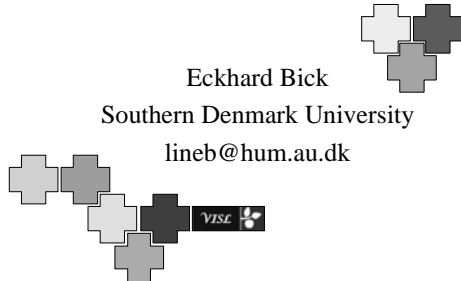
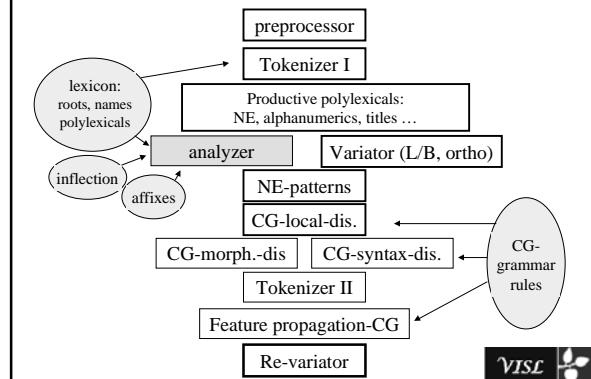


PALMORF's morpholymphical headaches



Eckhard Bick
Southern Denmark University
lineb@hum.au.dk

Distributed morphology



Program chains used in the Morpholymphics

Runcorp. analex:
prp.pre | preprocessor | prp.post | morfanalyse | cg2adapt
| samtrad | dis --grammar portcg.wordlist | mdis --grammar
portcg.wordlist | propagation.perl | joinall | cleanup.perl
| unjoin | samtrad | niceline.perl | uppercase.perl |
remove_secondary | dianafilter.post

Runwordlist:
perl -wnpe 's/s/\n/' | prp.pre | preprocessor | perl -wnpe
's/^(da|do)s*-\$/\$1A-/' | morfanalyse | cg2adapt | samtrad
| dis --grammar portcg.wordlist | mdis --grammar
portcg.wordlist | propagation.perl | cleanup.perl | samtrad
| niceline.perl | uppercase.perl | remove_secondary |
dianafilter.post | perl -wnpe 's/\s*\n/g'

Some problems in retrofixing an integrated tagger-disambiguator for multi-tagging

- Palmorf does some tokenization only after disambiguation:
dese. o que. nos. consigo. hyphenates. somename chains
- Most adjectives with potential nominal function are not tagged as nouns by the analyzer, but identified as np-heads by the syntactic grammar, and - if wanted - later marked with a secondary noun tag
- The analyzer itself uses some port-manteau tags (*-amos PR/PS -ista MF*) instead of 2 separate tag lines. The ambiguity is resolved after syntax by dependency propagation.
- Palmorf expects running text input, and its pattern matching filters can get confused by spaces before punctuation, non-standard quotes, non adjacent apostrophes etc.
- In a tagger-disambiguator, derivation overgeneration across word classes is no problem, since it can be contextually resolved, so the analyzer as such is not optimized for precision in this field

Why polylexicals (MWE) ?

- Recognizing polylexical "**adverbs**", "**prepositions**" etc. (*locuões*) is essential for context based disambiguation and syntax, since it makes context patterns less complex
- Recognizing polylexical "**nouns**" and **verb incorporation** (*cobra cascavel, estar com fome*) helps semantic disambiguation and MT.
(Verb incorporation is currently inactivated in PALAVRAS)
- Recognizing **name chains**, and treating them as units,
 - a) creates **simpler context** for disambiguation of other words/functions
 - b) allows **NE-subtyping**, e.g. <hum> recognition by Christian name first part
 - c) prevents **lower case name parts** (*de, von*) from receiving other word classes
 - d) allows a meaningful analysis of **name-integrated numbers and punctuation** (*car names, &, ., D. Fernando II, Nimbus 2000*)
 - e) avoids **unnecessary ambiguity** in uppercase words also lexicalized as non-names

Local disambiguation 1

Local word class disambiguation in hyphenates
MAP (\$ADD) TARGET (<stop> PROP) (-1 <hyfen> LINK 0 (<stop> PROP)); # EUA-África
MAP (\$ADD) TARGET (<stop> PROP) (0 <hyfen>) (1 (<stop> PROP)); # EUA-África
SELECT (PERS) (-1 <hyfen> + V);
REMOVE (N) (0 ("político" ADJ) LINK 0 <hyfen>) (1C ADJ); # político-social
SELECT (PRP) (0 <hyfen>) (-1 <hyfen>);
REMOVE NON-N-WORD (-1 <hyfen>) (NOT 0 <hyfen>) (-1C (PR 3S)); # guarda-avanga
REMOVE NON-V-WORD (0 <hyfen>) (1 PERS);
REMOVE NON-N-WORD (0 <hyfen> LINK 0 N-DYR OR N-HUM); # cobra-coral

Local word class disambiguation otherwise
REMOVE (PROP) (0 <art> OR PRP OR KS OR NUM OR <rel> OR <inter> OR <atemp>);
REMOVE (<doc> ADV) (0 ("ser")) (NOT 0 (3S IND));

Local case disambiguation on clitics
MAP (\$DAT) TARGET (<hyfen> PERS ACC/DAT) (1 PERS); deu-mos
MAP (\$DAT) TARGET (PERS ACC/DAT) (-1 <hyfen> LINK 0 <vd>); # ajudou-me
MAP (\$ACC) TARGET (PERS ACC/DAT) (NOT 0 <hyfen>) (-1 <hyfen>) (NOT -1 <vd> OR <vdt&>); # convidou-me
MAP (\$ACC) TARGET (PERS ACC/DAT) (-1 <hyfen>) (NOT 1 PERS) (NOT -1 <vd> OR <vdt&>); # lavar-se-iam

Local disambiguation 2

Local gender/number disambiguation in hyphenates

MAP (\$M) TARGET (<hyfen> M/F) (1C (ADJ M) OR (PCP M)); # comunista-adjunto, azul-claro

MAP (\$S) TARGET (<hyfen> S/P) (1C (ADJ M) OR (PCP S));

MAP (\$M) TARGET (M/F) (-1C (N M) OR (ADJ M) OR (PCP M) LINK 0 <hyfen>);

MAP (\$S) TARGET (S/P) (-1C (N S) OR (ADJ S) OR (PCP S) LINK 0 <hyfen>);

Local mode disambiguation in hyphenates

REMOVE (IMP 2S) (0 <hyfen>) (1 PERS) (NOT 1 (2S ACC)); # incorpora-se, ikke: lava-te

Overgenerating affixation

REMOVE (<DERS <DERS) (0 PROP); REMOVE (<DERP <DERP) (0 PROP);

REMOVE (<DERS -az [AU])> <DERS -ão [AU])>; # no double augmentatives

REMOVE (<DERS -aço [AU/PEJ])> <DERS -ão [AU])>; # no double augmentatives in competition with -ação/-ações

REMOVE (<ico [DIM]> N) OR (<im [DIM]> N) OR (<ela [DIM]> N); # these diminutives are not productive for N

REMOVE (<ito [DIM]> ADJ); # these diminutives are not productive for ADJ

REMOVE (<DERS <DERS -ia [ABSTR])>; # '-ia' not after other suffixes

REMOVE (<DERS <DERS -or [ABSTR])>; # '-or' not after other suffixes

Local disambiguation after NE-type-recognition

REMOVE (<DERS) OR (<DERP) (0 (<<xheur>>)); # after name recognition in samtrad

REMOVE (<foreign> PR 1/3S); REMOVE (<HEUR> PROP) (0 (<top> PROP) OR (<H> PROP));

Prepare for re-tokenizer (doesn't fuse if ambiguous)

SELECT (PROP) (-1 PROP); # SOS Ásia



Hopes for the future



Hopes for the future

- Evaluation of **disambiguated** PoS/morphology, *in context* and on *running text*.
- Evaluation of **syntactic** tagging in conjunction with PoS/morphology (would solve ADJ/N, ADV/CONJ, focus particle, and a number of other problems in inter-system comparison
 - though probably creating a load of new ones

