Mail „Korpus- und Serverupdate, automatische („datengetriebene“) Variantensuche“ von Don Tuggener (13. Mai 2016)

Liebe alle,

hier kommt nun das mehrfach angekündigte Update bzgl. unseres aktualisierten Korpus, des Projektservers und der datengetriebenen Variantensuche. Das Wichtigste in Kürze:

1. Das Korpus ist nun dependenzgeparst und hat vier neue Attribute (gf, new_lemma, new_morph, new_pos; s.u. die genaue Beschreibung).

2. Das aktualisierte Korpus (sowie vgfinal) sind nun auf unserem projekteigenen Server abgelegt. Der Zugang Erfolg über die Adresse: http://dssrv014.uzh.ch/CQPWeb/ (login wie bisher)
Die neue Version des Korpus (mit den neuen Attributen) heisst "VGFINAL2". Es unterscheidet sich nur im Bezug auf die vier neuen Attribute von "VGFINAL", ansonsten sind die beiden Korpora identisch (d.h. es wurde nichts an der bestehenden Annotation in vgfinal verändert; vgfinal ist aus historischen Gründen noch da. Ich empfehle aber die Benutzung von vgfinal2, besonders wegen der verbesserten Lemmatisierung, s.u.)

3. Es ist jetzt möglich, über dem geparsten Korpus Abfragen zu tätigen, die Dependenzrelationen beinhalten. Das geht leider nicht in CQPWeb, sondern muss in Absprache mit mir gemacht werden. Einfach bei mir anklopfen.

4. Wir haben mit einem datengetriebenen Ansatz bisher etwa ca. 30 neue Varianten entdeckt (s.u.).

Ausführlichere Informationen zu den obigen Punkten:

1. Neue Attribute in VGFINAL2:
Die neuen, zusätzlichen Attribute sind Tokenattribute wie die bisherigen (also wie z.B. "pos"). Folgende 4 haben wir hinzugefügt:

gf: grammatikalische Funktion (z.B. "subj" für Subjekte). Man kann z.B. nach einem Lemma in Subjektposition suchen und zusätzlich Numerus einschränken:
[lemma="Polster" & gf="subj" & new_morph=".*Sg"]
Es ist hier festzuhalten, dass der Parser natürlich nicht perfekt funktioniert und man immer fehlerhafte Analysen antreffen wird. Eine Liste der grammatikalischen Funktionen findet sich in der neuen Korpusdokumentation im Anhang und auf dem Wiki (unter Datenerhebung).

new_lemma: Anhand des Parsers war es möglich, abtrennbare Verbpräfixe zu

identifizieren und das Lemma des Verbs korrekt abzubilden. Bisher wurden Verblexeme, deren Präfix abgetrennt ist nicht voll lemmatisiert. Z.B.
"Sie bereitet den Vortrag vor."
-Lemmas in vgfinal: "Sie bereiten der Vortrag vor."
-Lemmas in vgfinal2: "Sie vorbereiten der Vortrag vor."
D.h. man kann nun Lemmas von Verben, die ein abtrennbares Präfix haben suchen und erhält Belege von Gebrauch mit und ohne Abtrennung des Präfixes, z.B.
[new_lemma="vorbereiten"]

new_morph: neue (hoffentlich) verbesserte Morphologieanalyse. Aufgrund des Parsers konnte in manchen Fällen u.a. die Bestimmung von Numerus verbessert werden. Wenn man weiss, dass das finite Verb im Plural steht, weiss man auch, dass das entsprechende Subjekt auch im Plural steht. D.h. v.a. im Bezug auf grammatikalische Subjekte sollte die Morphologieanalyse besser sein (s. "Polster"-Beispiel oben).

new_pos: Neue Wortartanalyse. Dieses Attribut weicht nur in wenigen Fällen von der bisherigen Wortartenerkennung (Attribut pos) ab. V.a. sind Konstruktionen betroffen, die Auxiliar- oder Modalverben beinhalten. In diesen wurde manchmal die Finitheit des Hauptverbs falsch getaggt (VVFIN vs. VVINF). Das sollte nun behoben sein in new_pos. Beispiel finite Formen von "vorbereiten" finden (mit neuer Lemmatisierung):
[new_pos="VVFIN" & new_lemma="vorbereiten"]

3. Abfragen über dem Dependenzparse:
Da in CQPWeb keine Dependenzrelationen abgebildet werden können, muss dies ausserhalb geschehen. Z.B. können wir aufgrund des Parses bestimmen, ob ein Nomen einen Artikel hat oder nicht usw. Diese Information ist in CQPWeb nicht abfragbar, über dem Parse geht das aber leicht. Auch können wir z.B. ermitteln, welche Präpositionen ein Verb subkategorisiert und wie oft etc. Es geht also um Relationen, die zwischen mehreren Wörtern bestehen und die nicht leicht über die lineare Wortsequenz abgefragt werden können. Wenn jemand solche Anfragen hat, einfach bei mir melden. Ich kann diese dann entsprechend kodieren über dem Parse laufen lassen.

4. Datengetriebene Variantensuche
Ein Ziel des Projekts ist es, automatisch und datengetrieben Varianten in unserem Korpus zu finden; im Idealfall solche, die bislang nicht bekannt waren. Ich habe im letzten Monat unterschiedliche Programme und Verfahren entwickelt, die dies anstreben. Grundlage für einen solchen Ansatz ist eine gewisse (genug grobkörnige) Heuristik. Leider hat die automatische Suche basierend auf N-Grammen keine neuen oder interessanten Varianten zu Tage gefördert, da die Wort- oder Wortartsequenz in unserem Fall keine gute solche Heuristik zu sein scheint. Dafür haben Dependenzrelationen zu Erfolg geführt. Z.B. kann man annehmen, dass Verben areal Variation zeigen in

Bezug auf die subkategorisierten Präpositionen. Ich habe ein Programm geschrieben, dass die Heuristik implementiert und dadurch z.B. herausgefunden, dass man in gewissen Regionen in Deutschland das Verb "durchsetzen" mit der Präposition "über" anstatt "gegen" verwendet: "In der Fußball-Regionalklasse setzt sich Grün-Weiß Tanna 2:0 über den TSV Ranis durch ." Oder in Österreich kann man "um etwas ansuchen": "1951 suchte Kohnhausers Witwe um eine Opferrente an ."
Insgesamt sind durch die datengetriebene Analyse und solche Heuristiken ca. 30 neue Varianten gefunden worden. Ausserdem wurden ca. 40 bereits in unserer Datenbank erfassten Varianten bestätigt. Nicole Zellweger führt die neuen Varianten in der Datenbank nach und versieht sie mit dem tag "data driven", sodass sie auffindbar sind.
Falls jemand weitere Heuristiken für die automatische Suche ausprobieren möchte: Gerne bei mir anklopfen; es ist schon ganz interessant, was man finden kann.

Ich bedanke mich für die Lektüre diese(r|s) langen Email(s)! Bei Fragen wie immer gerne bei mir melden.

Mit besten Grüssen
Don