Methoden zur Erstellung von ÜA
bei mehreren Einzelvarianten/Einzellemmata

In Folgenden geht es um die Erstellung von ÜA

a. wenn mehrere Einzelvarianten zu einem ÜA zusammengefasst werden sollen (Methode 1, Downsampling)

b. wenn die generelle Tendenz eines Phänomenbereichs dargestellt werden soll und die EA dem ÜA nachgeordnet sind (Methode 2)

a. Methode 1: Downsampling

Downsampling wird vorgenommen, um mehrere Einzelvarianten zu einem Phänomen zusammenfassen zu können – dies geschieht im Normalfall in Überblicksartikeln**.**

Da wir in dem Fall nicht mit Prozentzahlen rechnen können, müssen wir die absoluten Belegzahlen der einzelnen Varianten, die wir vergleichen wollen, aneinander angleichen, damit nicht eine Variante, die insgesamt sehr viel häufiger vorkommt, das Gesamtergebnis verfälscht (Bsp. Präposition *während* kommt sehr viel häufiger vor als *mitsamt*).

Es wird dementsprechend von der kleinsten Belegzahl (=Minimum) ALLER zu vergleichender Varianten (Fragestellung 1) bzw. Variantenpaare (Fragestellung 2) ausgegangen, auf dieses Minimum werden alle Varianten(paare) downgesampelt – das bedeutet konkret:

- Die Summe der absoluten Belegzahlen von Variante A+B/Minimum aller Varianten(paare) = Ergebnis E
- Die absoluten Belegzahlen jeder Einzelregion werden dann durch dieses Ergebnis E geteilt; (pro Variante handelt es sich hier also um ein anderes Ergebnis)
- Die neu errechneten Zahlen werden summiert und in Prozentzahlen umgerechnet

Alternative (s. Excel-Datei *eur-or-*): Eine Alternative wäre, einen groben Richtwert zu ermitteln, der ungefähr dem Minimum entspricht und dann (durch Ausprobieren) die vorhandenen Belegzahlen pro Variante durch einen Quotienten zu teilen, der dann am Ende dem Richtwert entspricht.

Hinweise zur Excel-Vorlage (Präpositionen) für das Downsampling

Die Tabelle ist derzeit ausgelegt für das semi-automatische Downsampling von bis zu **13 Einzelphänomenen** mit je **zwei Varianten**. Die **Ausgangswerte** werden **von links nach rechts** eingetragen, die **Rechenvorgänge** erfolgen **von oben nach unten**. Daher sollte es mit etwas Excel-Erfahrung relativ gut möglich sein, die Tabelle zu modifizieren, sodass etwa mehr (oder weniger) als zwei Varianten pro Einzelphänomen berücksichtigt werden können.

Die **absoluten Zahlen** werden in **Zeile 2 bis 18** eingetragen. Am einfachsten geht das m.E., wenn man die entsprechende Tabelle der statistischen Auswertung in ein separates Excel-File kopiert und die

Regionen alphabetisch sortiert (!!!). Danach kann man sie über *Daten* > *Text in Spalten* trennen und die relevanten Abschnitte kopieren. Für die Arbeit mit der Tabelle ist es empfehlenswert, wenn die Varianten, die man zusammenfassen möchte, jeweils links bzw. rechts stehen. In meinem Beispiel ist etwa Präposition mit Dativ-NP immer links, mit Genitiv-NP dagegen rechts. Ansonsten muss man beim letzten Berechnungsschritt vorsichtig sein (s.u.).

In **Z. 19 und 20** werden die absoluten Belege der Varianten zunächst einzeln und dann gemeinsam mit der jeweiligen Gegenvariante addiert. In der Beispieltabelle steht also **im Feld B19 die Summe der absoluten Belege für *während* mit NP im Dativ**, **in C19 die Summe für *während* mit NP im Genitiv** und in **C20 die Summe aus B19 und C19**.

In **Z. 21** wird das **Minimum** von Zeile 20 angezeigt, d.h. die absolute Belegzahl des Einzelphänomens mit den wenigsten Belegen (hier: *mitsamt* mit exakt 1000 Belegen). Die Summen aus Z. 20 werden in in **Z. 22** durch das Minimum dividiert. Die absoluten Belegzahlen werden im nächsten Schritt durch die daraus resultierenden Quotienten dividiert.

Die **Zeilen 25 bis 41** sind analog zu Z. 2 bis 18 aufgebaut. Die absoluten Zahlen wurden hier durch die in Z. 22 ermittelten Quotienten dividiert und so an das Niveau des Phänomens mit der geringsten Belegzahl angeglichen.

Von **Zeile 42 bis 45** habe ich einen kleinen „Kontrollbalken" eingebaut. Er zeigt sicher nicht alle möglichen Fehler an, aber falls es beim Löschen/Hinzufügen von Spalten zu Fehlern in den Formeln gekommen ist, sollten sie hier tendenziell erkennbar werden:

- Z. 43 überprüft, ob die Summe der downgesampelten Belege pro Phänomen jeweils dem Minimum entspricht. (Ist dies nicht der Fall, sind die entsprechenden Kästchen nicht grün gefärbt und die Schriftart wird rot.)
- Z. 44f. prüft, ob das Verhältnis von V2 zu V3 (über alle Regionen hinweg) in den absoluten und downgesampelten Belegen dasselbe ist. Nicht-Übereinstimmung wird wiederum durch rote Farbgebung angezeigt. (Zu sehen in Spalte X u. Y; ursächlich hierfür ist aber kein Fehler in der Formel, sondern Excel, das an irgendeiner Stelle mit gerundeten Zahlen weitergerechnet hat. Da wir sehen können, dass mind. die ersten sechs Stellen nach dem Komma identisch sind, sollte uns das aber nicht stören.)

In **Zeile 47 und 63** werden die downgesampelten Zahlen zusammengefasst. In den Spalten B und C werden die Belege für jeweils die linke bzw. die rechte Variante aufsummiert (in der Beispieltabelle: Präposition mit Dativ-NP links und Präp. mit Genitiv-NP rechts). **Je nach Zahl der zusammengefassten Phänomene sind hier manuelle Anpassungen nötig! Vorsicht ist außerdem geboten, wenn korrespondierende Varianten nicht einheitlich verteilt sind, wenn in der Beispieltabelle also nicht Präp. + Dativ-NP immer in der linken Spalte stehen würde.** (Änderungen können aber rasch erfolgen, indem man sie in den Feldern B49 und C49 durchführt und die Formeln dann „nach unten zieht".) In Spalte E werden dann die Werte aus Spalte B und C aufsummiert und in **Spalte F und G** schließlich wird **die relative Verteilung der downgesampelten, zusammengefassten Varianten in Prozent** angezeigt.

b. Methode 2: Ableitung der EA aus einem zuvor erstellten ÜA

Diese Methode kommt (etwa im Bereich der Verbalflexion) zur Anwendung, wenn in der Fachliteratur oder in unserer Datenbank nicht Einzellemmata als Varianten vermerkt sind (also beispielsweise *Kontrolleur* vs. *Kontrollor*), sondern wenn ein abstrakterer Phänomenbereich als variierend angenommen wird (also beispielsweise Perfektbildung mit *haben* vs. *sein*). In letzterem Fall geht es um die

systematische Analyse der Variation, etwa der Auxiliarverbselektion im Überblick. Die Frage ist hier abstrakter: Wo wird das Perfekt tendenziell eher mit *haben*/*sein* gebildet? Der analytische Fokus wird hier also in erster Linie auf den Phänomenbereich gelegt. In der Gesamtauswertung, die dem ÜA zugrunde liegt, finden sich alle Lemmata, die eine Variation in Bezug auf den behaupteten Phänomenbereich aufweisen (also beispielsweise bedeutungsgleich mit *haben* und *sein* verwendet werden). So kann eine generelle Tendenz ermittelt werden. In weiterer Folge werden dann die einzelnen Lemmata einer näheren Signifikanzprüfung unterzogen. Bei denjenigen Verben, die so frequent sind, dass sie einen signifikanten Unterschied aufweisen, können EA geschrieben werden.

Genau zu prüfen ist jedenfalls, ob sich unter den analysierten Einzellemmata hochfrequente „Ausreißer" befinden, die auf das Gesamtbild wirken. In diesem Fall wäre dieser Zustand gemäß dem Instruktionspapier zu vermerken und/oder zu überlegen, ob das betreffende Lemma aus der Gesamtauswertung genommen werden sollte.

Bei Methode 2 zur Generierung eines ÜA wäre also folgendes Vorgehen zu systematisieren:

- Ansetzen bei dem (behaupteten) variierenden Phänomenbereich: Ist eine signifikante, systematische Variation nachzuweisen?
- Bei den Daten, die dem ÜA zugrunde liegen, handelt es sich um eine Akkumulation all jener Einzellemmata, anhand derer ein Phänomen nachgewiesen werden kann (bei denen also eine bestimmte Variation vorliegt).
- In einem Zwischenschritt wäre zu prüfen, ob nicht einige wenige (hochfrequente) Einzellemmata eine so differente Variationsverteilung aufweisen, dass sie das Gesamtbild verzerren. In diesem Fall müsste eine gruppenbezogene Auswertung oder die Sinnhaftigkeit eines ÜA hinterfragt werden.
- Aus dem Datenmaterial des ÜA werden dann diejenigen Einzellemmata herausgenommen, die die betreffende Variation in signifikanter Häufigkeit illustrieren. Über diese Lemmata können dann EA geschrieben werden.