Pál Uzonyi/Viktória Dabóczy

Quantitative Korpusuntersuchungen zur Phrasenstruktur von deutschen und ungarischen linken Feldern

Abstract: Mit Hilfe von Korpusuntersuchungen wird die Konstituentenstruktur im deutschen und ungarischen präverbalen Feld quantitativ analysiert. Der Schwerpunkt liegt auf der Bestimmung des Anteils von Mehrfachbesetzungen, d.h. linken Feldern mit mehreren primären Satzkomponenten. Im Deutschen kommt es zumeist dann zu einer komplexen Vorfelddbesetzung, wenn infinite Konstituenten des Verbalkomplexes ihre Satelliten mit ins Vorfeld ziehen. Ihr Anteil liegt in präverbalen Bi-, Tri-, Tetra- und Pentagrammen des deutschen Wikipedia-Korpus etwa bei einem Promille. Wesentlich weniger Belege gibt es mit echter Mehrfachbesetzung. Demgegenüber beträgt der Anteil an Mehrfachbesetzungen in präverbalen Bi- und Trigrammen des ungarischen Wikipedia-Korpus ca. 50%. Infinite Verbformen stehen in 0,4% der ungarischen Bi- und Trigramme vor dem Finitum.

Abstract (ung.): Cikkünkben mennyiségi korpuszelemzés segítségével vizsgáljuk a ragozott ige előtt található összetevős szerkezetet a németben és a magyarban. A középpontban a ragozott ige előtti többszörös mondatrészelőfordulások arányának meghatározása áll. A németben többnyire akkor találkozunk összetett előmezővel, ha az igei csoport igenévi részei (főnévi igenév, participium) saját bővítményeiket magukkal viszik az előmezőbe. Ezek hányada a német Wikipedia-korpuszban a két, három, négy és öt szót tartalmazó előmezőkben nagyjából egy ezreléket tesz ki. Jóval kisebb ez az arány, ha minden szempontból különböző mondatrészekről van szó. A magyarban ezzel szemben a több mondatrész előfordulásának aránya lényegesen magasabb, a magyar Wikipedia-korpuszban előforduló két és három szövegszót tartalmazó előmezők esetén konkrétan mintegy 50%. Az igenevek előfordulása a ragozott ige előtt a magyarban 0,4%-ot tesz ki.

1 Einleitung

Unser Beitrag setzt sich das Ziel, bezüglich der Konstituentenstruktur von deutschen und ungarischen linken Feldern (fortan: LF) mit korpuslinguistischen Mitteln quantitative Aussagen machen zu können. Unter LF verstehen wir den Abschnitt eines Satzes vor dem ersten finiten Verb. Im Deutschen geht es dabei um das Vorfeld, das i.d.R. durch eine Konstituente (Satzglied) besetzt ist. Das

finite Verb steht an zweiter Stelle in der linken Satzklammer. Im Gegensatz dazu lässt sich die Position des finiten Verbs im Ungarischen dank einer freieren Wortstellung nicht so exakt bestimmen. Die komplexe Morphologie, die für agglutinierende Sprachen charakteristisch ist, ermöglicht eine pragmatisch gesteuerte Wortstellung und eine entsprechende funktionale Verschiebbarkeit des finiten Verbs im Satz. Die im LF positionierten Elemente können auch im Ungarischen eine Konstituente darstellen, mit einer sog. Mehrfachbesetzung (d.h. mehr als einer Konstituente im LF) ist jedoch öfter zu rechnen.

Im Mittelpunkt unserer Untersuchung steht die Erfassung der Mehrfachbesetzungen in deutschen und ungarischen LFn. Die Recherchen werden dabei auf sog. Bi- und Trigramme, d.h. aus zwei bzw. drei Textwörtern bestehende Ketten beschränkt. Neben dem quantitativen Vergleich beider Sprachen steht auch die Zusammensetzung der Mehrfachbesetzungen im Interesse. D.h. welche Wortarten neigen dazu, mehrere Konstituenten im LF zu bilden?

Zur Recherche verwenden wir die getaggten Korpora Wiki_DE_Öffentlich¹ bzw. Wiki_HU2-Öffentlich, die im Rahmen des EuroGr@mm-Projekts aufgrund der deutschen bzw. ungarischen Version von Wikipedia² zu Forschungszwecken erstellt worden sind. Da eine syntaktische, phrasenstrukturelle Annotation von keinem der beiden angewendeten Taggern (Deutsch: TreeTagger³, Ungarisch: HUMOR⁴ und TnT-Tagger⁵) geliefert wird, müssen wir mit den Suchanfragen und zusätzlichen Untersuchungen gewissermaßen auch einen Parser im Hintergrund simulieren. Dabei können wir auf syntaktische Forschungsergebnisse im Bereich der Konstituenten- und Phrasenstrukturen beider Sprachen zurückgreifen (u.a. Kiefer (Hg.) 1992; Bassola (Hg.) 1998). Einige Hinweise sind auch bei Augustin (in diesem Band) zu finden, wie z.B. der folgende⁶:

Die Frequenzbestimmung von Wortklassen-N-Grammen alleine liefert noch keine hinreichenden syntaktischen Informationen: Beispielsweise sind in der ungarischen Wikipedia Art+N+Adv und Art+N+N die frequentesten Trigramme. Auch Art+N+N kann mit zwei eigenständigen Phrasen besetzt sein, z.B. wenn das zweite N durch ein Adverbale in Form eines Nomens mit semantischem Kasusuffix (z.B. Berlinben 'in Berlin' realisiert wird, oder wenn das erste N als Attribut zum zweiten N fungiert und diese Zugehörigkeitsrelation durch ein Possessivsuffix am zweiten N ausgedrückt wird (z.B. A szezon végén... 'am Ende der Saison').

1 Im Augenblick nur für Projektteilnehmer von EuroGr@mm via COSMAS II zugänglich.

2 URL: http://en.wikipedia.org/wiki/Main_Page.

3 Das von COSMAS II verwendete Tagset siehe unter www.ids-mannheim.de/cosmas2/projekt/referenz/stts/morph.html.

4 Prószéky/Tihanyi (1996).

5 URL: www.coli.uni-saarland.de/~thorsten/tnt/.

6 Ein Abkürzungsverzeichnis befindet sich im Anhang.

Da es sich im letzteren Fall um keine Mehrfachbesetzung handelt, wollen wir auch zählen, wie hoch die Zahl derjenigen Art+N+N-Dreier ist, bei welchen das zweite Substantiv kein Possessivsuffix vor dem Suffix des semantischen Kasus hat, d.h. nicht als Bezugswort eines Attributs fungieren kann.

Mit diesem Verfahren kann man freilich – zumindest im Ungarischen – den realen Anteil von Mehrfachbesetzungen (i.e. mehreren Satzkonstituenten/Satzgliedern im präverbalen Feld) kaum genau bestimmen, aber da er im Deutschen nahe Null liegen soll, müssen schon relativ wenige Prozente eine tendenziöse, sprachtypologisch bedingte Abweichung eindeutig signalisieren.

Im Folgenden sollen zuerst die deutschen LF mit infiniten Komponenten des Verbalkomplexes untersucht werden. Im Abschnitt 3 stellen wir eine Kontrollzählung in einem virtuellen Korpus dar. Dies ist notwendig, um die Daten der deutschen und ungarischen Korpora direkt miteinander vergleichen zu können. Abschnitt 4 präsentiert die Recherchen in ungarischen LFn, während Abschnitt 5 einen kontrastiven Vergleich zwischen dem Deutschen und Ungarischen bietet.

2 Zählungen zu deutschen linken Feldern mit infiniten Komponenten des Verbalkomplexes

Mit dem Problem der Infinita im deutschen Vorfeld haben sich bereits mehrere Wissenschaftler auseinandergesetzt (u.a. Dürscheid 1989; Hoberg 1997; Müller 2005; Bassola 2009) und im Augenblick scheint die Version vorzuherrschen, nach der ein Infinitum und seine Dependienten im Vorfeld eine Einheit, d.h. keine mehrfache Besetzung darstellen. Da es aber eben ein Diskussionsthema war, ist und vermutlich bleiben wird, wollten wir diese Konstruktion bei unseren quantitativen Recherchen auf keinen Fall ausklammern (zu einer qualitativen Untersuchung des Themas siehe Bassola/Schwinn in diesem Band).

Im Korpus Wiki_DE_Öffentlich (öffentliche deutschsprachige Wikipedia-Artikel) wurden insgesamt 7.661.900 Bi- und Trigramme im Linken Feld gefunden, davon hatten 19.688, d.h. etwa 0,25%, die Struktur „(Wort +) Wort + Infinitiv/Partizip II“. Der reale Anteil ist gewiss noch geringer, da <sa> bei Weitem nicht immer am Satzanfang platziert war.

Einige Belege mit Bi- und Trigrammen:

Bedeutend eingeschränkt werden kann dieser Zeitraum durch archäologische Funde.

(http://de.wikipedia.org/wiki/Augusta_Treverorum, Stand 2011)

Legendär geworden ist ihr Auftritt beim Woodstock-Festival.
(http://de.wikipedia.org/wiki/Alvin_Lee, Stand 2011)

Mitglied werden kann jedes Mitglied der Aktivitas des Bundes, sobald die Kriterien zur Erlangung des Status Alter Herr erfüllt sind.
(<http://de.wikipedia.org/wiki/AHCC>, Stand 2011)

Zeitlich eingeordnet werden könnte der Bau in die Mitte des 8. Jahrhunderts, urkundliche Beweise fehlen jedoch.
(http://de.wikipedia.org/wiki/Abtei_Brauweiler, Stand 2011)

Exemplarisch genannt werden kann hier die Hamburger „Bildungsgesellschaft – Bildungsverein zur Hebung der arbeitenden Klasse“.
(<http://de.wikipedia.org/wiki/Arbeiterbildungsverein>, Stand 2011)

Tetra- und Pentagramme mit Infinitiv/Partizip II vor dem Finitum: 4.975 Treffer, aber nach ergänzenden Recherchen erwiesen sich viele davon als durch *W*-Wörter eingeleitete Nebensätze oder Infinitiv- bzw. Partizipialkonstruktionen, d.h., ihre reale Zahl ist kleiner, zumal der Tagger Punkte nach Abkürzungen oder Ordnungszahlen oft als Satzzeichen interpretiert hat. Es gibt 3.779.372 Treffer mit LF4 und LF5. Der Anteil von Vorfeldern, die Infinita mit ihren Satelliten enthalten, ist dementsprechend max. 0,13%. Auch bei höheren LF-Zahlen bleibt der Anteil in der Nähe von einem Promille.

Einige Belege mit Tetra- und Pentagrammen:

Selbständig als Arzt tätig werden darf man nur, wenn für drei Jahre im Rahmen des »Turnus« verschiedene (definierte) Disziplinen durchlaufen wurden und die Arbeit vom jeweiligen Abteilungsvorstand positiv bewertet wurde.
(<http://de.wikipedia.org/wiki/Arzt>, Stand 2011)

Zur Wahl notwendig gewesen wären 367 Stimmen.
(http://de.wikipedia.org/wiki/Abdullah_Gül, Stand 2011)

In Betrieb genommen werden konnte es jedoch erst im Jahre 1979, da zuvor das Ochtumsperrwerk und das Huntesperrwerk fertiggestellt werden sollten.
(http://de.wikipedia.org/wiki/Hochwasser_in_Bremen, Stand 2011)

Einem größeren Publikum bekannt geworden war Maral durch seine Darstellung des Metin in der ARD-Vorabendserie Türkisch für Anfänger, die internationale Fernsehpreise gewann.
(http://de.wikipedia.org/wiki/Adnan_Maral, Stand 2011)

Belege mit längeren n-Grammen:

Erstmals einem grösseren Kreis bekannt geworden ist er durch eine Grosslesung in Luzern zum Bloomsday 2004.

(http://de.wikipedia.org/wiki/Dominik_Riedo, Stand 2011)

Aber in der Gemeinde dem Volk vorgelesen werden kann er bis zum Ende der Tage nicht, weder unter den Propheten, deren Zahl abgeschlossen ist, noch unter den Aposteln.

(http://de.wikipedia.org/wiki/Der_Hirte_des_Hermas, Stand 2011)

Nicht ganz ihrer Favoritenrolle gerecht werden konnte sie bei der Weltmeisterschaft 1996 in der Sierra Nevada, wo sie im Riesenslalom Dritte wurde (nach dem ersten Lauf hatte sie noch auf dem 14. Platz gelegen).

(http://de.wikipedia.org/wiki/Martina_Ertl-Renz, Stand 2011)

3 Kontrollzählungen mit einem virtuellen Korpus

Trefferzahlen des deutschen und des ungarischen Korpus konnten nicht direkt verglichen werden, weil die ungarische Variante aus verschiedenen Gründen auf 10 Millionen Wörter eingeschränkt werden musste, während das deutschsprachige Korpus 551 Millionen Wörter enthält. Um uns einigermaßen zu vergewissern, dass die Umrechnung in Prozente an sich schon die richtige Vorgehensweise mit der erheblichen Umfangsdifferenz wäre, haben wir aus dem deutschen Wikipedia-Korpus ein ca. 10 Millionen Wörter enthaltendes virtuelles Korpus zusammengestellt und gewisse POS-Zählungen auch in diesem durchgeführt. Die hierbei erhaltenen Prozentwerte haben wir mit denen des Gesamtkorpus konfrontiert und Folgendes herausgefunden:

Testkorpus:

Tokens gesamt: 9.986.820

Anzahl der Texte: 32.127

Wikipedia:

Tokens gesamt: 551.090.404

Anzahl der Texte: 2.137.295

Tab. 1: LF2

häufigste POS-Kombinationen	absolute Trefferzahl im Testkorpus	relative Häufigkeit (Testkorpus)	absolute Trefferzahl in der Wikipedia	relative Häufigkeit (Wikipedia)
ART+N nn	37.983	38%	1.910.601	35%
N ne+N ne	6.868	7%	425.304	8%
AP pr+CARD	3.864	3,8%	232.940	4%
AP prart+N nn	4.071	4%	150.454	2,7%
PRON dem at+N nn	4.312	4,3%	212.616	3,8%
ADJ at+N nn	4.678	4,7%	247.086	4,5%
PRON pos at+N nn	2.915	2,9%	162.369	2,9%
AP pr+N nn	3.030	3%	150.454	2,7%
AP pr+N ne	2.363	2,4%	131.833	2,4%
N ne+N nn	3.374	3,4%	176.690	3,2%
PRON ind at+N nn	1.887	1,9%	88.393	1,6%
KON kom+N nn	1.595	1,6%	79.270	1,4%
ART+N ne	1.160	1,2%	62.797	1,1%

Den obigen Prozentwerten ist zu entnehmen, dass zwischen dem gesamten Korpus und dem reduzierten Testkorpus eigentlich nur bei der Kombination 'Präposition mit Artikel + Nomen' gravierende Unterschiede in der Häufigkeit beobachtet werden können. Es muss noch angemerkt werden, dass bei 'AP pr + N nn' mit dem Tag 'N nn' auch viele Eigennamen ausgegeben wurden.

Tab. 2: LF3

häufigste POS-Kombinationen	absolute Trefferzahl im Testkorpus	relative Häufigkeit (Testkorpus)	absolute Trefferzahl in der Wikipedia	relative Häufigkeit (Wikipedia)
ART+ADJ at+N nn	9.845	9,9%	499.278	9,1%
AP pr+ART+N nn	7.701	7,7%	405.143	7,4%
AP prart+N nn+CARD	6.101	6,1%	343.416	6,2%
AP pr+PRON dem at+N nn	3.227	3,2%	176.500	3,2%
AP prart+ADJ at+N nn	2.240	2,2%	130.417	2,4%
ART+N nn+N nn	3.403	3,4%	167.271	3%
AP pr+ADJ at+N nn	2.213	2,2%	109.130	2%
AP pr+PRON pos at+N nn	1.425	1,4%	86.882	1,6%
ART+N nn+N ne	2.534	2,5%	137.437	2,5%

häufigste POS-Kombinationen	absolute Trefferzahl im Testkorpus	relative Häufigkeit (Testkorpus)	absolute Trefferzahl in der Wikipedia	relative Häufigkeit (Wikipedia)
PRON pos at+ADJ at+N nn	957	0,96%	55.670	1%
N nn+ART+N nn	16.879	17%	858.556	16%
AP pr+PRON ind at+N nn	818	0,82%	40.251	0,73%
CARD+KON neb+CARD	413	0,41%	25.274	0,46%

Wie aus Tabelle 2 ersichtlich, sind die Unterschiede zwischen dem gesamten Korpus und dem reduzierten Testkorpus noch kleiner als bei LF2. Bei ‘ART+N nn+N ne’ gab es eine relativ hohe, bei ‘N nn+ART+N nn’ eine sehr hohe Fehlerquote.

Ingesamt lässt sich also feststellen, dass die Korpusreduzierung keinen gravierenden Einfluss auf die Häufigkeit von LFn im Vergleich zum gesamten Korpus hat.

4 Mehrfachbesetzungen in ungarischen LF2 und LF3

Die n-Gramme mit höheren n-Werten haben wir bislang wegen der stark anwachsenden Zahl der POS-Kombinationen nicht untersucht (auch laut Augustin, in diesem Band, „nimmt die Varianz der n-Gramme in Bezug auf die Linearisierung ihrer Elemente so stark zu, dass ohne ein syntaktisches Parsing nicht mehr ohne Weiteres Rückschlüsse auf die Phrasenstrukturen dieser Linken Felder getroffen werden können“). Trigramme sollen übrigens im Ungarischen die frequenteste Wortzahl im LF repräsentieren (vgl. Augustin in diesem Band), d.h., man kann von einem zahlenmäßig ausreichenden Muster sprechen, zumal die hinzukommenden Bigramme mit ihrer Tokenzahl nur geringfügig darunter bleiben. Unsere späteren Recherchen, bei denen wir nicht Interpunktionszeichen (SPUNCT), sondern den ‘hidden tag’ <s> </s> als Satzlimitator in Anspruch nahmen, ergaben für alle n-Gramme höhere Zahlen und auch die Rangordnung wurde modifiziert (2-Gramme: 25.197, 3-Gramme: 31.887, 4-Gramme: 33.653, 5-Gramme: 32.008).⁷

⁷ Die Abweichung rührt davon her, dass Kapitelüberschriften durch die zwei Anfragetypen unterschiedlich behandelt werden: Bei SPUNCT konnten wir die anhaftenden Überschriften nur mit der Kombination von CQP-Suchen und Excel-Operationen aus dem LF entfernen, was wahrscheinlich zur Aussortierung mancher valider Treffer führte. Außerdem setzt die Verwendung

Von den Bigramm- und Trigramm-Kombinationen (i.e. Types) haben wir zunächst diejenigen POS-Zweier und -Dreier gewählt, die mindestens 100 Treffer (i.e. Tokens) hatten. Mit diesen haben wir die Suche nach Mehrfachbesetzungen begonnen. Bei gewissen Kombinationen konnten wir mit einschränkenden Ergänzungen der Suchanfrage Teilmengen abtrennen, für die eine/keine Mehrfachbesetzung charakteristisch ist. Auf ähnliche Weise konnten wir aus der Menge von UNKNOWNTAG (d.h. den vom Tagger nicht erkannten Formen) zwei Teilmengen, und zwar Numeralia als Temporalbestimmungen mit Jahreszahlen bzw. Eigennamen gewinnen, die es im ursprünglichen ungarischen Tagset nicht gab, die jedoch im TreeTagger-Set als 'N ne' erscheinen. Letzteres haben wir als 'N(pr)' übernommen, in allen anderen Fällen nennen wir die Tags, wie sie im Korpus stehen.⁸

Wenn wir keine (weitere) einschränkende Suche durchführen konnten, haben wir ein aus 100 Belegen bestehendes Muster durchgesehen, um durch Extrapolation den Anteil der Mehrfachbesetzungen einschätzen zu können.

Die Ergebnisse fassen wir unten in den Tabellen 3 und 4 zusammen („Treffer insgesamt“ bedeutet alle Treffer mit der angegebenen Kombination von Wortarten, „Mehrfachbesetzungen ≈“ bedeutet die eingeschätzte oder genaue Zahl derjenigen Fälle, die mindestens zwei Satzglieder enthalten).

Tab. 3: LF2 im ungarischen Korpus

Nr.	POS-Bigramme	Treffer insgesamt	Mehrfachbesetzungen ≈
1.	Det+N	2.843	0
2.	Num+N	2.169	19
3.	A+N	1.732	0
4.	N+N	1.642	1.053
5.	N+Adv	1.251	846
6.	Adv+N	1.140	1.076
7.	Adv+Adv	899	719
8.	N(pr)+N(pr)	896	103
9.	N(pr)+Adv	800	800
10.	N+Num	676	562
11.	Num+N(pr)	646	646
12.	Pro+Adv	538	317

von SPUNCT bei der Suche voraus, dass ein Satz vorausgeht, was jedoch bei ersten Sätzen nicht der Fall ist.

⁸ Zum ungarischen Tagset s. http://mnsz.nytud.hu/sugo_eng.html#msdrendszer.

Nr.	POS-Bigramme	Treffer insgesamt	Mehrfachbesetzungen ≈
13.	Adv+N(pr)	514	514
14.	Pro+N	482	143
15.	Adv+Pro	375	338
16.	Adv+Num	369	340
17.	DIG+N	351	3
18.	Con+Adv	346	246
19.	N(pr)+Pro	306	306
20.	N+Pro	279	276
21.	Pro+Pro	260	247
22.	Det+N(pr)	250	0
23.	N+Con	238	74
24.	N+NU	237	2
25.	Adv+A	197	154
26.	MI.+N	196	0
27.	Con+N	191	191
28.	Adv+Con	188	89
29.	Pro+Con	182	69
30.	Pro+NU	175	0
31.	N+A	170	163
32.	Det+DIG	165	0
33.	N(pr)+Con	146	26
34.	Det+A	142	2
35.	N+N(pr)	141	141
36.	Con+Pro	135	135
37.	N(pr)+Num	130	111
38.	Num+Adv	128	103
39.	Pro+N(pr)	114	92
40.	N(pr)+A	106	78
41.	Con+N(pr)	101	101
Insgesamt:		21.846	10.085

21.846 (Bigramme über 100) / 25.197 (Bigramme insg.) = 86,7%

10.085 (Mehrfachbesetzungen) / 21.846 (Bigramme über 100) = 46,1%

Tab. 4: LF3 im ungarischen Korpus

Nr.	POS-Trigramme	Treffer insgesamt	Mehrfachbesetzungen $\approx$
1.	Det+N+N	2.013	886
2.	Det+N+Adv	1.749	1.714
3.	Det+A+N	1.155	0
4.	Pro+Det+N	1.149	0
5.	Det+N+N(pr)	765	711
6.	Adv+Det+N	560	560
7.	Det+N(pr)+N	544	131
8.	Adv+A+N	516	418
9.	N+N+N	457	112
10.	Num+A+N	448	58
11.	N+A+N	406	357
12.	Num+N+Num	379	0
13.	Det+N+Pro	374	363
14.	N(pr)+N(pr)+N(pr)	364	11
15.	N(pr)+N(pr)+N	358	276
16.	Adv+Num+N	357	221
17.	A+N+N	342	210
18.	Det+N+Num	315	315
19.	N+Num+N	309	309
20.	N+Det+N	306	152
21.	Det+N+NU	290	3
22.	A+A+N	281	37
23.	Det+Num+N	270	0
24.	Det+N+A	262	246
25.	Pro+A+N	242	150
26.	Det+MI.+N	238	0
27.	Det+N(pr)+Adv	223	218
28.	N+N+Adv	194	78
29.	Det+N+Con	190	95
30.	Det+N(pr)+N(pr)	181	11
31.	Num+N+Adv	180	122
32.	Con+Det+N	176	176
33.	Pro+N+N	161	90
34.	Adv+N+N	159	159

Nr.	POS-Trigramme	Treffer insgesamt	Mehrfachbesetzungen ≈
35.	N+Con+N	153	40
36.	Num+N+Con	152	12
37.	N+Adv+N	151	151
38.	N+N(pr)+N	150	113
39.	N+Adv+Adv	146	146
40.	Num+N+N	144	96
41.	A+N+Adv	138	137
42.	Pro+N+Adv	137	137
43.	Adv+N(pr)+N	129	129
44.	Det+Pro+N	120	6
45.	Adv+Adv+Adv	118	118
46.	Pro+Adv+Adv	118	118
47.	N(pr)+N(pr)+Adv	111	98
48.	N+N(pr)+N(pr)	110	80
49.	Num+Det+N	107	101
50.	Adv+Pro+N	102	96
	Insgesamt:	18.001	9.767

18.001 (Trigramme über 100) / 31.887 (Trigramme insg.) = 56,4%

9.767 (Mehrfachbesetzungen) / 18.001 (Trigramme über 100) = 54,2%

Wenn wir die untersuchten Tokens der beiden n-Gramme addieren, erhalten wir 39.847. Die Gesamtzahl der Mehrfachbesetzungen in den beiden n-Grammen beträgt 19.852, was 49,8%, d.h. etwa die Hälfte der Gesamtmenge ergibt. Anders formuliert: Wenn vor dem ungarischen Finitum zwei oder drei Wörter stehen, handelt es sich in jedem zweiten Fall um mehrere Satzglieder.

Eine derartige Verallgemeinerung ist insofern gerechtfertigt, als die untersuchten Kombinationen (Types) eindeutig die Mehrheit der Tokens abdecken. Es gab 181 Bigramm-Kombinationen mit mindestens einem Treffer, die Trefferzahl war bei 140 davon unter 100, d.h., 22,6% aller POS-Kombinationen lieferten 86,7% aller LF2-Treffer. Es gab 967 Trigramm-Kombinationen mit mindestens einem Treffer, die Trefferzahl war bei 917 davon unter 100, d.h., 5,1% aller POS-Kombinationen lieferten 56,4% aller LF3-Treffer.

Im deutschen 10-Mio-Testkorpus wurden 50.529 1-Wort-Linkefelder gefunden, außerdem 41.471 Bigramm-Tokens, 28.481 Trigramm-Tokens, 22.759 Tetragramm-Tokens usw. Das zeigt, dass in einem Korpus mit annähernd gleichem Umfang den ungarischen ca. 25.000 Bigramm-Tokens etwa 41.000 deutsche, den 32.000 ungarischen Trigrammen 28.000 deutsche gegenüberstehen.

0 und sehr kleine Zahlen waren meistens vorhersagbar aufgrund der typischen Strukturen von satzgliedwertigen Phrasen. Die drei frequentesten Bigramme Det+N, Num+N und A+N sind auch im Ungarischen prototypische NP-Strukturen. Num+N könnte wegen affigierter Numeralia in adverbialer Funktion anders sein, aber bei Gattungsnamen im Singular wäre meistens anstelle des attributiven Zahlwortes auch im Ungarischen ein Determinans nötig. Ein Beweis dafür ist einerseits, dass Num+Det+N (Trigramm Nr. 49) von 107 Vorkommen 101-mal zwei Satzglieder enthält, und andererseits, dass Num+N(pr), d.h., wenn es kein Gattungsname ist, hundertprozentig für Mehrfachbesetzungen steht.

1967-ben Izrael elfoglalta [...]

(<http://hu.wikipedia.org/wiki/Al-Axa-mecset>, Stand 2011)

‘1967 Israel okkupierte’

Bei den Trigrammen war das drittfrequenteste Det+A+N kaum anders als eine NP vorstellbar, was die Untersuchung mit 0 Mehrfachbesetzungen weitestgehend bestätigt hat. Bei dem nach Frequenz darauffolgenden Trigramm Pro+Det+N hätte man dagegen keine 0 prognostiziert, weil man für diese POS-Kombination auch Mehrfachbesetzungen konstruieren kann, aber in unserem Korpus stand Pro immer für die Begleitwörter *ez* und *az*, die mit dem Det kombiniert die Äquivalente von deutschem *dieser* und *jener* sind. Die 0 war auch bei Num+N+Num (Trigramm Nr. 12) kaum vorhersagbar, wenn man nur diese Sequenz von Wortarten in Betracht zieht, aber es hat sich herausgestellt, dass dies in unserem Korpus immer für Datumsangaben des Formats „Jahr, Monat, Tag“ steht (z.B. 1998. július 12-én). Das ist keine Akkumulation, weil außerhalb dieses Gefüges sowohl Jahreszahl als auch Monatsname eigene Endungen haben müssen (z.B. 1998-ban, júliusban). Det+Num+N und Det+MI.+N schließen sich dem Typ Det+A+N an (MI. mit dem Platzhalter-Punkt steht für verschiedene Partizipien).

Ein hoher Anteil an Mehrfachbesetzungen war zu erwarten, wo mindestens eine der Wortarten ein Adverb war, was die Suchen dann auch bestätigt haben.

Hamarosan szerepet kapott [...]

(http://hu.wikipedia.org/wiki/Andrew_Hamilton_Russell, Stand 2011)

‘Bald Rolle bekam-er/sie’

Da aber auch die Negationswörter *nem* und *sem* (‘nicht’ bzw. ‘auch nicht’) als Adverbien annotiert sind, mussten wir die Fälle aussortieren, wo sie mit einem oder zwei Wörtern in ihrem Skopus eine Phrase im LF darstellen (z.B. Adv+N, Adv+Pro, Adv+A+N usw., die eben deshalb nicht immer als Mehrfachbesetzungen realisiert werden).

Nem hívásra jöttek [...]

(<http://hu.wikipedia.org/wiki/Bojkók>, Stand 2011)

‘Nicht auf-Ruf kamen-sie’

Ein Sonderfall war es allerdings, wenn das Negationswort auf das Verb bezogen war (im Ungarischen steht es auch in diesem Fall links vom Bezugswort); es ist nämlich einerseits Teil einer anderen Phrase, andererseits ist diese Phrase der Verbalkomplex selbst, also ist die LF-Position nur dann stichhaltig, wenn die Delimitator-Funktion eindeutig auf den finiten Teil des Verbalkomplexes beschränkt wird.

Irina nem bánná [...]

(http://hu.wikipedia.org/wiki/Három_nővér, Stand 2011)

‘Irina nicht würde-es-bereuen’

Hundert Prozent an Mehrfachbesetzungen bekamen wir bei N(pr)+Adv, Num+N(pr), Adv+N(pr), N(pr)+Pro, Con+N, N+N(pr), Con+Pro, Con+N(pr), Adv+Det+N, Adv+N+N, N+Adv+N, N+Adv+Adv, Pro+N+Adv, Adv+N(pr)+N, Adv+Adv+Adv, Pro+Adv+Adv. Auch bei einigen anderen Kombinationen gab es nur sehr wenige Ausnahmen, z.B. Adv+N, Adv+Pro, Adv+Num, N+Pro, Pro+Pro, N+A, N(pr)+Num, Det+N(pr)+Adv (auf Verben bezogene Negationswörter haben wir jeweils als neue Satzglieder bewertet), A+N+Adv, N(pr)+N(pr)+Adv, Adv+Pro+N.

Vor Adjektiven, Numeralien standen die als Adverbien annotierten Wörter manchmal in Gradpartikel-Funktion.

Nagyon sokat jelent [...]

(http://hu.wikipedia.org/wiki/Emma_Watson, Stand 2011)

‘Sehr viel bedeutet’

Bei Det+N+N konnten wir mehr als die Hälfte mit einer Suche nach Possessivsuffixen als *ein* Satzglied bestimmen. Wegen des Pro-Drop-Typs des Ungarischen und der (begrenzten) Weglassbarkeit des Determinans vor dem Nomen mit Possessivsuffix sind aber diese Konstruktionen in der dritten Person ohne Kontext oft ambig. So hat z.B. der Satz *A kollégium feladatának tekinti, hogy tanulói keresz-tény szellemben [...] nevelje* (http://hu.wikipedia.org/wiki/Pápai_Református_Kollégium, Stand 2011) folgende Lesarten: ‘Das Internat sieht es als seine Aufgabe an, seine Schüler in einem christlichen Geist [...] zu erziehen’ bzw. ‘Er/Sie sieht es als Aufgabe des Internats an, dass es seine Schüler in einem christlichen Geist [...] erzieht’, wobei nur der erste Fall als Mehrfachbesetzung gilt.

N+Det+N ist nur dann ambig, wenn das erste Nomen im Nominativ steht, aber in beiden Fällen handelt es sich um eine Mehrfachbesetzung, z.B. *Apu a szemüvegét keresi* – ‘Vati sucht seine (eigene) Brille’ bzw. ‘Vati sucht seine/ihre Brille’ (der deutsche Satz *Vati sucht seine Brille* ist ebenso ambig, da Subjekt und Possessivpronomen nicht zwangsläufig koreferent sind). Zu einer attributiven Konstruktion, wo jemand Vatis Brille sucht, ist noch ein Dativsuffix nötig („*Apu-nak a szemüvegét keresi*“), das aber unser Korpustool leicht finden kann. Bei Det+N+N und ähnlichen Types muss man aber wegen der Lesarten mit und ohne Mehrfachbesetzung zusätzlich in den abgetrennten Teilmengen, wenn diese zu groß sind, stichprobenartig ein Muster untersuchen.

5 Kontrastiver Vergleich der deutschen und der ungarischen linken Satzperipherie in Hinsicht auf Mehrfachbesetzungen

Deutsche Vorfelder mit mehreren erststellenfähigen Phrasen enthalten meistens eine infinite Verbform des Verbalkomplexes, wobei im Durchschnitt nur ein bis zwei Belege von 1.000 diese Konstruktion vertreten (vgl. Abschnitt 2). Nach Müller (2005) ist es allerdings nur eine scheinbar mehrfache Besetzung des Vorfeldes (zu einer detaillierteren Darstellung der Problematik s. Bassola/Schwinn in diesem Band). Im Ungarischen, wo kein mit dem deutschen vergleichbares Vorfeld und keine Verbalklammer vorliegen, gilt eine Wortfolge „Infinitum Auxiliar“ als völlig unmarkiert.

Mindörökké *él*ni fog a szívünkben.

(http://hu.wikipedia.org/wiki/Janet_Jackson, Stand 2011)

‘Ewig leben wird (...)’

Ezzel is *provokálni* akarta a reformpárti klérust.

([http://hu.wikipedia.org/wiki/VIII._Gergely_\(ellenpápa\)](http://hu.wikipedia.org/wiki/VIII._Gergely_(ellenpápa)), Stand 2011)

‘Hiermit auch provozieren wollte (...)’

Egyik vége *rögzítve* van, a másik, elvékonyított vége szabadon ...

(<http://hu.wikipedia.org/wiki/Szaxofon>, Stand 2011)

‘Das-eine sein-Ende befestigt ist (...)’

Dass ihr Anteil mit 0,4% nicht um Größenordnungen höher ist als im Deutschen, mag daran liegen, dass es im Ungarischen wesentlich weniger analytische Formen gibt: Als Tempusform nur das Futur, als Genus Verbi nur ein Zustandspassiv

und ein morphologisch analoges Zustandsaktiv (vgl. Uzonyi 2010). In der Hungarologie werden diese Zustandskonstruktionen allerdings nicht als analytische Verbformen behandelt (vgl. z.B. Keszler/Lengyel 2008: 92); es gibt nur spärlich abweichende Versionen (z.B. Lengyel 2000 oder ProGr@mm kontrastiv 2013). Auch Infinitive bei Äquivalenten von deutschen Modalverben werden der Tradition nach teils als Objekte (z.B. *segíteni akarok* 'ich will helfen'), teils als Subjekte angesehen (z.B. *segítenem kell* 'ich muss helfen'), vgl. z.B. Keszler/Lengyel (2008).

Im Ungarischen lassen sich aufgrund der Ergebnisse von Abschnitt 4 diejenigen LF2- und LF3-Types bestimmen, für die eine signifikante Mehrfachbesetzung charakteristisch ist. Wir haben entsprechende deutsche Types daraufhin untersucht, welche syntaktischen Beziehungen in diesen zwischen denselben oder ähnlichen Kategorien bestehen. In manchen Fällen können wir deshalb nicht von identischen, sondern nur von ähnlichen Kategorien sprechen, weil einerseits die beiden Systeme von vorneherein nicht isomorph sind und andererseits die verwendeten Tagger nicht mit dem gleichen Tagset arbeiten. Ein direkter Vergleich der Zahlen der Tokens ist wiederum etwas irreführend, da die Kurven, die den Anteil der einzelnen n-Gramme abbilden, in den beiden Sprachen bei weitem nicht parallel verlaufen. Laut Augustin (in diesem Band: Abb. 1a und Abb. 1f) liegen deutsche Bigramme bei ca. 21%, während die ungarischen Pendantes kaum mehr als ca. 12% erreichen. Der Anteil an Trigrammen soll beinahe gleich sein, an dieser Stelle schneiden sich nämlich die beiden Kurven in der Höhe von etwa 15%. Nur bei Letzteren kann man daher Trefferzahlen bzw. Prozentwerte direkt miteinander vergleichen, bei Bigrammen muss man nämlich mit einbeziehen, dass in deutschen Texten etwa zweimal so viele LF2-Tokens zu finden sind wie in ungarischen Texten.

In der unten stehenden Tabelle werden ungarischen LF2- und LF3-Types deutsche Types gegenübergestellt, die bezüglich der POS-Kombination den ungarischen am nächsten stehen. Zu den ungarischen Types werden jeweils die Token-Zahl und die Zahl der Mehrfachbesetzungen angegeben. Die deutschen Tokens stammen aus dem virtuellen 10-Mio-Korpus (vgl. Abschnitt 3). Eventuelle sporadische deutsche Mehrfachbesetzungen sind den Kommentaren zu entnehmen, die unter jeder Zeile zu sehen sind und vornehmlich auf die deutschen Tokens Bezug nehmen. Die Tabelle ist in drei Zonen geteilt: Ungarische Types, bei denen (a) alle Tokens eine Mehrfachbesetzung aufweisen; (b) min. 90% (aber weniger als 100%) der Tokens eine Mehrfachbesetzung aufweisen; (c) der Anteil der Mehrfachbesetzungen zwischen 40% und 90% liegt, aber mit einer Type-Vorkommenshäufigkeit von min. 3% (100% ist die Gesamtzahl der entsprechenden n-Gramme).

Tab. 5a: 100% der ungarischen Belege stellen Mehrfachbesetzungen dar

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
N(pr)+Adv	800	800	N ne + ADV	164
Kommentar: ADV ist <i>selbst</i> in 80 Belegen, <i>hingegen</i> : 26, <i>dagegen</i> : 13, <i>jedoch</i> : 9, <i>aber</i> : 7 usw.				
Num+N(pr)	646	646	CARD + N ne	429
Kommentar: Das Zahlwort ist hierbei meistens eine Jahreszahl in einer chronologischen Aufzählung (mit oder ohne Doppelpunkt), d.h. keine Vorfeldkonstituente – in den ungarischen Belegen ist es dagegen immer eine Jahreszahl mit Adverbialsuffix, d.h. Temporaladverbial.				
Adv+N(pr)	514	514	ADV + N ne	166
Kommentar: ADV sind <i>auch</i> , <i>selbst</i> , <i>nur</i> usw. Fehler rühren zumeist von einem falsch platzierten Satzgrenzen-Tag her.				
N(pr)+Pro	306	306	N ne + PRON	33
Kommentar: 20 sind falsch, 13 richtig, Letztere mit <i>dagegen</i> als PRON.				
Con+N	191	191	KON + N nn	1888
Kommentar: Davon sind 1733 <i>als</i> – im Ungarischen entspricht diesem meist das Kasusuffix des Formalis: <i>-ként</i> .				
N+N(pr)	141	141	N nn + N ne	932
Kommentar: Mit hoher Fehlerrate: Von den ersten 100 sind nur 57 richtig, die Extrapolation aufs Ganze ergibt ≈530 richtige Belege. Darunter gibt es viele „enge Apposition + Name“, z.B. <i>Präsident Gorbatschow</i> , was jedoch im Ungarischen mit anderer Wortfolge ausgedrückt wird: <i>Gorbacsov elnök</i>				
Con+Pro	135	135	KON + PRON	449
Kommentar: 88 davon sind Nebensätze, es bleiben also 361 Bigramme, die mit koordinierender Konjunktion im Vorfeld – und z.T. im Vor-Vorfeld – stehen; Fehler rühren z.T. von als KON getaggtten Adpositionen wie <i>bis dahin</i> , <i>statt dessen</i> her.				
Con+N(pr)	101	101	KON + N ne	137
Kommentar: 21 davon sind Nebensätze, es bleiben also 116, kaum mehr als die ungarischen Belege. Alle 116 enthalten koordinierende Konjunktionen bzw. das Äquivalent des ung. Kasus Formalis, nämlich <i>als</i> .				
Adv+Det+N	560	560	ADV + ART + N nn	368
Kommentar: In den Belegen gibt es 232 <i>auch</i> als ADV, deren ung. Äquivalent <i>is</i> rechts vom Bezugswort steht; die anderen sind ebenfalls NP-bezogene Partikeln, z.B. <i>lediglich</i> , <i>nur</i> usw.				
Adv+N+N	159	159	ADV + N nn + N nn	19
Kommentar: Davon 3 falsch. ADV sind <i>auch</i> und andere Partikeln. In fast allen Belegen steht N nn für Eigennamen.				

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
N+Adv+N	151	151	N nn + ADV + N nn	30
Kommentar: 24 davon sind metasprachliches „x (auch y)“ oder „x, auch y“, oder statt <i>auch</i> je einmal <i>anfangs</i> , <i>ehemals</i> , <i>alias</i> , plus 3 falsche Treffer.				
N+Adv+ Adv	146	146	N nn + ADV + ADV	2
Kommentar: Beide Belege sind falsch.				
Pro+N+Adv	137	137	PRON + N nn + ADV	39
Kommentar: 6 sind falsch, sonst sind PRON Determinantien und Quantoren zu den N nn. ADV sind nachgestellte Attribute. Es gibt eine Ausnahme: <i>Dieser Überlegung zugrunde liegt die Annahme, dass ...</i> , wo ADV ein Verbzusatz, d.h. Teil einer anderen Primärkomponente im Vorfeld ist, was möglicherweise als Mehrfachbesetzung gilt.				
Adv+N(pr)+N	129	129	ADV + N ne + N nn	26
Kommentar: 2 sind falsch, sonst sind ADV NP-bezogene Partikeln wie <i>auch</i> , <i>selbst</i> , <i>insbesondere</i> usw.				
Adv+Adv+ Adv	118	118	ADV + ADV + ADV	33
Kommentar: 27 davon sind <i>Auch heute noch</i> , auch die restlichen 6 sind richtige Treffer, aber ohne Mehrfachbesetzung.				
Pro+Adv+ Adv	118	118	PRON + ADV + ADV	2
Kommentar: Die zwei Belege sind <i>Er selbst aber</i> und <i>Diese selbst aber</i> .				

Tab. 5b: Min. 90% der ungarischen Belege stellen Mehrfachbesetzungen dar

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
Adv+N	1.140	1.076	ADV + N nn	169
Kommentar: Etwa 5% Fehlerquote, sonst vorwiegend <i>auch</i> , <i>nur</i> , <i>besonders</i> und andere Partikeln; die Eigennamen <i>Anders</i> und <i>Dann</i> wurden falsch als ADV getaggt.				
Adv+Pro	375	338	ADV + PRON	196
Kommentar: Niedrige Fehlerquote; ADV sind wortbezogene Partikeln wie <i>auch</i> , <i>selbst</i> , <i>insbesondere</i> usw. bzw. Fügungen wie <i>bald darauf</i> , <i>erst danach</i> usw. Wenn <i>einhergehen</i> als lexikalische Einheit angesehen wird, dürfte folgender Beleg eine Mehrfachbesetzung darstellen: <i>Einher damit ging eine Umstellung auf populärere Inhalte und auf ...</i>				
Adv+Num	369	340	ADV + CARD	729
Kommentar: <i>bereits/schon/erst</i> usw. + Jahreszahl				
N+Pro	279	276	N nn + PRON	237
Kommentar: ca. 10% falsch, sonst zumeist Komplemente von Substantiven wie <i>Grund dafür</i> etc., seltener adversatives <i>dagegen</i> .				

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
Pro+Pro	260	247	PRON + PRON	90
Kommentar: ca. 20% eingeleiteter NS, weitere 2-3% falsch, der Rest heterogen, z.B. <i>einige davon, dies alles, all das</i> usw.				
N+A	170	163	N nn + ADJ	134
Kommentar: 81, d.h. ca. 60% sind falsch. Typische Konstruktionen mit lexikalischer Variation sind: <i>Monate später, Kohlenstoff (schwarz), Projekt 667B</i> . Weitere Konstruktionen: <i>Ski nordisch kommentiert er zusammen mit ...; Flughäfen weltweit sind damit nun in der Lage, ...</i> ; im letzteren Beleg tritt <i>weltweit</i> in adverbialen Gebrauch auf. ⁹				
Det+N(pr)+Adv	223	218	ART + N ne + ADV	13
Kommentar: <i>7 selbst, 3 wiederum, 2 hingegen, 1 seinerseits</i> .				
A+N+Adv	138	137	ADJ + N nn + ADV	32
Kommentar: 11 falsch, sonst sind ADV nachgestellte Attribute oder adversative Fügewörter wie <i>hingegen, jedoch, wiederum</i> .				
Adv+Pro+N	102	96	ADV + PRON + N nn	205
Kommentar: 2-3% falsch, sonst sind ADV Partikeln wie <i>auch, nur, sehr</i> usw., PRON sind Determinantien und Quantoren.				

Tab 5c: Anteil der Mehrfachbesetzungen im Ungarischen liegt zwischen 40% und 90%, mit einer Vorkommenshäufigkeit von min. 3%

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
N+N	1642	1053	N nn + N nn	908
Kommentar: Unter den ersten 100 waren nur 8, die der Suchanfrage in jeder Hinsicht gerecht werden: 5-mal „Ende + ‘Monatsname’“, zweimal Gattungsname als Apposition, einmal <i>Hunderte Sicherheitskräfte</i> . In den meisten Fällen war mindestens eines der zwei Nomina ein Eigenname: In 18 Fällen war es ein zweigliedriger Eigenname, in 29 weiteren Fällen war das erste Nomen ein Eigenname im Genitiv, es gab 10 Eigennamen als Appositionen, die restlichen 42 Treffer vom ersten Hundert waren aus verschiedenen anderen Gründen falsch.				

⁹ Diese Konstruktion zeigt, dass es kategoriell ein ADV ist, flektierte attributive Formen dürfen z.T. ein daraus konvertiertes ADJ darstellen (im Ung. gibt es nur ein ADV als Äquivalent in dieser lokalen Bedeutung: *világszerte* ‘in der ganzen Welt’, und es gibt kein adjektivisches Derivat; adjektivische Äquivalente wie *világméretű* haben eine andere Bedeutung, etwa ‘weltumfassend’).

¹⁰ Ein Richtungsadverb als Komplement des Substantivs ist zweifelsfrei Attribut: *Rollen vorwärts werden häufig als Sprung- oder Flugrollen ausgeführt*; Bestimmungen des Ortes drücken aber nicht unbedingt ein Merkmal des links stehenden Nomens, sondern möglicherweise die

ungarische Types	ung. Tokens	Mehr- fachb.	deutsche Types	deutsche Tokens
N+Adv	1251	846	N nn + ADV	134
Kommentar: 21 falsche Treffer, die „richtigen“ enthalten zu min. 90% Eigennamen als N nn annotiert. In der ADV-Position stehen 55 <i>selbst</i> , 14 <i>hingegen</i> , 13 <i>aber</i> , 9 <i>wiederum</i> , 7 <i>dagegen</i> , unter den restlichen 15 kommen noch vereinzelt <i>selber</i> , <i>jedoch</i> , <i>seiner-/ihrerseits</i> , <i>allein</i> , <i>etwa</i> , <i>allerdings</i> vor. 4 Belege enthalten Lokaladverbien, die nicht immer ohne Weiteres als Attribute zu deuten sind. ¹⁰				
Adv+Adv	899	719	ADV + ADV	707
Kommentar: Niedrige Fehlerquote; das erste ADV ist zumeist Partikel wie <i>auch</i> , <i>erst</i> , <i>etwas</i> , <i>immer</i> , <i>noch</i> , <i>schon</i> , <i>nur</i> usw. Es gibt wenige Belege zum akkumulativen Typ <i>hinten oben</i> , <i>darunter links</i> . Das zweite ADV ist manchmal <i>aber</i> , <i>jedoch</i> usw. Temporale und lokale Adverbien in einem Vorfeld sind in dieser Treffermenge nicht belegt, auch im Satz <i>Spätestens hier muss eine Behandlung mit Kortison einsetzen</i> hat <i>hier</i> eine temporal-situative Bedeutung.				
Det+N+N	2013	886	ART + N nn + N nn	3403
Kommentar: Meistens Appositionen, selten nachgestellte Genitive; die Fehlerquote war relativ hoch (vgl. Tab. 2), z.B. unter den ersten 100 Treffern waren 51 Appositionen Eigennamen, 16 fremde Fügungen, z.T. mit ADJ der Originalsprache (z.B. <i>Die Professional Version</i> , <i>Die American Football-Abteilung</i>), aber weitere 6 NP-s enthielten deutsche Adjektive (z.B. <i>Das Wetzlarer Kreuz</i>). In 16 Fällen handelte es sich um die Konstruktion „metasprachliches Wort + objektsprachliches Wort“, wo das zweite Wort in 6 Belegen wiederum ein Eigenname war, z.B. <i>Der Name Arundel</i> .				

6 Fazit

Ein wichtiges Charakteristikum der deutschen Wortfolge ist die Verbzweitstellung, die vor dem Finitum im Hauptsatz maximal ein Komplement oder Supplement des Verbalkomplexes zulässt. Falls die Struktur eines wohlgeformten Satzes von dieser prototypischen Regel abzuweichen scheint, muss der Linguist mit einer Erklärung aufwarten. Eine faktische Abweichung (sprich: in jeder Hinsicht mehrfache Vorfeldbesetzung) lässt sich in einer kaum messbaren Quantität beobachten (vgl. Speyer 2008) und eben daher ist sogar der Korpuslinguist geneigt,

Verortung des ganzen Sachverhalts aus, was ein Charakteristikum von Satzadverbialia ist: Hintergrund hier ist, dass Spiele der 8- oder 16-Bit Ära auf Hardware ...; Produzenten dort sind die General Motors Egypt wie in Lizenz auch ...; Sendetermin hier war immer sonntags um Mitternacht.; mit einem Artikel vor dem Nomen liegt die Interpretation als Attribut nahe, z.B.: Die Angaben hier beziehen sich weiterhin auf den kostenlosen download, ...

wieder einmal über Begriffe wie ‘abweichend’, ‘Norm’ und ‘Standard’ nachzudenken. In der germanistischen Fachliteratur wird jedoch diese Wortstellung zumeist ohne Problematisierung der Normativität exemplifiziert, bei Eisenberg (1999: 389) u.a. mit folgendem Satz: *Ihm den Stern hat Irene gezeigt*.

Umstritten ist, ob Komponenten, die mit einem Infinitum zusammen ins Vorfeld herausgelagert werden, eine Mehrfachbesetzung darstellen. Jedenfalls beläuft sich ihr Anteil auf eine schon „mit bloßem Auge sichtbare“ Quantität: Aufgrund unserer Korpusuntersuchungen (Korpus: Wiki_DE_Öffentlich – öffentliche deutschsprachige Wikipedia-Artikel) machen sie ca. 1 Promille der linken Felder aus. Im Ungarischen (Korpus: Wiki_HU2_Öffentlich) ist der Anteil der linken Satzperipherien mit Infinita nicht wesentlich höher: 0,4%. Dabei hat das Ungarische – ebenso wie die meisten anderen Sprachen – keine Verbalklammer mit Verbzweitstellung, außerdem ist seine Wortfolge ziemlich frei. Der Grund der nicht um Größenordnungen höheren Zahl mag u.a. der wesentlich kleinere Anteil der analytischen Formen im ungarischen Verbparadigma sein.

Wenn die ungarischen und die deutschen präverbalen Bi- und Trigramme verglichen werden, wird es niemanden verwundern, dass die Häufigkeit der Mehrfachbesetzungen im Ungarischen höher ist. Was eventuellen Erwartungen eher zuwiderlaufen kann, ist das Ausmaß dieser Häufigkeit, die nach unseren Recherchen ca. 50% beträgt. Mit anderen Worten: 2-3 Textwörter im linken Feld konstituieren in jedem zweiten Fall mehr als eine primäre Komponente des ungarischen Satzes.

Bei einem Vergleich derjenigen ungarischen Bi- und Trigramme, die immer oder sehr häufig mit Mehrfachbesetzung einhergehen, mit ihren deutschen Pendanten stellte sich Folgendes heraus: Es wurden insgesamt 29 Types untersucht. Bei 10 machten die deutschen Tokens weniger als 20% der ungarischen aus (bei zwei gab es je 2 Treffer). In 13 Fällen lagen die Frequenzen einander näher und bei 6 Types gab es auf der deutschen Seite – mitunter erheblich – höhere Werte. Dass darunter lediglich 5 „mehrfachbesetzungsverdächtige“ Belege zu finden waren, liegt v.a. an der Vielzahl von als ADV annotierten satzgliedinternen Partikeln und von einfach falschen Annotationen. Nominale Ketten erwiesen sich im Deutschen meistens als Phrasen mit Appositionen oder Genitivattributen. In der Funktion der deutschen wortbezogenen Konjunktion *als* stehen im Ungarischen ein Suffix und eine Konjunktion zur Verfügung; vielleicht gab es im Deutschen deshalb fast 10-mal so viele Treffer mit „KON + N nn“ wie im ungarischen Korpus.

Kontraste im System und im Gebrauch zweier Sprachen lassen sich schwerlich exakt messen. Mit den obigen zahlenmäßigen Angaben konnten wir jedoch vielleicht einigermaßen zur linguistischen Erfassung des untersuchten Gebiets beitragen, und zwar nicht nur im Vergleich, sondern auch in Bezug auf beide Sprachen im Einzelnen.

Anhang

Verzeichnis von Part-of-Speech-Tags, die nicht den allgemein bekannten Kategoriensymbolen entsprechen:

ADJ at	attributive Adjektive
AP pr	Präpositionen
AP prart	Präpositionen mit Artikel
Art	bestimmter/unbestimmter Artikel
CARD	Kardinalzahlen
Con	Konjunktionen
DIG	Zahlen mit Ziffern geschrieben
KON kom	Vergleichspartikeln
KON neb	nebenordnende Konjunktionen
MI. (für MIA/MIB/MIF)	Partizipien
N ne	Eigennamen
N nn	normale Nomina
N(pr)	Eigennamen
NU	Postpositionen
Num	Zahlwörter
PRON dem at	attribuierende Demonstrativpronomen
PRON ind at	attribuierende Indefinitpronomen
PRON pos at	attribuierende Possessivpronomen
SPUNCT	Satzzeichen

Weitere Abkürzungen

LF	linkes Feld
LFx	linkes Feld, das x Tokens enthält
<sa>	Satzanfang

Literatur

- Bassola, Péter (Hg.)(1998): Beiträge zur Nominalphrasensyntax. (= Acta Germanica 6). Szeged: JATE.
- Bassola, Péter (2009): Deutsche Wortstellung – grammatisch – pragmatisch. Prädikative Elemente im Vorfeld. In: Brdar-Szabó, Rita/Knipf-Komlósi, Elisabeth/Péteri, Attila (Hgg.): An der Grenze zwischen Grammatik und Pragmatik. Frankfurt a.M.: Lang. 35–52.
- Dürscheid, Christa (1989): Zur Vorfeldbesetzung in Verbzweit-Strukturen. Trier: Wissenschaft Verlag.
- Eisenberg, Peter (1999): Grundriß der deutschen Grammatik. Bd. 2: Der Satz. 3. Aufl. Stuttgart: Metzler.
- Hoberg, Ursula (1997): Topologie des Vorfeldes. In: Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. (= Schriften des Instituts für Deutsche Sprache 7). Berlin/New York: de Gruyter. 1577–1644.
- Keszler, Borbála/Lengyel, Klára (2008): Ungarische Grammatik. Hamburg: Buske.
- Kiefer, Ferenc (Hg.)(1992): Strukturális magyar nyelvtan [Strukturelle ungarische Grammatik]. I. kötet, Mondattan [Syntax]. Budapest: Akadémiai Kiadó.
- Lengyel, Klára (2000): Az igenevek helye a szófaji rendszerben [Zum Status der infiniten Verbformen im Wortartensystem]. (= Nyelvtudományi Értekezések 146). Budapest: Akadémiai Kiadó.
- Müller, Stefan (2005): Zur Analyse der scheinbar mehrfachen Vorfeldbesetzung. In: Linguistische Berichte 203. 297–330.
- ProGr@mm kontrastiv (2013): ProGr@mm kontrastiv/Aus ungarischer Sicht/Primäre Komponenten des Satzes/Verbalkomplex. Internet: http://hypermedia.ids-mannheim.de/call/public/gruwi.ansicht?v_typ=o&v_id=3308 (Stand: 18.7.2013).
- Prószéky, Gábor/Tihanyi, László (1996): Humor – a morphological sytem for corpus analysis. In: Proceedings of the first TELRI Seminar in Tihany. Budapest: Research Institute for Linguistics. 149–158.
- Speyer, Augustin (2008): Doppelte Vorfeldbesetzung im heutigen Deutsch und im Frühneuhochdeutschen. In: Linguistische Berichte 216. 455–486.
- Uzonyi, Pál (2010): Zustandspassiv und Zustandsaktiv im Deutschen und im Ungarischen. In: Czeglédy, Anita/Hess-Lüttich, Ernest W.B./Langanke, Ulrich H. (Hgg.): Deutsch im interkulturellen Begegnungsraum Ostmitteleuropa. (= Cross Cultural Communication 19). Frankfurt a.M.: Lang. 451–462.