Nutzungshinweise zu den Lemmalisten für das Teilkorpus DEU_L1_EV

September 2015

David Stoppel, Franziska Wallner

Einleitung

Die Lemmalisten liefern Häufigkeitsangaben für Wörter der deutschen gesprochenen Wissenschaftssprache. Sie bieten einen Überblick über die sprachlichen Einheiten des deutschsprachigen L1-Expertenkorpus (DEU_L1_EV). Diese sind in ihren jeweiligen Grundformen (Lemmata) angegeben, so wie sie in einem Wörterbuch zu finden wären. Verschiedene Wortformen können dabei auf ein gemeinsames Lemma zurückgeführt werden. So entsprechen die folgenden Wortformen mit ihren jeweiligen Häufigkeiten

seriöse *3*

seriösen *1*

einem Eintrag in der Lemmaliste mit der summierten Häufigkeit

seriös *4*

Typisch gesprochensprachliche Phänomene, wie beispielsweise Klitisierungen, wurden in ihren schriftsprachlichen Standard überführt (orthografisch normalisiert) und entsprechend lemmatisiert (*haste -> hast du -> haben du*).

Es können separate Lemmalisten für ausgewählte Wortarten ausgegeben werden. Eine Beschreibung der einzelnen Wortarten finden Sie im Abschnitt Wortarten und Konventionen.

Nutzungshinweise

Über die Menüpunkte oberhalb der Lemmaliste können Sie verschiedene Teillisten anwählen. Standardmäßig wird die Gesamtliste aller Lemmata des Teilkorpus angezeigt, zu der Sie immer durch Klick auf den Menüpunkt **alle Lemmata** zurückkehren können (Abbildung 1).

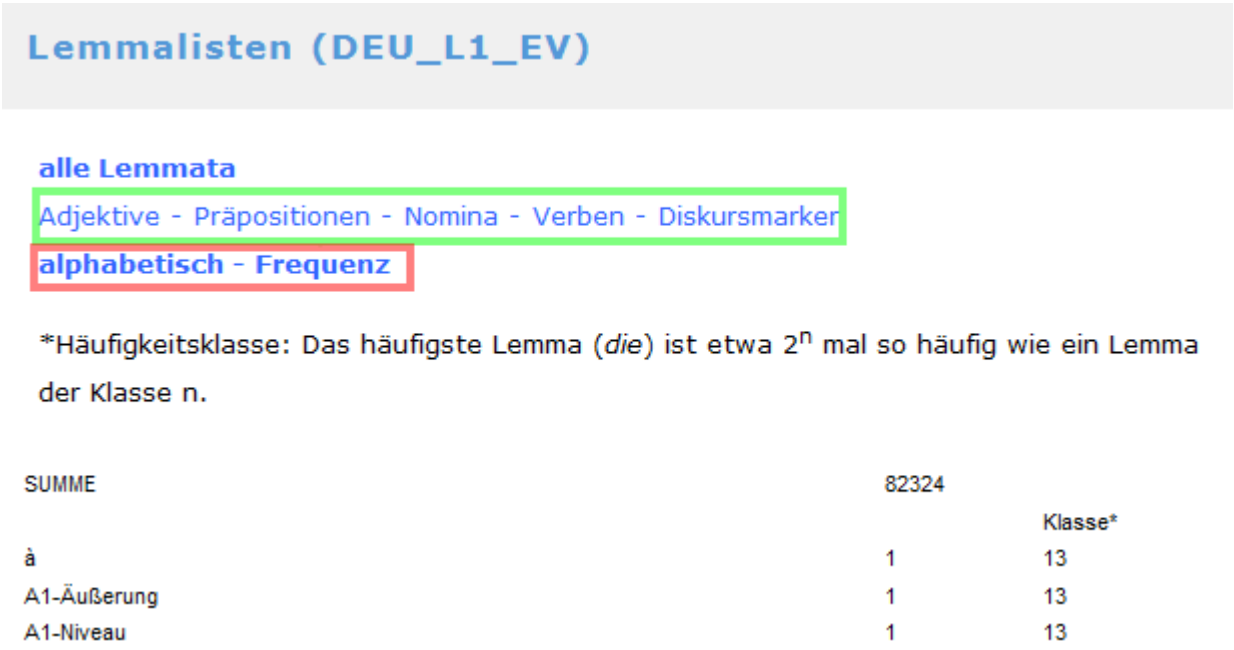

Abb. 1: Lemmaliste – Menü

Durch Klicken auf **alphabetisch** bzw. **Frequenz** (Abb. 1, rot umrahmt) können Sie die aktuell angewählte Liste alphabetisch bzw. nach Frequenz sortieren lassen. Durch Klicken auf eine der Wortarten (grün umrahmt) können Sie sich eine separate Lemmaliste für die jeweilige Wortart anzeigen lassen, die Sie wiederum sortieren können.

Die Liste führt zu jedem Lemma (Type) die Häufigkeit seines Vorkommens (Token) auf. Im Kopf der Liste sehen Sie die Gesamtsumme der Token in der aktuell gewählten Liste (Abbildung 2, grün umrahmt – beispielhaft für die Verben, die 12531 Token im Teilkorpus umfassen). Für jedes Lemma ist die Anzahl der Token innerhalb der angewählten Wortart zu sehen (rot umrahmt für *sein* – 2245 Token unter den Verben).

Abb. 2.: Lemmaliste – Angaben

Die Anzahl der Token innerhalb einer Wortart kann dabei von der Anzahl der Token in der Gesamtliste abweichen. Unter einem Lemma in der Gesamtliste können mehrere verschiedene Lemmata zusammengefasst sein, die nur oberflächlich identisch sind. Im obigen Beispiel tritt das Lemma *sein* 2245 Mal als Verb auf. In der Gesamtliste ist *sein* mit 2296 Belegen aufgeführt. Die Differenz ergibt sich daraus, dass das Possessivpronomen *sein* in der Gesamtliste unter demselben Lemma *sein* subsumiert wird.[1] Jedes Lemma ist zudem einer Häufigkeitsklasse zugeordnet (Abb. 2 – rechte Spalte). Dabei ist ein Lemma beispielsweise der Klasse 3 zugeordnet, wenn das häufigste Lemma der Gesamtliste (*die*) etwa 2^3=8 Mal so häufig auftritt.[2]

Bereinigung der Listen

Nicht alle Belege sind in die Listen mit eingeflossen. Dazu zählen jene Belegstellen in den Transkripten, bei denen keine eindeutige Transkription vorgenommen werden konnte – so in Fällen, in denen mehrere, voneinander abweichende Transkriptionsvorschläge gemacht wurden. Darüber hinaus wurden die folgenden Kategorien ausgeschlossen:

- **Eigennamen**,
- **fremdsprachliches Material** (Lemmata, die nicht im Duden (2013) oder im Duden Fremdwörterbuch (Duden 2010) gelistet sind),
- **Nichtwörter** (Einheiten der Transkription, die keiner sprachlichen Einheit direkt zuzuordnen sind (Platzhalterzeichen); Sprechpausen; Stottern; nicht-

[1] Es ist zu berücksichtigen, dass sich nicht in allen Fällen gleichlautende Lemmata unterschiedlicher Teillisten zur Gesamtzahl aufsummieren. Zum einen werden einige Wortarten, wie die Artikel, die Adverbien oder die Pronomen nicht in separaten Teillisten aufgeführt, werden aber trotzdem in der Gesamtliste erfasst. Zum anderen ist aufgrund der in gesprochener Sprache häufigen syntaktischen Abbrüche und Reformulierungen nicht immer eine eindeutige Zuweisung zu einer Wortart möglich. Beispielsweise kann *sein* je nach syntaktischem Kontext Verb oder Possessivpronomen sein. Wurde der Sprechbeitrag jedoch direkt nach der Äußerung von *sein* abgebrochen oder eine Reformulierung vorgenommen, ist keine eindeutige Zuordnung möglich. In diesem Fall würde *sein* ebenfalls in keiner der Teillisten für die einzelnen Wortarten erfasst werden.

[2] Die Berechnung der Häufigkeitsklasse eines Lemmas *lemma* folgt DEREWO (2013):
$K(lemma) = \lfloor \log_2(f(die) / f(lemma)) + 0{,}5 \rfloor$, wobei K(*lemma*) die Häufigkeitsklasse und f(*lemma*) die Frequenz von *lemma* angibt. Es ist zu beachten, dass die Berechnung der Häufigkeitsklasse immer auf die Häufigkeit des Lemmas *die* in der Gesamtliste bezogen ist, nicht auf die Häufigkeit innerhalb der aktuellen Teilliste. Für das Korpus DEU_L1_EV gilt damit f(*die*) = 7629.

rekonstruierbare Wortabbrüche; im Einklang mit dem STTS Nichtwörter wie Sonderzeichen oder Kombinationen aus Ziffern und Buchstaben („Aufgabe *1a*“)),

- **objektsprachliches Material** (umfasst beispielsweise Zitate aus literarischen Werken, die selbst Gegenstand der literaturwissenschaftlichen Analyse eines Vortrags sind sowie sprachliche Einheiten, die auf linguistischer Ebene analysiert werden).

Wortarten und Konventionen

Für einige Wortarten wurden separate Lemmalisten angelegt. Mit Ausnahme der Diskursmarker folgen die Zuordnungen weitgehend den im Stuttgart-Tübingen Tag Set (STTS) (Schiller et al., 1995) aufgeführten Kategorien. Berücksichtigung finden dabei diejenigen Wortarten, für die das POS-Tagging im L1-Expertenkorpus (DEU_L1_EV) bereits als weitestgehend abgeschlossen angesehen werden kann. Da die manuelle Korrektur des automatischen POS Taggings noch nicht abgeschlossen ist, können für einige Kategorien (wie bspw. Adverbien und Konjunktionen) noch keine separaten Listen erstellt werden.

Adjektive

(POS-Tags ADJD/ADJA im STTS): umfasst alle Adjektive, sofern diese nicht als Diskursmarker (s. u.) gebraucht wurden sowie Partizipien, die ohne Hilfsverb verwendet wurden (attributiv wie bspw. „*gesprochene Sprache*“ oder in elliptischen Konstruktionen wie bspw. „*wie gesagt*“).

Präpositionen

(Adpositionen im Sinne des STTS: POS-Tags APPR/APPRART/APPO/APZR): Umfassen neben eigentlichen Präpositionen auch Postpositionen (*zufolge*) und Zirkumpositionen (*um…willen*). Dabei werden einzelne Bestandteile von Zirkumpositionen separat in der Lemmaliste aufgeführt.

Nomina

(POS-Tag NN im STTS): umfasst sämtliche Nomina mit Ausnahme der Eigennamen (s. o.)

Verben

(POS-Tags beginnend mit V im STTS): diese umfassen Voll-, Hilfs- und Modalverben in finiten und infiniten Formen, im Imperativ sowie im Partizip Perfekt mit Hilfsverb („er wurde *geschlagen*"; „das ist nicht klar *geregelt*").

Trennbare Verben mit disloziertem Präfix („er *fand* sich damit *ab*") wurden wie folgt lemmatisiert: der Stamm inklusive Präfix (*abfinden*), das Präfix mit Bindestrich (*ab-*). Bei der Entscheidung welche Verben als trennbar gelten, wurde Duden (2013) gefolgt und in Zweifelsfällen der empfohlenen Variante der Vorzug gegeben (z. B. *zueinanderpassen* statt *zueinander passen* wurde als trennbares Verb behandelt; *klein schneiden* statt *kleinschneiden* wurde als Verb und Adjektiv behandelt).

Diskursmarker

Die Kategorie der Diskursmarker umfasst Elemente mit gesprächssteuernder und -strukturierender Funktion, die für die gesprochene Sprache charakteristisch sind und syntaktisch untypische Positionen besetzen. So werden beispielsweise *ja* und *so* in den Textbeispielen (1) und (2) als Diskursmarker analysiert, da sie entgegen dem schriftsprachlichen Standard im Vor-Vorfeld auftreten und eine diskurssteuerende Funktion übernehmen.

(1) „**ja** (0.71) vielen dank für die einladung auch hier vortragen zu dürfen" (EV_DE_104)

(2) „(1.39) **so** ich möchte ihnen gerne drei (.) beispiele aus unserem korpus vorstellen" (EV_DE_092)

Die Zuweisung zu den Diskursmarkern folgt hier im Wesentlichen den von Westpfahl (2014) vorgeschlagenen Ergänzungen und Anpassungen des STTS für gesprochene Sprache.[3]

[3] Ausführliche Informationen zur Kategorie Diskusmarker im GeWiss-Korpus finden sich in Wallner (in Vorb.) sowie in Fandrych/Meißner/Sadowski/Wallner (in Vorb.).

Literatur

DEREWO (2013): Korpusbasierte Wortgrundformenliste DEREWO, v-ww-bll-320000g-2012-12-31-1.0, mit Benutzerdokumentation, http://www.ids-mannheim.de/derewo. © Institut für Deutsche Sprache, Programmbereich Korpuslinguistik, Mannheim, Deutschland, 2013.

DUDEN (2010): Duden – Das Fremdwörterbuch, 10. Auflage. Mannheim/Zürich: Dudenverlag.

DUDEN (2013): Duden – Die deutsche Rechtschreibung, 26. Auflage. Berlin/Mannheim/Zürich: Dudenverlag.

Fandrych, Christian/Meißner, Cordula/Sadowski, Sabrina/Wallner, Franziska (in Vorb.): Gesprochene Wissenschaftssprache – digital. Verfahren zur Annotation und Analyse mündlicher Korpora. Tübingen: Stauffenburg.

Schiller, Anne/Teufel, Simone/Stöckert, Christine/Thielen, Christine (1999): Guidelines für das Tagging deutscher Textcorpora mit STTS (Kleines und großes Tagset). Institut für maschinelle Sprachverarbeitung, Stuttgart. http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf

Wallner, Franziska (in Vorb.): Diskursmarker in der gesprochenen Wissenschaftssprache. In: Kontutytė, Eglė / Žeimantienė, Vaiva (Hgg.): Sprache in der Wissenschaft. Germanistische Einblicke. Duisburger Arbeiten zur Sprach- und Kulturwissenschaft. Peter Lang.

Westpfahl, Swantje (2014): STTS 2.0? Improving the Tagset for the Part-of-Speech-Tagging of German Spoken Data. In: Lori, Levin/Stede, Manfred (Hgg.) Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop, Dublin, Irland, 1-10.