Summary of the Tagsets used in the Dutch SemEval coreference data

February 9, 2010

1 Introduction

The data set on which the annotations are performed is the KNACK-2002 corpus, which contains texts from the Flemish weekly magazine Knack. The full corpus contains 267 texts which will be split into a trial, training and test set. For the initial guidelines, we refer to http://webs.hogent.be/%7Evhos368/proefschrift/AppendixA.pdf. The annotation format has been adapted to the shared task.

Th data set contains different annotations layers, all of which have been automatically generated. The lemma, part of speech and named entity information was generated by the memory-based shallow parser for Dutch (Daelemans, Bucholz and Veenstra 1999).

2 Part of speech information

There has been a conversion of the Dutch fine-grained tagset (Daelemans, Bucholz and Veenstra 1999) to a more coarse-grained tagset (represented in Column 7). The **PoS tags** are:

1. VNW: pronoun
2. WW : verb
3. N: noun
4. ADJ: adjective
5. SPEC: special tokens
6. VZ: preposition
7. TW: number
8. BW: adverb
9. VG: conjunction
10. LID: determiner
11. LET: punctuation

The PoS features are:

1. NOUNS

type	common
	name
num	singular
	plural
case	standard
	genitive
	dative
degree	basis
	dimin
gender	mascfem
	neutral

2. ADJECTIVES

position	postnominal
	nominal
	free
degree	basis
	comparative
	superlative
	diminutive
inflection	no
	with-e
	with-s

3. VERBS

mood	finiteform
	infinitive
	pastparticiple
	gerund
tense	present
	past
	subjunctive
position	prenominal
	nominal
	free
number	singular
	plural

4. NUMBERS

type	cardinal
	ordinal
position	prenominal
	nominal
	free
degree	basis
	diminutive

5. PRONOUNS

type	personal
	personal_reflexive
	possessive
	reflexive
	reciprocal
	interrogative
	relative
	interrogative_relative
	exclamative
	demonstrative
	indefinite
pdtype	pron
	det
case	standard
	oblique
	nominative
	genitive
	dative
person	1
	2_familar
	2_polite
	2
	3_impersonalreferent
	3_personalreferent
	3_malereferent
	3_femalereferent
	3
number	singular
	plural
gender	masc
	fem
	neutral
position	prenominal
	nominal
	free
inflection	without
	with-e

6. ARTICLES

type	definite
	indefinite
case	standard
	genitive
	dative

7. PREPOSITIONS

type	initial
	final
	combi_art

8. CONJUNCTIONS

type	coord
	subord

3 Dependency information

The dependency information was provided by the Alpino parser (Bouma et al. 2000) which is available from http://www.let.rug.nl/vannoord/alp/Alpino/.

ROOT	sentence (ROOT)
app	apposition
body	body of subordinate clause
cnj	member of conjunction
crd	coordinator
det	determiner
hd	head
hdf	closing element of circumposition
ld	locational or directional complement
me	measure complement
mwp	any part of a multiword unit
mod	modifier
obcomp	comparative complement
obj1	direct object
obj2	secondary object
pc	prepositional object
pobj1	provisional direct or first object
predc	predicative complement
predm	secondary predicate
punct	punctuation
sat	satelite
se	obligatory reflexive object
su	subject
sup	provisional subject
svp	verbal particle
vc	verbal complement

References

Bouma, Gosse, Gertjan van Noord, Robert Malouf. *Alpino: Wide Coverage Computational Analysis of Dutch.* In: Computational Linguistics in the Netherlands CLIN 2000.

Daelemans, Walter, Sabine Buchholz, Jorn Veenstra, *Memory-Based Shallow Parsing*. In: Proceedings of CoNLL-99, Bergen, Norway, June 12, 1999, pp. 53-60.