingvorgang ist durch eine Linie abgetragen. Dabei kann man deutlich sehen, dass es gerade bei den *fremd*-Verben sehr große Unterschiede beim Sampling gibt, d.h. dass es methodisch gesprochen nicht robust ist.

Als Resümee aus den Analysen lässt sich also nur ziehen: Mit den *gegen-* und *fremd*-Verben liegen zwei produktive Wortbildungsmuster vor. Wann sie produktiv wurden, lässt sich mit den Korpusanalysen basierend auf DeReKo nicht zweifelsfrei beantworten. Fraglich ist dabei, ob das Phänomen in den Quellen und mit den hier angewandten Verfahren nicht zu messen ist oder ob es sich wirklich nicht um ein Sprachwandelphänomen handelt, denn drei Dinge passieren gleichzeitig: Die Korpusgröße wächst, die Korpusvielfalt wird größer und wir finden mehr *gegen-* und *fremd*-Verben. Ob es diese Verben aber nicht auch schon in den 1970er Jahren so häufig gegeben hat, wissen wir nicht, da aus dieser Zeit deutlich weniger Korpusmaterial vorliegt und fast keine Treffer gefunden wurden. Methodisch sollen diese Analysen verdeutlichen, dass die Antwort, obwohl ein sehr großes Korpus vorliegt und bereits entsprechende quantitative Verfahren zur Analyse des in Betracht kommenden linguistischen Phänomens entwickelt wurden, leider nicht immer so klar auf der Hand liegt wie erhofft.

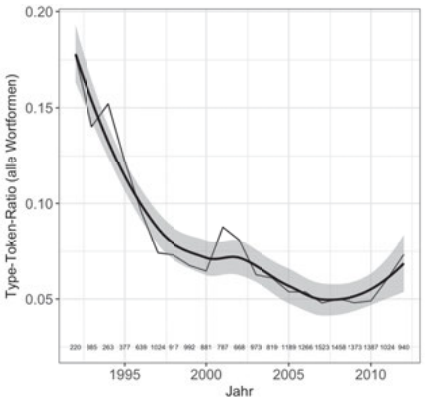
Die Forderung, dass linguistische Forschung empirisch abgesichert möglichst zeitnah neue Entwicklungen beschreiben soll, ist allerdings berechtigt. Deshalb könnte es durchaus lohnend sein, andere Quellen als die großen zeitungslastigen Korpora als Quelle heranzuziehen. In diesem Sinne wollen wir das Kapitel nicht beenden, ohne zumindest Ideen zu skizzieren, wie man den Innovationsstatus der *gegen-* und *fremd*-Verben weiter untersuchen könnte und wie man damit auch generell das Problem der „data sparseness" bei lexikalischen Innovationen zumindest teilweise auffangen könnte.



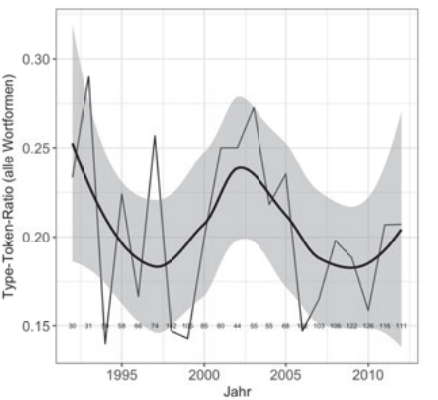
a) Entwicklung der „expanding producitivity" für *gegen-*



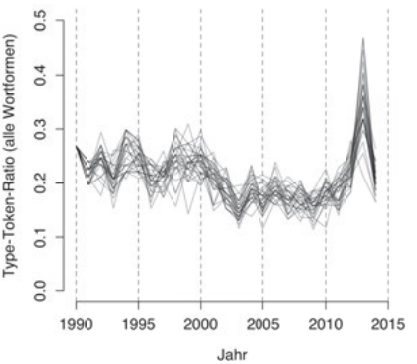
b) Entwicklung der „expanding producitivity" für *fremd-*



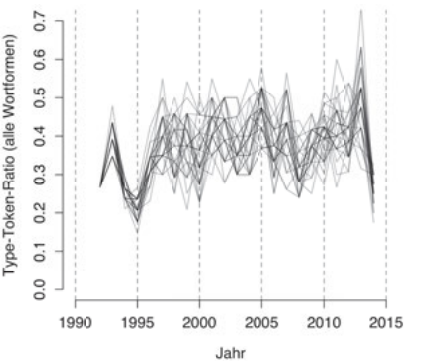
c) *gegen-*: Type-Token-Ratios über Jahre (1 Type = 1 Wortform, einmaliges Sampling, Smoothing über –/+ 2 Jahre)



d) *fremd-*: Type-Token-Ratios über Jahre (1 Type = 1 Wortform, einmaliges Sampling, Smoothing über –/+ 2 Jahre)



e) *gegen-*: Type-Token-Ratios über Jahre



f) *fremd-*: Type-Token-Ratios über Jahre

Abb. 3: Produktivitätsanalysen für *gegen-* (linke Spalte) und *fremd*-Verben (rechte Spalte). Erste Zeile: Entwicklung der „expanding productivity" nach Baayen (2009, S. 905–906). Zweite Zeile: Entwicklung der TTR über Jahre mit mehrmaligen Sampling und Smoothing +/– 2 Jahre. Dritte Zeile: TTR über Jahre mit abgebildetem mehrmaligen Sampling[11]

11 Sampling wird notwendig, da die TTR wie andere Maße zur Messung lexikalischer Vielfalt stark von der Korpusgröße abhängt (vgl. Tweedie/Baayen 1998). Um diesen Effekt zu kontrollieren, kann man sich an der kleinsten Subkorpusgröße in der Stichprobe orientieren und in den anderen Subkorpora (= Jahren) verhältnismäßig gleich viele Token sampeln. Man schrumpft sozusagen künstlich die größeren Korpusabschnitte, damit man sie mit dem kleinsten Abschnitt vergleichen kann (vgl. Koplenig 2015).

Eine Lösung, die zumindest in der Anglistik auch schon praktiziert wird (vgl. Kerremans/Stegmayr/Schmid 2012), ist die Analyse von Webkorpora zur Detektion von Neologismen. Dies ermöglicht die Analyse einer größeren Bandbreite von Sprachdaten. Gleichzeitig ist es bei der Analyse von Webkorpora schwieriger, Einzeltexte zeitlich sicher einzuordnen und weitere notwendige Metadaten verlässlich zu ermitteln. Bisherige Analysen wie in Würschinger et al. (2016) oder Kerremans (2015) zeigen, dass die Detektion von Neologismen und vor allem auch die Diffusionsprozesse verschiedener Formen besonders dann gut in Webkorpora untersucht werden können, wenn es sich um sogenannte „loaded words" handelt, d.h. Wörter, die emotional oder politisch aufgeladen sind. Es liegt nahe, dass genau solche Wörter gut im schriftsprachlichen Diskurs über soziale Medien zu verfolgen sind, da sie ja selbst Gegenstand oder Sinnbild einer virulenten Diskussion sind. Ob man allerdings in Webkorpora solche linguistischen Fragestellungen wie die Produktivitätsentwicklung einzelner Wortbildungsmuster, für die man u.a. verlässliches Wortartentagging benötigt, sinnvoll verwenden kann, ist aus unserer Sicht fraglich.

Eine andere Idee besteht darin, Korpusdaten durch andere Formen von Evidenz anzureichern. Für die *gegen-* und *fremd*-Verben könnte man beispielsweise einen Akzeptabilitätstest durchführen, mit dem man linguistisch nicht vorgebildete Personen die Akzeptabilität anderer potenzieller Verbbildungen mit *gegen-* und *fremd-* einschätzen lässt, die noch nicht in Korpora belegt sind (für ein vergleichbares methodisches Vorgehen mit Partikelverben sei auf den Beitrag von Sabine Schulte im Walde in diesem Jahrbuch verwiesen). Gleichzeitig könnte man die Teilnehmenden befragen, für wie alt sie das Verb halten (wenn sie es als akzeptabel einstufen). Auch das könnte zumindest Hinweise darauf geben, wie flexibel das Wortbildungselement einsetzbar ist und für wie neu das Verb gehalten wird. Für die Detektion von Neologismen, die sich noch nicht in der Schriftsprache finden, könnte man außerdem gezielt bestimmte Blogs etc. nach Diskussionen zu sprachlichen Themen absuchen, um mögliche Kandidaten sprachlicher Innovation zu finden.

Für den Moment lässt sich allerdings zu den anfangs gestellten Fragestellungen nur sagen:

a) Wir können quantitativ nicht nachzeichnen, dass *gegen-* in den letzten 25 Jahren ein produktives Wortbildungselement geworden ist. Das gleiche gilt für *fremd-*.
b) Beide Verbgruppen zeigen verschiedene Wortbildungsbedeutungen.
c) Die korpuslinguistische Evidenz der Daten, die wir zur Verfügung hatten, ist nicht groß genug, um die Frage der Produktivität der beiden verbalen Wortbildungselemente abschließend zu beantworten.

Die Studie zu den *gegen-* und *fremd*-Verben sollte deutlich machen, dass wir im Phänomenbereich der Produktivität von Morphemen in der quantitativen Analyse schon bei der Betrachtung von Einzelphänomenen wie bestimmten Verbgruppen schnell an die Grenzen der Datenlage und der eingesetzten Verfahren stoßen. Dementsprechend weit sind wir davon entfernt, das Phänomen an sich so klar zu operationalisieren, dass wir Morpheme, die gerade beginnen produktiv zu werden, mit quantitativen Verfahren detektieren könnten. Insgesamt ist es immer eine Herausforderung für quantitative Methoden, wenn Phänomene analysiert werden sollen, die im Bereich der seltenen Ereignisse liegen, was *per definitionem* für alle Formen lexikalischer Innovation der Fall ist.

Wir behandeln im folgenden Abschnitt eine völlig andere linguistische Fragestellung, die auf einer höheren Abstraktionsstufe liegt und losgelöst von sprachlichen Einzelphänomenen ist. Mit dem folgenden Thema können wir somit das andere Ende der Skala zeigen, auf der sich die quantitative Analyse lexikalischer Daten heute bewegt.

3 Der Zusammenhang von Wortstruktur- und Wortstellungsregularität

Die Studie, die wir im Folgenden vorstellen, ist bereits ausführlich in Koplenig et al. (2017) dargestellt. Wir werden sie deshalb hier eher zusammenfassend beschreiben.

Wie bereits eingangs erwähnt, wollen wir uns das sequenzielle Auftreten von Wörtern in Texten zunutze machen, um uns einer schon seit längerem in der linguistischen Typologie prominenten – überwiegend jedoch qualitativ formulierten – Hypothese über ein rein quantitatives Verfahren zu nähern. Diese Hypothese besagt, dass sich die Sprachen der Welt darin unterscheiden, wie bestimmte Arten der sprachlichen Information (so beispielsweise die grammatische Beziehung zwischen Wörtern) vermittelt werden. Im Rückgriff auf das Prinzip der Sprachökonomie (Köhler 2009; Zipf 2012) wird folgendermaßen argumentiert: Sprachen nehmen unterschiedliche Standpunkte ein, die sich bezüglich eines *Trade-offs* zwischen der Informationsvermittlung anhand von Regularitäten *zwischen* Wörtern und *innerhalb* von Wörtern definieren. Anders formuliert: Die Sprachen der Welt unterscheiden sich darin, ob sie Informationen eher über die Wortstellung oder die (interne) Wortstruktur transportieren. Das Prinzip der Sprachökonomie besagt hierbei: Kodiere Informationen so sparsam wie möglich, aber auch so aufwändig wie nötig. Der *Trade-off* kann dann folgendermaßen formuliert werden: Je wichtiger die Wortstellung bei der Informationsvermittlung in einer Sprache ist, desto weniger wichtig sollte die Wortstruktur sein und umgekehrt.

Parkvall (2008, S. 325) formuliert es folgendermaßen: „Languages with a more or less free word order tend to be those with an extensive morphological machinery. [...] By adding inflexional morphemes to the NPs, the listener is at no risk of confusing who did what to whom." Wir werden im Folgenden nicht von Morphologie und Syntax, sondern nur von Wortstellung und Wortstruktur sprechen. Das Ziel unserer Untersuchung ist auch nicht alle Aspekte von Syntax und Morphologie in die Untersuchung einzubringen. Unter anderem ist der Ansatz auch als Vorschlag zu verstehen, wie die Konzepte Wortstellung und Wortstruktur möglichst sparsam und voraussetzungsarm für quantitative Analysen operationalisiert werden können, um dem bisher hauptsächlich qualitativ begründeten Zusammenhang auf die Spur zu kommen.

Im ersten Schritt müssen wir die Hypothese in Form von Regularitäten ausdrücken. Sie besagt dann, dass je mehr inter-lexikalische Regularität in einer Sprache vorhanden ist, desto weniger intra-lexikalische Regularität ist vorhanden. Regularität bedeutet in diesem Zusammenhang, dass bestimmte sprachliche Strukturen immer wieder auftauchen und sie damit in ihrem Vorkommen redundant sind. So kann man beispielsweise aufgrund der Flexionsparadigmen deutscher Verben davon ausgehen, dass bestimmte Graphem- oder Phonemkombinationen immer wieder auftauchen. Das Präteritumsuffix für schwache Verben *-te* wird uns etwa in einem Text, der vorwiegend im Präteritum geschrieben ist, immer wieder begegnen (je nach Person mit weiteren Endungen wie *-test*, *-ten* oder *-tet*). Auch die verschiedenen Nominalisierungssuffixe in der deutschen Sprache wie *-heit* oder *-ung* werden sich häufig wiederholen.

All diese Redundanzen sind messbar. So können wir computergestützt einen Text von Anfang bis Ende zeichenweise traversieren und bei jedem Zeichen die Information extrahieren, wie lang die längste Übereinstimmung (*longest match-length*) im bereits gelesenen Text ist. Im echten Pangramm „‚Fix Schwyz!', quäkt Jürgen blöd vom Paß." werden wir nur sehr wenig Redundanz (nämlich lediglich die für die Leerzeichen) messen. In einem normalen Zeitungstext, in dem sich Wörter wiederholen, bestimmte Affixe häufig benutzt werden, Hauptsätze oft mit „und" koordiniert werden und weitere Wiederholungen auf Zeichenebene vorkommen, werden die gemessenen Redundanzwerte pro Zeichen weitaus größer sein. Wie wir in Koplenig et al. (2017) zeigen, lässt sich über die mittlere Redundanz die Entropie (der mittlere Informationsgehalt pro Zeichen) eines Texts schätzen. Die Entropie kann als eine Maßzahl dafür verstanden werden, wie viel Redundanz in einer Zeichenkette (hier: einem Text) herrscht (Cover/Thomas 2006).

Doch wie können wir uns nun Redundanz und die daraus abgeleitete Entropie zunutze machen, um unserer ursprünglichen Frage nach dem *Trade-off* zwischen Wortstellung und Wortstruktur auf die Spur zu kommen? Die Antwort liegt in einer gezielten „Zerstörung" der einen oder anderen Quelle an Information

(Juola 2008): Wir schätzen zunächst die Originalentropie eines Texts, in dem alle Wörter in Kleinschreibung überführt wurden.[12] Ein Beispieltext würde folgendermaßen lauten:

„am anfang schuf gott himmel und erde und die erde war wüst und leer“

Wir nennen die geschätzte Entropie für den Originaltext $\hat{H}_{original}$. Wir erzeugen dann zwei manipulierte Versionen desselben Textes. In einer Version wird die Reihenfolge der Wörter randomisiert:

„und schuf die und erde anfang und himmel erde gott war am leer wüst“

Wir messen erneut die Entropie des Texts und nennen diese $\hat{H}_{no\_order}$, da jegliche Wortstellungsinformation in diesem Text zerstört wurde. In einer weiteren Version ersetzen wir nun alle Wörter des Originaltexts durch zufällige Buchstabensequenzen der gleichen Länge (die Abfolge der Wörter bleibt jedoch erhalten):

„mf soapui islmö zjßs möpdkü bdq zimü bdq iwj zimü üiw gfvz bdq mxqw“

Wichtig ist dabei, dass ein Worttyp immer die gleiche Sequenz an Buchstaben erhält. Im Beispiel wird „erde“ immer durch „zimü“ ersetzt. Wir berechnen nun die Entropie $\hat{H}_{no\_structure}$ dieses Texts, in dem die Wortstellungsinformation zwar noch vorhanden ist, jegliche wortinterne Information jedoch getilgt wurde. Im letzten Schritt berechnen wir zwei Differenzwerte. Das ist einerseits der Wert für jene Information, die durch die Zerstörung der Wortstellungsinformation verloren gegangen ist: $D_{order} = \hat{H}_{no\_order} - \hat{H}_{original}$.[13] Andererseits wird noch der Wert für die Information berechnet, die durch die Zerstörung der Wortstrukturinformation verloren gegangen ist: $D_{structure} = \hat{H}_{no\_structure} - \hat{H}_{original}$.

Dieses Verfahren haben wir auf 1.559 Bibelübersetzungen in 1.196 Sprachen angewendet, die Bibelübersetzungen entstammen dem „Parallel Bible Corpus“ von Mayer/Cysouw (2014). Die Sprachen, die im Korpus enthalten sind, decken weltweit etwa sechs Milliarden Sprecherinnen und Sprecher ab. Danach wurden für jede Übersetzung der Bibel in jeder Sprache für jedes Buch die Werte D_{order} und $D_{structure}$ berechnet.

12 Dieses Vorgehen ist notwendig um die Einflüsse von Groß-/Kleinschreibung über Sprachen hinweg zu eliminieren. Grundsätzlich ist das vorgeschlagene Vorgehen aber auch mit Groß-/Kleinschreibung denkbar.

13 Da die Entropie eines Textes steigt, wenn potenzielle Regularität zerstört wird, sind die Entropiewerte für die manipulierten Texte jeweils die ersten Elemente in diesen Differenzbildungen (Montemurro/Zanette 2011).

Die Ergebnisse stützen die Hypothese: Es ergibt sich über alle Bücher der Bibel hinweg ein starker negativer Zusammenhang. Der Rangkorrelationskoeffizient Spearman's ρ liegt bei r_s = -0,73. Der negative Zusammenhang drückt aus, dass D_{order} umso kleiner ist, je höher $D_{structure}$ ist und umgekehrt.

Abbildung 4 zeigt den Zusammenhang für alle Übersetzungen aller Bücher. Es ist gut zu sehen, dass sowohl oben rechts (*viel* Information sowohl in Wortstellung als auch in Wortstruktur) als auch unten links (*wenig* Information in beiden) keine Datenpunkte liegen. Das ist aufgrund des oben formulierten Ökonomieprinzips auch nicht zu erwarten, denn es wäre unökonomisch, beide Wege der Informationsvermittlung gleichzeitig zu nutzen. Kodiert man Information aber weder in Wortstellung noch in Wortstruktur, wäre der kommunikative Erfolg in Gefahr. Interessant ist dabei, dass dieser wohlbekannte Zusammenhang durch dieses schlichte Verfahren, welches keinerlei Kenntnisse der Morphologie oder Syntax für die einzelnen Sprachen voraussetzt, nachgewiesen werden kann.

Alle Bücher aus deutschen Bibelübersetzungen befinden sich in dem Rechteck, das in Abbildung 4 eingezeichnet ist. Es ist demnach nicht so, dass sich alle Datenpunkte für deutsche Texte an einem ganz bestimmten Punkt im Graphen befinden, sondern es gibt durchaus intertextuelle Variation (zwischen den einzelnen Büchern der Bibel und den verschiedenen Varietäten des Deutschen, für die Bibelübersetzungen im Korpus vorhanden sind) für den Zusammenhang der beiden Maße. Ein bestimmter Text (also in diesem Fall eine bestimmte Übersetzung eines bestimmten Buchs aus der Bibel) kann sich an verschiedenen Stellen dieses Zusammenhanggraphen befinden, eben innerhalb bestimmter Variationsgrenzen, die von der jeweiligen Sprache aufgespannt werden.

In Abbildung 5 sind alle Datenpunkte, die zu einer Sprache gehören, über Mittelwertbildung aggregiert, der Rangkorrelationswert ändert sich dadurch nur unwesentlich (r_S = 0,73). Die Beschriftungen, die in den Plot eingefügt sind, befinden sich an der Position der jeweiligen Sprache, die über ihren ISO-Code identifiziert ist (ESK = Northwest Alaska Inupiatun; DEU = Deutsch; ENG = Englisch; CMN = Mandarin Chinese). Die Sprachen befinden sich sowohl relativ zueinander als auch absolut an den Positionen, die man unter typologischer Perspektive erwarten würde. So gelten Inuit-Sprachen als stark synthetisch, Mandarin-Chinesisch wird dagegen als isolierende Sprache beschrieben (vgl. u.a. Balles 2004, S. 51), was ebenfalls zur Position im Zusammenhangsgraphen passt. Es ist auch zu sehen, dass das Deutsche im Vergleich zum Englischen zwar *eher* dazu neigt, Information über die Wortstruktur zu vermitteln, im Gesamtvergleich aller beachteten Sprachen aber eher im Mittelfeld des *Trade-offs* angesiedelt ist, also einen Kompromiss aus beiden Strategien realisiert.

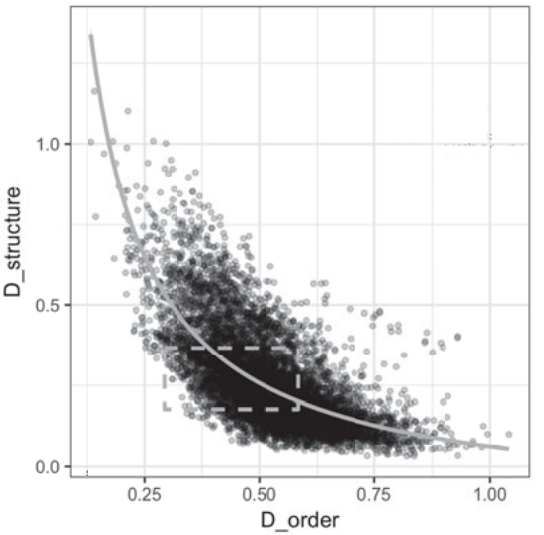


Abb. 4: Zusammenhangsgraph für D_{order} und $D_{structure}$. Ein Datenpunkt entspricht einer Übersetzung eines Buchs der Bibel. Die Anpassungslinie beschreibt einen reziproken Zusammenhang der beiden Variablen

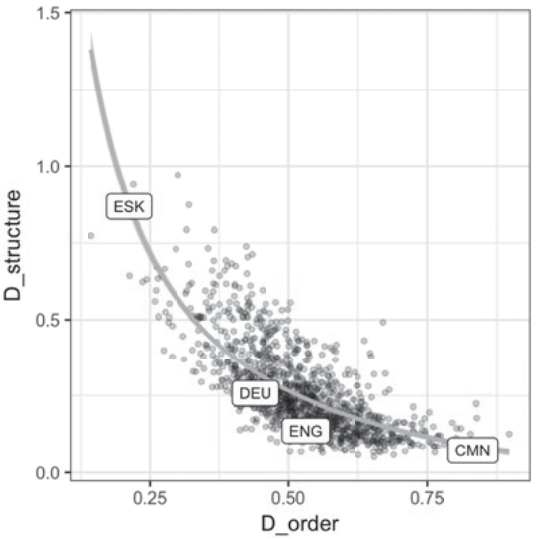


Abb. 5: Zusammenhangsgraph für D_{order} und $D_{structure}$. Ein Datenpunkt entspricht dem aggregierten Wert aller Übersetzungen aller Bücher für eine Sprache. Die Positionen einiger Sprachen im Graphen sind durch Labels hervorgehoben (ESK = Northwest Alaska Inupiatun; DEU = Deutsch; ENG = Englisch; CMN = Mandarin Chinese)

In Koplenig et al. (2017) gehen wir u.a. detaillierter darauf ein, wie sich die Entwicklung vom Altenglischen in die Gegenwart darstellt, die eine Bewegung des Englischen hin zu einer eher analytischen Sprache nachzeichnet. Auch diese in der linguistischen Forschung wohlbekannte sprachgeschichtliche Veränderung lässt sich demnach mit dem hier vorgestellten Verfahren in den verschiedenen historischen Übersetzungen der Bibel nachweisen. Durch weitere in Koplenig et al. (2017) genauer beschriebene Verfahren versuchen wir auszuschließen, dass es sich bei den hier gezeigten Ergebnissen um bloße Artefakte der methodischen Herangehensweise handelt.

Auf der multilingualen Plattform für lexikalisch-lexikografische Daten des IDS – OWID^{plus} – haben wir zwei Online-Apps zur Verfügung gestellt, anhand derer die auf Basis des „Bible Corpus" berechneten Daten weiter exploriert werden. Mit dem „Entropy explorer"[14] können mehrere Variablen (darunter D_{order} und $D_{structure}$) über verschiedene Regressionsmodelle in Zusammenhang gebracht werden. Außerdem können Sprachfamilien sowie die verschiedenen Bücher der Bibel selektiert werden. Alle angezeigten Daten stehen außerdem zum Download bereit. Mit der Web-App „Entropy data world map"[15] kann man sich ein Bild davon verschaffen, wie sich die Sprachen der Welt in räumlichem Zusammenhang bezüglich der berechneten Werte verteilen. Bei Bedarf können die Werte mehrerer Sprachen über Pop-Ups verglichen werden. Weitere Informationen zum Projekt, Auswertungscode sowie eine Java-Implementierung für die Berechnung der *longest match-length* können begleitend dazu[16] abgerufen werden.

Zusammenfassend können wir an dieser Stelle konstatieren, dass sich 1) die Sprachen der Welt hinsichtlich der Informationsvermittlung über Wortstellung und interne Wortstruktur in einer Ausgleichsbeziehung befinden und 2) dass die hier gezeigte Methode ein relativ voraussetzungsarmer (im Sinne eines sich nicht einer gewissen linguistischen Theorie verschreibender) Ansatz ist, um die sequenzielle Natur von Sprache für die Untersuchung typologischer und ggf. stilistischer Phänomene heranzuziehen.

Das Verfahren ist somit unseres Erachtens nach ein Beispiel für ein quantitatives Verfahren, welches eine interessante Ergänzung zum *close reading* der typologischen Forschung darstellt. Als Einschränkung ist hier nur zu nennen, dass eine solche Untersuchung ein besonderes Korpus erfordert. Damit wir diesen Zusammenhang nachweisen konnten, benötigten wir ein vielsprachiges Korpus,

14 www.owid.de/plus/eebib2016/ (Stand: 30.5.2017).

15 www.owid.de/plus/wmbib2016/ (Stand: 30.5.2017).

16 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/8KH0GB (Stand: 31.5.2017).

in dem die gleichen Texte in möglichst vielen Sprachen vorliegen. Wir würden die Untersuchung gerne an anderen Textsorten wiederholen, da die Bibel in ihrer Versstruktur ein sehr spezieller Text ist. Wir suchen allerdings noch nach geeigneten Texten, die lang genug sind, damit das entropiebasierte Verfahren hinreichend gut funktioniert, die urheberrechtsfrei sind und für die gute Übersetzungen in möglichst vielen Sprachen vorliegen (die UN-Menschenrechtscharta ist recht kurz und wurde in wesentlich weniger Sprachen übersetzt als die Bibel). Insofern ist diese Studie auch ein Beispiel dafür, dass interessante quantitative Analysen meist hohe Anforderungen an die Datenbasis stellen.

4 Schlussbemerkung

Anfangs haben wir herausgestellt, dass pragmatisch aufgebaute Textkorpora ein geeignetes Mittel sein können, um überhaupt eine quantitative empirische Erforschung der Sprache zu ermöglichen. Insofern ist der Aufbau großer Zeitungskorpora der richtige Weg gewesen, große Mengen an Sprachdaten in den Blick zu nehmen. Es scheint uns aber wichtig, die Beschränkungen, die mit dieser Art der Daten einhergehen, deutlich in den Blick zu nehmen. Zum einen benötigen wir nach unserer Auffassung eine breitere Korpusbasis: Es wäre wichtig, mehr diachrone Daten und breiter gefächerte schriftsprachliche Daten für die quantitative Forschung zur Verfügung zu haben. Die anglistische korpuslinguistische Forschung ist da schon etwas weiter, was natürlich auch an der Dominanz der englischen Sprache und der entsprechenden Mittel, die zu ihrer Erforschung bereitstehen, zusammenhängt. Hier ist mit dem „Corpus of Historical American English" ein Korpus aufgebaut worden, welches eine ganze Reihe neuer quantitativ ausgerichteter diachroner linguistischer Forschungen ermöglicht hat (vgl. z.B. Allan/Robinson 2012; Hilpert 2013; Hilpert/Perek 2015). Der Programmbereich Korpuslinguistik des IDS hat mit der Akquisition aller Ausgaben des „Spiegel" und der Wochenzeitung „Die Zeit" schon einen wichtigen Schritt gemacht, um die Lücke fehlender diachroner Texte zumindest ansatzweise für die Nachkriegszeit zu schließen. Auch soll die neue Korpusanalyseplattform KorAP es ermöglichen, flexibler mit den Rohdaten quantitativ arbeiten zu können. Zum anderen brauchen wir für ein besseres empirisches Fundament mehr Arten von Sprachdaten, die quantitativ analysiert werden können. Dies wird in der linguistischen Forschung meist unter dem Schlagwort der „konvergierenden Evidenz" diskutiert (vgl. u.a. Arppe et al. 2010; Gries/Hampe/Schönefeld 2005; Schmid 2010). Die Idee dabei ist, korpuslinguistische Evidenz mit anderen Arten von Evidenz, insbesondere solche von psycholinguistischer und neurolinguistischer Natur, zu vergleichen. Das Ziel

muss darin bestehen, Konvergenzen und Divergenzen aufzudecken, um ein vollständiges Bild der sprachlichen Realität zu zeichnen.

Wenn die linguistische Forschung sich in diese beiden Richtungen weiterentwickelt, kann die quantitative linguistische Forschung – so denken wir – dem Ziel, aus einer Vogelperspektive auf Sprachdaten neue Erkenntnisse abzuleiten, auch für lexikalische Daten noch deutlich näher kommen als bislang.

Literatur

Allan, Kathryn/Robinson, Justyna A. (2012): Current methods in historical semantics. (= Topics in English Linguistics (TiEL) 73). Berlin/Boston.

Arppe, Antti/Järvikivi, Juhani (2007): Take empiricism seriously! In support of methological diversity in linguistics. A commentary of Geoffrey Sampson 2007. Grammar without Grammaticality. In: Corpus Linguistics and Linguistic Theory 3, 1, S. 99–109.

Arppe, Antti et al. (2010): Cognitive Corpus Linguistics: Five points of debate on current theory and methodology. In: Corpora 5, 1, S. 1–27.

Baayen, R. Harald (2009): Corpus Linguistics in Morphology: Morphological productivity. In: Lüdeling, Anke/Kyto, Merja (Hg.): Corpus Linguistics. An international handbook. (= Handbücher zur Sprach- und Kommunikationswissenschaft/Handbooks of Linguistics and Communication Science (HSK) 29, 2). Berlin, S. 900–919.

Balles, Irene (2004): Die Tendenz zum analytischen Sprachtyp aus der Sicht der Indogermanistik. In: Hinrichs, Uwe (Hg.): Die europäischen Sprachen auf dem Weg zum analytischen Sprachtyp. (= Eurolinguistische Arbeiten 1). Wiesbaden, S. 33–53.

Bybee, Joan (2015): Language change. (= Cambridge Textbooks in Linguistics). Cambridge.

Bybee, Joan (2006): From usage to grammar: The mind's response to repetition. In: Language 82, 4, S. 711–733.

Calzolari, Nicoletta et al. (Hg.) (2014): Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14). Reykjavik.

Clauset, Aaron/Shalizi, Cosma Rohilla/Newman, Mark E.J. (2009): Power-law distributions in empirical data. In: SIAM (Society for Industrial and Applied Mathematics) Review 51, 4, S. 661–703.

Cover, Thomas A./Thomas, Joy A. (2006): Elements of information theory. 2. Aufl. Hoboken.

Engelberg, Stefan (2015): Quantitative Verteilungen im Wortschatz. Zu lexikologischen und lexikografischen Aspekten eines dynamischen Lexikons. In: Eichinger, Ludwig M. (Hg.): Sprachwissenschaft im Fokus. Positionsbestimmungen und Perspektiven. (= Jahrbuch des Instituts für Deutsche Sprache 2014). Berlin u.a., S. 205–230.

Evert, Stefan (2006): How random is a corpus? The library metaphor. In: Zeitschrift für Anglistik und Amerikanistik 54, 2, S. 177–190.

Fleischer, Wolfgang/Barz, Irmhild (2012): Wortbildung der deutschen Gegenwartssprache. 4. Aufl. Berlin.

Gilquin, Gaëtanelle/Gries, Stefan T. (2009): Corpora and experimental methods: A state-of-the-art review. In: Corpus Linguistics and Linguistic Theory 5, 1, S. 1–26.

Gries, Stefan T. (2015): Quantitative Linguistics. In: Wright, James D. (Hg.): International Encyclopedia of the Social & Behavioral Sciences. Bd. 19. 2. Aufl. Oxford, S. 725–732.

Gries, Stefan T./Hampe, Beate/Schönefeld, Doris (2005): Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions. In: Cognitive Linguistics 16, 4, S. 635-676.

Gulordava, Kristina/Baroni, Marco (2011): A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In: Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics. Edinburgh, S. 67–71.

Hilpert, Martin (2013): Constructional change in English. Developments in Allomorphy, Word Formation, and Syntax. Cambridge.

Hilpert, Martin/Perek, Florent (2015): Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. In: Linguistics Vanguard 1, 1, S. 339–350.

Juola, Patrick (2008): Assessing linguistic complexity. In: Miestamo, Matti/Sinnemäki, Kaius/Karlsson, Fred (Hg.): Language complexity: Typology, contact, change. (= Studies in Language Companion Series (SLCS) 94). Amsterdam/Philadelphia.

Keibel, Holger/Hennig, Sophie/Perkuhn, Rainer (2011): Effiziente halbautomatische Detektion von Neologismuskandidaten. Technical Report. Mannheim. Internet: www1.ids-mannheim.de/fileadmin/kl/dokumente/ids-kl-2010-01.pdf (Stand: 31.5.2017).

Kerremans, Daphné/Stegmayr, Susanne/Schmid, Hans-Jörg (2012): The NeoCrawler: Identifying and retrieving neologisms from the internet and monitoring ongoing change. In: Allan, Kathryn/Robinson, Justyna A. (Hg.): Current methods in historical semantics. (= Topics in English Linguistics (TiEL) 73). Berlin/Boston, S. 59–96.

Kerremans, Daphné (2015): A Web of New Words. (= English Corpus Linguistics 15). Frankfurt a.M.

Koplenig, Alexander (2015): Using the parameters of the Zipf–Mandelbrot law to measure diachronic lexical, syntactical and stylistic changes – a large-scale corpus analysis. In: Corpus Linguistics and Linguistic Theory 2015. Internet: www.degruyter.com/view/j/cllt.ahead-of-print/cllt-2014-0049/cllt-2014-0049.xml.

Koplenig, Alexander (2016): Analyzing lexical change in diachronic corpora. Mannheim. [Dissertation]. Internet: https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/4890 (Stand: 31.5.2017).

Koplenig, Alexander et al. (2017): The statistical trade-off between word order and word structure – large-scale evidence for the principle of least effort. In: PLoS ONE 12, 3: e0173614, S. 1–25.

Klosa, Annette (2003): *gegen*-Verben – ein neues Wortbildungsmuster. In: Sprachwissenschaft 28, 4, S. 467–494.

Köhler, Reinhard (2009): System theoretical Linguistics. In: Theoretical Linguistics 1, 2–3, S. 241-257.

Kupietz, Marc/Lüngen, Harald (2014): Recent Developments in DeReKo. In: Calzolari et al. (Hg.), S. 2378-2385. Internet: www.lrec-conf.org/proceedings/lrec2014/pdf/842_Paper.pdf (Stand: 31.5.2017).

Lansdall-Welfare, Thomas et al. (2017): Content analysis of 150 years of British periodicals. In: Proceedings of the National Academy of Sciences 114, S. E457–E465.

Mayer, Thomas/Cysouw, Michael (2014): Creating a massively parallel Bible corpus. In: Calzolari et al. (Hg.), S. 26-31. Internet: www.lrec-conf.org/proceedings/lrec2014/pdf/220_Paper.pdf (Stand: 31.5.2017).

Montemurro, Marcelo A./Zanette, Damián H. (2011): Universal Entropy of Word Ordering Across Linguistic Families. In: PLoS ONE 6, 5: e19875, S. 1–9.

Parkvall, Mikael (2008): Limits of language. Almost everything you didn't know about language and languages. Wilsonville.

Rohrdantz, Christian et al. (2012): Lexical semantics and distribution of suffixes: A visual analysis. In: Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, S. 7–15. Internet: www.aclweb.org/anthology/W12-0202 (Stand: 31.5.2017).

Schmid, Hans-Jörg (2010). Does frequency in text instantiate entrenchment in the cognitive system? In: Glynn, Dylan/Fischer, Kerstin (Hg.): Quantitative methods in cognitive semantics: Corpus-driven approaches. (= Cognitive Linguistics Research (CLR) 46). Berlin, S. 101-134.

Schmid, Hans-Jörg (2015): EnerG - English Neologisms Research Group. Internet: www.neocrawler.anglistik.uni-muenchen.de/crawler/html/index.php?abfrage=about (Stand: 31.5.2017).

Sinclair, John (1991): Corpus, concordance, collocation. Oxford.

Szmrecsanyi, Benedikt (2016): About text frequencies in historical linguistics: Disentangling environmental and grammatical change. In: Corpus Linguistics and Linguistic Theory 12, 1, S. 153–171.

Templin, Mildred C. (1957): Certain language skills in children. 4. Ausg. (=Monograph Series/ Institute of Child Welfare, University of Minnesota 26). Minneapolis.

Tweedie, Fiona J./Baayen, R. Harald (1998): How variable may a constant be? Measures of lexical richness in perspective. In: Computers and the Humanities 32, 5, S. 323–352.

Weinreich, Uriel/Labov, William/Herzog, Marvin (1968): Empirical foundations for a theory of language change. In: Lehmann, Winfried/Malkiel, Yakov (Hg.): Directions for Historical Linguistics. A symposium. Austin, S. 97–195.

Würschinger, Quirin et al. (2016): Using the web and social media as corpora for monitoring the spread of neologisms. The case of ‚rapefugee', ‚rapeugee', and ‚rapugee'. In: 10th Web as Corpus Workshop (WAC-X). Annual Meeting of the Association for Computational Linguistics (ACL). Berlin. Internet: www.aclweb.org/anthology/W16-2605 (Stand: 31.5.2017).

Zipf, George Kingsley (2012): Human behavior and the principle of least effort. An introduction to human ecology. Mansfield Center. [Nachdruck der Ausgabe 1949].