10.20378/irb-59990

Tanja Anstatt (Bochum)

Wortfrequenz und Textsorten

1. Einleitung[1]

Mit der rasanten Entwicklung der Korpusdaten erleben Methoden der quantitativen Linguistik wie die Frequenzforschung in den letzten Jahren einen Boom. Für das Russische steht seit 2009 mit dem Neuen Frequenzwörterbuch (*Novyj Častotnyj Slovar' Russkoj Leksiki*, Ljaševskaja/Šarov online, im weiteren NČSl, http://dict.ruslang.ru/freq.php) eine Wortfrequenzliste einer neuen quantitativen Dimension online zur Verfügung. Sie basiert auf russischen Texten im Umfang von 92 Mio. laufenden Wortformen, die einen ausgewählten Ausschnitt des Russischen Nationalkorpus (*Russkij nacional'nyj korpus,* RNK) darstellen.

Sebastian Kempgen befasste sich allerdings mit quantitativen Ansätzen und statistischen Methoden in der Linguistik, bevor alle diese Materialien so bequem zugänglich waren – er ist einer der Vorreiter dieses Feldes in der Slavistik und leistete hier schon lange vor dem großen Boom wichtige Beiträge. Quantitativ-statistische Analysen bilden eine Art roten Faden durch einen Großteil seiner Arbeiten, die er beispielsweise auf Wortartenklassifikation (1981/2008), Verbmorphologie (1995b, 2007), Phonologie (2004) und Sprachtypologie (2004, Kempgen/Lehfeldt 2004) anwandte. Von bleibendem Wert sind auch die breit angelegten Übersichtsarbeiten (Kempgen 1995a/2007, Kempgen 1999), in denen Sebastian Kempgen den Forschungsstand der quantitativen linguistischen Forschung zum Russischen präsentiert und die nach wie vor eine Fundgrube an Anregungen bieten.

Mit diesem Artikel möchte ich an einige dort präsentierte Aspekte anknüpfen und fragen, wie sich die Frage nach dem Zusammenhang von Wortfrequenz und Textsorte, die in seiner Überblicksmonographie (Kempgen 1995a/2007) mehrfach angesprochen wird, aktuell anhand des

[1] Für kritische Kommentare zu einer früheren Version dieses Artikels danke ich Christina Clasmeier (Bochum). Mein Dank für statistische Beratung gilt Johannes Herrmann (Gießen). Selbstverständlich liegen alle etwaigen Unzulänglichkeiten allein in meiner Verantwortung.

im Vergleich zu den 1990er Jahren stark angewachsenen Frequenzmaterials darstellt. Mein besonderes Interesse gilt dabei der Frage, inwieweit die Aussagen zum Gesamtkorpus des Russischen auch für die mündliche Sprache zutreffen. In Abschnitt 2 beleuchte ich zunächst den Zusammenhang von sprachlichen Frequenzerscheinungen und Textsorten. Der 3. Abschnitt gibt einen Überblick über die Informationen, speziell die textsortenbezogenen, die sich dem *Novyj Častotnyj Slovar' Russkoj Leksiki* entnehmen lassen. In Abschnitt 4 stelle ich schließlich zwei exemplarische Auswertungen zu Textsortenunterschieden vor, die sich auf die die Frequenz von Präpositionen und die Häufigkeit von Wortarten beziehen.

2. Frequenz und Textsorten

In der systemlinguistischen Forschung steht in Bezug auf die Frequenz sprachlicher Einheiten – von Wörtern, aber auch Elemente und Kategorien aller anderen sprachlichen Ebenen – ihr Zusammenhang mit den strukturellen Eigenschaften dieser Einheiten und mit ihrer historischen Entwicklung im Zentrum. Einen Forschungsüberblick über frequenzbezogene Eigenschaften des Russischen bis Ende des 20. Jh.s gibt wie eingangs erwähnt Kempgen (1995a/2007 und 1999), einen Überblick über die russische Theorieentwicklung bietet Kelih (2008). Jüngere Studien zu sprachstrukturellen Fragen sind etwa Kempgen (2007) oder Kopotev (2008), um nur einige Beispiele zu nennen. In jüngerer Zeit erfährt die Frequenzforschung aus einer weiteren Richtung intensive Aufmerksamkeit, nämlich der kognitiv und psycholinguistisch orientierten Sprachforschung: Die Frequenz, mit der sprachliche Einheiten auftreten, wirkt sich auf ihre Verarbeitung aus. Hochfrequente Wörter werden im Vergleich zu selteneren schneller und korrekter verarbeitet und früher erworben (einen Überblick gibt Ellis 2002, neue psycholinguistische Evidenz zum Russischen s. jüngst Vlasova/Sinitsyn/Pechenova 2015 zur Verarbeitung, Grigoriev/Oshhepkov 2013 zum Erstspracherwerb). Aus dieser Perspektive ist Frequenz eine Variable von zentraler Wichtigkeit, die – was in der Psycholinguistik natürlich schon seit vielen Jahrzehnten gut bekannt ist – in Studien beispielsweise zur Wortverarbeitung stets kontrolliert werden muss (weiterführende Literatur s. Anstatt 2016, Clasmeier/Anstatt/Ernst/Belke 2016).

Wenn für psycholinguistisch orientierte Forschungen also ein großer Bedarf an Frequenzinformation besteht, so stellt sich die Frage, wie und woher diese gewonnen werden kann. Mit den neuen, großen Textmengen der Nationalkorpora stehen erstmals valide Daten zur Ermittlung dieser Variablen zur Verfügung. Allerdings drängt sich hier im nächsten Schritt die Frage auf, ob die dort präsentierten Sprachdaten auch dem entsprechen, womit die Sprecher/innen bei ihrer Sprachverwendung tatsächlich zu tun haben. Während die Textkorpora aus leicht nachvollziehbaren technischen Gründen zu weit überwiegenden Teilen auf schriftlichen Texten basieren, dürfte die Sprachproduktion und -rezeption des Menschen im Schnitt mindestens zur Hälfte – selbstverständlich unterschiedlich je nach individuellen Lebensbedingungen – aus gesprochener Sprache bestehen.

Textsorten[2] können sich hinsichtlich der in ihnen verwendeten sprachlichen Elemente erheblich unterscheiden – dies ist eine seit Langem gut bekannte Tatsache. Und dass der Graben hinsichtlich der gesamten Verwendungsbedingungen und Strukturen zwischen gesprochener (bzw. genauer: konzeptionell mündlicher oder nähesprachlicher) Sprache einerseits und geschriebener (bzw. konzeptionell schriftlicher oder distanzsprachlicher) Sprache andererseits besonders tief ist, hat die Forschung der letzten Jahrzehnte ebenfalls klar nachgewiesen. Eine wichtige theoretische Begründung lieferten Koch/Oesterreicher (1985): Die gesprochene Sprache unterliegt völlig anderen Produktions- und Rezeptionsbedingungen als die geschriebene. Bereits die Forschungen der 1970er und 80er Jahre zur russischen Standardumgangssprache, der *Russkaja Razgovornaja Reč'*[3] (s. z. B. Zemskaja (Hrsg.) 1973, 1983, Zemskaja/Kitajgorodskaja/Širjaev 1981) wiesen anhand empirischer Daten

---

[2] Aus praktischen Gründen verwende ich den Terminus hier als Übersetzung des russischen traditionellen Begriffs *funkcional'nyj stil'*, denn er wird in ähnlicher Form auch im NČSl gebraucht. In der russischen Tradition werden die fünf Großgruppen Amtsstil („oficial'no-delovoj stil'"), belletristischer, publizistischer, wissenschaftlicher und umgangssprachlicher Stil unterschieden (vgl. z. B. Valgina 2003). Im NČSl werden aber wissenschaftlicher und Amtsstil mit einigen weiteren Stilen zur „sonstigen nichtbelletristischen Literatur" zusammengefasst.

[3] Zum Terminus s. Zemskaja (1973, 5). Wichtig ist, dass mit der *Russkaja Razgovornaja Reč'* die mündliche Form des *literaturnyj jazyk*, also der Standardsprache gemeint ist; Zemskaja grenzt diesen Terminus von *Ustnaja Reč'* ab, mit der sie jede Sprache in

nach, dass es in vielen Parametern des Wortschatzes deutliche Unterschiede zwischen dieser mündlichen Form und der geschriebenen russischen Sprache gibt. Hier finden sich auch schon zahlreiche Hinweise auf Unterschiede in der quantitativen Verteilung der sprachlichen Einheiten. Besonders systematisch wurden diese in den von Sirotinina herausgegebenen Bänden (1983a/2003, 1992/2003) untersucht. Dort wurden Daten zur *Razgovornaja Reč'* aus Saratov im Umfang von 100.000 laufenden Wortformen ausgewertet und mit Frequenzdaten zur geschriebenen russischen Sprache aus dem Frequenzwörterbuch von Zasorina (1977) verglichen. Auch verschiedene der früheren quantitativen Arbeiten zur Russistik, die Kempgen (1995a, 1999) im Überblick präsentiert, thematisieren die Textsortenunterschiede. Beispielsweise ermittelte Markov (1966, zit. nach Kempgen 1995a, 51) zwischen mündlicher Sprache und Belletristik eine Übereinstimmung von nur zwei Dritteln des Grundwortschatzes. Große Unterschiede gibt es auch in der Textdeckung: Für einen Text von Puškin werden für eine Textdeckung von 95% 8.000 Lemmas benötigt (Kempgen 1995a, 55), für die Umgangssprache sind es erheblich weniger (Kempgen 1995a, 51). Eine ganze Reihe von Arbeiten beschäftigt sich mit Unterschieden grammatischer Kategorien in Bezug auf Frequenz. Viele interessante Arbeiten hierzu sind in Kempgen (1995a und 1999) zusammengefasst, eine jüngere Arbeit ist z. B. Kopotev (2008) zum Kasus.

Trotz dieser bekannten Unterschiede verwenden aktuelle psycholinguistische Arbeiten zum Russischen (z. B. Vlasova/Sinitsyn/Pechenova 2015 oder Grigoriev/Oshhepkov 2013) aus praktischen Gründen zur Kontrolle der Variable Wortfrequenz Daten zum Gesamtkorpus des NČSl. Ich möchte daher einige Aspekte der o. g. Arbeiten aufgreifen und in zwei Fallstudien untersuchen, ob und inwieweit sich die früher an kleineren Datenmengen beobachteten Verteilungen anhand der nun vorliegenden größeren Korpora bestätigen lassen. Damit möchte ich die Frage diskutieren, ob die anhand des Gesamtkorpus gewonnenen Frequenzdaten überhaupt als repräsentativ für die tatsächliche Sprachverwendung gelten

mündlicher Form – von wissenschaftlichen Vorträgen bis zu dörflichen Dialekten – meint.

können, wenn wir davon ausgehen, dass die gesprochene Sprache in dieser eine erheblich größere Rolle spielt als ihr Anteil im Korpus dies abbildet.

3. Informationen zur Wortfrequenz und Textsorten im *Novyj Častotnyj Slovar' Russkoj Leksiki* (NČSl)

Die Erstellung von Wortfrequenzlisten blickt im Russischen bereits auf eine längere Tradition zurück; Kempgen (1995a, 46–51) stellt die Werke des 20. Jh.s ausführlicher vor.

Wie eingangs erwähnt, repräsentiert das *Novyj Častotnyj Slovar' Russkoj Leksiki* aufgrund seiner breiten Textgrundlage und seiner elektronischen frei verfügbaren Recherchierbarkeit eine neue Generation von Frequenzlisten. Es wurde von den Autoren Ljaševskaja/Šarov (2009) in Printform unter dem Titel *Častotnyj Slovar' sovremennogo russkogo jazyka* publiziert und steht online zur Verfügung (http://dict.ruslang.ru). Alle Aussagen dieses Artikels beziehen sich auf die Online-Version mit Stand August 2016, ich zitiere sie als Ljaševskaja/Šarov (2009/2016).[4] Hintergrundinformationen zum Frequenzwörterbuch liefern insbesondere Šarov/Ljaševskaja (2009/2016).[5]

3.1. Überblick über die Informationen des NČSl

Das NČSl basiert auf einem Ausschnitt aus dem Russischen Nationalkorpus (http://ruscorpora.ru). Dieser Ausschnitt umfasst Texte von 1950–2007 und enthält 92 Mio. laufende Wortformen. Die Wortfrequenzdaten werden in einer ganzen Reihe von unterschiedlichen Listen präsentiert:

---

[4] Die letzten Aktualisierungen der Frequenzlisten in der Online-Version erfolgten 2010 (Ljaševskaja, 17.8.2016, Auskunft per E-Mail).

[5] Dieses Dokument mit dem Titel „Vvedenie k Novomu častotnomu slovarju" steht online als pdf auf der Seite des NČSl zur Verfügung. Es handelt sich dabei um das Vorwort zur Printversion von 2009; allerdings wurde die Online-Version offenbar mehrmals leicht aktualisiert, denn die zu verschiedenen Zeitpunkten von mir gespeicherten Versionen weisen kleine Unterschiede auf. Eine kurze Übersicht über die wichtigsten Grundlagen findet sich außerdem auf der Internetseite „Kak pol'zovat'sja slovarem" (http://dict.ruslang.ru/freq_faq.html). Die Angaben stimmen manchmal in Details nicht überein, so werden etwa in den FAQs 100 Mio. laufende Wortformen als Korpusgrundlage angeben, Šarov/Ljaševskaja 2009/2016 geben hingegen 92 Mio. an (a. a. O., ii).

1. Alphabetische Liste der 49.720 häufigsten Lemmas („Alfavitnyj spisok lemm“);
2. nach Frequenzrang angeordnete Liste der 20.004 häufigsten russischen Lemmas („Častotnyj spisok lemm“);
3. a: nach Alphabet geordnete Liste der häufigsten rund 5.000 Lemmas für vier verschiedene Textsorten („Raspredelenie lemm po funkcional'nym stiljam“) (pro Liste zwischen 4.927 und 5.018 Einträge);
   b: zusätzlich findet sich für jede der Textsorten eine Übersicht der 1.000 häufigsten charakteristischen Lemmas („Slovar' značimoj leksiki“);
4. nach Alphabet geordnete Frequenzliste der häufigsten 19.762 Wortformen des Korpus („Alfavitnyj spisok slovoform“);
5. Ranglisten nach Wortarten: Nach Frequenzrang angeordnete Liste der jeweils zwischen 480 und 1.000 häufigsten Vertreter der Wortarten Substantive, Verben, Adjektive, Adverbien und Prädikative, Pronomen, Numeralia und „Hilfswortarten“ (Konjunktionen, Präpositionen, Interjektionen, Partikeln);
6. Hilfstabellen:
   a: Liste der Häufigkeit der Wortarten nach Tokens;
   b: nach Alphabet geordnete Frequenzliste der Grapheme des russischen Alphabets;
   c: nach Alphabet geordnete Frequenzliste der häufigsten 694 russischen Zweierkombinationen von Graphemen;
   d: Alphabetische Liste der häufigsten 2.418 Eigennamen und Abbreviaturen.

Als Maß der Frequenz wird der *ipm*-Wert (instances per million words), also eine relative Größe angegeben; dies ist die Häufigkeit, mit der das gegebene Lemma auf eine Million laufende Wortformen vorkommt. Die absoluten Werte werden nicht genannt; sie können bei Bedarf aus der *ipm*-Angabe und der Gesamtzahl der Tokens, die das Korpus enthält, ermittelt werden.

Neben *ipm* und Rang finden sich in den unter 2. und 5. genannten Listen auch Informationen zur Verteilung der Lemmas im Gesamtkorpus. Die Spalte *teksty* gibt an, in wie vielen Texten das betreffende Lemma auftritt. Für R und den Koeffizienten D wurde das Korpus in 100 gleich große Segmente zerlegt: R (*range*) benennt, in wie vielen dieser Segmente

das gegebene Lemma auftritt und informiert somit über die Breite der Verteilung (s. Šarov/Ljaševskaja 2009/2016, vi–vii): Wenn ein Lemma in allen Segmenten vorkommt, liegt R bei 100.[6] Der Koeffizient D bildet die Gleichmäßigkeit der Verteilung in den Segmenten ab: Wenn ein Lemma etwa in wenigen Segmenten sehr häufig ist, in den anderen hingegen nur selten vorkommt, ist D niedrig (vgl. Šarov/Ljaševskaja 2009/2016, vi–vii).

Die Listen 4 (Wortformen) und 5 (Wortarten) wurden auf der Grundlage eines kleineren Ausschnittes des RNK erstellt, nämlich des Teils mit manuell beseitigter Homonymie. Liste 6a, die die Häufigkeit der Wortarten insgesamt anführt, wurde auf beiden Grundlagen ermittelt, entsprechend werden jeweils die absolute Frequenz und die prozentuale Häufigkeit einmal auf der Basis des bereinigten Korpus (M-Werte) und einmal auf Grundlage des gesamten Korpus (T-Werte) angegeben.[7]

3.2. Informationen zu Textsorten im NČSl

Für die Analyse der Frequenzunterschiede zwischen Textsorten steht also Liste 3 mit zwei Unterlisten zur Verfügung. Die alphabetischen Listen (3a) umfassen die jeweils rund 5.000 häufigsten Wörter und reichen bis zu einem *ipm* von 19 oder 20, bei den mündlichen Texten bis 10. Sie schließen somit die hoch- und mittelfrequente Lexik ein, nicht jedoch die seltene.[8]

Dies ist eine beträchtliche und bereits sehr aussagekräftige Menge: Die häufigsten 2.000 Wörter haben laut Kempgen (1999, 16) eine durchschnittliche Textdeckung von 76%. Für das NČSl geben Šarov/Ljaševskaja an, dass die ersten 1.000 Einträge der Gesamtliste 61% aller Tokens abdecken. Allerdings gibt es dabei große Unterschiede zwischen den Textsorten, wie eingangs erwähnt. Diese Unterschiede reflektieren auch die relativen Frequenzwerte: Die seltensten der rd. 5.000 gelisteten Lemmas in der mündlichen Sprache erreichen einen *ipm* von 10, in den anderen Textsorten sind es etwa 20 *ipm*. In den letzteren gibt es also viel mehr weitere, seltenere Wörter.

---

6 R ist nicht zu verwechseln mit *Rang* in Liste 5: Dieser Ausdruck in den Wortartenlisten bezieht sich auf die Position in der Gesamtfrequenzliste.

7 Ljaševskaja 17.8.2016, E-Mail-Auskunft.

8 Brysbaert/New (2009) bezeichnen Wörter mit einem *ipm*-Wert von unter 10 als niedrigfrequent.

In den oben unter 3b. genannten Listen werden die für die jeweilige Textsorte charakteristischen Wörter aufgeführt, die sich statistisch signifikant häufiger in der betreffenden Textsorte im Vergleich zum Gesamtkorpus finden. Über die Größe des Unterschiedes informiert der Log-likelihood-Wert („LL-score“).[9] Für die mündliche Sprache werden in dieser Liste beispielsweise 780 signifikant häufigere Lemmas aufgeführt.

Textsorte[10]	**Anteil (gerundet)**	**Anzahl Tokens**[11] **(gerundet)**	**Anzahl Texte**
1. Belletristik	39%	35,2 Mio.	2.418
2. Publizistik	42%	39,7 Mio.	27.390
3. Sonstige nichtbelletristische Literatur[12]	17%	15,5 Mio	7.495
4. Mündliche nichtöffentliche Sprache	0,9%	0,8 Mio.	1.005
5. Anderes	0,9%	0,8 Mio.	61
Gesamtes Korpus des NČSl	100%	92 Mio.	38.369

Tabelle 1: Textsorten und ihre Anteile am *Novyj Častotnyj Slovar' Russkoj Leksiki* (Angaben nach Šarov/Ljaševskaja 2009/2016, iii)

Tabelle 1 gibt eine Übersicht über den Umfang der Daten, die den Textsortenlisten zugrunde liegen. Sie alle fließen in das Gesamtkorpus des NČSl ein, das die Grundlage für die allgemeinen Frequenzdaten (oben

9 In die Liste wurden nur signifikante Werte aufgenommen, deren LL-score über 15,31 liegt. In diesem Fall kann mit 99%iger Wahrscheinlichkeit davon ausgegangen werden, dass der Unterschied zwischen den beiden Werten nicht zufällig ist (Šarov/Ljaševskaja 2009/2016, viii; zuerst Rayson/Berridge/Francis 2004).

10 Die Aufteilung der Textsorten unterscheidet sich etwas von derjenigen Zasorinas (1977).

11 Orthographische Wörter.

12 Unter diesem Grobtyp (im Weiteren mit „Nichtbell.“ abgekürzt) fasst das NČSl mehrere traditionell getrennte Textsorten und -untersorten zusammen: Den weitaus größten Anteil bilden mit 65% hier wissenschaftliche Texte; jeweils 10% entfallen auf Behördensprache, elektronische Kommunikation und kirchlich-liturgische Literatur; kleinere Anteile mit je 3,5–1,5% haben Werbung, Alltagstexte und industriell-technische Literatur (vgl. Šarov/Ljaševskaja 2009/2016, iii). Seiner Struktur nach ist er also sehr heterogen, was die Aussagekraft etwas in Frage stellt, jedoch dominieren ingesamt Texte sehr deutlich, die sich maximal von der mündlichen Sprache unterscheiden.

unter 1. und 2. genannt) bildet. Wichtig ist, dass sich diese Korpusauswahl nicht vollständig mit den im Russischen Nationalkorpus enthaltenen Daten deckt. Zum einen wurden wie erwähnt nur Texte ab 1950 ausgewertet. Zum anderen basieren die Frequenzangaben zur gesprochenen Sprache gegenwärtig ausschließlich auf einem Teil des mündlichen Subkorpus, nämlich der „mündlichen nichtöffentlichen Sprache"; es umfasst „Alltagsgespräche, Mikrodialoge im Geschäft, Erzählungen von Träumen, Streitgespräche u. a."[13] (Šarov/Ljaševskaja 2009/2016, xvi, Fn. 9). Nicht in der Auswertung der mündlichen Wortfrequenz enthalten sind die Texte der öffentlichen gesprochenen Sprache[14] und die Filmtexte.[15] Die mündlichen Daten des NČSl decken sich auf diese Weise in etwa mit den Kriterien für die russische Standardumgangssprache, die *Razgovornaja Reč'*. Zu beachten ist schließlich, dass die Summe der vier Textsorten nicht ganz identisch mit dem gesamten Korpus des NČSl ist, da letzteres darüber hinaus noch das Material aus der Rubrik „Anderes" enthält.

3.3. Einschränkungen und Probleme

Ein noch weitgehend ungelöstes generelles Problem ist die Polyfunktionalität von sprachlichen Einheiten, also Homonymie und Polysemie. Im NČSl wurden Wortartenhomonyme (etwa *stali* zum Verb *stat'* und zum Substantiv *stal'*) automatisch beseitigt und für die 20.000 häufigsten Wörter von Hand nachbearbeitet. Homonyme derselben Wortart und Polysemien werden hingegen nicht getrennt (Šarov/Ljaševskaja 2009/2016, xiii). Welche Anteile beispielsweise an der Häufigkeit von *lico* also auf die polyseme Bedeutung ‚Gesicht' und welche auf ‚Person' entfallen oder wie sich die Frequenz von *kosa* ‚Zopf' zum homonymen *kosa* ‚Sense' verhält, kann nicht ermittelt werden.

---

[13] „bytovye razgovory, mikrodialogi v magazine, pereskazy snov, spory i t.p.".

[14] Šarov/Ljaševskaja (2009/2016, xvi, Fn. 9) erläutern, dass die öffentliche gesprochene Sprache für die Frequenzlisten ausgeschlossen wurde, da sie zu viele Übereinstimmungen mit der Publizistik enthielt.

[15] Auf der Internetseite (http://dict.ruslang.ru/freq_faq.html, letzter Abruf am 9.1.2017) wird hingegen angegeben, die mündlichen Daten enthielten neben den Aufzeichnungen authentischer mündlicher Sprache auch Filmtranskripte; Letzteres trifft jedoch nicht zu (Ljaševskaja 17.8.2016, E-Mail-Auskunft).

In Bezug auf die Textsortenfrage ist zum einen die Einschränkung zu nennen, dass Frequenzen in Textsorten nur für Zitierformen (Infinitive bei Verben, Nominativ Singular bei Substantiven) angegeben werden. Die Liste mit Frequenzen der Wortformen bezieht sich auf das gesamte Korpus, hier lassen sich deswegen keine Unterschiede zwischen gesprochener und geschriebener Sprache ermitteln.

Die Korpusgrößen für die vier Textsorten sind darüber hinaus, wie die Zahlen in Tabelle 1 zeigen, sehr heterogen. Insbesondere umfasst die Textgrundlage für die gesprochene Sprache nur einen Bruchteil der Tokens von den anderen Textsorten. Natürlich wurden hier dennoch im Vergleich zu früheren Frequenzlisten ebenfalls enorme Fortschritte in der Aufbereitung dieser bearbeitungsintensiven Textsorte erzielt.[16] Allerdings ist das Korpus zur mündlichen Sprache noch zu klein, um tatsächlich repräsentativ zu sein, die folgenden Aussagen müssen also erneut überprüft werden, wenn weitere Daten vorliegen.

4. Textsortenunterschiede anhand der Frequenzdaten des NČSl

4.1. Vorgehen

Im Folgenden möchte ich einige Schlaglichter darauf präsentieren, welche Unterschiede zwischen den Textsorten sich mit dem vorgestellten Material untersuchen lassen. Leitfrage dabei ist, inwieweit die Angaben zum Gesamtkorpus verlässliche Daten aus psycholinguistischer Perspektive bieten. Die gesprochene Sprache stellt im gesamten Korpus nur einen winzigen Anteil (weniger als 1%), während für die menschliche Sprachverwendung die gesprochene Sprache sicherlich die Hälfte ausmacht. Die Frage ist also, ob und wie stark sich die Frequenzdaten zum Gesamtkorpus in verschiedenen Parametern von den Werten für die mündliche Sprache unterscheiden. Ich stelle daher in den folgenden Abschnitten zwei exemplarische Auswertungen vor.

Datengrundlage dafür sind die Listen der häufigsten rund 5.000 Lemmas der vier Textsorten (in der Übersicht in Abschnitt 3.1. unter 3.a. genannt). Darüber hinaus wird zum Vergleich die Frequenzliste für das gesamte Korpus herangezogen. Um das Material besser vergleichbar zu

[16] Für statistische Vergleiche ist dieser quantitative Unterschied jedoch weniger problematisch, da zumindest der Chi-Quadrat-Test die Korpusgrößen einbezieht und gegenüber Unterschieden robust ist (Rayson/Berridge/Francis 2004, 8).

halten, habe ich in allen Fällen die 4.927 häufigsten Lemmas einbezogen, denn dies ist der Umfang der kürzesten Liste (derjenigen zur mündlichen Sprache). Die im NČSl auf einzelnen Internet-Seiten angeführten und alphabetisch geordneten Listen zu den Wortarten sowie die Liste zum ganzen Korpus habe ich dafür in ein gemeinsames Dokument überführt. In dieses sind Lemma, Frequenz in *ipm*, Textsorte und Wortart sowie absoluter Umfang des jeweiligen Korpus bzw. Teilkorpus eingeflossen. Darüber hinaus wurden die absoluten Frequenzwerte der Lemmas anhand von *ipm* und Korpusumfang berechnet. Die Daten wurden mit SPSS[17] ausgewertet.[18]

4.2. Frequenz einzelner Lemmans in Textsorten am Beispiel von Präpositionen

Welche Wörter statistisch gesehen charakteristisch für die Textsortenkorpora sind, kann den Listen der bedeutsamen Lexik im NČSl entnommen werden. Es wäre denkbar, diese Informationen in die Frequenzinformationen für psycholinguistische Untersuchungen einzubeziehen. Unter den signifikant häufigeren Lemmas des mündlichen Korpus finden sich bis zur Position 50 beispielsweise die Verben *znat', govorit', značit', kupit', pojti, chodit'* und *zvonit'* oder das Substantiv *smech*. Auffällig ist hier, dass unter diesen 50 charakteristischen Lemmas auch zahlreiche Funktionswörter vertreten sind: Insbesondere Partikeln und Personalpronomina *(nu, da, vot, tam, ty, ugu, ja)* zählen zu den mündlich signifikant häufigeren Lemmas.

Für Präpositionen könnte angenommen werden, dass sie aufgrund ihrer grammatischen Funktion wenig Wahlmöglichkeit und daher geringe Varianz zwischen den Textsorten aufweisen. Die Liste der charakteristischen Wörter der mündlichen Sprache weist allerdings auf Rang 11 mit *u* auch eine Präposition auf. Gibt es für diese Wortart weitere Unterschiede zwischen den Textsorten?[19]

---

[17] IBM SPSS Statistics for Macintosh, Version 23.0. Armonk, NY: IBM Corp, IBM Corp. Released 2015.

[18] Bei den im Folgenden genannten statistischen Angaben sehe ich Zusammenhänge auf dem 5%-Niveau als signifikant an (gekennzeichnet mit * für $p < 0{,}05$ und ** für $p < 0{,}01$).

[19] Vorüberlegungen hierzu habe ich bereits in Anstatt (2016) angestellt.

Zahlreiche Beobachtungen zu dieser Frage im Hinblick auf die gesprochene Sprache finden sich in der Forschung zur russischen Standardumgangssprache, der *Razgovornaja Reč'* (RR). So nennt Zemskaja (1983, 113) insgesamt 14 Präpositionen, die in der RR die gebräuchlichsten seien: *iz, s, do, dlja, ot, u, posle, nasčet, po, k, v, na, za* und *s*. Prokurovskaja (1983/2003) bestätigt dies im Wesentlichen anhand des Materials zur Saratover RR. Sie weist allerdings darauf hin, dass die Häufigkeit von *v, na, s, k, za* und *iz* in der Belletristik recht ähnlich sei wie in der RR. Vergleichen wir diese Aussagen mit den Daten des NČSl, so sind kleine Unterschiede zu vermerken: Die häufigsten 14 Präpositionen im mündlichen Teilkorpus des NČSl (vgl. auch Tab. 2) sind *v*, *u*, *na*, *s*, *po*, *k*, *za*, *iz*, *do*, *ot*, *dlja*, *o*, *pro* und *bez*. *Posle* folgt erst auf Platz 17 und *nasčet* mit großem Abstand auf Platz 36 unter allen Präpositionen des mündlichen Teilkorpus. *O, pro* und *bez* nehmen jedoch vordere Ränge ein, diese hatte wiederum Zemskaja nicht vermerkt.

In der RR-Forschung wurde mehrfach darauf hingewiesen, dass Präpositionen in der Umgangssprache generell seltener aufträten als in anderen Textsorten (Zemskaja 1983, 113, Prokurovskaja 1983/2003, 159). Dies bestätigen die Daten des NČSl: Auch hier sind im mündlichen Teilkorpus die Präpositionen in der gesprochenen Sprache allgemein auf niedrigeren Rängen angesiedelt als in den anderen Textsorten, wo sie weit vorne rangieren. Die Präposition *v* ‚in' steht im Gesamtkorpus und in der Belletristik unter allen Lemmas (also nicht nur Präpositionen, sondern Lemmas aller Wortarten) auf Rang 2, in den Korpora der Publizistik und der sonstigen nichtbelletristischen Literatur findet sie sich sogar auf dem ersten Rang. Im Korpus der mündlichen Sprache folgt *v* hingegen erst auf Rang 5, und sie ist die einzige Präposition unter den häufigsten 10 Wörtern, während sich in den anderen Textsorten jeweils mindestens drei Präpositionen auf den obersten 10 Rängen finden.

	NČSl ges		Mündlich		Belletristik		Publizistik		Nichtbell.		Chi-Quadrat (df=4)	Cramérs V
	Rang	ipm	Rang	ipm	Rang	ipm	Rang	ipm	Rang	ipm		
v / vo	1	32.542,0 [a]	1	20.005,0 [b]	1	26.558,3 [c]	1	37.539,7 [e]	1	35.644,5 [d]	79.551,38**	,021
na	2	15.867,3 [a]	3	11.776,4 [b]	2	16.688,4 [c]	2	16.596,9 [e]	2	13.509,4 [d]	9.023,51**	,007
s / so	3	12.463,2 [a]	4	9.386,4 [b]	3	12.674 [c]	3	12.649,1 [c]	3	12.348,0 [d]	784,27**	,002
po	4	5.786,7 [a]	5	3.612,2 [b]	6	4.736,9 [c]	4	6.541,6 [e]	4	6.838,9 [d]	14.112,36**	,009
k / ko	5	5.628,7 [a]	6	3.214,1 [b]	4	5.711,6 [c]	5	5.617,5 [a]	5	5.590,5 [a]	838,06**	,002
iz	6	4.314,1 [a]	8	1.707,2 [b]	8	4.114,2 [c]	7	4.772,9 [d]	9	3.849,3 [e]	4.236,63**	,005
u	7	4.306,1 [a]	2	11.807,9 [b]	5	5.314,9 [c]	9	3.906,7 [d]	11	3.372,5 [e]	22.466,30**	,011
o / ob	8	4.121,6 [a]	11	1.314,1 [b]	10	3.086,3 [c]	6	4.792,1 [d]	7	4.680,3 [e]	16.220,2**	,009
za	9	3.904,1 [a]	7	2.707,6 [b]	7	4.474,3 [c]	8	3.991,4 [d]	12	2.702,6 [b]	8.968,9**	,007
ot	10	3.672,5 [a]	10	1.338,0 [b]	9	3.615,3 [c]	11	3.651,8 [a, c]	8	3.955,0 [d]	1.503,71**	,003
dlja	11	3.229,3 [a]	12	1.096,1 [b]	12	1.818,1 [c]	10	3.883,7 [d]	6	5.044,8 [e]	43.728,93**	,015
do	12	2.061,1 [a]	9	1.606,4 [b]	11	1.881,9 [c]	12	2.255,6 [d]	13	2.111,4 [e]	1.359,03**	,003
pri	13	1.550,8 [a]	18	289,8 [b]	18	759,0 [c]	13	1.577,5 [d]	10	3.460,2 [e]	51.133,85**	,017
pod	14	1.126,0 [a]	17	442,2 [b]	13	1.406,4 [c]	15	1.063,6 [d]	17	748,8 [e]	4.839,10**	,005
posle	15	1.080,1 [a]	15	609,8 [b]	16	860,9 [c]	14	1.349,3 [d]	14	993,6 [e]	4.459,63**	,005
bez	16	1.018,8 [a]	14	788,7 [b]	14	1.140,5 [c]	16	1.024,5 [a]	16	793,2 [b]	1.316,58**	,003
čerez	17	805,4 [a]	16	580,8 [b]	15	896,3 [c]	17	801,5 [a]	18	626,2 [b]	1.026,1**	,002
pered	18	626,7 [a]	19	244,4 [b]	17	760,3 [c]	19	575,1 [d]	20	373,1 [e]	2.955,99**	,004
meždu	19	607,5 [a]	20	236,9 [b]	21	447,2 [c]	18	631,3 [d]	15	942,8 [e]	4.553,09**	,005
nad	20	537,0 [a]	24	118,4 [b]	19	696,7 [c]	20	492,5 [d]	22	279,4 [e]	3.994,74**	,005
pro	21	392,1 [a]	13	884,5 [b]	20	609,4 [c]	26	282,4 [d]	25	214,9 [e]	7.063,04**	,006
krome	22	343,6 [a]	23	126,0 [b]	24	263,2 [c]	21	406,4 [d]	19	391,4 [d]	1.319,24**	,003
sredi	23	314,9 [a]	26	30,2 [b]	23	267,1 [c]	22	372,4 [d]	21	282,8 [e]	919,96**	,002
iz-za	24	289,6 [a]	21	229,3 [b]	22	348,6 [c]	25	285,3 [a]	26	197,6 [b]	876,49**	,002
protiv	25	236,5 [a]	25	49,1 [b]	25	147,0 [c]	23	327,3 [d]	24	217,2 [e]	2712,19**	,004
okolo	26	228,1 [a]	22	178,9 [b]	26	125,3 [c]	24	322,2 [d]	23	252,6 [d]	3.179,82**	,004

Vorhergehende Seite: Tabelle 2: Rang unter den Präpositionen und relative Frequenzen (*ipm*) für die im Gesamtkorpus des NČSl häufigsten 26 Präpositionen und Informationen zur Signifikanz des Unterschiedes pro Präposition[20]

Tab. 2 zeigt die häufigsten 26 Präpositionen des Gesamtkorpus des NČSl im Vergleich mit den anderen Textsorten.[21] Für diese habe ich den Rang unter allen Präpositionen und die Frequenzwerte in den verschiedenen Teilkorpora ermittelt.

Die Spalten mit den Rangangaben lassen erkennen, dass sich die Rangfolge der Präpositionen in den verschiedenen Textsorten hier und da unterscheidet. Eine Berechnung der Korrelationen zeigt, dass die Rangfolgen der Präpositionen in den Teilkorpora erwartungsgemäß meist eine fast perfekte Korrelation aufweisen. Auffällige Abweichungen von der Rangfolge der anderen Textsorten zeigt aber das mündliche Teilkorpus. Entsprechend ist die Korrelation der Rangfolge zwischen dem Gesamtkorpus und dem mündlichen Teilkorpus auch die niedrigste.[22]

In den Spalten mit den *ipm*-Werten lässt sich ablesen, dass die Häufigkeit der Präpositionen zwischen den Korpora meist divergiert. Ein Chi-Quadrat-Test erbrachte für die meisten Frequenzwerte untereinander statistisch hochsignifikante Unterschiede. Bei der Interpretation ist allerdings eine gewisse Vorsicht angebracht.[23] Die Betrachtung der *ipm*-Werte

---

[20] Pro Zeile gilt: gleicher Buchstabe = kein signifikanter Unterschied, unterschiedliche Buchstaben = signifikanter Unterschied. Der Wert in der Spalte „Chi-Quadrat" bezieht sich auf die gesamte Zeile. Die Signifikanz wurde mit einem Chi-Quadrat-Test auf der Grundlage der absoluten Frequenzen der Tokens sowie der Tokens im Gesamtkorpus (Bonferroni-korrigiert für multiples Testen) berechnet. Aus Gründen der Übersichtlichkeit fasse ich die Ergebnisse mit der Tabelle zusammen, die die *ipm*-Werte darstellt, die nicht die Grundlage des Tests waren.

[21] Ausgangspunkt waren zunächst die häufigsten 30 Präpositionen. Unter diesen fanden sich aber einige rein lautlich bedingte Varianten (*o* und *ob*, *v* und *vo*, *s* und *so*, *k* und *ko*); diese habe ich zusammengefasst, sodass sich 26 verschiedene Präpositionen ergaben.

[22] Die Korrelationen (nach Pearson) im Einzelnen:

	Publiz.	**Nichtbell.**	**Mündl.**	**NČSl ges.**
Belletr.	,979**, p < ,001	963**, p < ,001	,940**, p < ,001	,989**, p < ,001
Publiz.		,995**, p < ,001	,906**, p < ,001	,998**, p < ,001
Nichtbell.			,889**, p < ,001	,991**, p < ,001
Mündl.				,920**, p = ,007
NČSl ges.				

[23] In der korpuslinguistischen Forschung wurde mehrfach darauf hingewiesen, dass große Datenmengen sehr schnell zu signifikanten Unterschieden führen (Kilgariff

des mündlichen Teilkorpus zeigt jedoch einige sehr stark hervortretende Auffälligkeiten. Zu diesen gehört insbesondere eine sehr niedrige Frequenz der Präpositionen im mündlichen Korpus generell, vgl. etwa *v/vo, o/ob* oder *dlja* und *pri,* aber auch *meždu, krome, sredi* oder *protiv*. Bemerkenswert ist weiterhin, dass die Vorkommenshäufigkeit in der Belletristik meist deutlich näher bei der mündlichen Sprache liegt als die Werte der anderen Textsorten und des Gesamtkorpus. Hier bietet sich also eine gute Ausgleichsmöglichkeit.

Es gibt aber auch zwei Fälle, in denen die Frequenz der Präposition in der mündlichen Sprache beträchtlich höher liegt als in den anderen Korpora: Dies sind *u* sowie *pro*. Die oben bereits angesprochene Präposition *u* ist eine besonders herausstechende Ausnahme in Bezug auf die generell niedrigere Frequenz von Präpositionen in der gesprochenen Sprache: Die Frequenz von *u* im mündlichen Korpus liegt 2–3mal höher als in den anderen Textsorten.

Wie sind die Besonderheiten im Auftreten von Präpositionen in der mündlichen Sprache zu erklären? Hier können wir wieder auf die Forschungen zur RR zurückgreifen. Das generell seltene Auftreten von Präpositionen erklärt Zemskaja (1983, 113) zum einen dadurch, dass auch Substantive als diejenige Wortart, die von Präpositionen begleitet wird, in der russischen Umgangssprache seltener seien als in anderen Textsorten (s. dazu den nächsten Abschnitt). Außerdem dominierten in der *Razgovornaja Reč'* der Nominativ und Wortfügungen per Adjunktion.[24]

Ein anschauliches Beispiel für die generell niedrigere Frequenz von Präpositionen liefert der unter (1) angeführte Beleg aus der *Razgovornaja Reč'* (Zemskaja/Kitajgorodskaja/Širjaev 1981). Es demonstriert gleichzei-

---

1996, Oakes 1998, 29). Die Effektstärke des Unterschiedes in der letzten Spalte von Tab. 2 (Cramérs V) weist darauf hin, dass die Unterschiede nicht sehr stark sind. Ein nennenswerter Effekt wird nur bei *v, u, o/ob, dlja* und *pri* erreicht. Auch dieses Maß sollte mit Vorsicht interpretiert werden, weil es sich als anfällig gegen unterschiedliche Randverteilungen erweist. Letztlich bleibt eine Diskussion der *ipm*-Werte vorerst die beste Lösung. Eine noch bessere Deutung würden die logarithmierten Werte ermöglichen, da der Frequenzanstieg nicht linear ist; davon sehe ich hier aus Platzgründen ab.

[24] Ein gewisser Widerspruch dazu ist die von Zemskaja (1983, 96) konstatierte Ausweitung der Funktionen der Präpositionen in der russischen gesprochenen Sprache, die die Autorin in Zusammenhang mit der oft für die RR beobachteten Tendenz zum Analytismus stellt.

tig, wie es zur auffällig hohen Frequenz von *u* in der gesprochenen Sprache kommt: Der Vergleich mit der ‚Übersetzung' der Autoren in eine explizite Form (1b) illustriert, wie die elliptischen Strukturen der gesprochenen Sprache zum Wegfall der Substantive samt Präposition führen und *u nas* ‚bei uns' als Ersatz für eine komplexe Substantivgruppe eintritt:

(1) *Chleb u nas segodnja ničego / my ne chodili* ‚Das Brot bei uns ist heute okay / wir sind nicht gegangen'

(1b) *Chleb v našem bližajšem chlebnom magazine segodnja neožidanno ničego (chorošij, mjagkij), i poėtomu my ne chodili v magazin za školu, gde chleb vsegda mjagkij* ‚Das Brot in unserem nächsten Brotladen ist heute unerwartet okay (gut, weich), und deswegen sind wir nicht in den Laden hinter die Schule gegangen, wo das Brot immer weich ist'

(Zemskaja/Kitajgorodskaja/Širjaev 1981, 220)

Prokurovskaja (1983/2003, 160) verbindet die hohe Frequenz von *u* mit der ‚mestoimennost'', der Vorliebe für Pronomina der mündlichen Sprache – rund 80% der Personalpronomina träten gemeinsam mit der Präposition *u* auf. Ein zweiter Aspekt sei die Präferenz für den Themenkreis *a u nas, a u vas* ‚und so ist es bei uns, und so ist es bei euch' bzw. ‚und wir haben..., und ihr habt...' in der gesprochenen Sprache. Aus übereinzelsprachlicher Perspektive gesehen spiegelt sich hier natürlich auch ein charakteristischer Zug des Russischen, nämlich seine Konstruktion *u* XGEN *est'* ‚bei X ist' = ‚X hat, besitzt'.

Die zweite Präposition, die in der mündlichen Sprache deutlich häufiger belegt ist als in den anderen Textsorten, ist *pro* ‚über, um'; dieses Ergebnis ist in den oben genannten Arbeiten zur RR nicht dokumentiert. Die Präposition gilt generell als umgangssprachlich und konkurriert in der Funktion der Markierung des Gesprächsgegenstandes mit *o* ‚über'; die letztere Präposition ist entsprechend in den anderen Textsorten häu-

figer. Dass *pro* überhaupt unter den häufigsten 26 Präpositionen des Gesamtkorpus genannt wird, dürfte auf den Anstieg umgangssprachlicher Lexik im schriftlichen Bereich zurückgehen.[25]

4.3. Wortartenverteilung in Textsorten

Wie im letzten Abschnitt diskutiert, treten im mündlichen Korpus deutlich weniger Präpositionen auf als in den anderen Textsorten und im gesamten Korpus. Dies führt uns zu der Frage, inwiefern sich die Häufigkeiten der Wortarten in den im NČSl repräsentierten Textsorten unterscheiden. Diese Frage kann mit den im NČSl gegebenen Informationen gut analysiert werden, denn die Frequenzlisten geben neben Rang und *ipm* auch jeweils die Wortart an. Sie sollen daher als zweites Fallbeispiel diskutiert werden. Hier ist allerdings zwischen Frequenzen nach Types, also der Zahl verschiedener Lemmas einer Wortart, und Tokens, den tatsächlich vorkommenden Formen, zu unterscheiden.

Mit Blick auf die hier besonders interessierende Frage nach Unterschieden zwischen der gesprochenen Sprache und anderen Textsorten möchte ich zunächst wieder auf die Forschung zur russischen Standardumgangssprache, der *Razgovornaja Reč'*, zurückgreifen. Sirotinina (1983b/2003, 7) fasst zusammen, dass der größte Teil der Lexik der russischen Standardumgangssprache zwar neutral sei und sich mit der allgemeinen russischen Sprache decke. Dennoch sei die Verteilung der Wortarten in der RR spezifisch. Anhand der Daten in Zasorina (1977) und der Daten der Saratover RR stellt sie auf Token-Ebene folgende Größenverhältnisse fest: Im Russischen insgesamt verzeichnet Sirotinina 27% Substantive, 13% Pronomina und 1% Partikeln, denen in der RR 15% Substantive, 17% Pronomina und 15% Partikeln gegenüberstehen. Für Verben ermittelt sie hingegen 17% in beiden Datengruppen.

Stoljarova (1992/2003) präsentiert eine noch detailliertere Untersuchung der Wortartenverteilung anhand der Saratover RR. Auf der Ebene der Tokens benennt sie als eines der hervorstechendsten Merkmale die starke Dominanz von substantivischen Pronomen (z. B. *on* ‚er'). Sie sind

---

[25] Die Verteilung der Vorkommen dieser Präposition nach Jahren, die das Russische Nationalkorpus unter dem Link „raspredelenie po godam" ausgibt, zeigt entsprechend eine gleichmäßig niedrige relative Frequenz im Korpus von 1800 bis etwa 2005 und einen anschließenden sehr starken Anstieg.

mit der starken Situationsgebundenheit und Subjektivität der mündlichen Umgangssprache zu erklären (ebd., 6). Ihr steht die relative Seltenheit von Substantiven gegenüber, die Zemskaja (1983) insbesondere mit zahlreichen Ellipsen erklärt.

Beide Forscherinnen stellen dabei erheblich mehr Substantiv-Ellipsen als Verb-Ellipsen fest. Stoljarova (1992/2003, 13) nennt hierfür als Erklärung beispielsweise die höhere Eindeutigkeit der Dinge und Ersetzbarkeit durch deiktische Gesten, die im Alltagsgespräch relevant sind (*papa tebe nal'et* statt *papa tebe nal'et vina* ‚Papa wird dir [Wein] einschenken'). Weiterhin können Substantive in häufig vorkommenden Kollokationen leicht ausgelassen werden (*on sdaet* statt *on sdaet ėkzamen* ‚er legt [das Examen] ab') (ebd., 17). Ein weiterer Aspekt der geringeren Häufigkeit von Substantiven seien für die mündliche Sprache typische Wortbildungsprozesse, bei denen Substantive durch ihre begleitenden Attribute ersetzt, letztere aber noch nicht als substantiviert aufgefasst würden (z. B. *kontrol'naja* < *kontrol'naja rabota* ‚kontrollierende [Arbeit], Prüfung') (ebd., 14).

Auch Adjektive sind in der gesprochenen Sprache seltener zu verzeichnen, die Gründe sind hier nach Stoljarova (1992/2003, 17ff.) zum Teil in der niedrigeren Frequenz der Substantive, aber auch in der geringeren Explizitheit der gesprochenen Sprache generell zu sehen. Adverbien und Partikeln sind hingegen als typische expressive Ausdrucksmittel in der gesprochenen Sprache sehr häufig.

Anders verhält es sich in Bezug auf Types: Hier beobachtet Stoljarova (1983/2003, 21) ein Dominieren der Substantive über die Verben. Während bei den Verben also eine hohe Token-Frequenz mit einer eher geringen Type-Frequenz verbunden ist, verhält es sich ihr zufolge bei den Substantiven umgekehrt.

Kempgen präsentiert eine Untersuchung von Markov (1960), der die Tokens in einem Korpus von 12.000 laufenden Wortformen auswertete, und zwar getrennt nach „Autorenrede" und „Personenrede" (Kempgen 1995a, 37, vgl. auch 1999, 535). Diese kontrastiert Kempgen mit einer eigenen Auswertung von Wortarten nach Types eines rückläufigen Wörterbuches. In den Tokens der Autorenrede, also den als geschriebensprachlich konzipierten Textteilen, ermittelt Markov 28% Substantive gegenüber 19% Substantiven in der Personenrede, also in den als mündlich

konzipierten Textteilen. Bei den Verben stehen sich 16% in der Autorenrede und 20% in der Personenrede gegenüber. Weitere auffällige Unterschiede sind erheblich geringere Anteile der Partikeln in der Autorenrede (3%) gegenüber der Personenrede (10%) und umgekehrt mehr Präpositionen in der Autorenrede (12%) als in der Personenrede (8%).

In Bezug auf den Vergleich von Types und Tokens sind besonders die starken Unterschiede zwischen Types und Tokens bei den Funktionswörtern hervorzuheben, was natürlich aus strukturellen Gründen (geschlossene, eher kleine Klasse mit hoher Verwendungshäufigkeit) erwartbar ist (vgl. entsprechend Kempgen 1995a, 37).

Wenden wir uns nun den neuen und umfangreicheren Daten zu, wie sie das NČSl präsentiert.

4.3.1. Wortarten nach Types

Kommen wir zunächst zur Type-Frequenz, also der Frage, wie viele unterschiedliche Substantive, Verben etc. unter den 4.927 häufigsten Lemmas genannt werden. Tab. 3 gibt eine Übersicht über die Type-Anteile der Wortarten in den Daten des NČSl auf die Textsorten bezogen. Anhand dieser Werte bestätigt sich die Beobachtung von Stoljarova, dass Substantive in Bezug auf die Zahl unterschiedlicher Lemmas die quantitativ dominierende Wortart in allen Textsorten sind – auch in den mündlichen Texten. Die Zahl ist hier niedriger als in der Publizistik und der sonstigen nichtbelletristischen Literatur, aber höher als in der Belletristik und liegt nahe beim Wert für das gesamte Korpus.

Der Type-Wert für Verben ist allerdings im Vergleich zu den anderen Textsorten relativ hoch; hier unterscheiden sich die Daten von den Beobachtungen Stoljarovas. Nur die Belletristik hat einen höheren Verb-Anteil unter den Types, die anderen Textsorten und auch das Gesamtkorpus liegen niedriger. Der Anteil der Adjektive in der gesprochenen Sprache ist – wiederum übereinstimmend mit Stoljarovas Daten – niedrig. Überraschenderweise gilt dies auch für das belletristische Korpus; einen höheren Adjektiv-Anteil weisen Publizistik und die sonstigen nichtbelletristischen Texte auf. Nicht überraschend ist der höhere Anteil an verschiedenen Types von Partikeln und Interjektionen im mündlichen Korpus im Vergleich zu den anderen Textsorten und dem Gesamtkorpus. Während wir bisher schon häufig deutliche Ähnlichkeiten zwischen belletristischem und mündlichem Korpus beobachten konnten, zeigt sich

aber bei den Interjektionen ein klarer Unterschied zwischen diesen beiden Textsorten. Die Werte für das Gesamtkorpus des NČSl stehen hier wiederum bei unterschiedlichen Ausprägungen der Häufigkeiten meist etwa in der Mitte zwischen den Randpositionen.

	Sign.[26]	NČSl ges.	Mündlich	Belletristik	Publizistik	Nichtbell.
s	**	45,6%	43,1%	40,6%	46,5%	48,6%
v	**	24,8%	28,9%	30,9%	23,0%	19,4%
pr	n.s.	1,3%	1,0%	1,2%	1,1%	1,2%
a	**	15,9%	12,1%	12,5%	17,4%	20,2%
conj	n.s.	0,8%	0,7%	0,8%	0,8%	0,7%
spro	n.s.	0,6%	0,7%	0,6%	0,6%	0,6%
adv	**	7,0%	7,0%	8,6%	6,8%	5,7%
apro	n.s.	0,8%	0,8%	0,9%	0,8%	0,8%
part	n.s.	1,1%	1,5%	1,1%	1,0%	1,0%
advpro	n.s.	1,0%	1,3%	1,3%	1,0%	0,8%
num	n.s.	0,8%	1,0%	0,9%	0,7%	0,6%
anum	*	0,2%	0,5%	0,2%	0,2%	0,2%
intj	**	0,2%	1,3%	0,3%	0,1%	0,3%
Gesamt		100,0%	100,0%	100,0%	100,0%	100,0%

Tabelle 3: Prozentuale Anteile der Wortarten (Types) unter den 4.927 häufigsten Wörtern[27]

[26] Die Chi-Quadrat-Tests für die einzelnen Wortarten wurden anhand der absoluten Häufigkeiten berechnet. Sie ergaben: Substantive: $\chi 2(4) = 41{,}729$, $p < ,001$; Verben: $\chi 2(4) = 163{,}407$, $p < ,001$; Präpositionen: $\chi 2(4) = 1{,}387$, $p = ,846$; Adjektive: $\chi 2(4) = 145{,}694$ $p < ,001$; Konjunktionen: $\chi 2(4) = ,460$, $p = ,977$; Substantivpronomina $\chi 2(4) = ,675$, $p = ,954$; Adverbien: $\chi 2(4) = 29{,}680$, $p < ,001$; Adjektivpronomina: $\chi 2(4) = ,422$, $p = ,981$; Partikeln $\chi 2(4) = 8{,}447$, $p = ,077$; Adverbialpronomina $\chi 2(4) = 8{,}511$ $p = ,075$; Numeralia $\chi 2(4) = 6{,}071$, $p = ,194$; Adjektivnumeralia $\chi 2(4) = 9{,}493$ $p = ,05$; Interjektionen $\chi 2(4) = 103{,}574$, $p < ,001$.

[27] Die Wortartenzuweisung im RNK folgt im Wesentlichen und in allen Zweifelsfällen den Prinzipien von Zaliznjak (1977). Es werden die folgenden Wortarten unterschieden (gemeinsam genannte werden als eine Klasse behandelt): s (Substantiv), v (Verb), pr (Präposition), a (Adjektiv), conj (Konjunktion), spro (Substantiv-Pronomen), adv, praedic und parenth (Adverb, Prädikativ und „vvodnoe slovo"), apro (Adjektiv-Pronomen), part (Partikel), advpro und praedicpro (Adverb-Pronomen [nicht bei Zaliznjak] und prädikative Pronomen), num (Numerale), anum (Adjektiv-Numerale), intj (Interjektion), nonlex und com (Sonstiges). Reflexive und nichtreflexive Verben sowie perfektive und imperfektive Verben mit derselben Wurzel werden getrennt gezählt (vgl. Šarov/Ljaševskaja 2009/2016, xi).

4.3.2. Wortarten nach Tokens

Die Ermittlung der Frequenzinformationen zu den Tokens ist etwas schwieriger, da sie im NČSl nicht für die einzelnen Textsorten angegeben werden.[28] Sie können aber aus den relativen Frequenzdaten (*ipm*) und den Angaben zum Korpusumfang errechnet werden. Die prozentualen Angaben beziehen sich hier wie auch bei den Types nur auf die 4.927 häufigsten Wörter und nicht auf das Gesamtkorpus. Die so ermittelten Werte sind in Tab. 4 angeführt.[29]

Von den einzelnen Textsorten zeigen Belletristik, Publizistik und Nichtbelletristik eine annähernd perfekte Korrelation mit den Werten für das gesamte Korpus. Auch die Werte für die Textsorten untereinander sind sehr stark korreliert, etwas geringer Belletristik mit den sonstigen nichtbelletristischen Texten und Publizistik. Einzig die Textsorte Mündlich fällt hier wieder heraus. Mit dem Gesamtkorpus zeigt sie eine deutlich geringere Korrelation, die Korrelationen mit den anderen Textsorten sind meist noch deutlich niedriger. Eine Ausnahme macht hier nur die Korrelation zwischen mündlichem Korpus und Belletristik, die stark ist. Insgesamt ergibt sich im Wesentlichen dasselbe Bild, das auch Sirotinina (1983b/2003) und Stoljarova (1992/2003) für Umgangssprache und die Vergleichswerte zum Russischen insgesamt zeichnen: In den mündlichen Texten sind äußerst wenige Substantiv- und Adjektiv-Tokens zu verzeichnen, dafür aber deutlich mehr Substantiv-Pronomina, Partikeln und

---

[28] Das NČSl bietet in einer der sog. Hilfslisten, nämlich „Dannye o častotnosti častereč-nych klassov“ ‚Angaben zur Frequenz der Wortartenklassen‘, zwar Angaben zur Wortartenverteilung unter den Tokens; aber sie bezieht sich auf das gesamte Korpus und ist nicht nach Textsorten aufgegliedert.

[29] Während in Bezug auf die Types eine Korrelation keine ergiebigen Informationen erbrachte, da die Wortarten im Wesentlichen in derselben Rangfolge auftreten, ist diese Methode für die Tokens gut geeignet. Sie hat zudem den Vorteil, dass sie Aussagen über das Verhältnis zwischen den Textsorten liefert. Die Korrelationen (nach Pearson) im Einzelnen:

	Publiz.	**Nichtbell.**	**Mündl.**	**NČSl ges.**
Belletr.	,919**, p < ,001	,833**, p < ,001	,858**, p < ,001	,956**, p < ,001
Publiz.		,982**, p < ,001	,631*, p = ,021	,994**, p < ,001
Nicht-bell.			,507, p = ,077	,957**, p < ,001
Mündl.				,706** p = ,007

Interjektionen. Der Anteil der Verben im mündlichen Teilkorpus unterscheidet sich nicht stark vom Gesamtkorpus.

Wortart	NČSl ges.	Mündlich	Belletristik	Publizistik	Nichtbell.
s	25,2%	13,12%	19,82%	27,41%	33,97%
v	15,18%	17,2%	18,2%	14,19%	12,06%
pr	13,76%	8,79%	12,56%	14,26%	14,01%
a	7,68%	3,8%	5,67%	8,78%	10,88%
conj	9,68%	8,62%	10,19%	9,27%	8,49%
spro	8,78%	15,57%	11,51%	7,3%	5,07%
adv	5,26%	7,12%	6,08%	5,11%	4,16%
apro	5,74%	4,66%	5,38%	5,95%	5,24%
part	4,97%	12,25%	6,11%	4,31%	3,48%
advpro	2,48%	5,87%	3,15%	2,14%	1,59%
num	0,82%	1,39%	0,87%	0,84%	0,62%
anum	0,38%	0,44%	0,34%	0,42%	0,34%
intj	0,06%	1,17%	0,12%	0,02%	0,08%
Gesamt	100,0%	100,0%	100,0%	100,0%	100,0%

Tabelle 4: Prozentuale Anteile der Wortarten (Tokens) unter den 4.927 häufigsten Wörtern

Im Gegensatz zu den Werten bei der Verteilung der Wortarten nach Types, in denen mündliches Korpus und Belletristik nahe beieinander lagen, treten bei den Tokens stärkere Unterschiede zwischen diesen beiden Textsorten hervor. Hier möchte ich an die oben erwähnte Untersuchung zur Wortartenverteilung in Autoren- und Personenrede von Markov (1960) erinnern. Es zeigt sich, dass die Verteilung der Wortarten nach Tokens in der Personenrede in der von Markov untersuchten Belletristik der mündlichen Sprache recht zutreffend nachempfunden ist und insofern als Ersatz herangezogen werden kann. Sicherlich geht die im Vergleich zu den anderen Textsorten größere Nähe der Belletristik zur mündlichen Sprache auf die Anteile imitierter gesprochener Sprache zurück. Die Wortartenverteilung nach Tokens im Gesamtkorpus unterscheidet sich hingegen teilweise sehr stark von der gesprochenen Sprache und bildet diese also nur schlecht ab.

5. Fazit

Frequenzeigenschaften von Wörtern sind eine Größe, die in der jüngeren Zeit in der empirischen Forschung eine wichtige Rolle spielt; insbesondere in der Psycholinguistik kommt ihnen eine wichtige Funktion als Kontrollvariable zu. Um verlässliche Informationen zur Wortfrequenz zu gewinnen, sind möglichst große Datenmengen notwendig. Mit dem online verfügbaren NČSl (*Novyj Častotnyj Slovar' Russkoj Leksiki* ‚Neues Frequenzwörterbuch der russischen Lexik', Ljaševskaja/Šarov 2009/2016, online) steht seit einigen Jahren für das Russische eine Informationsquelle mit einem sehr breiten Datenfundament zur Verfügung. Allerdings basiert das Korpus überwiegend auf schriftlichen Texten; im menschlichen Umgang mit Sprache hat hingegen die mündlich verwendete Sprache einen großen Anteil. Es stellt sich daher die Frage, inwieweit die Wortfrequenzdaten repräsentativ für die Sprachform sind, mit der Menschen tatsächlich zu tun haben.

In diesem Artikel habe ich die Unterschiede zwischen den Frequenzdaten zu verschiedenen Textsorten schlaglichtartig an zwei Beispielen beleuchtet, wobei die Besonderheiten des mündlichen Teilkorpus im Fokus standen. 1. Der Vergleich der Frequenz von Präpositionen im Gesamtkorpus und in den unterschiedlichen Textsorten mit derjenigen in der mündlichen Sprache erbrachte teilweise deutliche Unterschiede. Dabei sind die meisten Präpositionen in der mündlichen Sprache erheblich seltener als im Gesamtkorpus, die Präpositionen *u* und *pro* hingegen signifikant häufiger. 2. In der Wortartenverteilung werden insbesondere bei den Tokens große Unterschiede zwischen mündlicher Sprache und den anderen Textsorten deutlich. In erster Linie sind dies ein erheblich niedrigerer Anteil an Substantiven und Adjektiven, etwas mehr Verben und deutlich mehr Substantivpronomina, Partikeln und Interjektionen.

In mancher Hinsicht zeigen das mündliche und das belletristische Korpus starke Ähnlichkeiten, sodass die Belletristik – vermutlich insbesondere durch ihre größeren Anteile an direkter Rede, die der mündlichen Sprache nachempfunden ist – durchaus dazu geeignet ist, die Unterrepräsentanz der mündlichen Sprache in gewissem Umfang auszugleichen. Die Ähnlichkeit hat jedoch Grenzen; dies wurde sowohl bei den Präpositionen als auch in Bezug auf die Wortartenverteilung deutlich. Die

Beobachtungen der früheren quantitativen Forschungen zu Textsortenunterschieden im Russischen, wie sie Kempgen (1995a/2007, 1999) im Überblick präsentiert, sowie die Untersuchungen zur *Russkaja Razgovornaja Reč'* ließen sich im Wesentlichen bestätigen, sie haben ihre Tragfähigkeit insofern bewiesen und können als wichtige ausgleichende Informationsquelle herangezogen werden.

Literatur

Anstatt, Tanja (2016): „Subjektive Frequenz als Forschungsmethode." In: *Wiener Slawistischer Almanach* 77. 7–35.

Brysbaert, Marc/New, Boris (2009): „Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English." In: *Behavior Research Methods* 41 (4). 977–990.

Clasmeier, Christina/Anstatt, Tanja/Ernst, Jessica/Belke, Eva (2016): „Are *Schalter* and *šapka* good competitors? Searching for stimuli for an investigation of the Russian-German bilingual mental lexicon." In: Anstatt, Tanja/Clasmeier, Christina/Gattnar, Anja (Hrsg.), *Slavic Languages in Psycholinguistics. Chances and Challenges for Empirical and Experimental Research*. Tübingen. 191–224.

Ellis, Nick (2002): „Frequency effects in language processing. A Review with Implications for Theories of Implicit and Explicit Language Acquisition." In: *Studies in Second Language Acquisition* 24 (2). 143–188.

Grigoriev, Andrei/Oshhepkov, Ivan (2013): „Objective age of acquisition norms for a set of 286 words in Russian: Relationships with other psycholinguistic variables." In: *Behavior Research Methods*. https://www.researchgate.net/publication/235717084_Objective_age_of_acquisition_norms_for_a_set_of_286_words_in_Russian_Relationships_with_other_psycholinguistic_variables (letzter Aufruf 16.8.2016).

Kelih, Emmerich (2008*): Geschichte der Anwendung quantitativer Verfahren in der russischen Sprach- und Literaturwissenschaft*. Hamburg (Studien zur Slavistik 19).

Kempgen, Sebastian (1981/2., elektron. Auflage 2008) *„Wortarten" als klassifikatorisches Problem der deskriptiven Grammatik. Historische und systematische Untersuchungen am Beispiel des Russischen*. Bamberg.

Kempgen, Sebastian (1995a/2., elektron. Auflage 2007): *Russische Sprachstatistik: Systematischer Überblick und Bibliographie*. München.

Kempgen, Sebastian (1995b): „Codierung natürlicher Sprache auf morphologischer Ebene." In: *Die Welt der Slaven XL* 1. 52–57.

Kempgen, Sebastian (1999): „Quantitative Aspekte des Russischen." In: Jachnow, Helmut (Hrsg.), *Handbuch der sprachwissenschaftlichen Russistik und ihrer Grenzdisziplinen*. Wiesbaden. 525–550.

Kempgen, Sebastian (2004): „Greenbergs phonologische Universalien und das Russische." In: Lehmann, Volkmar/Udolph, Ludger (Hrsg.), *Normen, Namen und Tendenzen in der Slavia. Festschrift für Karl Gutschmidt zum 65. Geburtstag*. München. 191–194.

Kempgen, Sebastian (2007): „Zur Zeitoptimierung der russischen Verbalmorphologie." In: Köhler, Reinhard/Grzybek, Peter (Hrsg.), *Exact Methods in the Study of Language and Text. Dedicated to Gabriel Altmann on the Occasion of his 75th Birthday*. Berlin/New York. 281–286.

Kempgen, Sebastian/Lehfeldt, Werner (2004): „Quantitative Typologie." In: Booij, Geert et al. (Hrsg.), *Morphologie/Morphology. Ein internationales Handbuch zur Flexion und Wortbildung*. Berlin. 1235–1246 (Handbücher zur Sprach- und Kommunikationswissenschaft 17.2).

Kilgarriff, Adam (1996): „Comparing word frequencies across corpora: Why chi-square doesn't work, and an improved LOB-Brown comparison." In: *Proceedings from ALLC-ACH '96*. 169–172.

Koch, Peter/Oesterreicher, Wulf (1985): „Sprache der Nähe – Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungfeld von Sprachtheorie und Sprachgeschichte." In: *Romanisches Jahrbuch* 36. 15–43.

Kopotev, Michail (2008): „K postroeniju častotnoj grammatiki russkogo jazyka. Padežnaja sistema po korpusnym dannym." In: Mustajoki, Arto et al. (Hrsg.), *Instrumentarij rusistiki: Korpusnye podchody*. Helsinki. 136–151 (Slavica Helsingiensia 34).

Lagerberg, Robert (2011): *Variation and Frequency in Russian Word Stress*. München/Berlin.

Ljaševskaja, Ol'ga/Šarov, Sergej (2009): *Častotnyj slovar' sovremennogo russkogo jazyka: Na materialach nacional'nogo korpusa russkogo jazyka*. Moskva.

Ljaševskaja, Ol'ga/Šarov, Sergej (online): *Novyj častotnyj slovar' sovremennogo russkogo jazyka: Na materialach nacional'nogo korpusa russkogo jazyka*. http://dict.ruslang.ru/freq.php (letzter Aufruf 20.1.2017).

Markov, Jurij (1960): „K voprosu o častotnosti grammaticeskich kategorij." In: *Russkij jazyk v nacional'noj škole* 4. 19–20.

Markov, Jurij (1966): „Nekotorye aspekty razgovornoj reči s točki zrenija leksičeskoj statistiki." In: *Russkij jazyk v nacional'noj škole* 5. 21–28.

Oakes, Michael P. (1998): *Statistics for Corpus Linguistics*. Edinburgh.

Prokurovskaja, N. A. (1983, unveränd. Nachdruck 2003): „Neznamenatel'naja leksika." In: Sirotinina, O. B. (Hrsg.), *Razgovornaja reč' v sisteme funkcional'nych stilej sovremennogo russkogo literaturnogo jazyka. Leksika*. Moskva. 157–187.

Rayson, Paul/Berridge, Damon/Francis, Brian (2004): „Extending the Cochran rule for the comparison of word frequencies between corpora." In: *7th International Conference on Statistical analysis of textual data (JADT 2004*). Louvain-la-Neuve, Belgium. 926–936.

Šarov, Sergej/Ljaševskaja, Olga (2009/2016): *Vvedenie k novomu častotnomu slovarju russkoj leksiki*. http://dict.ruslang.ru/freq.php (letzter Aufruf 20.1.2017).

Sirotinina, O. B. (Hrsg.) (1983a, unveränd. Nachdruck 2003): *Razgovornaja reč' v sisteme funkcional'nych stilej sovremennogo russkogo literaturnogo jazyka. Leksika*. Moskva.

Sirotinina, O. B. (Hrsg.) (1992, unveränd. Nachdruck 2003): *Razgovornaja reč' v sisteme funkcional'nych stilej sovremennogo russkogo literaturnogo jazyka. Grammatika*. Moskva.

Sirotinina, O. B. (1983b, unveränd. Nachdruck 2003): „Obščaja charakteristika leksiki razgovornoj reči.“ In: Sirotinina, O. B. (Hrsg.), *Razgovornaja reč' v sisteme funkcional'nych stilej sovremennogo russkogo literaturnogo jazyka. Leksika*. Moskva. 6–10.

Stoljarova, Ė. A. (1983, unveränd. Nachdruck 2003): „Suščestvitel'nye.“ In: Sirotinina, O. B. (Hrsg.), *Razgovornaja reč' v sisteme funkcional'nych stilej sovremennogo russkogo literaturnogo jazyka. Leksika*. Moskva. 21–48.

Stoljarova, Ė. A. (1992, unveränd. Nachdruck 2003): „Časti reči.“ In: Sirotinina, O. B. (Hrsg.), 4–47.

Valgina, Nina (2003): *Teorija teksta: Učebnoe posobie*. Moskva.

Vlasova, Roza/Sinitsyn, Valentin/Pechenova, Ekaterina (2015): „The Effect of Word Frequency on the Brain Correlates of Object Naming in Russian.“ In: *The Russian Journal of Cognitive Science* 2 (1). 24–40.

Zaliznjak, A. A. (1997): *Grammatičeskij slovar' russkogo jazyka: Slovoizmenenie*. Moskva.

Zasorina, Lidia (1977): *Častotynj slovar' russkogo jazyka*. Moskva.

Zemskaja, Elena/Kitajgorodskaja, Margarita/Širjaev, Evgenij (1981): *Russkaja razgovornaja reč'. Obščie voprosy. Slovoobrazovanie. Sintaksis*. Moskva.

Zemskaja, E. A. (Hrsg.) (1973): *Russkaja razgovornaja reč'*. Moskva.

Zemskaja, E. A. (Hrsg.) (1983): *Russkaja razgovornaja reč'. Fonetika. Morfologija. Leksika. Žest*. Moskva.