XeroX
Dutch Grammar Report
Preliminary Version 1.0

Xerox Research Center Europe
France, Meylan
September 18th, 2007
Author: Joeri Honnef
Supervision: Anne Schiller

Table of Contents

Introduction

This document gives reports on of the grammar development work during five and a half months (April-September 2007) of the internship carried out by Joeri Honnef under the supervision of Anne Schiller. The first part will focus on the Volkskrant corpus (Dutch newspaper) and gives some very practical examples of problems in this corpus and the preliminary results obtained by the grammar. The second part of this evaluation report will focus on a more specialized corpus, namely the pocket engineer corpus. Translations of all given examples can be found in the notes on top of the examples, indicated in yellow in OpenOffice.

Dutch Grammar performance on *Volkskrant*

When we talk about the overall performance of the Dutch grammar we mainly focus on the recall and precision of NP's, PP's, Subjects and objects.

The Corpus (De Volkskrant)

The Volkskrant corpus is constructed from online articles found at: http://www.volkskrant.nl. Since the website is subdivided into categories we also chose to subdivide the articles into categories, namely Foreign(Buitenland), Sport(Sport), Economy(Economie). In each directory of the main corpus directory[1] you will find a number of articles named with the title of the article and a date stamp. Besides these articles you will find one file which contains all the separate articles. In this file the date stamps are removed together with headings which only have one word, the journalist who wrote the article and the location stamp. This file makes it easier to copy-past the entire set of sentences into the Xerox Incremental Parser (XIP).

[1] '/home/jhonnef/parSem/xip/dev/GRAMMARS/DUTCH/'

Disambiguation

Words can have different part of speech tags in different contexts. A lexicon provides us with these part of speech tags but we have to select the right part of speech tags according the specific features or the context which is surrounding the word. Next, I will describe some frequent ambiguities.

A main ambiguity problem is the noun/verb ambiguity. Plural or singular forms of verbs can often also occur as singular or plural nouns. Examples (which are treated) are (notes are translation):

- naar buiten *trekken*
- naar binnen *duwen*
- ... , *wil* precies weten wat er is gebeurd.
- geen *berichten*
- de ruzie ontaardde in fel anti-Aziatisch *geweld*
- Herstellen van beschadigde gebouwen is lastig in oorlogstijd.

Other frequent ambiguities are verb-particles/prepositions and adverbs/adjectives. The next examples show different usages of this word. In the first example the 'bij' is a bee(insect). In the second example it is the preposition 'at' denoting a place and in the last example it is the separable prefix of the verb 'bijwerken' which in this sentence means 'keeping up with his homework'.

- De *bij* vloog zeer hoog in the lucht.
- *Bij* het meer stond ook een huis.
- Hij werkte zijn huiswerk *bij.*

Another example of this kind of ambiguity is the word *'in'.*

- Ik *in* het geld zodra ik de mogelijkheid heb.
- *In* de toillet vond ik mijn sleutels.
- Hij voerde de gegevens *in.*

Chunking

Chunking aims to retrieve the right constituents in a sentence. Constituents are formed by taking the right part of speech tags into a constituent which is also often referred to as a chunk. An well known constituent is the noun phrase which could for example exist of a determiner followed by an adjective and a noun. The main difficulty in this task was finding the right boundaries for different sorts of subclauses within sentences. Especially the start of these phrases is hard to detect since some cues could also function as prepositions. When the phrases begin with a noun phrase it becomes especially hard to decide if this is the actual beginning of a subclause. In general, noun phrases and prepositional phrases were not that difficult given that the correct part of speech tags were assigned in the tagging process before. Examples:

- Hij deed dat speciaal *om haar vaker te kunnen zien*.
- Het is een trend waarmee bedrijven zich proberen te profileren.
- Zij verbaasde het publiek *door het zelfs nog sneller te doen*.

Some relative clause examples :

- Hanna heeft een broer *die met zijn vrouw en dochtertjes in België woont*.
- Luuk is het jongetje *dat in de dierentuin is kwijtgeraakt*.

Subordinate/subclauses always start with the subordinating conjunction and end with the finite verb. Sometimes prepositional occur after the finite verb but we chose not to include them since it is not always clear if they belong to the subclause or to the next constituent. Examples[2]:

- 's Maandags ga ik laat naar de universiteit *omdat ik dan college heb.*
- Ik ga vaak naar de universiteit *omdat ik bijna elke dag college heb.*
- *Als ik geen college heb*, ga ik naar de universiteit om te studeren.

[2] Examples taken from http://www.ucl.ac.uk/dutch/grammatica/subordinate_clauses.htm

Dependencies

Apart from finding the dependencies between determiners and nouns, mainverbs, auxiliaries, etc. the main focus of the dependency task was on finding the right subject(s) and object(s). Generally, the subject occurs in front of the main verb and the object after it. But of course there are a lot of examples where this is not the case. I follow with some examples. A lot more examples can be found in the grammar itself.

Subject-Verb-Object examples :

- *Ik* ga vaak naar de universiteit.
- *Pakistaanse stammen* hebben deze week met hulp van het Pakistaanse leger *driehonderd buitenlandse militanten* gedood.
- *Wij* hebben *dat* niet gezegd.

Subject-Object-Verb examples:

- ... , dat je Parijs-Roubaix eerst moet verliezen om ...
- ... ,dat Wolfowitz nooit toestemming heeft gekregen.

Verb-Subject-Object example:

- Zonder die opleiding is het winnen van de klassiekers een eitje.
- Is de ronde tafel geen mythe?

Evaluation

Since there are no correct analysis' for the articles we used, we analysed the articles first with XIP and then manually corrected wrong or missing output of XIP. We both used the Foreign and Economy categories of the collected Volkskrant articles to develop the Dutch grammar. The Foreign articles were mainly used for the chunking and disambiguation tasks whereas the Sport category was also used to develop the Dependency rules.

Performance of the chunking grammar was mainly measured using recall and precision on noun phrases, prepositional phrases. Statistics on other chunks can also be retrieved using the evaluation scripts in the EVAL directory of the main Dutch grammar directory. Performance of the dependeny rules was also measured by comparing the discovered subject(s) and object(s) with the manually annotated corpus and calculating the recall and precision.

Results

The following results are obtained by developing the grammar rules for this specific corpus.

Foreign, 67 sentences	***Precision***	***Recall***
NP	99.2	98.9
PP	99.3	100
Subject	85.7	80
Object	82.4	64.6

These results are obtained by runinng the grammar on the Economy corpus without developing/ adjusting the rules for this corpus.

Economy, 70 sentences	***Precision***	***Recall***
NP	95.5	97.2
PP	97.4	98.3
Subject	86.8	51.9
Object	51.1	38.1

After the previous test we developed rules further and specified them to meet certain errors and we obtained these results:

Economy, 70 sentences	***Precision***	***Recall***
NP	95.4	95.7
PP	96.6	98.3
Subject	83.8	86.1
Object	73.8	75.0

Dutch Grammar & *Pocket Engineer*

Corpus

This corpus is a set of sentences explaining the user of a printer what to do when certain warnings, errors or other messages appear on the printer screen. Often sentences start with an imperative since they give commands to the user. Another frequent phenomenon is the usage of names of certain windows/tabs which are not always as easy as they seem to detect because they are not consistently annotated with the same HTML-tags. Sometimes they are denoted in italic and sometimes in bold. However differences in HTML tags are not that a big problem as long as they just a certain tag. A bigger problem are the names that do not have a HTML-tag at all. Here we need to focus on capitalization and hope for consistent capitalization. If this is not the case we are unable to extract the correct NP chunks and have to switch to more lexicon orientated resources.

Evaluation

For this corpus we did not do any real evaluation yet since the specifications of what kind of information to be extracted is not yet fully developed and as it can change we will not define these beforehand. Preliminary analysis of the parser on the pocket engineer sample file can be found in the pocket engineer directory.