Parsing and evaluating the French Europarl corpus

Eckhard Bick
Institute of Language and Communication
University of Southern Denmark
eckhard.bick@mail.dk

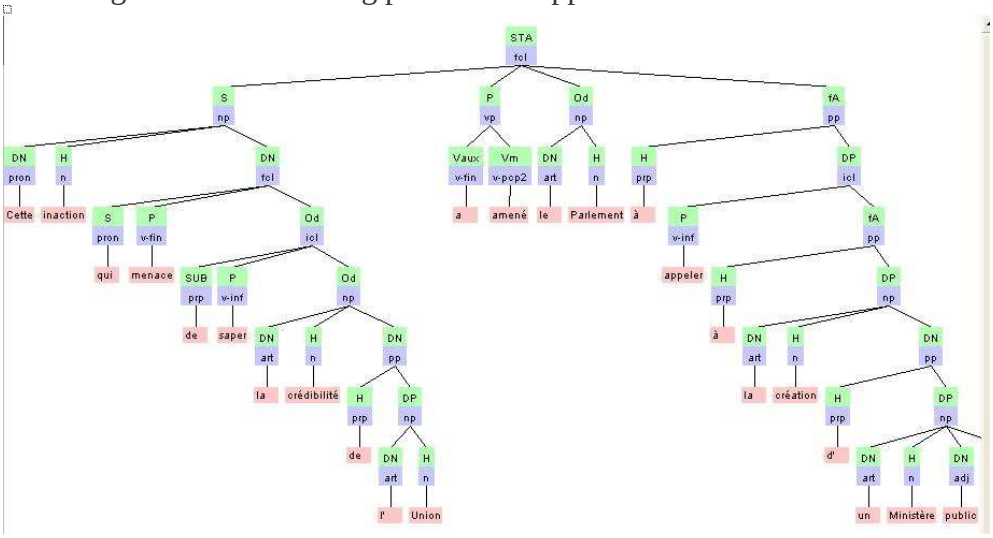
1. Introduction

This paper describes tools and preliminary results from an automatic corpus annotation project. The hybrid Constraint Grammar parser FrAG (Bick 2003 and http://beta.visl.sdu.dk) was run on the French version of the parallel Europarl corpus (http://www.isi.edu/~koehn/europarl/), consisting of 29 million words of original and translated debate transcripts from the European Parliament, and linguistic features were statistically evaluated across the different source languages (SL). The project is part of a larger initiative to create freely accessible and grammatically annotated corpora for French, the FReeBank project (Salmon-Alt et.al. 2004).

2. The parser

The core of the FrAG system *(French Annotation Grammar)* is a sentence scope Constraint Grammar (CG), with linguist-written rules. However, unlike traditional CG, the system uses hybrid techniques on both its morphological input side and its syntactic output side. Thus, FrAG draws on a pre-existing probabilistic Decision Tree Tagger (DTT, Schmid 1994) before and in parallel with its own lexical stage, and feeds its output into a Phrase Structure Grammar (PSG) that uses CG syntactic function tags rather than ordinary terminals in its rewriting rules. The CG stage itself addresses 3 main tasks successively, using context, valency patterns (for 6.000 verbs as well as some nouns and adjectives) and semantic prototype (for 14.000 nouns, about 30% of the lexicon):

1. Correcting the DTT input: Normally, a CG receives input from a fully lexicon based morphological analyser, and handles all disambiguation in a rule based way. With unambiguous probabilistic input, however, additional morphological readings have to be added and evaluated, and PoS errors corrected with special replacement rules.
2. Syntactic analysis: FrAG has about 1200 syntactic rules that map and disambiguate word based tags for syntactic function, in combination with shallow dependency markers.
3. Attachment markers: In order to reduce overgeneration and time-space complexity, a special CG-stage adds tags for close and long postnominal pp-attachment and coordination matches.



The last, PSG-based, level of analysis, adds structural depth, while profiting from the classical robustness of the underlying CG-system. A tree-chooser program, using a structure and tag based weighting system, is activated if more than one well-formed tree is found for a given sentence. At

the treebank level, FrAG's inventory of grammatical categories follows a cross-language standardisation scheme (http://beta.visl.sdu.dk/visl2/cafeteria.html) used for teaching treebanks in 22 languages at the University of Southern Denmark. GUI tools and format filters are available for end-users, among them TIGER-treebank XML and PENN-treebank bracketing formats.

In order to measure tagging accuracy with relation to the corpus at hand, a chunk of 1.790 words from the tagged Europarl corpus was automatically analysed in a small pilot study and manually evaluated at the CG-level with the following results:

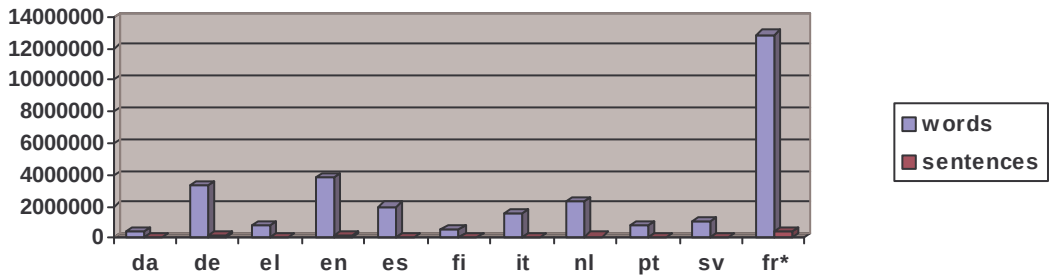
	Recall	**Precision**	**F-score**
Part of speech[1]	98.7 %	98.7 %	98.7
Syntactic function[2]	93.7 %	92.5 %	93.1

Table 1: DTT+CG Performance

On running text, with uncorrected input, the PSG-stage produces around 40% complete PSG trees for entire sentences, though of course the vast majority of individual noun phrases or subclauses may well be correctly chunked even in trees with incomplete global analyses. However, no real evaluation of structural accuracy at the tree level has been performed so far.

3. The corpus

The Europarl corpus consists of parallel language sections of 20-30 million words each. The author is currently parsing several of these subcorpora, making the annotated versions searchable through a user-friendly interface on the internet (http://corp.hum.sdu.dk), but only the French part will be discussed here. The text language (TL) distribution is as follows:



At the time of writing, the complete 29 million word corpus has been analysed at the syntactic CG-level, while about half has also been processed in syntactic tree format[3]. Though it must be borne in mind that the annotation was unrevised, error rates indicate that general statistical conclusion are still valid. Thus, with an ideal, category-neutral, error-distribution, a PoS category with a frequency of 10%, given its share of 1.3% PoS-errors, would have an error margen between 9.87% and 10.13%. Even in the unlikely event that *all* errors were related to this one category, the saftety margen would still be 8.7-11.3%. Also, since languages are not compared directly, but in French translation, parsing errors can optimally be regarded as "noise" affecting all SL's equally, and thus not reducing the significance of *relative* differences across languages.

From a linguistic perspective, comparing languages in the Europarl corpus has the disadvantage of translations not being "natural language" in lexical and structural terms. On the other had, using French as a kind of neutralizing filter has the advantage of allowing a more direct

[1] Separately counting tenses, participles, infinitive.

[2] Including subclause function, but without making a distinction between free and valency bound adverbials.

[3] While a complete CG-run of the Europarl corpus takes only 3 hours on a centrino based linux machine, it takes about 100 times as long to run a Phrase Structure Grammar on the result, at least with VISL's present, not yet optimised, cg2tree compiler – even though it processes function tags, not word forms as terminals.

comparison independent of, e.g., traditional differences in word class definition, or even – to a certain degree – of differences in the use of tense and mood. Note that French itself was marked xx/fr in the data, since the transcripts cover both native speakers, and non-natives adapting to the international parliament setting. The French numbers may thus be atypical, and the same might be true of English (though most non-native chosers of English probably come from neighbouring Germanic languages.

Table 2 compares the frequencies (in % of words) for a number of form and function categories across all 11 source languages (SL). GER is the average for germanic SL-speakers, ROM the average of Romance SL-speakers with the exception of French.

	da	sv	de	en	nl	GER	xx/fr	es	it	pt	ROM	fi	el
words per sentence	25.5	25.1	25.3	25.7	**23.1**	**24.9**	**27.8**	32.1	32.9	**33.2**	**32.7**	25.3	31.0
finite subclauses	3.81	3.75	3.47	3.47	3.30	3.56	**3.16**	**4.04**	3.68	3.52	3.75	**3.00**	3.72
relative clauses	1.95	2.05	1.68	1.70	1.58	**1.79**	**1.72**	**2.16**	2.10	2.07	**2.11**	**1.50**	2.09
direct object clauses	**1.11**	1.04	1.02	1.03	0.95	**1.03**	**0.85**	1.10	0.90	0.81	**0.94**	**0.78**	0.94
adverbial clauses	0.63	0.54	0.67	0.61	0.63	0.62	**0.52**	**0.70**	0.63	0.55	0.63	0.57	0.62
participial adverbial subclauses (log-5)	2.92	**2.15**	3.20	4.35	4.52	**3.43**	**3.96**	3.82	4.09	4.71	**4.21**	3.31	**4.78**
auxiliary chain parts	**3.46**	3.35	3.34	3.36	3.13	**3.33**	2.89	2.98	2.99	2.52	**2.83**	3.02	**2.77**
passive pcp2	**0.47**	0.45	0.42	0.45	0.44	0.45	0.41	**0.33**	0.34	0.39	0.35	0.44	0.39
active pcp2	1.17	1.14	1.15	**1.33**	1.07	1.17	1.12	1.22	1.20	**0.95**	1.12	1.04	1.17
infinitive	**1.43**	1.38	1.39	1.21	1.25	1.33	**0.99**	1.12	1.11	**0.93**	1.05	1.20	0.89
subjunctive/vfin	4.99	**5.58**	4.76	4.53	4.40	4.85	**4.19**	4.76	4.26	4.79	4.60	5.55	4.35
conditional	0.56	0.56	0.56	**0.62**	0.43	**0.55**	0.43	0.49	0.43	**0.40**	**0.44**	0.56	0.39
vocative	**0.04**	**0.04**	0.06	0.05	0.06	0.05	0.05	0.06	**0.07**	**0.04**	0.06	0.05	0.05
attributive	**6.70**	6.98	7.02	7.01	7.29	**7.00**	**7.26**	7.37	7.64	**8.13**	**7.71**	7.65	7.62
common nouns	**20.90**	21.26	21.00	21.33	21.35	21.2	22.07	21.37	21.09	22.14	21.5	**22.66**	21.71
finite verbs	**8.94**	8.59	8.48	8.29	8.49	**8.56**	7.57	8.18	7.78	**7.23**	**7.73**	7.83	7.86
coordinating conjunction	2.67	2.48	2.80	2.68	2.56	**2.64**	**2.74**	3.20	**3.16**	3.28	**3.21**	**2.40**	3.20
subordinating conjunct.	2.33	2.16	2.22	2.17	2.13	2.20	**1.84**	**2.35**	2.01	1.87	2.08	1.88	2.06
demonstrative	**1.96**	2.14	**2.34**	2.17	2.24	2.17	**1.99**	2.17	1.98	2.02	2.06	1.82	1.81

Table 2: Category distribution in the French Europarl corpus

The table shows that sentence length is an obvious dividing parameter between Germanic SL (24.9 w/s) and Romance SL (32.7 w/s) languages, with Finnish joining the Germanic camp, and Greek the Romance one. There is a pattern to subclause distribution, too, with Romance winning on relative clauses (lack of genitives?) and Germanic on direct object clauses. However, 2 individual languages spike across this pattern: Spanish has a record-high in subclauses in general, and Finnish scrapes bottom (prepositional constructions?). It is also clear that Germanic languages compensate for a relative lack of verb inflection features with (tense, mood) with an increased use of modal auxiliaries with infinitives, strikingly visible even through the "translation filter". Auxiliary passive constructions are also a trace of Germanic, being a means of creating distance and neutral statements, which does not disappear in translation, since the passive in modern Romance has to be auxiliary rather than inflexional. Both language groups have conditional forms, either inflexional or complex, but Germanic seems to use it more (possibly due to the fact that subjunctives are less of an option), and this usage is apparingly retained in translation. That the subjunctive incidence as such is *higher* for Germanic SL, too (4.85 vs. 4.60), appears odd at first glance, but can probably be explained by translational forces (subclause subjunctive etc.) – the originals need not have had subjunctives at all.

In terms of word class, Romance uses more adjectives and adjectival participles (7.71%) than Germanic (7.00%), while Germanic has a higher incidence of finite verbs (8.56% vs. 7.73). Among the structurally important word classe, coordinating conjunctions are most unbalanced, with Romance winning out with a 25% higher usage (3.21% vs. Germanic 2.64% and Finnish 2.40%).

Some languages appear to be more equal than others. Thus, Danish and Spanish seem to be, in certain ways, extreme exponents of their respective language families, especially in the area of subclause types and verb chains. In a league of its own is Parliament French, the only language not translated, but on the other hand influenced by non-native speakers whose mother tongue syntactic and stylistic preferences may shine through even clearer than would be the case with a professional translation filter. In this vein, it is not clear if the short "Germanicoid" sentences of xx-French and an incidence of subclauses that is lower than *both* the Romance *and* Germanic average, can be explained by French being the most Germanic of Romance languages, or indeed by the fact that people speaking in a foreign tongue tend to limit the complexity of their utterances. Even in absolute terms, across all languages, the xx-data show record lows for complexity indicators like conjunctions and subjunctive ratio.

Two features were chosen to evaluate a more pragmatic angle, too. Thus, syntactic vocative – though a difficult function to ascertain automatically – was quantified to measure "deference" og "politeness". For what it's worth, Italian SL'ers ranked highest, and Scandinavian ones lowest, with a difference ratio of almost 2:1. Demonstratives were chosen to measure deixis or "discourse immediacy", but with a difference of only 15% between the highest ranking SL'ers (German) and lowest (Danish), no significant conclusions can be drawn.

4. Corpus access

The annotated Europarl corpus is freely accessible for internet based searches at http://corp.hum.sdu.dk, in both CG and treebank formats (L'Arboratoire). The former, word based format easily lends itself to fast and flexible CQP[4]-based search structures, while the latter, through the intermediate PENN-treebank-format, allows Tgrep2[5]-based tree-searches (Fig. 3). For the CG-versions of FrAG-annotated corpora, a special menu-based search interface (Fig. 2) has been built targeting "non-technical" users with a linguistic interest only.

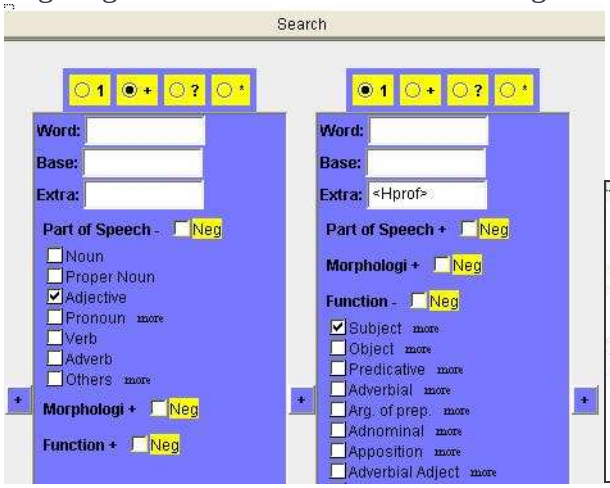


Fig.2: menu-based corpus searches

Fig. 3: Tgrep2-treebank-search: non-finite object clause as daughter of relative clause

#446 ID=ep-fr-00-01-17-sp59-xx-p3-3 Nous espérons que ceux **qui envisagent de voter contre ces amendements** pourront donner de bonnes raisons de leur choix au Parlement et aux citoyens qui sont à la recherche d'un emploi. #A1/2
#589 ID=ep-fr-00-01-17-sp69-xx-p2 Nous proposons en particulier un impôt unifié sur le capital, **qui permettrait d'alimenter des fonds utilisés pour accompagner l'harmonisation des systèmes de protection sociale et la réduction du temps de travail au niveau européen**. #A6/8
#762 ID=ep-fr-00-01-17-sp83-es-p1 Monsieur le Président, le groupe du parti des socialistes européens de ce Parlement est d'accord avec le rapport que vient d'exposer M. Berend et félicite son auteur, tant pour la qualité de ses conclusions que pour sa flexibilité, **qui a permis d'introduire en commission des amendements provenant des différents groupes**. #A2/2

References

Bick, Eckhard. 2003. "A CG & PSG Hybrid Approach to Automatic Corpus Annotation". In: Kiril Simow & Petya Osenova: *Proceedings of SProLaC2003*, pp. 1-12. Corpus Linguistics 2003, Lancaster

Salmon-Alt, Susanne & Eckhard Bick & Laurent Romary & Jean-Marie Pierrel. 2004. "La FReeBank: Vers une base libre de corpus annotés". In: *Proceedings of TALN2004.* Fes, Marocco. (forthcoming)

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees". In *Proceedings of the International Conference on New Methods in Language Processing*, September 1994. Manchester,UK.

[4] The Corpus Query Processor (CQP) has been developed at the Institut für Maschinelle Sprachverarbeitung (Universität Stuttgart). For references on the IMS Corpus Workbench, cp. http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/.

[5] The Tgrep2 tool was programmed by Douglas L. T. Rohde, for references cp. http://tedlab.mit.edu/~dr/Tgrep2/