Benchmark voor parsers van Nederlandse historische morfologie

Folgert Karsdorp & Tom Kenter

Instituut voor Nederlandse Lexicologie

{folgert.karsdorp,tom.kenter}@inl.nl

2 mei 2011

1 Introductie

Voor de workshop op 27 mei over parsers van historische morfologie is een dataset opgesteld waarmee de deelnemers hun morfologische parsers kunnen testen op diachroon morfologisch materiaal. Met deze dataset hopen we een beter beeld te krijgen van welke parsers er voor handen zijn en hoe goed de parsers presteren voor de verschillende fasen van het Nederlands. Op basis van de resultaten van deze benchmark, krijgen we hopelijk meer inzicht in de hoeveelheid onderzoek die nog nodig is op het gebied van parsing van historische morfologie.

In dit stuk bespreek ik de opbouw van de dataset (§2) en de taakstelling voor de benchmark (§3).

2 Dataset

Aangezien de morfologische parsers ingezet zullen worden bij de ontwikkeling van een groot diachroon corpus, is de dataset zo opgebouwd dat deze min of meer representatief is voor de Nederlandse morfologie vanaf de zesde eeuw tot nu. De dataset bevat 4897 geanalyseerde woordvormen[1], gelijk verdeeld over de tijd. De 4897 woordvormen zijn uit de verschillende historische woordenboeken van het INL gehaald: het ONW, VMNW, MNW en WNT.

De selectie van de woorden is als volgt gegaan. Voor elk lemma in de historische woordenboeken zijn alle citaten met een datering geëxtraheerd. De dateringen zijn gegroepeerd per tien jaar. Voor elk decennium zijn er willekeurig circa 60 woordvormen gekozen.

[1] Er is gekozen voor woordvormen in plaats van lemma's, omdat met name de lemma's in de oudere fasen van het Nederlands enkel reconstructies zijn en geen geattesteerde vormen.

2.1 Opmaak van de dataset

De velden in de dataset zijn gescheiden door puntkomma's. In de eerste kolom staat de ID van de woordvorm. De tweede kolom geeft de orthografische representatie van het lemma. De derde kolom geeft de woordvorm. In de vierde kolom staat het jaartal van de bron waarin de woordvorm is geattesteerd. In sommige gevallen zijn er meerdere morfologische analyses van een woord mogelijk. Daarom is in de vijfde kolom de ID van de morfologische analyse gegeven. De laatste kolom geeft de morfologische analyse van de woordvorm.

De morfologische analyse van een woord is weergegeven in een boomstructuur, waarbij telkens het linker argument van een haakjespaar de woordsoort aangeeft van het rechter argument. Het rechter argument kan een complex morfologisch segment zijn of een lexicaal element.

3 Taakstelling

De deelnemers analyseren de complete dataset. Hierbij gaat het om een volledige hiërarchische analyse van de woordvormen (dus niet van de lemmata). Er is bewust voor gekozen om geen periodisering in de dataset aan te brengen, zodat de deelnemers zelf een voor hun systeem zinvolle indeling kunnen maken. Het spreekt voor zich dat te analyseren testwoorden niet mogen voorkomen in het trainingsmateriaal. Aangezien de dataset vrij klein is, wordt de deelnemers geadviseerd om gebruik te maken van methoden van cross-validation (bijvoorbeeld *leave-one-out cross-validation* of 10 *fold cross-validation*).

Er mag gebruik gemaakt worden van alle mogelijke externe bronnen. Daarbij kunnen we denken aan morfemenlexica of aan lexicale databanken voorzien van morfologische informatie, zoals de databank CELEX en het lexicon van het *Corpus Gesproken Nederlands* (CGN).

Bij elke analyse moeten de volgende elementen als output gegeven worden:

1. het ID van het testitem;
2. de te analyseren woordvorm;
3. de gemaakte analyse;
4. de geobserveerde analyse; en
5. het jaartal (of jaartallen) van het citaat waarin het woord is geattesteerd.

Als er meerdere morfologische analyses van een woord als output gegeven worden, worden die op een afzonderlijk regel gezet. De verschillende velden

in de output moeten gescheiden worden door een puntkomma. Hieronder is een voorbeeld gegeven van hoe de output opgemaakt moet worden:

```
1;versnel;(VRB (PRE ver) (ADJ snel));(VRB (PRE ver) (ADJ snel));1956
```

3.1 Evaluatie

De resultaten van de analyse moeten voor 20 mei opgestuurd worden naar Folgert Karsdorp (`folgert.karsdorp@inl.nl`) die de gegevens verder zal verwerken. De evaluatiematen die gebruikt worden, zijn: de *precision*, *recall* en *F-score*. De scores worden voor verschillende periodes in de diachronie berekend en vervolgens in een grafiek geplaatst.

De resultaten van alle systemen worden gecombineerd in één grafiek. Op deze manier hopen we een beeld te krijgen van de periodes waarin een bepaalde parser goed presteert en in welke minder. Ook kunnen we een beeld krijgen welke systemen voor welke periode het beste ingezet kunnen worden en voor welke periodes er nog meer onderzoek gedaan zal moeten worden.