VG-Korpus

Klaus Rothenhäusler (ergänzt von Don Tuggener)

12. Mai 2016

1. Annotationstypen

In der Corpus Workbench[1] (CWB) wird unterschieden zwischen **positionalen** und **strukturellen** Attributen. Letztere entsprechen grob dem, was gewöhnlich mit XML-Elementen annotiert wird. Ihnen können weitere Eigenschaften zugeordnet werden, von denen normalerweise als Attribute gesprochen wird. Wir halten die Benennung deshalb für ein wenig unglücklich und verwenden an Stelle von „strukturellem Attribut" die Bezeichnung **strukturelle Einheit**.

- Positionale Attribute sind Annotationen auf Tokenebene: *Jedem einzelnen* Token (= jeder Position) im Korpus ist ein entsprechender Attributwert zugeordnet.

  Als Sonderfall von positionalen Attributen lassen sich **feature set** Attribute behandeln, die ein Bündel von Merkmalen oder alternative Annotationen für eine Korpusposition erfassen. Feature set Attribute verhalten sich genauso wie positionale Attribute und können auch so abgefragt werden, für eine sinnvolle Suche sind aber spezielle Suchoperatoren definiert. In den Korpora für das Variantengrammatikprojekt ist das einzige positionale Attribut dieser Art die morphologische Analyse, wobei die Elemente des Attributs (die Elemente der Merkmalsmenge) die alternativen Analysen sind, die Morphisto für einen Token liefert.

- Strukturelle Einheiten können sich über einen oder aber auch über *mehrere* Token erstrecken.

  Die CWB kann mit Eigenschaften von strukturellen Einheiten nicht direkt umgehen, da einer einzelnen Einheit maximal ein Wert zugeordnet sein kann. Es ist deshalb beispielsweise nicht möglich alle konjunktivischen subjungierten Gliedsätze mit einer Anfrage der Form

  ```
  * <c relation="sub" infl=".*Subj.*"> []; % führt zu Syntax Error!!
  ```

[1] http://cwb.sourceforge.net/

zu finden. Stattdessen werden von der CWB für die Eigenschaften struktureller Einheiten (also deren Attribute, s.o.) automatisch zusätzliche strukturelle Einheiten eingeführt, die sich über genau die gleichen Token erstrecken und denen zudem der entsprechende Attributwert zugeordnet ist. Ihr Name ergibt sich aus dem Namen der strukturellen Einheit gefolgt von einem Unterstrich (`_`) und dem Attributnamen. So wird beispielsweise für das Inflektionsattribut der Nebensatz-Annotation (`c`) die zusätzliche strukturelle Einheit `c_infl` geschaffen, der jeweils einer der Werte aus Tabelle 6 zugeordnet ist. Die korrigierte Abfrage, um die gewünschten Gliedsätze zu erhalten, sieht daher so aus:

```
<c_relation="sub"> <c_infl=".*Subj.*"> [];
```

Da die CWB keine rekursiven Einbettungen verarbeiten kann, werden für verschachtelte strukturelle Einheiten vom gleichen Typ ebenfalls neue Einheiten geschaffen, deren Name durch Anhängen der Rekursionstiefe an den Ausgangsnamen gebildet wird, d.h. ein Nebensatz innerhalb eines anderen Nebensatzes wird mit `c1` markiert, ist darin ein weiterer Nebensatz eingebettet, wird er als strukturelle Einheit vom Typ `c2` gekennzeichnet usw. Die entsprechenden strukturellen Einheiten für deren Attribute werden gleichfalls durchnummeriert, also `c_infl1`, `c_infl2`, etc.

2. Positionale Attribute

2.1. Suchsyntax

Grundsätzlich werden positionale Attribute innerhalb von Tokenbeschreibungen abgefragt, die durch eckige Klammern markiert sind:

```
[ATTRIBUTNAME="ATTRIBUTWERT"]
```

Der Attributwert wird dabei als regulärer Ausdruck[2] spezifiziert. Abfragen, die unterschiedliche positionale Attribute betreffen, können mit `&` (Konjunktion) und `|` (Disjunktion) verknüpft werden. Außerdem können die Operatoren mit `!` (Not) negiert werden: `!=`, `&!`, `|!`. So lässt sich beispielsweise eine Abfrage formulieren, die Token zeigt, bei denen sich die beiden Tagger uneinig sind, ob es sich um ein normales Nomen handelt:

```
[(pos="NN" & rfpos!="N.Reg.*") | (pos!="NN" & rfpos="N.Reg.*")];
```

2.2. Typen

Die positionalen Attribute umfassen:

[2] s. beispielsweise http://kitt.cl.uzh.ch/clab/regex/index.jsp

1. **word**: Die Wortform, so wie sie im Text erscheint. Als Default-Attribut kann für Tokenbeschreibungen, die nur das **word** Attribut betreffen eine vereinfachte Syntax ohne eckige Klammern verwendet werden.

   ```
   ".*ung";
   ```

   statt

   ```
   [word=".*ung"];
   ```

2. **pos**: Die Wortart, wie vom TreeTagger[3] geliefert, d.h. in Form eines von 54 STTS Tags (Stuttgart-Tübingen Tagset)[4]. Eine komplette Auflistung findet sich im Anhang in Tabelle 5.

3. **rfpos**: Die morphosyntaktisch ausgeprägte Wortart, wie vom RFTagger[5] geliefert. Es gibt über 700 Ausprägungen, die von den Wortartannotationen im TIGER-Korpus abgeleitet sind. Dabei handelt es sich um komplexe Tags, deren Komponenten durch einen "`.`" voneinander getrennt werden. Das erste Glied spezifiziert jeweils die Wortartkategorie und entspricht grob den STTS Tags. Die Wortartkategorie legt auch Anzahl und Inhalt der folgenden Glieder fest, die den morphosyntaktischen Merkmalen der Wortform entsprechen. Bleibt ein Merkmal unspezifiziert, erhält es den Wert `*`. Alle auftretenden Merkmale sind in Tabelle 1 zusammengefasst.

   Sie verteilen sich folgendermaßen auf die Wortarten

   - `ADJA`: Das attributiv gebrauchte Adjektiv hat die **4** Merkmale KOMPARATION, KASUS, NUMERUS und GENUS.
   - `ADJD`: Das adverbial oder prädikativ gebrauchte Adjektiv hat **1** Merkmal nämlich KOMPARATION.
   - `ADV`: Adverbien haben keine weiteren Merkmale.
   - `APPO`: Postpositionen haben **1** Merkmal KASUS, wobei `Nom` hier vernünftigerweise nicht vorkommt.
   - `APPR`: Präpositionen und der linke Teil einer Zirkumposition haben **1** Merkmal PREP, das unvernünftigerweise für die Präposition `neben` den Wert `Nom` annimmt. Der Wert `?` wird immer und ausschließlich für die Präposition `über` vergeben.
   - `APPRART`: Kontraktionen aus Präposition und Artikel haben die **3** Merkmale KASUS, NUMERUS und GENUS, wobei KASUS lediglich die Werte `Acc` und `Dat` annehmen kann.

[3] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
[4] Das komplette Tagset ist beschrieben auf:
http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html
[5] http://www.ims.uni-stuttgart.de/projekte/corplex/RFTagger/

Merkmal	Wertebereich
Komparation	{Comp, Pos, Sup}
Kasus	{Nom, Gen, Dat, Acc, *}
Numerus	{Sg, Pl, *}
Genus	{Fem, Masc, Neut, *}
Person	{1, 2, 3, -}
Tempus	{Past, Pres}
Determination	{Def, Indef}
Prep	{Nom, Gen, Dat, Acc, Als, An, Außer, Auf, Bis, Hinter, In, Je, Namens, Per, Pro, Unter, Vor, -, ?}
Conj	{Comp, Coord, SubFin, SubInf}
ConjCoord	{Aber, Als, Bis, Denn, Doch, Noch, Wie}
Part	{Ans, Deg, Neg, Verb, Zu}
Pro	{Dem, Indef, Pers, Refl, Rel}
Proadv	{Dem, Inter}
Subst	{Subst, Attr}
Sym	{Other, Paren, Pun, Quot}
SymSpec	{Aster, Auth, XY, Left, Right, Colon, Comma, Cont, Hyph, Sent, Slash}
Partes	{Adj, Noun, Verb}
Verb	{Aux, Full, Haben, Mod, Sein}
Zu	{zu, -}
Participium	{Psp}

Tabelle 1: Die unterschiedlichen Merkmale der Tags des RFTagger

- `APZR`: Der rechte Teil einer Zirkumposition hat keine weiteren Merkmale.
- `ART`: Artikel haben die **4** Merkmale DETERMINATION, KASUS, NUMERUS, GENUS.
- `CARD`: Kardinalzahlen haben keine weiteren Merkmale.
- `CONJ`: Umfasst sowohl neben- (`KON` im STTS) und unterordnende (`KOUS` und `KOUI` im STTS) Konjunktionen als auch Vergleichskonjunktionen (`KOKOM` im STTS). Von seinen **2** Merkmalen gibt das erste, CONJ, Aufschluss um welchen Typ es sich handelt, während das zweite (COORDCONJ) die nebenordnende Konjunktion spezifiziert und wo nicht anwendbar den Dummy-Wert - annimmt.
- `FM`: Fremdsprachliches Material wird nicht weiter differenziert.
- `ITJ`: Interjektionen ebensowenig.
- `N`: Eigennamen (`NE` im STTS) und normale Nomen werden zusammengefasst. Eine Differenzierung ist über das Merkmal NOMEN möglich. Wenig überraschend sind des weiteren KASUS, NUMERUS und GENUS soweit wie möglich spezifiziert. Insgesamt gibt es also **4** Merkmale.
- `PART`: Partikel haben lediglich **1** Merkmal und zwar PART, dessen Wert `Ans` für die Antwortpartikel (z.B. `nein`, `bitte`) Verwendung findet. Die anderen Werte sollten selbsterklärend sein.
- `PRO`: Sämtliche Pronomina werden hier vereint. Es gibt **6** Merkmale: PRO, SUBST, PERSON, KASUS, NUMERUS, GENUS. PRO zeigt den Pronomentyp an, SUBST unterscheidet zwischen attribuierenden und substituierenden Pronomen.
- `PROADV`: Pronominaladverbien werden durch **1** Merkmal (PROADV) als Zusammensetzung aus einem Interrogativ- oder Demonstrativpronomen identifiziert.
- `SYM`: Interpunktions- und andere Nicht-Wortzeichen sind durch die **2** Merkmale SYM und SYMSPEC bedeutend detaillierter bestimmt als im STTS. Die einzige Merkmalsausprägung, die eventuell der Klärung bedarf, ist vermutlich der Wert `Other` in SYM, der stets mit einem der SYMSPEC Werte `Aster`, `Auth` oder `XY` kombiniert wird. Dabei wird `Aster` ausschließlich für den `*` vergeben, `Auth` ist wohl dadurch motiviert, dass das TIGER-Korpus aus Zeitungstexten zusammengestellt wurde, und soll Autorenkürzel anzeigen, während `XY` als Restkategorie sämtliche anderen Nicht-Wortzeichen abdeckt.
- `TRUNC`: Kompositionserstglieder haben **1** Merkmal, PARTES, das dessen Wortart verrät.
- `VFIN`: Finite Verbformen haben **5** Merkmale, deren erstes VERB ist und die Formen einteilt nach Vollverb und einer Reihe von Nicht-Vollverbklassen. Bei den übrigen Merkmalen handelt es sich um PERSON, NUMERUS, TEMPUS und MODUS.

- `VIMP`: Markiert Imperativformen und hat die **3** Merkmale VERB, PERSON (nur `2` und `3`) und NUMERUS.
- `VINF`: Die **2** Merkmale von Infinitiven sind VERB und ZU, das nur bei Infinitiven, die die Partikel `zu` in die Wortform integrieren, den Wert `zu` annimmt und ansonsten den Dummy-Wert (`-`) erhält.
- `VPP`: Das Partizip Perfekt hat zwar **2** Merkmale (VERB und PARTCIPIUM), von denen das zweite aber immer den Wert `Psp` hat, was als *past participle* zu lesen und insofern vermutlich eine redundante Kodierung sein dürfe.

4. **lemma**: Die Grundform des Token, wie vom TreeTagger geliefert.

5. **morph**: Die morphologische Analyse liefert Morphisto[6][7]. Da allerdings keine Fugenelemente markiert werden, haben wir sie in den Analysen rekonstruiert und eine zusätzliche Morphkategorie `<FUG>` eingeführt.

   Morphisto liefert alle möglichen Analysen für ein Eingabewort. Eine Disambiguierung mit Hilfe der Taggerausgaben wird nicht mehr vorgenommen. Bei Analysen, die die maximale Länge (4095 Zeichen) für Attribute in der CWB überschreiten, werden die komplexesten Einzelanalysen entfernt, d.h. diejenigen, die aus den meisten Morphen bestehen. Ein Beispiel zeigt Abbildung 1.

   Die Abbildung stimmt vielleicht ein wenig nachdenklich, wie brauchbar die morphologische Analyse überhaupt ist. Als feature set Attribut können für **morph** aber die zusätzlichen Suchoperatoren „contains" und „matches" verwendet werden, mit denen sich für manche Phänomene eventuell doch brauchbare Abfragen formulieren lassen. So sollte

   ```
   [morph contains ".*<V><SUFF><\+NN>.*"];
   ```

   helfen Verlaufsformen zu finden. Dabei werden mehr Treffer zurückgeliefert, als durch die Abfrage

   ```
   [morph matches ".*<V><SUFF><\+NN>.*"];
   ```

   weil letztere nur Token zurückliefert, wo alle generierten Analysen Verbableitungen sind.

[6] http://www1.ids-mannheim.de/lexik/TextGrid/morphisto/

[7] Für weitergehende Informationen zu den von Morphisto generierten Analysen muss hier auf http://www.ims.uni-stuttgart.de/projekte/gramotron/PAPERS/LREC04/smor.pdf und http://www.textgrid.de/fileadmin/TextGrid/konferenzen_vortraege/presentation_Morphisto_-_FSMNLP.pdf verwiesen werden.

```
  Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>Verbraucher<NN>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bund<NN>es<FUG>verbrauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
* Bund<NN>es<FUG>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verb<NN>rauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>Verbraucher<NN>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>ver<PREF>brauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<NN>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Mini<OTHER>Ster<NN>in<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>In<SUFF><+NN><Fem><Nom><Sg>
  Bundes<NN>verbrauchen<V>er<NN><SUFF>Schutz<NN>Minister<NN>in<SUFF><+NN><Fem><Nom><Sg>
```

Abbildung 1: Entfernung der komplexesten Analysen (mit * gekennzeichnet). Anders als in den Korpora wurde die Anzahl der Analysen durch teilweise Disambiguierung mittels des RFTagger Outputs reduziert, allein um die Präsentation zu ermöglichen.

2.3. Neue Positionale Attribute (Don Tuggener)

Aufgrund des Dependenzparses mit *ParZu*[8] des Korpus konnten folgende vier neue Attribute eingefügt werden.

1. **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."
   -**lemma** in vgfinal: "Sie bereiten der Vortrag vor."
   -**new_lemma** in vgfinal2: "Sie vorbereiten der Vortrag vor."
   D.h. man kann nun Lemmas von Verben suchen, die ein abtrennbares Präfix haben und erhält Belege von Gebrauch mit und ohne Abtrennung des Präfixes, z.B.

   ```
   [new_lemma="vorbereiten"]
   ```

2. **gf**: grammatikalische Funktion (z.B. ßubjfür 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. Die vorhandenen grammatikalischen Funktionen und deren Abkürzungen für die Verwendung in der Suche sind in Anhang A.1 gelistet.

3. **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).

4. **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"]
   ```

[8] https://github.com/rsennrich/ParZu

3. Strukturelle Einheiten

3.1. Suchsyntax

Für die Suche nach strukturellen Einheiten braucht es eine sehr viel umfangreichere Suchsyntax, um alle Anwendungsfälle abdecken zu können. Grob verkürzt gibt es folgende Möglichkeiten:

1. Die Verwendung des Namens einer strukturellen Einheit innerhalb einer Tokenbeschreibung (zwischen `[]`) sorgt dafür, dass ein Token sich mit der Suchabfrage nur deckt, wenn er innerhalb der entsprechenden strukturellen Einheit vorkommt, z.B. findet

   ```
   [pos="NE" & heading];
   ```

   nur Eigennamen innerhalb von Überschriften.

2. Die direkte Suche mittels Anfangs- und Ende-Tag. Dabei wird der Name der strukturellen Einheit in spitze Klammern eingeschlossen, wie ein XML-Element, z.B. liefert

   ```
   <s> []* </s>;
   ```

   alle Sätze. Die Anfangs- und Ende-Tags müssen nicht notwendigerweise gepaart verwendet werden, so dass Satzanfänge und Satzenden mit

   ```
   <s> [];
   ```

   bzw.

   ```
   [] </s>;
   ```

   gefunden werden können. Um die einfache Suche nach strukturellen Einheiten zu erleichtern ist das Makro `/region[]` definiert. Sätze lassen also auch mit

   ```
   region[s];
   ```

   suchen.

   Des weiteren ist es möglich, innerhalb des Start-Tags den Attributwert einer strukturellen Einheit abzufragen. So lassen sich in einer ersten Näherung konjunktivische Nebensätze mit folgender Abfrage untersuchen:

   ```
   <c_infl=".*Subj.*"> []* </c_infl>;
   ```

Tatsächlich müssten jedoch alle Einbettungsniveaus einzeln abgefragt werden, um tatsächlich alle entsprechenden Nebensätze zu finden, was zu recht unschönen Abfragen führen kann, etwa in der Form:

```
(<c_infl=".*Subj.*">|<c_infl1=".*Subj.*">|<c_infl2=".*Subj.*">) []* (</c_infl>|</c_infl1>|</c_infl2>);
```

3. Die Suche innerhalb einer strukturellen Einheit mittels des `within` Operators. Im Gegensatz zur vorherigen Möglichkeit stellt der `within` Operator lediglich sicher, dass der vorangehenden Suchausdruck die Grenzen der strukturellen Einheit nicht überschreitet. Die einzelnen Suchergebnisse brauchen aber nicht die gesamte strukturelle Einheit zu umfassen. Die Ergebnisse der Suchabfrage

   ```
   "um|für" []+ [word="besorgt" & pos="VVPP"] within c;
   ```

   sind deshalb nicht vollständige Nebensätze, sondern nur die Token zwischen `um` oder `für` und `besorgt`. Im Gegensatz zur direkten Tag-Suche

   ```
   <c> []* "um|für" []+ [word="besorgt" & pos="VVPP"] []* </c>;
   ```

   spart das nicht nur Tipparbeit sondern bringt auch sehr viel übersichtlichere Ergebnisse. Nebenbei wird die Sucheffizienz erheblich gesteigert.

4. Es können Labels verwendet werden, um für eine bestimmte Position innerhalb einer Suchanfrage den Wert eines Attributs einer strukturellen Einheit zu prüfen. Diese Möglichkeit ist immer dann zu gebrauchen, wenn das Start-Tag einer strukturellen Einheit, deren Attributwert von Interesse ist, sehr weit von dem eigentlichen Suchstring entfernt auftreten kann. Ein typisches Beispiel sind Suchen die auf Dokumente mit bestimmten Metaeigenschaften beschränkt werden sollen:

   ```
   a:"um|für" []+ [word="besorgt" & pos="VVPP"] :: a.doc_date=".*2011" within c;
   ```

   Hier wird das Label `a` an die erste Position in der Abfrage geheftet (in diesem Fall wäre natürlich auch ein beliebiger anderer Token möglich), für die in der globalen Randbedingung, die nach dem `::` folgt, geprüft wird, ob sie sich innerhalb einer strukturellen Einheit `doc_date` befindet, der ein Datum aus dem Jahr 2011 zugeordnet ist.

3.2. Typen

Die folgenden Abschnitte entsprechen drei Klassen von strukturellen Einheiten, die inhaltlich motiviert sind.

Attribut	Beschreibung
`crawl_date`	Tag an dem der Text heruntergeladen wurde. Muss nicht mit Erscheinungsdatum identisch sein.
`source`	Zeitung, aus der der Artikel stammt, die möglichen Werte sind Tabelle 7 zu entnehmen
`country`	Das Land, aus dem der Artikel stammt (s. Tabelle 8)
`region`	Die Region, aus der der Artikel stammt (s.Tabelle 9)
`area`	Das Gebiet, aus der dem Artikel stammt; unsauber extrahiert
`author`	Autor des Textes; unsauber extrahiert
`date`	Erscheinungsdatum; unsauber extrahiert
`url`	Internetadresse, von der der Artikel heruntergeladen wurde
`title`	Überschrift eines Artikels
`subtitle`	Kurzzusammenfassung, Lead In
`section`	Ressort; unsauber extrahiert

Tabelle 2: Dokument-Metainformation

Attribut	Beschreibung
`category`	Der Nebensatztyp sofern bestimmbar. Folgende Werte kommen vor: kopulativ, adversativ, kausal, disjunktiv, relativ, temporal, modal, final, konsekutiv, konzessiv, konsekutiv|modal, konditional
`relation`	Unterscheidet unter- von nebengeordneten Nebensätzen: `sub`, `co`
`construction`	erhält nur für Infinitivsätze mit `zu` den Wert `zu+inf`
`infl`	Beschreibt die Inflektionsmerkmale des Nebensatzes. Die möglichen Werte sind Tabelle 6 im Anhang zu entnehmen.

Tabelle 3: Nebensatzattribute

3.2.1. Textstrukturelle Einheiten

<text> Auf textstrukturell höchster Ebene angesiedelt ist die Textannotation, die einem einzelnen Zeitungsartikel entspricht. Bei jeder Suche sollte, notfalls durch `within doc`, sichergestellt sein, dass Textgrenzen nicht überschritten werden, da CWB-intern das gesamte Korpus am Stück repräsentiert ist. Der Text ist auch die Einheit, für die Metainformation vorhanden ist in Form der Eigenschaften, wie sie in Tabelle 2 aufgeführt sind.

<title> die Artikelüberschrift. Auch als Dokument-Metadatum vorhanden.

<body> markiert den eigentlichen Textkörper eines Artikels.

3.2.2. Syntaktische Einheiten

<s> Sätze sind die höchste annotierte syntaktische Einheit. Es gibt keine Einbettungen.

<c> Für Nebensätze sind die Attribute in Tabelle 3 verfügbar. In den gecrawlten Korpora kommen bis zu 11 Einbettungen vor.

Attribut	Beschreibung
`entityType`	Die Werte sind • `PER` für Personen • `LOC` für Orte • `ORG` für Organisationen und • `MISC` beispielsweise für Adjektive und Nationalitäten, die von Eigennamen abgeleitet sind

Tabelle 4: Einteilung von Named Entities

<field> steht für topologische Felder und hat lediglich das Attribut `category`, das Werte aus {`VF`, `LSK`, `MF`, `RSK`, `NF`} nimmt. Die Einbettungstiefen stimmen in beiden Korpora mit denen für Nebensätze überein.

3.2.3. Lexikalisch semantische Einheiten

<kommunikationsverb> Kommunikationsverben wurden an Hand einer Liste annotiert und werden durch das Attribut `semFeature` weiter unterteilt (s. Tabelle 10). Da ein Verb unterschiedlichen Klassen zugeordnet sein kann ist das `semFeature`-Attribut ein feature set und kann abgefragt werden wie für `morph` unter Punkt 5 beschrieben. Einbettungen kann es natürlich hier nicht geben.

<ne> Eigennamen („Named Entities") wurden mit Hilfe des Stanford Named Entity Recognizers (Finkel et al., 2005) annotiert unter Verwendung der Modelle für das Deutsche wie in Faruqui & Padó (2010) beschrieben. Auch hier keine Einbettungen. Es existiert ein Attribut, das den Entitätstyp angibt wie in Tabelle 4 beschrieben.

A. Anhang

Tag	Beschreibung	Beispiele
ADJA	attributives Adjektiv	das *große* Haus
ADJD	adverbiales oder prädikatives Adjektiv	er fährt *schnell*, er ist *schnell*
ADV	Adverb	*schon*, *bald*, *doch*
APPR	Präposition; Zirkumposition links	*in* der Stadt, *ohne* mich
APPRART	Präposition mit Artikel	*im* Haus, *zur* Sache
APPO	Postposition	ihm *zufolge*, der Sache *wegen*
APZR	Zirkumposition rechts	von jetzt *an*
ART	bestimmter oder unbestimmter Artikel	*der*, *die*, *das*, *ein*, *eine*
CARD	Kardinalzahl	*zwei* Männer, im Jahre *1994*
FM	Fremdsprachliches Material	Er hat das mit *,, A big fish* '' übersetzt
ITJ	Interjektion	*mhm*, *ach*, *tja*
KOUI	unterordnende Konjunktion mit zu und Infinitiv	*um* zu leben, *anstatt* zu fragen
KOUS	unterordnende Konjunktion mit Satz	*weil*, *daß*, *damit*, *wenn*, *ob*
KON	nebenordnende Konjunktion	*und*, *oder*, *aber*
KOKOM	Vergleichskonjunktion	*als*, *wie*
NN	normales Nomen	*Tisch*, *Herr*, das *Reisen*
NE	Eigennamen	*Hans*, *Hamburg*, *HSV*
PDS	substituierendes Demonstrativpronomen	*dieser*, *jener*
PDAT	attribuierendes Demonstrativpronomen	*jener* Mensch
PIS	substituierendes Indefinitpronomen	*keiner*, *viele*, *man*, *niemand*
PIAT	attribuierendes Indefinitpronomen ohne Determiner	*kein* Mensch, *irgendein* Glas
PIDAT	attribuierendes Indefinitpronomen mit Determiner	ein *wenig* Wasser, die *beiden* Brüder
PPER	irreflexives Personalpronomen	*ich*, *er*, *ihm*, *mich*, *dir*
PPOSS	substituierendes Possessivpronomen	*meins*, *deiner*
PPOSAT	attribuierendes Possessivpronomen	*mein* Buch, *deine* Mutter
PRELS	substituierendes Relativpronomen	der Hund *,* *der*
PRELAT	attribuierendes Relativpronomen	der Mann *,* *dessen* Hund
PRF	reflexives Personalpronomen	*sich*, *einander*, *dich*, *mir*
PWS	substituierendes Interrogativpronomen	*wer*, *was*
PWAT	attribuierendes Interrogativpronomen	*welche* Farbe, *wessen* Hut
PWAV	adverbiales Interrogativ- oder Relativpronomen	*warum*, *wo*, *wann*, *worüber*, *wobei*
PAV	Pronominaladverb	*dafür*, *dabei*, *deswegen*, *trotzdem*
PTKZU	zu vor Infinitiv	*zu* gehen
PTKNEG	Negationspartikel	*nicht*
PTKVZ	abgetrennter Verbzusatz	er kommt *an*, er fährt *rad*
PTKANT	Antwortpartikel	*ja*, *nein*, *danke*, *bitte*
PTKA	Partikel bei Adjektiv oder Adverb	*am* schönsten, *zu* schnell
TRUNC	Kompositions-Erstglied	*An-* und Abreise
VVFIN	finites Verb, voll	du *gehst*, wir *kommen* an
VVIMP	Imperativ, voll	*komm* !
VVINF	Infinitiv, voll	*gehen*, *ankommen*
VVIZU	Infinitiv mit zu, voll	*anzukommen*, *loszulassen*
VVPP	Partizip Perfekt, voll	*gegangen*, *angekommen*
VAFIN	finites Verb, aux	du *bist*, wir *werden*
VAIMP	Imperativ, aux	*sei* ruhig !
VAINF	Infinitiv, aux	*werden*, *sein*
VAPP	Partizip Perfekt, aux	*gewesen*
VMFIN	finites Verb, modal	*dürfen*
VMINF	Infinitiv, modal	*wollen*
VMPP	Partizip Perfekt, modal	*gekonnt*, er hat gehen *können*
XY	Nichtwort, Sonderzeichen enthaltend	*3:7*, *H2O*, *D2XW3*
$,	Komma	*,*
$.	Satzbeendende Interpunktion	*. ? ! ; :*
$(	sonstige Satzzeichen; satzintern	*-] ( )*

Tabelle 5: Die Tabelle des Stuttgart Tübingen Tagset wurde kopiert von http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/stts-table.html

Annotation	Beschreibung
Full.FutII.Ind.Act Full.FutII.Ind.Pass Full.FutI.Ind.Act Full.FutI.Ind.Pass Full.Past.Ind.Act Full.Past.Ind.Pass Full.Past.SubjI.Act Full.Past.SubjII.Act Full.Past.SubjI.Pass Full.Perf.Ind.Act Full.Perf.Ind.Pass Full.PPerf.Ind.Act Full.PPerf.Ind.Pass Full.Pres.Ind.Act Full.Pres.Ind.Pass Full.Pres.SubjI.Act Full.Pres.SubjII.Act Mod.FutII.Ind.Act Mod.FutII.Ind.Pass Mod.FutI.Ind.Act Mod.FutI.Ind.Pass Mod.FutI.SubjI.Act Mod.FutI.SubjI.Pass Mod.Past.Ind.Act Mod.Past.Ind.Pass Mod.Past.SubjI.Act Mod.Past.SubjI.Pass Mod.Perf.Ind.Act Mod.Perf.Ind.Act.nonstandard Mod.Perf.Ind.Pass Mod.PPerf.Ind.Act Mod.PPerf.Ind.Pass Mod.Pres.Ind.Act Mod.Pres.Ind.Pass Mod.Pres.SubjI.Act INF	Die komplexen Annotationen für die Inflektionsmerkmale von Nebensätzen spezifizieren der Reihe nach: 1. ob es sich beim finiten Verb um ein Vollverb oder ein Modalverb handelt 2. das Tempus 3. den Modus 4. die Diathese Einzig Infinitivsätze haben die atomare Annotation INF.

Tabelle 6: Inflektionsmerkmale von Nebensätzen

D-Osnabrücker Zeitun
D-Oberbayerische Vol
D-Nürnberger Nachric
D-Nordkurier
D-Neue Westfälische
D-Mitteldeutsche Zei
D-Mittelbayerische Z
D-Märkische Oderzeit
D-Märkische Allgemei
D-Leipziger Volkszei
D-Lausitzer Rundscha
D-Kölner Stadt-Anzei
D-Kieler Nachrichten
D-Hessische/Niedersä
D-Heilbronner Stimme
D-Hannoversche Allge
D-Hamburger Morgenpo
D-Freie Presse
D-Fränkischer Tag
D-Frankenpost
D-Dresdner Neue Nach
D-Der Tagesspiegel
D-Darmstädter Echo
D-Badische Zeitung
D-Augsburger Allgeme
CH-Südostschweiz.ch
CH-St. Galler Tagbla
CH-Neue Luzerner Zei
CH-Bernerzeitung.ch
CH-Basellandschaftli
CH-Aargauer Zeitung
CH-1818 - Das Oberwa
BELG-Grenz-Echo
Badische Zeitung
A-Wirtschaftsblatt
A-Wiener Zeitung
A-Vorarlberger Nachr
A-Tiroler Tageszeitu
A-Salzburger Nachric
A-Salzburger Fenster
A-Oberösterreichisch
A-Niederösterreichis
A-Kurier (Wien
A-Kurier (Oberösterr
A-Kurier (Niederöste
A-Kurier (Burgenland
A-Krone
A-Kleine Zeitung
A-Der Standard

Tabelle 7: Werte des `<text>`-Attributs `source`. Leider sind die Werte tatsächlich momentan nach 20 Zeichen abgeschnitten. Im nächsten Release wird das behoben (KR).

Belgien
Deutschland
Liechtenstein
Luxemburg
Österreich
Schweiz
Südtirol

Tabelle 8: Werte des `<text>`-Attributs `country`

A-Mitte
A-Ost
A-Süd
A-West
BELG-Eupen
CH-Ost
CH-Süd
CH-West
D-Mittelost
D-Mittelwest
D-Nordost
D-Nordwest
D-Südost
D-Südwest
LIE-Vaduz
LUX-Letzebuerg
STIR-Südtirol

Tabelle 9: Werte des `<text>`-Attributs `region`

Tabelle 10: Liste der Kommunikationsverben und ihre Einteilung

`semFeature`-Werte	Verben
Deklarativa	abdanken, aberkennen, abordnen, abrreren, abserzen, anklagen, anzeigen, befreien, beglaubigen, beichten, belangen, benedeien, berufen, bestellen, beurlauben, definieren, degradieren, einsetzen, einstellen, endassen, entlassen, entziehen, ernennen, fesdegen, feuern, freilassen, freisprechen, kundtun, lossprechen, nominieren, proklamieren, ratifizieren, rauswerfen, schassen, segnen, suspendieren, taufen, verdonnern, verklagen, verknacken, verkünden, versetzen, weihen, zurückbeordern, zurücktreten
Direktiva	abblasen, abfragen, abhören, abkommandieren, abmahnen, abonnieren, absagen, abschlagen, abverlangen, abwürgen, anbetteln, anflehen, anleiten, anlernen, anmahnen, annullieren, anordnen, anraren, anserzen, anweisen, appellieren, auffordern, aufgeben, aufrufen, auftragen, ausfragen, aushorchen, ausquetschen, ausweisen, autorisieren, beantragen, beauftragen, befehlen, befehligen, befragen, beibringen, beordern, berechtigen, beschwatzen, besrärken, betteln, bevollmächtigen, bewilligen, billigen, bitten, bitten zu, breitschlagen, einarbeiten, einberufen, einfordern, eingreifen, einladen, einschreiten, einweisen, einziehen, erbetteln, erflehen, erfragen, erlauben, ermächtigen, ermahnen, ersuchen, examinieren, flehen, fordern, fragen, gebieten, gemahnen, genehmigen, gestatten, gewähren, herumfragen, herumkriegen, instruieren, interviewen, kommandieren, konsultieren, laden zu, löchern, mahnen, mobilisieren, nachfragen, nachhaken, nachsuchen, ordern, prüfen, raten, rekrutieren, reservieren, rückfragen, rumfragen, rumkriegen, überreden, überweisen, unterbinden, untersagen, unterweisen, veranlassen, verbieten, verfügen, verhören, verlangen, vernehmen, verordnen, verschreiben, verwehren, verweisen, vorbestellen, vorladen, vorschlagen, wünschen, zitieren, zuraten, zureden, zurückfragen, zurückpfeifen
Expressiva	anfauchen, angeifern, angifren, anherrschen, anprangern, anscheißen, anschimpfen, anschnauzen, anschwärzen, aufschneiden, ausschelten, ausschimpfen, beanstanden, bedauern, befürworten, beglückwünschen, begrüßen, beleidigen, bemängeln, beschimpfen, beschönigen, beurteilen, bewerten, blamieren, bloßstellen, danken, diffamieren, diskreditieren, diskriminieren, ehren, einschätzen, einstufen, fluchen, frohlocken, geifern, giften, grantein, gratulieren, grüßen, hänseln, herabsetzen, herabwürdigen, höhnen, honorieren, huldigen, jammern, jauchzen, jubeln, jubilieren, klassifizieren, klatschen, kompromittieren, kondolieren, kritisieren, lamentieren, lästern, loben, lobpreisen, mäkeln, meckern, missbilligen, monieren, mosern, motzen, murren, necken, nörgeln, poltern, prahlen, preisen, protzen, rüffeln, rügen, rühmen, schelten, scherzen, schimpfen, schlecht machen, schmähen, schmeicheln, schönfärben, schönreden, schwärmen, spotten, tadeln, urteilen, veralbern, veräppeln, verfluchen, verherrlichen, verhöhnen, verklären, verleumden, verspotten, verunglimpfen, vorhalten, vorwerfen, wehklagen, werten, wettern, witzeln, würdigen, zeihen, zujubeln, zurechtweisen
Gesprächs- und Themenstrukturierende Verben	abschweifen, andeuten, anfügen, anführen, anknüpfen an, anmerken, anreißen, anschließen, anschneiden, antworten, aufgreifen, aufnehmen, beantworten, bemerken, betonen, dagegenhalten, dazwischenrufen, dazwischenwerfen, differenzieren, eingehen, einhaken, einschieben, einwenden, einwerfen, entgegnen, ergänzen, erwähnen, erwidern, fortfahren, fortführen, fortsetzen, herausstellen, hervorheben, hinweisen, hinzufügen, kommentieren, rekurrieren, repetieren, resümieren, unterscheiden, unterstreichen, vorbringen, wiederholen, zurückgreifen, zurückkommen, zusammenfassen
Kommissiva	ablehnen, abmachen, abschwören, anbieten, androhen, aushandeln, ausmachen, bedrohen, bürgen, drohen, entsagen, garantieren, geloben, gewährleisten, offerieren, protestieren, schwören, übereinkommen, verabreden, vereinbaren, versprechen, verzichten, zusagen, zusichern
Kommunikationseröffnende Verben	anreden, kontaktieren
Mediale Kommunikationsverben	annoncieren, anrufen, aufsagen, aufschreiben, chatten, deklamieren, emailen, faxen, funken, inserieren, korrespondieren, krakeln, kritzeln, morsen, niederschreiben, posten, rezitieren, schreiben, simsen, stenografieren, telefaxen, telefonieren, telegrafieren, texten, vorlesen, vortragen

Fortsetzung auf nächster Seite

Tabelle 10 – Fortsetzung von vorheriger Seite	
Modale Kommunikationsverben	anbrüllen, anschreien, brabbeln, brüllen, brummen, flüstern, grölen, herumbrüllen, herumgrölen, herumkreischen, herumschreien, herumstammeln, herumstottern, herunterbeten, herunterleiern, herunterrasseln, krächzen, kreischen, leiern, lispeln, murmeln, näseln, nuscheln, plappern, quäken, raunen, rumbrüllen, rumgrölen, rumkreischen, rumschreien, rumstammeln, rumstottern, runterbeten, runterleiern, runterrasseln, schnattern, skandieren, stammeln, stottern, tuscheln, überschreien, übertönen, verfassen, wispern, zischen, zubrüllen, zuflüstern, zuraunen, zurufen, zuschreien
Redesequenzverben	beratschlagen, besprechen, debattieren, diskutieren, disputieren, erörtern, klönen, plaudern, plauschen, quatschen, ratschen, schnacken, schwatzen
Repräsentativa	abstreiten, anflunkern, ankündigen, anlügen, anschwindeln, anvertrauen, anzweifeln, argumentieren, aufdecken, ausplaudern, ausrichten, beflunkern, beipflichten, bekräftigen, belügen, benachrichtigen, berichten, beschreiben, beschwindeln, bestreiten, beteuern, bezweifeln, darlegen, demenrieren, dementieren, eingestehen, einlenken, einräumen, entgegenkommen, enthüllen, entkräften, erflunkern, erläurern, erlügen, erschwindeln, erzählen, flunkern, hinterbringen, informieren, insinuieren, klar machen, konstatieren, kontern, leugnen, lügen, mitteilen, nachgeben, nahe bringen, näher bringen, preisgeben, prophezeien, richtig stellen, rumflunkern, rumlügen, rumschwindeln, schildern, schwindeln, suggerieren, überbringen, übermitteln, überzeugen, unken, veranschaulichen, verbreiten, verdeudichen, verlautbaren, vermitteln, verneinen, verraten, verständigen, voraussagen, vorflunkern, vorhersagen, vorlügen, vorschwindeln, warnen, weissagen, wetten, widerlegen, widerrufen, widersprechen, zugeben, zurückziehen, zutragen
Verba dicendi	schreien
Deklarativa ∨ Direktiva	abberufen, anberaumen, festsetzen, freistellen, kündigen, vorschreiben, zurückrufen
Deklarativa ∨ Expressiva	anerkennen, beschuldigen, bestimmen, bezichtigen, klagen, verurteilen
Deklarativa ∨ Kommissiva	absprechen
Deklarativa ∨ Repräsentativa	bekannt geben, bekennen, bekunden, bestätigen, einweihen, erklären, feststellen, gestehen, melden, offenbaren
Deklarativa ∨ Repräsentativa ∨ Direktiva	bestellen
Deklarativa ∨ Repräsentativa ∨ Gesprächs- und Themenstrukturierende Verben	nennen
Direktiva ∨ Expressiva	empfehlen, gutheißen
Direktiva ∨ Kommissiva	anfragen, einwilligen, verweigern
Direktiva ∨ Mediale Kommunikationsverben	diktieren
Direktiva ∨ Modale Kommunikationsverben	rufen
Direktiva ∨ Redesequenzverben	beraten, bereden
Direktiva ∨ Repräsentativa	beharren, beschwören, bestehen, erinnern, pochen, unterrichten, zugestehen, zurücknehmen, zustimmen
Expressiva ∨ Gesprächs- und Themenstrukturierende Verben ∨ Repräsentativa	angeben
Expressiva ∨ Repräsentativa	bejahen, versichern, zurückweisen
Gesprächs- und Themenstrukturierende Verben ∨ Kommunikationseröffnende Verben	ansprechen
Mediale Kommunikationsverben ∨ Kommunikationseröffnende Verben	anschreiben
Redesequenzverben ∨ Verba dicendi	reden, sprechen
Repräsentativa ∨ Verba dicendi	behaupten, sagen

A.1. Grammatikalische Funktionen

ACHTUNG: die Abkürzungen hier sind gross geschrieben, in der Suche müssen sie aber kleingeschrieben werden!

```
The dependency labels implemented in ParZu are described in:
Killian A. Foth. 2005. Eine umfassende Constraint-Dependenz-Grammatik
des Deutschen. University of Hamburg.
Here is a short summary

ADV - ADVERB
------------
Adverbial function (includes adverbial adjectives, negations, and similar)
    Er kommt *spät*

APP - APPOSITION
----------------
includes both close and true appositions
close apposition:
    Präsident *Obama*;
    Elton *John*
true apposition:
    Peter, der *Besitzer* des Ladens

ATTR - ATTRIBUTE
----------------
attributive adjectives or numbers
    Der *kleine* Mann

AUX - AUXILIARY
---------------
auxiliary verb relation
(note that the auxiliary verb is the head of the relation)
    Ich bin *angekommen*

AVZ - ABGETRENNTES VERB-PRÄFIX
------------------------------
separated verb particle
    Er kommt morgen *an*

CJ - CONJOINED ELEMENT
----------------------
last element in a conjunction
    Peter und *Lisa*
```

```
DET - DETERMINER
----------------
    *der* Mann

EXPL - EXPLETIVE "ES"
---------------------
"es" as a syntactic placeholder for a clausal object/subject
    *es* ist gefährlich, alleine zu gehen

GMOD - GENITIVE MODIFIER/ATTRIBUTE
----------------------------------
genitive noun phrase that modifies another noun phrase.
    Der Krieg der *Sterne*

GRAD - DEGREE
-------------
noun phrase that expresses a magnitude/degree
    Er ist drei *Jahre* alt

KOM - COMPARATIVE
-----------------
comparative conjunction
    hart *wie* stahl
    schöner *als* Ferien

KON - COORDINATION
------------------
all members in a coordination (except for the last one, which is CJ)
both the conjoined elements and the conjunctions are marked as KON
    Peter, *Susi* *und* Mark

KONJ - CONJUNCTION
------------------
subordinating conjunction
    Ich hoffe, *dass* es klappt

NEB - NEBENSATZ (SUBORDINATE CLAUSE)
------------------------------------
subordinate clause that is adjunct
    Wenn alles *klappt*, komme ich morgen

OBJA - ACCUSATIVE OBJECT
------------------------
    Ich sehe den *Mann*
```

```
OBJA2 - 2nd ACCUSATIVE OBJECT
-----------------------------
is mapped to OBJA

OBJC - CLAUSAL OBJECT
---------------------
subordinate clause that is complement
    Ich hoffe, dass es *klappt*

OBJD - DATIVE OBJECT
--------------------
    Er vertraut *ihr*

OBJG - GENITIVE OBJECT
----------------------
    Er verdächtigt den Mann des *Mordes*

OBJI - INFINITIVE OBJECT
------------------------
    Er versuchte, ihr zu *helfen*

OBJP - PREPOSITIONAL OBJECT
---------------------------
prepositional phrase that is complement
    Er ging mit ihr *ins* Kino

PAR - PARENTHETICAL EXPRESSIONS
-------------------------------
used for some tricky constructions in which one clause is interrupted by another one.
    "Ich", *sagte* er, "komme morgen"

PART - PARTICLE
---------------
mostly used for "zu" and some fixed expressions:
    nicht *zu* glauben
    *ein* wenig besser

PN - PREPOSITION COMPLEMENT
---------------------------
    Er liegt in der *Sonne*

PP - PREPOSITIONAL PHRASE
-------------------------
```

prepositional phrase that is adjunct
    Er ging *mit* ihr ins Kino

PRED - PREDICATIVE NOUN

    Er ist *tot*, Jim

REL - RELATIVE CLAUSE
----
connects head of relative clause to its head.
the relative pronoun receives the label that corresponds to its function in the relat
    der Spion, der mich *liebte*

S - SENTENCE

Mostly used for reported speech
    Er sagte, er *sei* unschuldig

SUBJ - SUBJECT

    Er sagte, *er* sei unschuldig

SUBJC - CLAUSAL SUBJECT

    Es freut mich, dass ihr alle hier *seid*

VOK - VOCATIVE

    Er ist tot, *Jim*

ZEIT - TEMPORAL NOUN PHRASE
----
noun phrase (without preposition/conjunction) with temporal function
    2011 ging die Firma bankrott
    Anfang Februar begann der Aufschwung

Literatur

Faruqui, M. & Padó, S. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS 2010*, Saarbrücken, Germany.

Finkel, J. R., Grenager, T., & Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, (pp. 363–370)., Stroudsburg, PA, USA. Association for Computational Linguistics.