XIP German Grammar for Pocket Engineer

(august 2006, S. Maurel)

1. Introduction

The purpose of this document is to present the work done for the XIP German grammar of Pocket Engineer (by august 2006).

2. Utilized files

Grammar files:

`german_PE.grm`	in GRAMMAR/GERMAN/GRMFILES
`german_PE.xip`	in GRAMMAR/GERMAN/GRMFILES
`disamb_agr.xip`	in GRAMMAR/GERMAN/DISAMB
`disamb_lemma.xip`	in GRAMMAR/GERMAN/DISAMB
`disamb_rank.xip`	in GRAMMAR/GERMAN/DISAMB
`disamb_tagger.xip`	in GRAMMAR/GERMAN/DISAMB

Added files:

a) Rule files:

`localgram_PE.xip`	in GRAMMAR/GERMAN/LOCALGRAMS
`chunker_PE.xip`	in GRAMMAR/GERMAN/CHUNKING
`dependencies_PE.xip`	in GRAMMAR/GERMAN/DEPENDENCIES
`normalize_PE.py`	in GRAMMAR/GERMAN/DEPENDENCIES

b) Lexicon:

`lexicon_PE.xip`	in GRAMMAR/GERMAN/LEXICONS

c) Declarations:

`features_PE.xip`	in GRAMMAR/GERMAN/FEATURES
`categories_PE.xip`	in GRAMMAR/GERMAN/FEATURES
`controls_PE.xip`	in GRAMMAR/GERMAN/FEATURES
`functions_PE.xip`	in GRAMMAR/GERMAN/DEPENDENCIES

3. Creation of PEUnits

The file `dependencies_PE.xip` creates PEU nodes that get a dependency PEUNIT at the end. The intermediate added dependencies are COMBINE, PEUNIT (and PEU, but this is not needed, it contains only the PEU, and not the normalized form that is in PEUNIT) and MLD (for the message texts, in German or English).

Simple PEU are NPs, verbal chunks, PPs and APs.
Grouped PEU are:
- a genitive NP that follows a nominative (nominalized) NP

  PEU{NP{aus^=wechseln} NP{das heften^#klammern^#Magazin}}}
- a PP that follows a NP which contains the lemma "Anweisung"

  PEU{NP{Anweisung} PP{zu NP{fest^=legen}} NP{die Standard^#Einstellung}}}
- abbreviations (in brackets) are put together with the preceding PEU

  PEU{PP{mit NP{das NOUN{Kodak poly^#Chrom Graphics}}} (NP{KPG})}
- a verb and its object

  PEU{VFIN{werden} PP{von NP{Vorlage^#Einzug}} beschädigen}}
- a verb and its predicate

  PEU{VFIN{^=ziehen} NP{die Vorlage} schräg ein}
- a verb and its prefix (they are put together at the normalization anyway, even when the prefix is far from the verb)

  PEU{VFIN{^=führen} NP{kein Papier} zu}
- negation

  PEU{VFIN{werden} nicht akzeptieren}
  PEU{VFIN{können} nicht}
  PEU{VFIN{erkennen} NP{das Material^#Format} nicht}
- product/printer names separated by / or |

  PEU{PP{über NP{ein Fiery VERSION{X12 / XP12 / EX12}}}
- coordinated nouns

  PEU{NP{AP{schwarz} Linie} und NP{Streif}}

NB: the word "Meldung" in italic forms a unique PEU. Everything before (the text of the message) is one single PEU.

PEU{" NP{Schließe} " VFIN{betätigen} , dann NP{Funktion^/s^#Auswahl} überprüfen ...} PEU{NP{Meldung}}

4. Normalization

The file `normalize_PE.py` creates the normalized form of the PEUNIT.

- are deleted in the normalized form:
determinants, conjunctions, pronouns, punctuation, prepositions, auxiliary and modal verbs, particles and prefixes.
- determinants/particles with the feature NEG are replaced by the word **NOT**

- verbprefixes are attached to their verb
- other words get the lemma form

5. Tags and Features

Tags of the corpus:
All html-tags except of the italic tags were deleted from the corpus (→ `ed_x_main_de_training_ohneHTML.txt`). The tags <i> and </i> give the features italic to the words they are surrounding.

Special features:
`peuplus`: for the word “Anweisung”
`mld`: for the word “Meldung”

6. Corpus

The corpus (in GRAMMAR/GERMAN/CORPUS/PE) was separated in two parts, every second line of the original file (`ed_x_main_de.txt`) was put in the file `ed_x_main_de_test.txt` for testing purpose. The rest (`ed_x_main_de_training.txt`) was used to develop the rules.
First the html-tags were deleted (except of the italic ones) to create the file `ed_x_main_de_training_ohneHTML.txt`. The file `ed_x_main_de_training_i+tab.txt` still contains the tabulating tags, but at the moment no rules were created to handle the tables correctly.
The results of the analyses are in the folder GRAMMAR/GERMAN/PETest.

7. Problems

The corpus is not very well translated and formatted. Especially the messages aren’t coded in a coherent way. Often only the word “Meldung” is in italic, and not the real message. Messages in English cannot be analyzed by the German grammar. Messages which contain more than one sentences are divided by XIP, and the rest of the German sentence will not be analyzed with the beginning, this can cause erroneous analyzes.
Error codes are for the most time read as adjectives, this can create strange APs and NPs.