In: Dalmas, Martine/Fabricius-Hansen, Cathrine/Schwinn, Horst (Hrsg.): Variation im europäischen Kontrast. Untersuchungen zum Satzanfang im Deutschen, Französischen, Norwegischen, Polnischen und Ungarischen. Berlin/Boston: de Gruyter, 2016. (Konvergenz und Divergenz 5), S. 1-8.

Martine Dalmas/Cathrine Fabricius-Hansen/Horst Schwinn

Einleitung
Kontrastivität/Satzanfang/Korpus

1 Vorgeschichte

Die vorliegende Publikation zur kontrastiven Untersuchung des Satzanfangs ist im Zusammenhang mit dem Forschungsprojekt EuroGr@mm entstanden. EuroGr@mm war ein Netzwerkprojekt zur typologisch und kontrastiv vergleichenden grammatischen Erforschung und Beschreibung des Deutschen auf europäischer Ebene. Das Forschungsnetzwerk wurde vom Institut für Deutsche Sprache (IDS) in Mannheim in Kooperation mit fünf wissenschaftlichen Forschergruppen aus dem europäischen Ausland gebildet. Im Einzelnen handelte es sich um Forschergruppen aus Deutschland (Mannheim), Frankreich (Paris), Italien (Genua, Neapel, Palermo, Salerno), Norwegen (Oslo), Polen (Wrocław) und Ungarn (Budapest, Szeged). Inhaltlich orientierte sich das Forschungsprojekt an der am IDS erarbeiteten Grammatik der deutschen Sprache[1] und an der daraus entwickelten Online-Grammatik ProGr@mm[2].

Gemeinsames Ziel der Forschergruppe war zunächst die Analyse der Grammatik des Deutschen unter einem kontrastiven und didaktischen Gesichtspunkt und die Implementierung der Forschungsergebnisse in die Online-Plattform ProGr@mm. Dabei war die deutsche Sprache sowohl die zu beschreibende als auch die Beschreibungssprache. Ergänzend zur Untersuchung der deutschen Grammatik flossen relevante Aspekte aus kontrastsprachlicher Perspektive in die unterschiedlichen Analyseebenen und auch in die integrierten Übungen ein. Ein zentraler Gegenstandsbereich der Kooperation in der ersten gemeinsamen Arbeitsphase war u.a. die typologisch und kontrastiv vergleichende Erforschung der Flexionsmorphologie des Deutschen, deren Ergebnisse in einem gemeinsamen Sammelband erschienen sind.[3] In einer zweiten Phase stand die korpusgestützte vergleichende Erforschung der grammatischen Variation im standardnahen Deutsch im Vordergrund. Hierbei lag der Schwerpunkt der Betrachtung auf

1 Zifonun, Gisela/Hoffmann, Ludger/Strecker, Bruno (1997): Grammatik der deutschen Sprache. 3 Bde. Berlin/New York: de Gruyter.

2 hypermedia.ids-mannheim.de/programm/.

3 Vgl.: Augustin, Hagen/Fabricius-Hansen, Cathrine (Hgg.) (2012): Flexionsmorphologie des Deutschen aus kontrastiver Sicht. Tübingen: Groos.

der linken Satzperipherie des Deutschen und seiner Kontrastsprachen unter Berücksichtigung insbesondere morphosyntaktischer und informationsstruktureller Aspekte. Dieser Sammelband ist nun ein Ergebnis der zweiten Phase des Projekts.

2 Zur Terminologie

Die zahlreichen Schriften zur Informationsstruktur von Äußerungen, die in den letzten Jahrzehnten aus unterschiedlichen theoretischen Perspektiven entstanden sind, haben sich als Nährboden für eine Fülle von Begriffen und Termini erwiesen, die uns allen heute zwar geläufig sind, jedoch stets der Klärung bedürfen, zumal wenn sie auf die Beschreibung typologisch unterschiedlicher Sprachen angewendet werden. Diese Erfahrung blieb auch unseren sechs Forschergruppen nicht erspart! Die Zusammenarbeit führte von Anfang an zu Diskussionen über Begriffe wie THEMA, TOPIK, FOKUS usw. und zu den notwendigen theoretischen Präzisierungen, aber als es dann darum ging, den Gegenstand des gemeinsamen Forschungsprojekts und der Publikation zu benennen, nahm das Problem andere Ausmaße an: Die typologisch recht unterschiedlichen Sprachen bzw. Sprachfamilien lassen sich, was eine rein topologische Aufteilung des Satzes betrifft, bei weitem nicht problemlos unter einen Hut bringen.

Vorfeld

Zugleich Hilfe und Hürde war bei unserem Vorhaben die gemeinsame Kontrastsprache Deutsch. Die klare topologische Struktur des deutschen Aussagesatzes mit seinem finiten Verb als ‚Grenzstein' und dem dadurch nie zu verfehlenden VORFELD diente zur Orientierung, war aber nicht ohne Weiteres auf die anderen untersuchten Sprachen übertragbar. Was im Deutschen – und im Norwegischen – auf Anhieb als Vorfeld identifiziert wird, weil es rechts durch das finite V2 abgegrenzt wird und vor allem weil es – bis auf wenige Ausnahmen[4] – obligatorisch ist und nur durch eine syntaktische Einheit besetzt werden darf, kann in Sprachen, in denen am Satzanfang relativ viel und vor allem syntaktisch Heterogenes verpackt werden kann oder in denen das Verb überhaupt spät im Satz erscheint, nicht wiedergefunden werden. Den Terminus VORFELD haben wir selbstverständ-

4 S. die sog. Mehrfachbesetzung.

lich für das Deutsche und das Norwegische beibehalten, beide V2-Sprachen, in denen die selegierten Vorfeld-Elemente aus informationsstruktureller Sicht wichtige Faktoren des Textzusammenhalts (Kohärenz) und der Textentwicklung (Progression) sind. Bei den anderen Sprachen erwies sich der Terminus als unbrauchbar. Die Bezeichnung, die sich dann anbot, ist zwangsläufig vager, sie hat aber den Vorteil, dass sie sowohl für das Deutsche angemessen ist, wo links vom Vorfeld noch Raum frei bleibt, der durch Konstituenten mit bestimmten Funktionen besetzt werden kann, als auch für die anderen hier untersuchten Sprachen, bei denen es keine feste Regelung für den Raum vor dem finiten Verb gibt.

Linkes Feld

Was wir hier als LINKES FELD bezeichnen, entspricht also im Deutschen und im Norwegischen sowohl dem Vorfeld (falls vorhanden) als auch dem Raum für weitere Konstituenten, die sich links davon befinden und entweder syntaktisch-semantisch mit der Vorfeldeinheit oder nur semantisch mit einer anderen Konstituente oder aber pragmatisch mit der ganzen Äußerung zusammenhängen. Der Terminus LINKES FELD subsumiert in diesem Sinne sowohl das Vorfeld als auch, was von manchen Autoren als „Vorvorfeld" und/oder „linkes Außenfeld" bezeichnet wird. Bei den anderen Sprachen, die keine vergleichbaren Stellungsfelder mit festen Besetzungsregularitäten haben (vgl. Augustin in diesem Band), ermöglicht die Wahl dieses Begriffs als tertium comparationis den Zugriff auf den linken Satzbereich, sie erspart bei der Korpuszusammenstellung weitere Unterscheidungen und erleichtert somit den Sprachvergleich.

Die weiteren Bezeichnungen, die von den Autoren dieses Bandes benutzt werden, zeigen deutlich, wie unterschiedlich ‚großzügig' andere Sprachen mit dem Bereich vor dem finiten Verb umgehen: Fürs Französische, eine SVO-Sprache, geht es – je nachdem, ob nur eine Konstituente (NP in Subjektfunktion) vor dem finiten Verb steht oder mehrere mit unterschiedlichen Funktionen – entweder um den „Satzanfang" oder den mehr oder weniger ‚aufgeblähten' „Initialbereich", der sich gern durch die sog. additive Akkumulation auszeichnet. Beim Polnischen wird hier der Begriff LINKES FELD benutzt und damit der mehrfach besetzte Raum zwischen Satzanfang und finitem Verb bzw. dessen funktionalem Äquivalent gemeint. Fürs Ungarische wird einerseits – aus informationsstruktureller Perspektive – von „Topikbereich" bzw. „Topikfeld" gesprochen und andererseits – aus topologischer Perspektive – von „linkem Feld" mit häufiger Mehrfachbesetzung.

Satzanfang

Vor diesem Hintergrund der unterschiedlichen topologischen Sprachstrukturen und Perspektiven haben wir für den Titel des Bandes den neutralen Terminus SATZANFANG gewählt, um den gemeinsamen Ausgangspunkt und Gegenstand zu verdeutlichen, der von den meisten Autoren sprachvergleichend unter die Lupe genommen wird.

Die vergleichende Herangehensweise an fünf verschiedene Sprachen mit Deutsch als Kontrastsprache erweist sich in dreifacher Hinsicht als besonders aufschlussreich:

- Die quantitativ ausgerichteten Korpusanalysen ermöglichen einen genauen Einblick in die unterschiedlichen Strukturmerkmale der betreffenden Sprachen sowie in sprachübergreifende Textmerkmale;
- die qualitativen Untersuchungen zeigen Ähnlichkeiten und Abweichungen bei bestimmten Verfahren, die sich morphosyntaktisch niederschlagen und besonders am Satzanfang relevant sind, so zum Beispiel die Determinationsmarkierung, die Rahmensetzung oder die eventuell kontrastierende Hervorhebung;
- insgesamt erlauben uns die hier durchgeführten Untersuchungen, Hypothesen zur topologisch markierten Informationsstruktur und zu Präferenzen in den jeweiligen Sprachen aufzustellen, aber auch zu möglichen Konstanten und Gemeinsamkeiten, was – auf differenziertere Korpora erweitert – für die Bereiche Sprache und Kognition sowie computergestützte Übersetzung ein großer Gewinn sein dürfte.

3 Wikipedia-Korpora

Mehr als die Hälfte der Beiträge des Sammelbandes stützt sich bei der Analyse des Satzanfangs auf Wikipedia-Texte. Wikipedia-Texte haben den Vorteil, dass sie für alle Kontrastsprachen des Forschungsnetzwerks zur Verfügung stehen, leicht zugänglich sind und keinen Urheberrechten unterliegen. Außerdem bestehen die Wikipedia-Texte aus zwei Subtextsorten, den eigentlichen Wikipedia-Artikeln und den Wikipedia-Diskussionen zu den jeweiligen Wikipedia-Artikeln, die in ihren sprachlichen Charakteristika sehr unterschiedlich sind. Während die Wikipedia-Artikel den allgemeinen Merkmalen geschriebener Sprache entsprechen, formellen Charakter aufweisen und in der Regel normgerecht verfasst sind, finden sich bei den Wikipedia-Diskussionen Merkmale verschriftlichter mündlicher Sprache, außerdem weisen die Texte teilweise normabweichenden Gebrauch im

Bereich der Orthografie und der Syntax auf. Gerade diese Bipolarität der beiden Subtextsorten stellt eine gute Basis dar, sollen synchrone Variationen im Sprachgebrauch ausfindig gemacht werden.

Für die quantitative als auch die qualitative Analyse des Satzanfangs des Aussagesatzes war es nötig, die Wikipedia-Daten der unterschiedlichen Sprachen aufzubereiten, um recherchierbare (Teil-)Korpora erstellen zu können. Einerseits sollten die Grenzen des Satzanfangs respektive des Vorfelds elektronisch erfassbar sein, andererseits sollten der Umfang wie auch die morphologische Zusammensetzung des Satzanfangs automatisch bestimmbar sein. Beides wurde ermöglicht durch eine POS-Annotation der Korpora. Geht man davon aus, dass linear betrachtet der Satzanfang der Bereich vor dem finiten Verb ist, das Finitum also die exklusive Grenze des Satzanfangs darstellt, dann ist mit einer Wortartenannotation sowohl diese Grenze elektronisch präzise ermittelbar als auch eine quantitative Satzanfangsanalyse möglich.[5]

Für die Aufbereitung der Wikipedia-Texte mussten zunächst bereinigte Texte erstellt werden, sodass Annotations- und Rechercheinstrumente problemlos darauf anwendbar wären. Das heißt im Einzelnen: Die hypertextuellen Auszeichnungen mussten aus den ursprünglichen Wikipedia-Texten entfernt werden, der bereinigte Text wurde in ein XML-Zwischenformat konvertiert und für eine Strukturierung bzw. Hierarchisierung der Texte erfolgte deren Annotation ohne Beeinflussung ihrer Syntax. Weitere Schritte waren die Überführung des so gewonnenen XML-Formats in eine XCES-Struktur, die alphanumerische Anordnung der Artikel und eine Metadatenannotation, die die Identifizierung der Artikel für Referenzzwecke und die Trennung von Artikel und Diskussion erlaubte. Nach diesen vorbereitenden Prozeduren konnten die Wortarten den Einzelausdrücken der Texte annotiert werden. Dazu wurde die deutsche, französische und italienische Version der Wikipedia-Texte mit dem Stuttgarter Tree-Tagger getaggt. Die polnische Wikipedia-Version wurde mit dem Morfeusz-Tagger bearbeitet. Die ungarische (stark reduzierte) Version konnte mit Hilfe eines Teams des Ungarischen National-Korpus annotiert werden und schließlich die norwegische Version mit dem Oslo-Bergen-Tagger.[6] So entstanden folgende nach Wortarten annotierte Wikipedia-Teilkorpora:

5 Zu Einzelheiten s. auch Augustin (in diesem Band).

6 Vgl. zur Konvertierung: Bubenhofer, Noah/Haupt, Stefanie/Schwinn, Horst (2011): A Comparable Wikipedia Corpus: From Wiki Syntax to POS Tagged XML. In: Working Papers in Multilingualism, 96 B. 141–144.

Sprache	Wikipedia-Artikel Anzahl der Wortformen	Wikipedia-Diskussionen Anzahl der Wortformen
Deutsch	551.090.404	246.028.026
Französisch	526.944.992	101.893.579
Italienisch	329.063.792	42.171.984
Norwegisch	70.377.449	1.400.276
Polnisch	190.046.721	15.470.942
Ungarisch (Stichprobe)	11.285.752	27.810

Die einzelnen Wikipedia-Korpora waren für interne Zwecke über das Corpus Search, Management Analysis System COSMAS II[7] recherchierbar. Nach einer Evaluationsphase der ersten Version steht mittlerweile eine aktualisierte öffentliche Version für Recherchezwecke über COSMAS II zur Verfügung. Bei dieser neuen Variante handelt es sich um die sechs aufbereiteten Fassungen der Wikipedia-Texte und Wikipedia-Diskussionen aus dem Jahr 2015.

4 Beiträge

Die Reihenfolge der Beiträge orientiert sich an den Spezifika der Sprachen sowie an den jeweiligen Fragestellungen. Einleitend in das Gesamtthema des Bandes ist die sprachübergreifende quantitative Analyse von Hagen Augustin (Mannheim), der anhand der annotierten Wikipedia-Korpora die sprachspezifischen Ausprägungen von Vorfeldern bzw. Satzanfängen in den sechs Kontrastsprachen aufzeigt. Ebenfalls quantitativ ausgerichtet ist die kontrastive Untersuchung von Pál Uzonyi (Budapest) und Viktória Dabóczi (Gießen), die sich mit der Konstituentenstruktur der deutschen und ungarischen präverbalen Felder befassen und die hohe Frequenz der Mehrfachbesetzung im Ungarischen feststellen, der gegenüber die deutschen Korpora eher nur komplexe Vorfeldbesetzungen aufweisen.

Der zweite Teil der Beiträge enthält zwei kontrastive Aufsätze zur Indefinitheit am Satzanfang. Für ihren deutsch-norwegischen Vergleich stützt sich Cathrine Fabricius-Hansen (Oslo) auf eine quantitative Untersuchung, legt aber dann den Akzent auf qualitative Aspekte und beschreibt die (kon)textuellen informationsstrukturellen Bedingungen für eine Besetzung des Vorfelds durch eine indefi-

7 https://cosmas2.ids-mannheim.de/cosmas2-web/.

nite Nominalphrase im Deutschen und im Norwegischen. Der Aufsatz von Séverine Adam und Cécile Delettres (Paris) befasst sich sprachvergleichend mit Komplement-NPs im Vorfeld, stellt präferierte Kombinationen im Deutschen und im Französischen vor und fokussiert textbezogene Funktionen (Textverknüpfung, Serialisierung, Einführung eines neuen Teilthemas, Rahmensetzung) für eine solche Vorfeldbesetzung.

Die nächste Gruppe von Beiträgen behandelt hauptsächlich unterschiedliche Aspekte der Topikfunktion. In ihrem ausschließlich dem Ungarischen gewidmeten Beitrag geht Beáta Gyuris (Budapest) auf informationsstatusbasierte Funktionen im Topikfeld ein und kommt zu neuen Erkenntnissen, was die syntaktische Struktur des ungarischen Satzes bzw. die Konstituentenabfolge im Topikfeld betrifft. Hélène Vinckel-Roisin und Gottfried Marschall (Paris) geht es in erster Linie um die unterschiedlichen sprachstrukturellen Möglichkeiten im Französischen und im Deutschen. Anhand des erweiterten Begriffs RAHMENSETZUNG weisen die beiden Autoren auf Gemeinsamkeiten zwischen beiden Sprachen hin. Im Beitrag von Péter Bassola (Szeged) und Horst Schwinn (Mannheim) werden markierte Vorfeldbesetzungen im Deutschen diskutiert. Anhand von Belegen mit infiniten Teilen des Verbalkomplexes (und evtl. einem Satzglied) im Vorfeld werden informationsstrukturelle Besonderheiten beschrieben, außerdem werden Fälle der besonders markierten Mehrfachbesetzung des Vorfelds aufgespürt. Um Variation geht es auch im Aufsatz von Edyta Błachut (Wrocław), diesmal allerdings bezogen auf die nominale Initialphrase des Satzes. Ausgehend von der breiten Variationsakzeptabilität des Polnischen im Vergleich zu den hier begrenzten Möglichkeiten des Deutschen, überprüft die Autorin die Faktoren für unterschiedliche Anordnungsfolgen in jeder Sprache.

Der Band schließt mit einem Aufsatz zu einem Spezifikum des gesprochenen Polnisch. Lesław Cirko (Wrocław) setzt sich hier mit der – nach den Regeln des heutigen Sprachsystems unbegründeten – Überpräsenz von Demonstrativpronomina im Linken Feld auseinander und stellt eine Hypothese über die Existenz eines entstehenden Artikelsystems auf.

Paris/Oslo/Mannheim, Oktober 2015

Martine Dalmas, Université Paris-Sorbonne (CeLiSo, Centre de linguistique en Sorbonne), Cathrine Fabricius-Hansen, Universitetet i Oslo (Institutt for litteratur, områdestudier og europeiske språk), Horst Schwinn, Institut für Deutsche Sprache, Mannheim