PaNoLa: Integrating Constraint Grammar and CALL applications for Nordic languages

by

Eckhard Bick
Institute of Language and Communication, Southern Denmark University
lineb@hum.au.dk, http://visl.hum.sdu.dk

1. Project status

The PaNoLa project (Parsing Nordic Languages) has been funded by The Nordic Council of Ministers' Language Technology Research Programmefor a 2-year period (2002-2003), and aimed at integrating on the one hand Nordic Constraint Grammar research, and on the other internet based grammar teaching based on such research. Participants were the University of Southern Denmark (Eckhard Bick, project leader, and John Dienhart), Oslo University (Janne Bondi Johannessen and Kristin Hagen), Helsinki University (Fred Karlsson) and Göteborg University (Torbjörn Lager). Though described here in terms of a 2-year status, it must be born in mind that the research and development activities addressed by PaNoLa are not limited to this period, and have both a historic foundation and a future continuity at the institutions mentioned above. Also, at least two of the participating CG-research groups (Danish and Norwegian) were and are involved in other Language Technology Programme supported joint efforts with different, but overlapping time lines (Nomen Nescio and the Treebank Network), thus guaranteeing a certain synergy and continuity.

Where not referenced in the bibliography, technical and comparative details mentioned below originate from personal project communications, mainly from presentations and discussions at the four PaNoLa workshops held in Odense, Ustaoset, Reykjavik and Göteborg.

2. Constraint Grammar

CG systems are rule and lexicon based robust tools for grammatical analysis of running text (Karlsson et. al. 1995), and share basic conceptual and notational conventions, if not their language dependent rules as such - allowing productive research transfer. However, differences arise in terms of (a) rule type and rule compilers, (b) lexical base and (c) levels or focus of analysis.

2.1. Rule formalism and architecture

Thus, Lingsoft's Swedish and Finnish CGs (http://www.lingsoft.fi) and the Norwegian Oslo-Bergen-Tagger (Johannessen et. al. 2000) use the cg1 compiler, while Danish DanGram uses cg2 and the compatible vislcg[1], that differ in allowing set rules and, in the latter case, substitution rules. A very different alternative, researched by Torbjörn Lager during the PaNoLa project and implemented as the μ-TBL system, is to automatically learn and order CG-rules from tagged corpora, rather than write them by hand. These different formal approaches obviously have an influence on the grammars as such: For instance, cg1 grammars will need more rules than set-targeting cg2 grammars, vislcg grammars will be more effective at correcting probabilistic input in hybrid systems, and while automatically learned grammars are cost effective and will optimize interplay between their rules better than your average human linguist, they cannot accomodate for long-context rules, linked rules and certain other complexities for reasons of mathematical complexity. As part of the PaNoLa exchange there have been Norwegian experiments with vislcg and μ-TBL.

2.2. Lexical base

While all non-probabilistic PaNoLa systems use computational lexica, there are differences in technique and the information type encoded. Lingsoft's CGs use highly efficient, but hard-to-maintain TWOL-systems (two-level-morphology) with special sublexica, while the Danish and Norwegian systems use a core lexicon and morphological analyzers for inflexion and composition. μ-TBL's "lexicon" varies, of course, with the training corpus used.

---

[1] The cg1 and cg2 compilers were developed by Pasi Tapanainen, The Oslo-Bergen-Tagger is developed in Lisp by Paul Meurer (The University of Bergen) and vislcg is developed by Martin Carlsen (VISL). The latter is open source under GNU.

Valency potential and semantic information can be integrated into CG-grammars in two different ways: Either sets of lexemes sharing a given feature or class are defined in the grammar itself (so they can be referenced by context conditions in the rule body), or the information is entered directly into the lexicon, to be used as a secondary tag by CG-rules. The latter solution is typically chosen for features common to many lexemes, while the former is used for smaller sets or experimental categories with a few core lexemes. Valency features are used by all non-probabilistic PaNoLa-systems for verbs, while they are more experimental for other word classes. Semantic information was introduced into DanGram during a 3-year project on semantic CGs (1999-2001) specifying, for instance, 200 semantic types for the 65.000 word noun lexicon, as well as certain selection restrictions for verbs, which have been useful for rule writing in the current project. Otherwise, only the Norwegian CG-group has started on semantic work, translating the Danish SIMPLE lexicon for their purposes.

2.3. Levels and focus of analysis

The common level of analysis for all PaNoLa systems is word class/morphological tagging and disambiguation. On top of this, the linguist-written systems have syntactic mapping and disambiguation grammars of varying complexity. Danish CG, for instance, has added attachment direction markers, close attachment markers and the form and function of subclauses, as for all other CG-parsers at VISL[2]. These features also facilitate syntactic tree-generation with add-on PSG-rules, a module developed earlier to generate live analyses of VISL teaching sentences and now used to compile treebanks (cp. the Arboretum-article in the Treebank network section of this volume). The treebank perspective is now shared by the Norwegian team, which has started to use VISL's open source PSG-compiler[3], and is itself working on the XML-adaptation of existing tree editing and search tools (Redwood, TIGER-search)[4] to VISL-format input.

Other focus areas of CG-modules are the Norwegian CG grammar checker and, for Danish, Case role-CG and full (numbered) dependency CG, both developed by a VISL Ph.D. student, Søren Harder, during the PaNoLa period. For both languages, CG-modules for named entity recognition have recently been developed (cp. Bick 2003-2 and Johannessen 2003).

---

[2] The 5 languages are Portuguese (PALAVRAS), Danish (DanGram), French (FrAG), English and German.

[3] The compiler was programmed by Martin Carlsen to VISL specifications.

[4] Adaptation work by Lars Nygaard, Redwood by Stephan Oepen.

3. VISL teaching system

PaNoLa's other leg has been the integration and strengthening of Nordic languages in SDU's VISL teaching system. This system offers grammar teaching tools for 22 languages on the internet (http://visl.sdu.dk), using a uniform system of grammatical categories and structural analysis, as well as color codes and symbolic notation, with a systematic focus on grammatical form and function (Dienhart 2000 and Bick 2001). Since 1996, VISL has developed interactive tools for the inspection and construction of syntactic trees, as well as a number of grammar games, like the Paintbox and Shooting Gallery games for teaching word classes, the Balloon Ride morphology game, or the PostOffice and Syntris for teaching syntactic function.

Under PaNoLa, starting from Danish and English models, all tools and games have been made available for Bokmål, Nynorsk, Swedish and Finnish, and paedagogically structured corpora of teaching sentences have been built, using XML-markup to encode, for instance, teaching topic and didactical progression. Since Finnish and Swedish sentences were modelled on Danish and Norwegian example files, these language pairs in particular could be used for comparative grammar teaching.

	sentences	words	words pr. sentence
Danish	1121	12.029	10.7
Bokmål	766	5.629	7.3
Nynorsk	766	5.888	7.7
Swedish	106	1.153	10.9
Finnish	102	545	5.3

Graded complexity filters allow VISL-material to be used not only at universities, but also for introductory courses and in schools. Thus, several school projects[5] have been active for both Danish and Norwegian during the PaNoLa-period, with synergistic effects working both ways. Supporting teaching material is now available online or in printed form in Danish, English and Norwegian (for instance, Bick 2002), and a number of teacher training courses have been and will be held in Denmark and Norway.

[5] The publically funded projects can be mentioned: VISL-GYM and VISL-HHX for Danish, and GREI for Norwegian (cp. PaNoLa-article in NorFa's yearbook 2002)

As an alternative to preanalysed teaching corpora, the VISL-system allows live input to some of the teaching tools, - a feature only feasible, of course, if supported by a working CG-parser for the language in question. Here, too, PaNoLa aimed at integrating existing Nordic CG-expertise. Thus, a "CGI-contact" was established between the VISL-server at SDU and the Oslo-Bergen-Tagger at the Text Laboratory in Oslo, allowing the former to pipe input through the latter and use its output in VISL-interfaces and -games. For Swedish and Finnish, CGs were licensed from Lingsoft and installed at the VISL-server for similar purposes, though these modules had not yet been VISL-interfaced at the time of writing. Future work will have to address notational compatibility issues and question of adding additional layers of analysis, like subclause function and phrase structure grammar.

4. Evaluation

CG-grammars can be improved incrementally - in principle, forever, and will also profit from improved lexica or improved rule formalisms. Therefore, evaluation is less definite than for probabilistic systems, where results are meant to be optimised, for a given architecture, method and training corpus, "once and for all". Also, evaluation results will depend on the granularity of the tag set used, the level of analysis and the method of evaluation (inspection, bench mark, number of reviewers etc.). Nevertheless, repeated evaluation provides an indication of development progress, and can function as an indicator for which kind of applications a given parser will be able to support.

When last evaluated (Bick 2003), the Danish CG achieved F-Scores of 98.65 for part of speech (including verbal subcategories), and 94.9 for syntactic function categories, on mixed news text. In an NER-evaluation, with 6 types, there were ca. 5% typing errors, plus ca. 2% chunking and proper name recognition errors. Lexical F-scores for Bokmål and Nynorsk were reported as 97.2% and 96.2%, respectively (Hagen and Johannessen 2003).. The Swedish point of departure, SWECG 1.0, for comparison, claimed a recall of 99.7% at a precision of 95% (www.sics.se/humle/projects/svensk/projectPlan.html, accessed 23.12.98), but no new values were available at the time of writing. Torbjörn Lager (1999) reports 98.1% lexical accuracy for a μ-TBL CG-grammar trained on a benchmark corpus for Swedish, when allowing for 1.04 tags per word.

Another type of evaluation was performed by the Oslo group on 104 school children, measuring positive effects of VISL tools on grammatical performance. The experiment was conducted in the framework of GREI and is described in Johannessen and Hagen in this volume).

5. Comparing VISL teaching grammar across the Nordic spectrum

Though one of the strengths of the VISL notational system is its relatively unified approach towards different languages, and a fixed set of categories and abbreviations for easy computational handling, it is still possible to express structural diffences between languages as well maintain certain differences in linguistic descriptive tradition. Thus, an analogous token with the same function in two Nordic languages can be assigned two different categories, if one so wishes. For instance, verb integrated particles like "i", "på", "om" are treated in Danish as prepositions if governing a noun phrase ("i huset", "på taget"), as opposed to the adverb reading they receive as verb-incorporated particles ("klappe i", "tage på", "skrive noget om". Norwegian, on the other hand, tags the latter as prepositions, too. Likewise, articles are tagged as determiner pronouns, and what is an infinitive marker in Danish, becomes a subordination conjunction in Norwegian. To maintain compatibility, however, these differences have been marked, so output can be filtered for cross language use.

Certain such differences are easily accomodated as subcategories of common supercategories, like the valency distinction made in Danish and Finnish, but not for Norwegian and Swedish, between adjunct adverbials and (valency bound) argument adverbials (with a nexus relation to either subject or object).

A more structural difference arises from the fact that word based grammars like Constraint Grammar aren't fond of zero constituent that only introspection can provide, while Chomskyan Constituent Grammar tradition favours the concept to explain deep structures. In the PaNoLa project, the conflict became apparent, when Swedish sentences were manually tagged according to the *English* VISL model, while other PaNoLa languages, notably the ones based on automatic CG analysis, did not employ this concept in their teaching corpora.

Finally, there are, of course, purely language dependent differences, not least between the Scandinavian language group on one side and Finnish on the other. For instance, the Scandinavian sentences have a preposition token frequency of 11% (Bokmål), 11.4% (Danish), 13.4% (Nynorsk), while there were only three instances of Finnish prepositions (0.5%).

References:

Bick, Eckhard (2001-1), "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, Århus

Bick, Eckhard (2001-2), "VISL - Tværsproglig computerbaseret grammatikundervisning", i *Sproglig mangfoldighed (Gymsprogs konferenceskrift)*, pp. 39-45, GYMSPROG 2001

Bick, Eckhard (2002), *Grammy i Klostermølleskoven: VISL-lite, Tværsproglig sætningsanalyse for begyndere.* Mnemo: Århus.Bick, Eckhard (2003-1), A CG & PSG Hybrid Approach to Automatic Corpus Annotation, in Kiril Simow & Petya Osenova: *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster), pp. 1-12

Bick, Eckhard (2003-2), Named Entity Recognition for Danish, in *Nordisk Sprogteknologi, Årbog 2002*. pp. 331 -350. Museum Tusculanums forlag, Københavns Universitet.

Dienhart, John (2000), "VISL-projektet: Om anvendelse af IT i sprogundervisning og -forskning" i Simon Heilesen (ed.), *At undervise med IKT*, Narayana Press: Gylling

Hagen, Kristin and Johannessen, Janne Bondi (2003), Parsing Nordic Languages (PaNoLa) – norsk versjon, in *Nordisk Sprogteknologi, Årbog 2002*. pp. 89 - 96. Museum Tusculanums forlag, Københavns Universitet.

Johannessen, Janne Bondi (2003), Nomen Nescio – Nettverk for en automatisk navnegjenkjenner for norsk, svensk og dansk, in *Nordisk Sprogteknologi, Årbog 2002*. pp. 327 -330. Museum Tusculanums forlag, Københavns Universitet.

Johannessen, Janne Bondi & Kristin Hagen & Anders Nøklestad (2000). A Constrain-based Tagger for Norwegian. In: Lindberg, Carl-Erik og Steffen Nordahl Lund (red.): *17th Scandinavian Conference of Linguistics. Odense Working Papers in Language and Commnication 19*, pp. 31-48, SDU, Odense

Lager, Torbjörn (1999), The μ-TBL System: Logic Programming Tools for Transformation-Based Learning, Paper presented at the *Third International Workshop on Computational Natural Language Learning* (CoNLL'99), Bergen

Karlsson, Fred & Atro Voutilainen & Juka Heikkilä & Arto Anttila (1995). *Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text*, Mouton de Gruyter: Berlin