PaNoLa: The Danish Connection

Eckhard Bick

Institut for Sprog og Kommunikation, Syddansk Universitet

lineb@hum.au.dk, http://visl.hum.sdu.dk

1. The project

PaNoLa (Parsing Nordic Languages) is a two-year Nordic Constraint Grammar research project involving the following participating institutions, with research team leaders in parentheses:

- Institute of Language and Communication, Southern Denmark University (Eckhard Bick)
- Tekstlaboratoriet, Oslo University (Janne Bondi Johannessen)
- Department of Linguistics, Helsinki University (Fred Karlsson)
- Department of Linguistics, Uppsala/Göteborg University (Torbjörn Lager)

Resources are equally distributed across the participating countries, with the exception of system development done centrally by the VISL team at Southern Denmark University, and software licensed from the Finnish firm Lingsoft.

PaNoLa's project goal is to develop, improve and applicationally integrate existing Constraint Grammars (CG) for the major Nordic languages. The computational tools that are developed during the project, are being made available through the internet, with a special focus on computer aided learning and corpus annotation.

2. The Danish section

One major activity in the Danish section has been the grammatical annotation of two large Danish quote corpora, Korpus90 and Korpus2000 (about 60 million words), both compiled by DSL (Det Danske Sprog- og Litteraturselskab) from a larger body of mixed genre texts by randomized sentence extraction. All texts were annotated with word based morphosyntactic tags, using a CG based multi-level parser (DanGram, Bick 2001), and made available for searching at http://corp.hum.sdu.dk, as well as http://www.dsl.dk . These corpus data, both raw and annotated, have provided valuable insights with regard to both descriptive and methodological issues. A limited subcorpus is being manually revised with the goal to correct errors from the automatic annotation and detect systematic patterns of annotation errors, as well as descriptively problematic areas of Danish grammar.

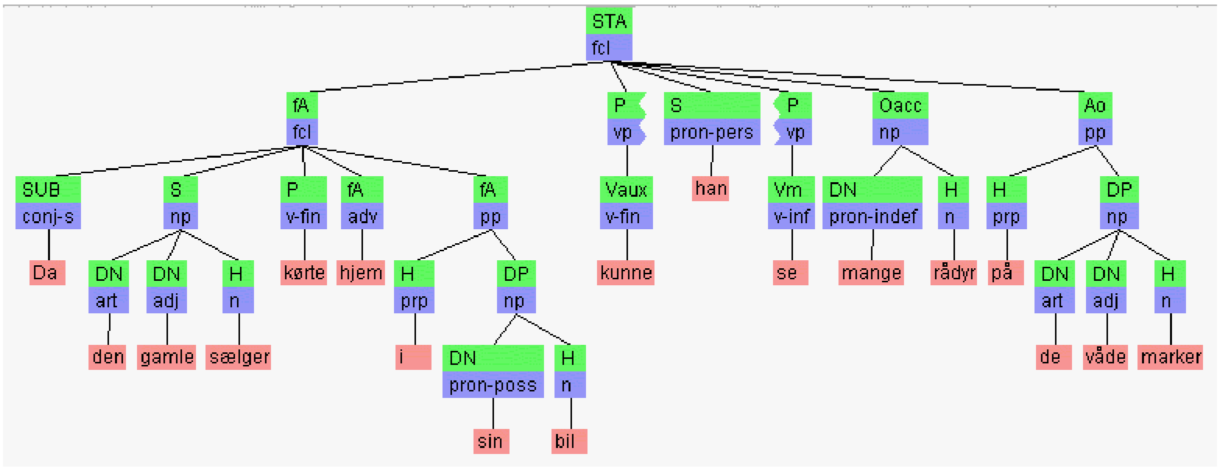
The annotation scheme used covered the full set of Danish word classes (15) and inflexion categories (about 25 features) and 104 syntactic categories (45 major categories, 15 dependency variants, 16 finite clause functions, 23 non-finite clause functions, 5 averbal clause functions). The main 20 categories, without dependency and subclause variants, are listed below.

@SUBJ	subject	**@ADVL**	free adverbial
@ACC	direct (accusative) object	**@PRED**	free prædikativ
@DAT	indirect (dative) object	**@APP**	apposition
@PIV	prepositional object	**@>N**	prenominal dependent
@SC	subject complement	**@N<**	postnominal dependent
@OC	object complement	**@>A**	adverbial pre-dependent
@SA	argument adverbial, subject related	**@A<**	adverbial post-dependent
@OA	argument adverbial, object related	**@P<**	argument of preposition
@MV	main verb	**@INFM**	infinitive marker
@AUX	auxiliary	**@VOK**	vocative

Completing the dependency description and allowing add-on constituent grammar analysis, subclause function was also included, using the first verb in every clause as function tag carrier. The example below contains a finite subclause (FS) functioning as "left adverbial" (@FS-ADVL>).

Da	[da]	**KS**	@SUB
den	[den]	**ART** UTR S DEF	@>N
gamle	[gammel]	**ADJ** nG S DEF NOM	@>N
sælger	[sælger]	**N** UTR S IDF NOM	@SUBJ>
kørte	[køre]	<mv> **V** IMPF AKT	**@FS-ADVL>**
hjem	[hjem]	**N** NEU P IDF NOM	@<ACC
i	[i]	**PRP**	@<ADVL
sin	[sin]	<poss> <refl> **DET** UTR S	@>N
bil	[bil]	**N** UTR S IDF NOM	@P<
,			
så	[se]	<mv> **V** IMPF AKT	@FMV
han	[han]	**PERS** UTR 3S NOM	@<SUBJ
mange	[mange]	<quant> **DET** nG P NOM	@>N
små	[lille]	**ADJ** nG P nD NOM	@>N
dyr	[dyr]	**N** NEU P IDF NOM &ACI-SUBJ	@<ACC
på	[på]	**PRP**	@<OA
de	[den]	**ART** nG P DEF	@>N
våde	[våd]	**ADJ** nG P nD NOM	@>N
veje	[vej]	**N** UTR P IDF NOM	@P<

Improving VISL's internet based grammar teaching tools (http://visl.sdu.dk) is another applicative aspect of the PaNoLa project, so attention was also paid to a special hybrid Phrase Structure Grammar (PSG), that uses CG tags, not lexicon entries, as terminals of its rewriting rules. This add-on PSG allows the automatic generation of syntactic trees from CG input.



However, due to the all-or-nothing character of rewriting rules, it is difficult to have a PSG cover all variants of natural language, and small errors in the CG analysis may propagate into structural errors or rule incompatibilities in the PSG. Thus, at present many automatically generated trees are still "partial trees". For limited corpus and teaching purposes this can, of course, be remedied by proof-reading analyses at the CG-level (i.e. PSG-input), or by controlling the tree format itself, and the developmental concept of PaNoLa is to use such gold standard data in turn to test and calibrate the parsing system.

Both formats allow, of course, simple word class recognition exercises (like VISL's ShootingGallery, WordFall and PaintBox games), while at the syntactic level each format has different applications, e.g. CG-format for word base cross-and-circle exercises (PostOffice game) and PSG-format for structural exercises (tree visualiser and manipulation program, SpaceRescue and SynTris games). The range and flexibility of these tools was enhanced as part of the PaNoLa project, and PaNoLa exploits most existing VISL-tools in a synergistic way by providing parallel data from the other Nordic languages, which in the first instance take the shape of manually established "closed" teaching corpora, but will later also cover automatically generated analyses.

In the following, a number of special descriptive problems from the Danish Constraint Grammar will be discussed.

3. Dative constructions

It is generally accepted, that as a case morpheme, the category of dative is all but extinct in Danish, with etymologically dative forms surviving only in fixed expressions *(til huse, på færde, i sinde).* Modern Danish nouns inflect for nominative and genitive, while personal pronouns retain a third case, generally regarded as accusative (and subsuming the original dative). It might be tempting, therefore, for the *function* of dative (@DAT), to use an exclusively structural definition as *indirect* object (preconditioning a *direct* object in the same clause), a notion common in English grammars:

(a1) Han gav **pigen** @<DAT æblet @<ACC.
(a2) Han fortæller **pigen** @<DAT en hemmelighed @<ACC

It seems reasonable to retain the dative valency of "fortælle" even without a direct/accusative object present:

(b) **Hvem** @<DAT fortæller han om @<PIV hemmeligheden?

Again concluding from English, one might assume that the lack of a morphological case marker would force a fixed word order on @ACC @DAT sequences, with @DAT either preceding @ACC (a1), or turning into a prepositional phrase (pp) with "til", here marked as prepositional object (prepositive, @PIV).

(c) Han gav æblet @<ACC **til pigen** @<PIV.

However, though useful for automatic disambiguation rules, object order does not appear to be all-decisive in Danish, in particular where "til" is part of the verb form:

(d1) Har du tilmeldt dig @<ACC **folkeregistret** @<DAT
(d2) Vi har kunnet tilpasse os @<ACC **de forandringer** @<DAT, som ...

As a general rule, Danish CG assigns dative function, where the Fillmorean case role of *benefactive* applies in a wider sense:

(e) Han købte konen @<DAT en ny bil.

In some cases, the benefactive case is restricted to reflexive pronouns:

(f1) Forestil **dig** @<DAT en verden uden Ekstrabladet!
(f2) Han tager **sig** @<DAT god tid.

As with "til" for ordinary datives, benefactive datives can often be substituted with a more isolatable "for"-pp:

(g1) Det er **mig** @<DAT ubegribeligt at han bliver ved. (ubegribeligt for mig)
(g2) I dag står det **byen** @<DAT frit @<SC at vælge @ICL-<SUBJ ...
(g3) Stolen er forbeholdt **direktøren** @<DAT (or @A<DAT dependent).
(g4) Det gik ham @<DAT godt ADV @<SA

For 2-object sentences, Danish allows passive-constructions with both subject-turned accusatives and subject-turned datives. The parser's descriptive strategy is here to retain the original function tag of the surviving (not subject-turned) object:

(h1) **Præsidenten** @<SUBJ blev tildelt prisen @<ACC.
(h2) Prisen @<SUBJ blev tildelt **præsidenten** @<DAT.
(h3) *aktiv:* De @<SUBJ tildelte **ham** @<DAT det @<ACC.

One might argue that (h2) only differs from (h1) in terms of constituent order, and that "præsidenten" should be kept subject, thus avoiding an indirect without a direct object. However, pronoun substitution shows that the last constituent in (h2) is non-nominative, ruling out the subject reading:

(h2') Prisen @<SUBJ blev tildelt **ham** @<DAT
(h1') **Han** @<SUBJ blev tildelt prisen @<ACC

Apart from pronoun case, coordination can be used to prove that the first constituents in (h) are true subjects:

(h2") Prisen var en middag for to og blev tildelt ...
(h1") Præsidenten tabte valget, men blev tildelt ...

This case test is positive even for s-passives and verbs without "til":

(h4) **De** @<SUBJ sikres **bøgerne** @<ACC.
(h5) De sikres **dem** @<DAT. (Kasus-test: akkusativ)

Sometimes the syntactic difference between @<ACC and @<DAT can carry discriminatory semantic weight, as in (i) where "skyldes" in (i2) is a "passive lexeme" with a different meaning - "be due to", not "owe", as in (i1).

(i1) Hun skylder Peter @<DAT 100 kroner.
(i2) Nederlaget skyldtes **vejret** @<ACC.

In sentences where only one of two objects can be turned subject, the accusative object seems to be the prototypical candidate:

(j1) Han blev tilmeldt folkeregistret @<DAT
(j2) *Folkeregistret blev tilmeldt ham @<ACC.
(j3) *aktiv:* Han har tilmeldt sig @<ACC folkeregistret @<DAT.

Last, the @DAT tag appears for want of a better tag where the parser, and - in a sense - a grammarian, can rule out other options:

(k1) Det går **mig** @<DAT på.
(k2) Det er **mig** @<DAT inderligt imod @<SC. (alternativ: postposition?)
(k3) at være **én** @DAT>A? hørig

The verbs in all three sentences cannot normally have direct objects, while inverted pp-readings can be ruled out in (k1) due to the change in meaning ("go on top of") and in (k2) due to the intensivier "inderligt". Rather, both "på" and "imod" should be read as adverbs, forcing the only remaining object category on "mig". (k3) is a special case, an alternative to the object reading being *argument of adjective,* with "én" becoming a dependent of "hørig".

4. Fronted nominal objects

The most frequent example of fronted objects in Germanic languages - apart from utterance objects before quote verbs - are relative and interrogative pronouns, followed by accusative marked and demonstrative pronouns ("This I could not do"). Unlike English, Danish does in principle allow fronting of *nominal* objects, too, even in ordinary speech, but being less of a free word order language than its case-inflecting relative German, these constructions are rare and either quite marked or part of frozen expressions.

The DanGram parser was used to corroborate linguistic intuition about the general distribution of major direct object form types in a 1.1 million word chunk of the annotated Korpus2000, and then specifically to extract the rarer fronted object candidate np's (and their contexts) for closer inspection. In all, 7.1% of all function tags (78.081 tokens) were direct object tagged, falling into the following subclasses:

form type	**fronted (ACC>)**	**right of main verb (<ACC)**
finite clause (FS)	5.2 % (quotes!)	12.8 %
non-finite clause (ICL)	0.0 % (1 case)	5.3 %
nouns (N)	0.3 % (checked)	53.8 %
proper nouns (PROP)	0.0 % (12 cases)	3.4%
relative pronouns	1.9 %	-
interrogative pronouns	0.5 %	- (4 adverbs)
personal pronouns	1.0 %	12.0 %
others	0.4 %	4.4 %
all	8.3 %	91.7 %

As can be seen, only 3-4% of non-clausal direct objects are fronted, and only about a tenth of these are nouns/np's, a finding that raises certain methodological questions. In general, a cautious, mainly REMOVE rule based, Constraint Grammar will tend to overrepresent a rare category, because of the "last surviving reading-maxime", preventing the last reading from being discarded, even *if* a relevant rule would apply. Let's assume, the REMOVE rule subset concerning a given frequent syntactic category has (a) a false positive error probability of 1% (i.e. removing where it shouldn't remove), and the corresponding REMOVE rule subsets for other major categories have (b) a false negative error probability of 3% each (i.e. being over-cautious and not removing what it should remove). Then, if there are 10 major categories, and the correct (major) reading has been removed (a), there will be a dependent probability of 0.97 to the tenth degree, i.e. a roughly 75% chance, that the other major category readings will be - correctly - removed (b). Thus, if there are 10 minor categories with equal chances of - wrongly - profiting from this, each will get a distribution of 0.75% even *without* counting the real cases, just because some reading *has* to survive.

Therefore, a traditional Constraint Grammar may be a good tool for extracting *candidates* for rare grammatical categories, but not for measuring their absolute frequency in relative terms. In the case of fronted nominal direct objects (@ACC>), DanGram "overgenerates" by about 30-50%, but since all cases were to be subtyped manually anyway, false positives did not have a bearing on the results. Not counting 23 non-clausal quotes, 272 correct @ACC> nouns were found, corresponding to 0.025% of all function tokens. With an automatic analysis alone it is hard to tell how many @ACC> escaped the CG-rules (false negatives), but a manually corrected test chunk from the same corpus, with 15.000 function tokens, though statistically usatisfying for such a rare feature, at least did not indicate false negatives as a major problem, since it yielded almost the same frequency for @ACC> (0.027%).

With 7 subtypes, distribution was as shown in the table below. Proper nouns had the same overall relative ACC> frequency as nouns, but occurred only in two subclasses, *focus* (7 instances) and *verb chain* (5 instances).

Subtype	**n**	**frequency**	**definition**
interrogative	79	29.0 %	at se, **hvilken interesse** kineserne skulle *have*
topic	74	27.2 %	**Denne interesse** *overførte* han på virksomheden **De problemer** *har* jeg slet ikke.
focus	55	20.2 %	**Blot 6-7 kr.** vil sparekassen *se* som betaling **Sin spillefilmsdebut** *fik* han i 1962 med ...
fronted in verb chain	43	15.8 %	... få **tyvekosterne** *bragt* hjem ... får man **billeder** at *se* gratis ... at lære **de nødvendige redskaber** at *kende*
raised	12	4.4 %	**Den slags** er vi jo nogle stykker der kan *lide*
fixed	7	2.6 %	Hvad **udvalget af værker** *angår*, har ...
vp-internal	2	0.7%	... at min søn **ingen huller** *havde* ... hun har **ingen kage** *bagt*

The largest class, interrogative, is in fact a variant of the interrogative pronoun case, and covers np's with interrogative modifiers ("hvilken/hvis bil", "hvor lang tid"), some in regular questions (7 cases), but most in interrogative object clauses. More interesting are the 3 marked structure categories (topic, focus and raised), which together amount to over half of all cases. The distinction between topic and focus was based on the semantic distinction between previously mentioned information (definite-anaphoric topic) and previously not mentioned information (newly introduced focus), but the borderline between these two categories was somewhat uncertain due to the fact that Korpus2000 is a quote corpus. The third marked structure category, main clause fronting of subclause constituents, is common in informal Danish, but more so for pronouns, and in particular, 'det' and the fronting of arguments of prepositions.

5. Gender fluctuation

With regard to nouns, Danish is a two-gender language. In general, the distinction between neuter gender (neutrum) and common gender (utrum) is a purely grammatical one, without semantic implications. However, neuter gender is also productively being employed to express the semantic feature of +mass:

(a1) Øllet var stærkt og mørkt.
(a2) De drak en øl hver.
(b) Det var noget godt mad, du lavede.
(c) Vejen får ekstra meget trafik om morgenen.

In (b) and (c), the neuter gender of a quantifier *(noget, meget)* forces a +mass reading of its dependency head, a common gender noun. (a) is one of the rare examples, where the distinction has been lexicalized in the noun itself. Note that in (b), neuter gender also extends to a modifying adjective. In order to test the hypothesis of semantic use of neuter gender and examine the phenomenons lexical distribution, word pairs with a neuter gender premodifier and a common gender noun were automatically extracted from the CG-tagged version of the Korpus2000. Note that the meget/megen cases (ca. 3700) have a 50/50 distribution in absolute terms, while the noget/nogen-cases (ca. 7300) have a 1:6 distribution, making "noget+UTR" a more marked structure than "meget+UTR". It appears that only neuter quantifiers can be used to project +mass onto Danish nouns, while the neuter articles *det* and *et* are disallowed, maybe because they imply "individuality" (i.e. +count) rather than +mass. Thus, a corresponding search extracted only spelling errors (heuristic nouns are mostly tagged UTR), English nouns, grammar-errors and special constructions like *det nye Hafnias kollaps,* where "det" *is* a prenominal dependent, but of another modifier *(Hafnias)* rather than the target word *kollaps.*

left context noget/nogen @>N (prenominally) (frequency > 3 and > 10)	NEU- %	**left context** meget/megen @>N (prenominally) (frequency > 4 and > 6)	NEU- %
aftensmad, ballade, creme, energi, fodbold, frugt, honning, ild, juice, kaffe, kriminalitet, morgenmad, musik, olie, selvtillid, vin	100	benzin, fodbold, føde, kaffe, sex, væske	100
		fejl*, medicin	91-92
		pris*, strøm, alkohol, frihed, mad, gang*, motion, larm, luft, søvn	81-90
		fart, olie, forstand, suverænitet, mælk, underholdning	71-80
opmærksomhed	50	lyst, vægt, sport, støj, spilletid, humor	60-75
plads (benplads)	30	plads, trafik, sol, tid, skade, magt, energi	51-60
tid	21	prestige, umage, musik, støtte	41-50
erfaring, viden, lovgivning	16-17	forskning, glæde, respekt, opmærksomhed,uro, regn, debat, indflydelse, kontakt, spalteplads, træning, kritik	31-40
usikkerhed, udvikling, debat	10-11	erfaring, fritid, tale, hjælp, diskussion, fantasi, nytte, kærlighed, mening, ros,	21-30
fremtid, succes, *trussel*	4	sympati, smerte, tvivl, (alvor, omsorg, vilje, forståelse, opbakning, smag), viden, virak, omhu	11-20
forskel, mulighed	1-2	medieomtale,inspiration, snak, omtale	1-10
aftale, anelse, art, chance, effekt, fare, *forbindelse,* garanti, *grad, grund, hemmelighed, hindring,* hjælp, *ide,* interesse, *katastrofe, konflikt, løsning, måde, nyhed ...*	0	(blæst, interesse, lidelse), læsning, (modgang, munterhed), møje, (omtanke, opmuntring), polemik, skepsis	

*The cases in question were *tage meget fejl, sætte meget pris på, der var meget gang i ...,* , where "meget" should be read as an adverb rather than a premodifier, since "fejl" alternates with verbal particles *(passe 'på, tage 'fejl),* rather than objects, and "meget" can *not* be substituted with "megen":

Den bensin han købte ... - *Den fejl han tog
... at han ofte købte meget/megen bensin - ... *at han ofte tog megen fejl

Since all lexemes in the *meget/megen* half of the table by necessity are mass nouns (since they allow a mass quantifier), the question is why some have a high and others a low incidence of neuter gender marking. In fact, for *meget/megen* a cline can be observed from mostly concrete mass nouns (bensin, føde, mad, væske) at the top of the table towards mostly abstract mass nouns at the bottom. The *noget/nogen* half of the table shows a corresponding concentration of concrete mass nouns at the top (vin, frugt, honning, ild), but the distribution gradient is much sharper than for *meget/megen,* simply because *nogen* is used with +count nouns, too (words italics in the zero-cell of the table). The negative *ikke nogen* is, in fact, a synonym of ingen, which is used with both count and mass nouns, and even in the plural (*nogen* is usurping this function, but officially remains *nogle* in the plural). Only the non-negative singular is +mass-forcing. Compare the following:

(a1) Jeg har ikke noget glas.
(a2) Jeg har ikke noget krus.
(b1) ?Jeg har noget glas.
(b2) *Jeg har noget krus. (--> Jeg har et krus)
(c1) Jeg har lavet noget (nogen?) aftensmad.
(c2) Han har vist nogen (noget?) interesse for forslaget.
(c3) *Han har nogen bil.

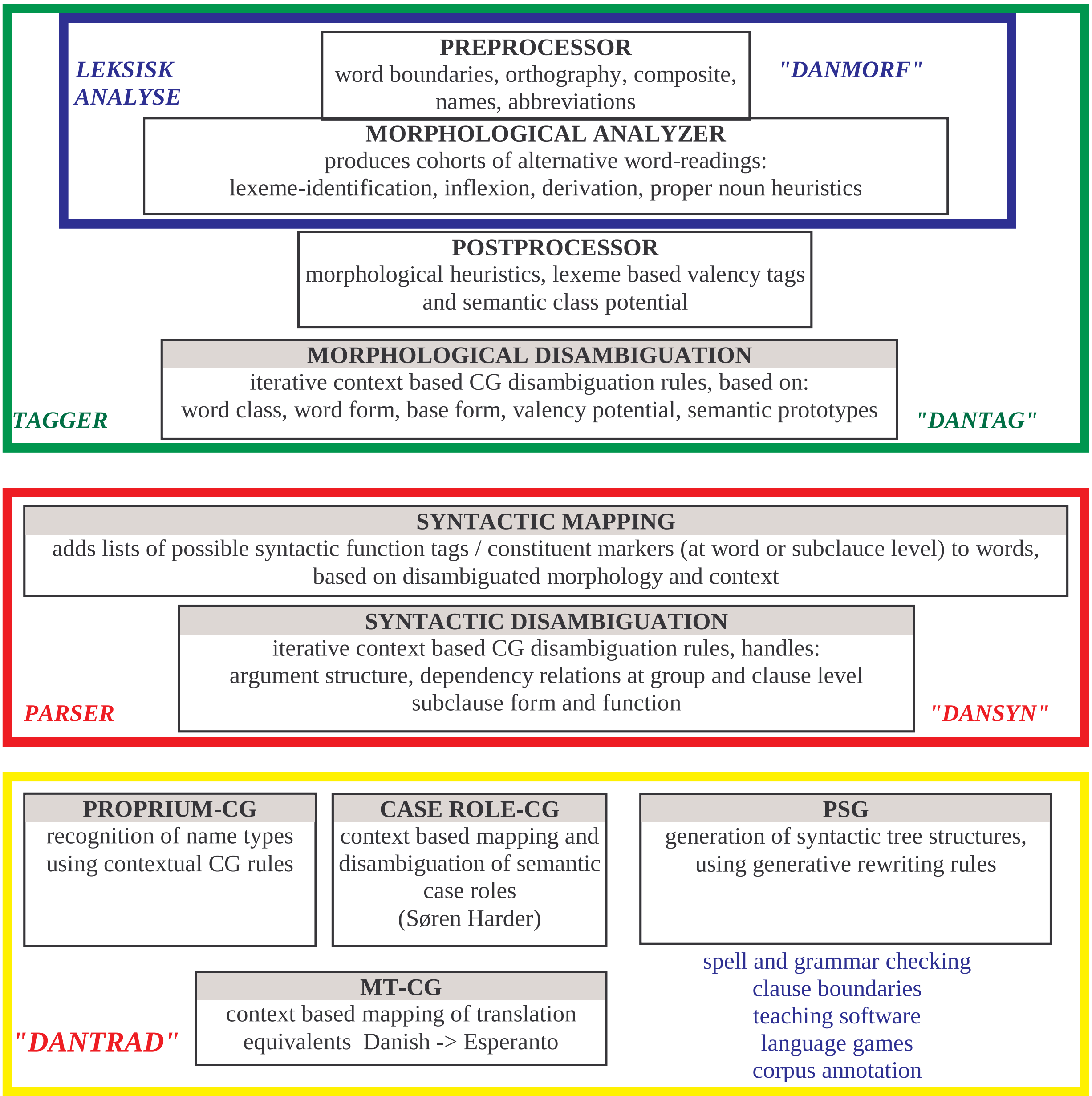
While allowing +count nouns in negative cases (a), *noget* is limited in usage in the affirmative versions (b), where there is a clash between the +mass forcing "noget" and the +count noun. In the case of "glass", an unintended mass reading is forced (glass as a material). When restricted to non-negative singular cases, *noget/nogen* shows the same cline as *meget/megen,* from (c1) concrete mass *(noget)* to (c2-3) abstract mass *(nogen).*

References:

Bick, Eckhard, *The Parsing System ‘Palavras’ - Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework*. Århus: Aarhus Universitetsforlag, 2000

Bick, Eckhard, "En Constraint Grammar Parser for Dansk", in Peter Widell & Mette Kunøe (eds.): *8. Møde om Udforskningen af Dansk Sprog, 12.-13. oktober 2000,* Århus Universitet, 2001

Appendix 1: A progressive level parser - modules of DanGram



LEKSISK ANALYSE

"DANMORF"

PREPROCESSOR
word boundaries, orthography, composite, names, abbreviations

MORPHOLOGICAL ANALYZER
produces cohorts of alternative word-readings:
lexeme-identification, inflexion, derivation, proper noun heuristics

POSTPROCESSOR
morphological heuristics, lexeme based valency tags and semantic class potential

MORPHOLOGICAL DISAMBIGUATION
iterative context based CG disambiguation rules, based on:
word class, word form, base form, valency potential, semantic prototypes

TAGGER

"DANTAG"

SYNTACTIC MAPPING
adds lists of possible syntactic function tags / constituent markers (at word or subclause level) to words, based on disambiguated morphology and context

SYNTACTIC DISAMBIGUATION
iterative context based CG disambiguation rules, handles:
argument structure, dependency relations at group and clause level
subclause form and function

PARSER

"DANSYN"

PROPRIUM-CG
recognition of name types using contextual CG rules

CASE ROLE-CG
context based mapping and disambiguation of semantic case roles
(Søren Harder)

PSG
generation of syntactic tree structures, using generative rewriting rules

spell and grammar checking
clause boundaries
teaching software
language games
corpus annotation

"DANTRAD"

MT-CG
context based mapping of translation equivalents Danish -> Esperanto

Appendix 2: The DanGram system in current numbers

Lexemes in morphological base lexicon: 146.342
(equals about 1.000.000 full forms), of these:
proper names: 44839 (experimental)
polylexicals: 460 (+ names and certain number expressions)
Lexemes in the valency and semantic prototype lexicon: 95.308
Lexemes in the bilingual lexicon (Danish-esperanto): 36.001

Danish CG-rules, in all: 6.233
morphological CG disambiguation rules: 2.678
syntactic mapping-rules: 1.701
syntactic CG disambiguation rules: 1.854
(plus 429 bilingual rules in separate MT grammars, and a smaller number of semantic case-role and proper name-rules in the semantics and name grammars)

Danish PSG-rules: 490 (for generating syntactic tree structures)

Performance:

At full disambiguation (i.e., maximal precision), the system has an average correctness of 99% for word class (PoS), and about 96% for syntactic tags (depending, on how fine grained an annotation scheme is used)

Speed:
full CG-parse: ca. 400 words/sec for larger texts (start up time 3-6 sec)
morphological analysis alone: ca. 1000 words/sec