

Live use of Corpus data and Corpus annotation tools in CALL: Some new developments in VISL

by

Eckhard Bick

1. Introduction

The VISL-project, also mentioned in this and earlier yearbooks in connection with PaNoLa, Nomen Nescio and the Nordic Treebank Network, is a 2-thronged cross-language teaching and research project, with a strong emphasis on Natural Language Parsing, corpus linguistics and internet based grammar teaching. Advocating a unified system of grammatical analysis (Dienhart 2000 and Bick 2002) across different languages and different teaching levels, VISL has always striven to integrate its NLP-tools with its CALL-applications (Bick 1997, 2004). This article, spawned by a couple of talks at NorFa seminars¹, first presents some recent developments in this area, the exercise building tool KillerFiller and the grammatical text-evaluator TextPainter, then discusses the live use of corpora in grammatical language awareness teaching.

2. KillerFiller: Turning corpora into slot-filler exercises

One of the most well-known types of CALL-applications are slot-filler exercises, used in almost all language teaching CD ROM systems, and renowned for their easy evaluability. Since in most cases, a simple automatic template comparison is sufficient for result grading, slot-filler exercises have also been a backbone of

¹ Nordisk Sprogteknologiseminar 2004, Copenhagen, July 17-18, and the CALL-conference at CBS, Copenhagen, september 30th - oktober 1st ("IT og sprogindl ring: Fremgangsm der og praktiske applikationer").

placement tests in mixed level language courses. However, though supported by free software like HotPotatoes (also earlier used in VISL), the individual exercises usually are hand-made one by one for a given course or purpose, and statistical evaluation is handled by the individual teacher.

A new VISL tool, christened KillerFiller, draws on existing annotated corpora to create slot-filler exercises ad hoc, and offers statistical evaluation of student performance in a systematic way. By registering the user, and maintaining a constant performance log, KillerFiller addresses at the same time (a) the researcher's need for evaluation (VISL is fun, but does it make a difference?) and (b) the teachers need for feed-back: Is a given student progressing in a given area of grammatical training or not? Can class-wide tendencies or problems be observed?

Please login to your VISL-game account
 If you do not have an account, create a new one by clicking [here!](#)

Username
 Password

Sentence collection
 Word class

Which language do you want to train?

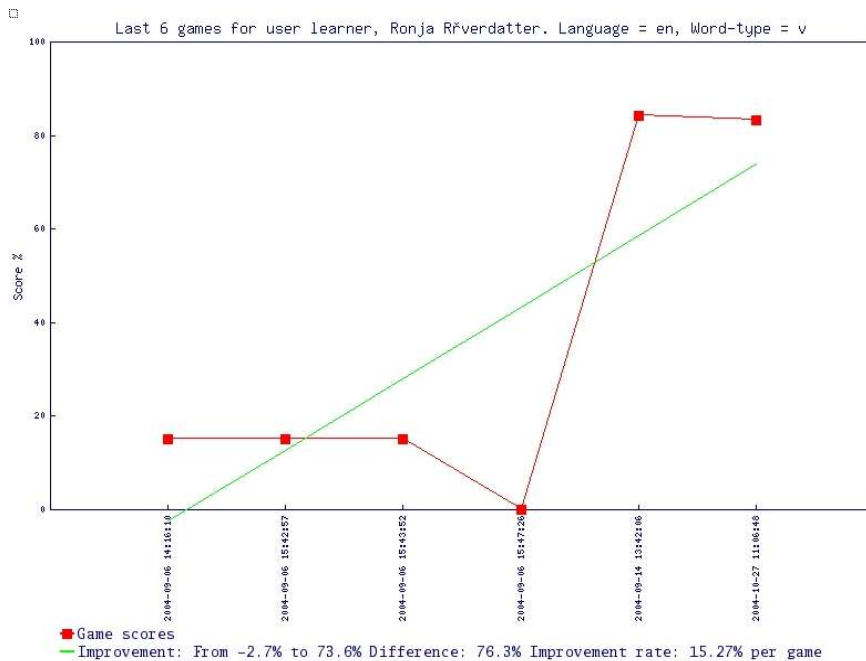


The tool itself is basically an IT-based pandemonium of slot-filler exercises, drawing random sentences with grammatical annotation not only from VISL's teaching corpora (22 languages), but also from its research data bases (7 languages, Bick 2003). The program replaces a given category of words (e.g. all prepositions or verbs) with blanks, that the user has to fill back in.

Kasparow zu (besiegen) (müssen -pr-) für den
 Computer ein Genuß (sein) (sein)

Depending on the language, for some word classes grammatical categories or base forms will be provided to avoid ambiguities regarding, for instance, tense and number. So far, only manually revised data have been used, but given the low PoS error rate of mature Constrained Grammar systems, slot-filler exercises for teacher-provided live texts could be envisioned for the future, at least for VISL's major NLP languages.

Based on login ID's and passwords, a server-side database stores the dates and results from every run, sorted by language, user and exercise type. After each run (defined as 10 sentences), improvement statistics and graphs are shown, and teachers and evaluators can access historical overview pages for relevant sections of the stored data. Teachers can create language- or level-specific groups, and students are registered as members of one or more groups. Thus, the system not only allows to grade the individual user, but also to quantify, say, the average improvement (or even improvement *rate*) of one's 10th grade French class after a VISL based grammar course. Below, an early test case example for English verbs, plotting scores from 6 different dates.



3. TextPainter: Grammatical evaluation of user texts

Many grammatical CALL exercises focus on topic or feature at a time, such as a word class, an inflectional problem or relative clauses, and if a teacher can't find the topic he is looking for, he is not usually allowed to adapt and change an existing exercise. On the other hand, full-analysis exercises like VISL's tree builder module may be too complex and not focused enough for a certain teaching stage or purpose. The problem is especially relevant in an "awareness building" phase, where a teacher wants students to work out for themselves what the characteristics, distribution and rules of usage are for a given grammatical feature. Here, corpus linguistics offers promising tools and easy access to flexible examples (cp. later chapters). In a new approach, we have tried to integrate corpus annotation, text grading and grammatical exercises, allowing topic-/feature-specific mark-up of user texts annotated "on the fly".

Thus, the *TextPainter* offers live analysis of cut-and-paste text in 7 languages. Results can be highlighted for a given category or category combination, say *objects*, *subjects*, *verbs* or *adjectives*. Thus, a sample text can be used to grade a novel as a verb-heavy action text or as an adjective heavy descriptive text.

The screenshot shows the TextPainter web interface. At the top, there is a language selection bar with radio buttons for Danish, English, Esperanto, French, German, Portuguese, and Spanish. Below this, there are two dropdown menus for selecting categories: 'subjects' (with options: direct/accusative objects, adverbials (free or bound), indirect/dative objects) and 'nouns' (with options: proper nouns, adjectives, adverbs). Between these menus are radio buttons for 'OR' and 'AND'. To the right of the 'nouns' menu is a text input field labeled 'or insert category label:'. Below the category selection is a text input field labeled 'Enter text to parse:' containing the text: 'Text Painter er et redskab til at analysere tekst på mange sprog. Resultaterne kan blive markeret mht. subjekter,'. Below the text input are 'Go!' and 'Reset' buttons. At the bottom, there are two dropdown menus: 'Parser:' set to 'Standard Parser' and 'Visualization:' set to 'Selected category highlight'.

For reasons of robustness, and in order to keep error rates as low as possible, all parsing is performed with Constraint Grammar parsers (http://beta.visl.sdu.dk/visl2/constraint_grammar.html), and all information is marked on words. Complex syntactic functions will thus be marked on constituent heads, in dependency grammar style. Subclause markers, like the relative clauses in the example, will be carried by the first verb in the clause:

Categories: @CL-N< ... OR... NONE

Den del af planloven, der **begrænser** nye dagligvarebutikker til højst 3.000 kvm og reelt **forhindrer** et Bilka i Horsens, skal væk og erstattes af noget mere fleksibelt og tidssvarende. ¶ Det mener både kommuner og eksperter, der **betegner** loven som alt=for restriktiv og firkantet og **peger** på, at maksimumsgrænsen gør det umuligt at gennemføre en fornuftig planlægning for detailhandelen

In interactive mode, users have to *find*, say, all objects themselves. Feedback is given in the shape of red and green beavers, for wrong and correct answers, respectively, and performance is evaluated in terms of an integrated recall/precision measure, the F-score.

check!

Den del af planloven, der begrænser nye dagligvarebutikker 
 til højst 3.000 kvm og reelt forhindrer et Bilka  i
 Horsens, skal væk og erstattes af noget mere fleksibelt
 og tidssvarende. ¶ Det mener både kommuner  og
eksperter , der betegner loven  som alt=for restriktiv og
firkantet og peger på, at maksimumsgrænsen gør det umuligt
 at gennemføre en fornuftig planlægning  for detailhandelen

All in all, there were 60 words in the text.

For the category/categories in question, you found 4 out of 5 possible words, and had 2 false positives. This equals a recall of 80 % and a precision of 66.66 %, combining into an **F-score of 72.72 %**. To compare your own with the computer's opinion, check the highlights below!

Den del af planloven, der begrænser nye **dagligvarebutikker** til højst 3.000 kvm og reelt forhindrer et **Bilka** i Horsens, skal væk og erstattes af noget mere fleksibelt og tidssvarende. ¶ **Det** mener både kommuner og eksperter, der betegner **loven** som alt=for restriktiv og firkantet og peger på, at maksimumsgrænsen gør det umuligt at gennemføre en fornuftig **planlægning** for detailhandelen

4. Language awareness: Examples of corpus based exercises

Language awareness has become an important key-word in Danish high school level language teaching, and has been declared instrumental in a new cross-language teaching subject, *Almen Sprogforståelse*:

En nyskabelse i gymnasireformen er et samarbejde mellem de almene gymnasiers sprogfag i forløbet Almen Sprogforståelse. Som en del af den nye gymnasieuddannelses grundforløb skal Almen Sprogforståelse styrke elevernes teoretiske sprogforståelse, samspillet mellem sprogene og studiekompetencen ... Forløbet Almen Sprogforståelse vil indgå i grundforløbet for nye elever på det almene gymnasium fra sommeren 2005.

Formålet er at vække og styrke elevernes sproglige viden, bevidsthed og opmærksomhed ... (Undervisningsministeriet - Nyhedsbrev nr. 9 - 2004)

To a much higher degree than traditional language teaching, these new objectives call for an empirical method and true authenticity of language data - in other words, a school-accessible corpus search interface. Thus, VISL's corpus site (<http://corp.hum.sdu.dk>, illustration below) is currently part of the curriculum in a number of teacher training courses (<http://beta.visl.sdu.dk/visl2/urkas.html>). The use of a unified cross-language annotation system is an additional asset with regard to the Ministry's secondary objective ("*samspillet mellem sprogene*"). In the following, a number of didactic examples will be discussed of how VISL's corpus tools can be made to serve these ends.

□



4.1 Animal metaphors

The distinction between literal and metaphorically extended usage of a word is important both from a point of view of stylistic variation and as a potential stumbling stone in translation exercises. Acknowledging the prominence of animal words in this area, students could be asked to examine to which degree animal names occur with non-literal meanings in newspaper text. For instance, based on VISL's part of speech and semantic class annotation, a search for adjectives accompanying animal names can be formulated². *Adjective*, being a standard closed class category, can be chosen from a search menu, while *animal*, or here the subclass of *earth_animal* <Azo>, has to be typed in:

² Internally, the search engine uses the Corpus Query Processor (CQP), developed at the Institut für Maschinelle Sprachenverarbeitung, Stuttgart (Christ 2004, <http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench/>)

The image shows two side-by-side screenshots of a linguistic analysis tool. Both panels have a top navigation bar with buttons labeled '1', '+', '?', and '*'. The left panel has input fields for 'Word:', 'Base:', and 'Extra:'. Below these is a 'Part of Speech' section with a 'Neg' checkbox. The list includes: Noun, Proper Noun, Adjective (checked), Pronoun, Verb, Adverb, and Others. Below the list are 'Morphologi +' and 'Function +' sections, each with a 'Neg' checkbox. The right panel has the same input fields, but 'Extra:' contains the text 'Azo'. In its 'Part of Speech' section, 'Morphologi +' is checked. The 'Function +' section also has a 'Neg' checkbox.

About half of the results from a Korpus90/2000 search will, in fact, be non-literal:

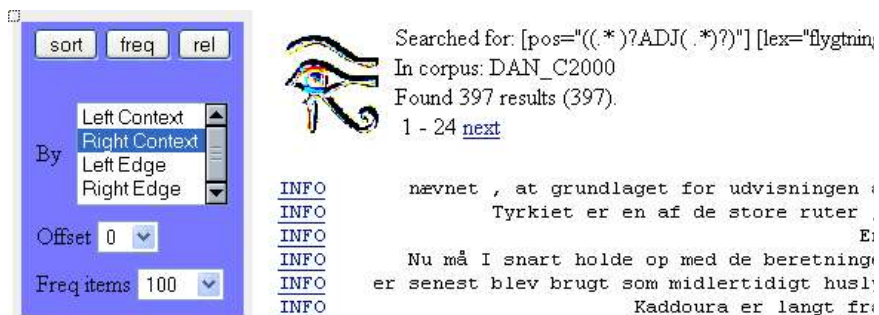
appellerer til den **politiske ræv**, fordi det kan spilles uden altid at tænke i strategi
 en ekstra jurypris med **tilhørende sølvbjørn** til den tyskfødte instruktør...
 ikke kan få den **russiske bjørn** til at gungre med
 Gys, gru og store **stygge ulv**.
 styrken ved at være i kontakt med den **indre abe** er i_hvert_fald ...
 og prøve at følge med de **unge løver**, der vil køre stærkt.
 Kun de allerbedste er i_stand til at tæmme den **olympiske vildhest** 49er.
 de **unge skakløvers** forslag blev fulgt af flertallet
 til forveksling kan ligne dem fra de mest ideologiske **unge løver** i Venstre

In a next step, the animal terms extracted could be compared with their equivalents in other languages, comparing "metaphoricity" as such, and - possibly - variations in which adjectives associate with which animal. For the German ECI-corpus, for instance, VISL's corpus site reveals a similar density, and also a similar core spectrum of metaphorical meaning (below), suggesting - in a quantitatively quite unscientific, yet discussion-worthy and language-awareness-raising way - that Danish might have certain animal metaphors in common with German. However, though transparent, not all of these metaphors translate as easily as the *clever fox* and the *strong bear*. Thus, the Danish *young lions*, and the German *party bear* (2nd example) are more of a challenge to the translating student's language awareness.

Der **Berliner Bär** lächelte nur einen Tag
 Ansonsten war dort der **sprichwörtliche Bär** los: Jubel, Trubel, Heiterkeit

Damit hat sich der **schlaue Fuchs** das Monopol der Regierung ...
 Gies, der **alte Fuchs**, steigt in den AGF-Verwaltungsrat auf
 vor der Oberhauswahl berief der **finanzpolitische Fuchs** für Freitagabende eine ...
 den Herkules aus Pennsylvania beendete der **schwarzhaarige Bär** aus Teheran
 auch die türkischen Neofaschisten um den **grauen Wolf** Alparslan Türkeş, ...
 Sollte sich einmal mehr Döring als **cleverer Fuchs** erwiesen und ...
 schliesslich muss er eigentlich den **einsamen Wolf** mimen ...

While the animal-metaphor exercise invites the student himself to interactively surf the corpus interface, it is also possible to construct ready-made language awareness tests from corpus data, as in the following example, where a teacher has extracted attributive left context for three near-synonyms, *indvandrere*, *udlænding*, *flygtning*, and the point of the exercise would be to guess which sets of attributes goes with which head word. In order to extract the lists of adjectives, the lemma-feature was used, lumping, for instance, *udlænding*, *udlændinge*, *udlændingen*, *udlændingene* etc. into one search, and the resulting concordance was frequency sorted using *relative* frequencies, i.e. frequencies *in context* divided by *lexical* frequencies (in the language as such).



The screenshot shows a corpus search interface. On the left is a control panel with buttons for 'sort', 'freq', and 'rel'. Below these are dropdown menus for 'By' (Left Context, Right Context, Left Edge, Right Edge) and 'Offset' (0). At the bottom of the panel is a 'Freq items' dropdown set to 100. On the right, there is a search query: 'Searched for: [pos="((*)?ADJ(.*)?")][lex="flygtning"]'. Below the query, it says 'In corpus: DAN_C2000' and 'Found 397 results (397)'. A link '1 - 24 next' is visible. The search results are displayed in a list with 'INFO' labels on the left and text snippets on the right. The snippets include words like 'nævnet', 'Tyrkiet', 'Nu må I snart holde op med de beretninger', 'er senest blev brugt som midlertidigt husly', and 'Kaddoura er langt fra'.

Search for: [pos="((*)?ADJ(.*)?")][lex="flygtning"]
 In corpus: DAN_C2000
 Found 397 results (397).
[next](#)

... , at grundlaget for udvisningen af en **iransk flygtning** for 16 måneder s
 Tyrkiet er en af de store ruter , som **illegale flygtninge** fra Asien bru
 En del **kosovo-albanske flygtninge** i Danm
 I snart holde op med de beretninger om **stakkels flygtninge** , som kommer
 t blev brugt som midlertidigt husly for **bosniske flygtninge** .
 Kaddoura er langt fra den **eneste flygtning** eller indvandrer

In a comparative table, it becomes clear that Danish texts subclassify fugitives according to nationality, immigrants according to (il)legality and age, and foreigners according to their crime and education record. In the table, *freq* is the local percentage (out of all hits), i.e. directly calculable from *num* (instances), while *rel* is the square of *freq*, divided by the tokens lexical frequency.

indvandrer/e/ne	udlænding/e/ne	flygtning/e/ne
DAN_C2000 (315)	DAN_C2000 (97)	DAN_C2000 (397)
<i>frequencies:</i> <i>rel freq num</i>	<i>frequencies:</i> <i>rel freq num</i>	<i>frequencies:</i> <i>rel freq num</i>
illegale 4941110 18.7 [59]	narkodømte 9565309 3 [3]	somaliske 2052303 7.3 [29]
unge 238271 29.2 [92]	velkvalificerede 1366472 3 [3]	bosniske 1705722 8.5 [34]
nordafrikanske 113378 0.9 [3]	herboende 781479 5.1 [5]	kosovo-albanske 1015170 1 [4]
illegal 100781 1.5 [5]	kriminelle 571063 12.3 [12]	rwandiske 518159 1.7 [7]
højtuddannede 41228 0.9 [3]	uønskede 298333 4.1 [4]	palæstinensiske 407880 7.5 [30]
kriminelle 30459 2.8 [9]	voksne 12961 3 [3]	hjemvendte 333937 2.5 [10]
uønskede 28289 1.2 [4]	sådanne 12753 3 [3]	Kosovo-albanske 285516 0.7 [3]
muslimske 11944 1.2 [4]	unge 10687 6.1 [6]	illegale 258260 4.2 [17]
arbejdsløse 10586 1.5 [5]	dollarstærke 10628.12 1 [1]	albanske 238863 4 [16]
tyrkiske 9655 1.2 [4]	narkodømt 10628.12 1 [1]	kosovaalbanske 142758 0.7 [3]
ikke-franske 8062.48 0.6 [2]	Tossedet 10628.12 1 [1]	tibetanske 79310 1.2 [5]
Veluddannede 8062.48 0.6 [2]	danskfødt 10628.12 1 [1]	nyankomne 78090 1 [4]

Instead of a sociolinguistic focus, similar corpus based correlation exercises can be constructed in order to teach the differences in usage for word pairs like *høj* - *stor* (high - big) or *stærk* - *kraftig* (strong - ?). In a Danish course for foreigners, the data may simply be used to craft yes/no multiple choice exercises, but at a higher level of education, one might address category awareness instead, nudging students towards an understanding of *selection restrictions* and *semantic categories*. Thus, the example data suggest that *høj* is used for measuring along a cline of discrete units, while *stor* is more descriptive than measuring, applying to objects and non-countable abstracta. Absolute frequencies for Korpus2000 are given in parentheses:

høj/t ... *grad* (434), *kvalitet* (148), *niveau* (141), *pris* (57), *prioritet* (47), *tempo* (42), *fart* (40), *indhold* (36), *humør* (35), *alder* (34), *kurs* (33), *hastighed* (31), *klasse* (28), *arbejdsløshed* (24)

stor/t ... *del* (1214), *betydning* (432), *succes* (297), *forskel* (267), *antal* (245), *interesse* (205), *vægt* (201), *problem* (196), *flertal* (156), *indflydelse* (154), *gruppe* (142), *rolle* (141), *glæde* (141)

In the second example pair, animates (humans, organisations, animals) prototypically seem to ask for *stærk*, while acts and events ask for *kraftig*. Note the metaphorical human-hood of *position*, *ønske* and *økonomi*.

stærk/t ... hold (28), pres (22), mand (21), vilje (15), position (15), ønske (15), vækst (15), kontrast (14), modstander (14), leder (14), økonomi (13)

kraftig/t ...stigning (31), vækst (30), kritik (22), jordskælv (13), afstand (9), advarsel (9), pres (8), opfordring (8), forbedring (8), fald (8), slag (7), mistanke (7)

The word *vækst*, occurring in both lists, shows that *relative* frequencies can be useful. In fact, in relative terms, in a search for ADJ + *vækst*, corpus data show *kraftig* on rank 2, while *stærk* is down on rank 11.

5. Propedeutic corpus uses: Preparing for an educated comma

Of late, there has been considerable interest in putting VISL's CALL tools to indirect uses - either chaining existing tools and games with additional texts and link structures into actual course materials, or simply using them eclectically to ascertain and train grammatical categories and analytical skills necessary for a more complex, primary teaching task to be addressed later. Thus, knowledge of certain "delimiter" word classes (conjunctions, relatives, interrogatives), as well as analytical skills concerning subject-predicator structure, could be made instrumental in preparing the ground for the teaching of Danish punctuation rules³, in this case the use of the clause-delimiting comma (Dansk Sprognavn 2004). Such "propedeutic" knowledge could be made explicit in the following ways:

- extract corpus-sentences with the word "at", pasting them into TextPainter and have students' choose between conjunction-'at' (with comma) and infinitive marker-'at' (without comma) by clicking (only) on the latter - red beavers meaning the 'at' in question should have been a conjunction.
- Use KillerFiller to familiarise students with the word class of subordinating conjunctions.
- compare examples of *subject - "ikke" - verb* sequences with *subject - verb - "ikke"* examples. The latter constitutes a main clause test in Danish, while the former characterises a subclause (with an obligatory comma in the end, and an optional comma in the beginning). A elaborating language awareness task would be to find

³ The example is inspired by didactic needs expressed by participants at a VISL-course currently run for teacher training colleges (<http://beta.visl.sdu.dk/visl2/vislsem.html>) .

out *which other adverbs* are allowed between subject and verb (predicator). Use the menu based CorpusEye interface described above.

- learn to distinguish the different meanings of "som" - comparative "prepositional conjunction" (*stærk som en bjørn*), comparative "subordinating conjunction" (*som jeg har sagt tidligere*) and relative pronoun (*som ikke kan vente*). While the last two types ask for clausal comma, the first one does not. Find examples from VISL's sentence collections ("closed corpora").
- use a treebank corpus (<http://beta.visl.sdu.dk/visl2/treebanks.html>) to extract (a) adverbial subclauses (fA:fcl)⁴ and (b) relative subclauses (DN:fcl), since the former always have clausal comma, while the latter are "comma-ambiguous" - they can occur without comma when parenthetical.

The last two items address fairly complex structural issues and would profit from graphical visualisation of corpus search results. In fact, both VISL's pedagogical teaching corpora (representing most Nordic languages, e.g. <http://visl.sdu.dk/da/>) and the large research treebanks like the Danish Arboretum (http://corp.hum.sdu.dk/tgrepeye_da.html) allow first to find all sentences with a given feature ("som" or "DN:fcl"), then to link to graphical representations of individual examples, and finally to interactively inspect (and rebuilt) the structures in question:

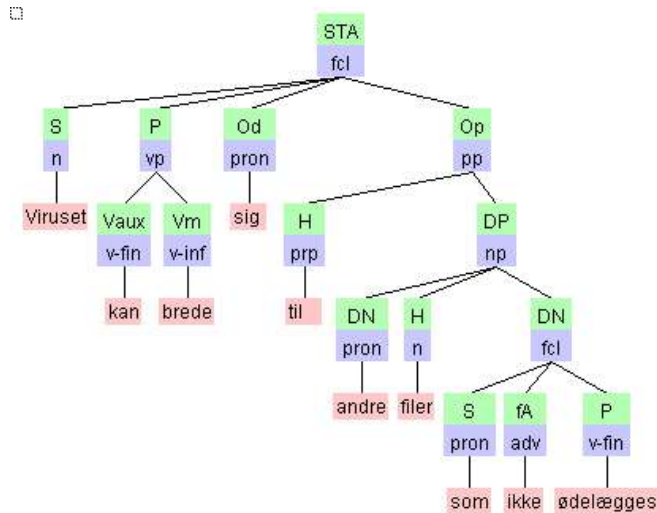
□

search results for 'DN+fcl << /som/' in *citat2000.t2c* :

For graphical tree inspection, click on ID-code

#10 ID=[adrvbmig](#) Japansk virksomhed overtager alle rettigheder til det industrielle spildevandsanlæg, **som EnvoTech i Sønderborg har udviklet**. #A1/2
#18 ID=[aphzklum](#) Naturligvis vil han blive spurgt om, hvordan han vil øremærke den danske støtte til ganske bestemte formål, når den nu engang skal kanaliseres gennem et enevældigt regimes hænder og derefter - måske - videre til politi og dommere, **som er dele af et effektivt kinesisk undertrykkelsessystem**. #A12/12
#24 ID=[agsshzw](#) "Men ud af ca. 500 møbelproducenter er der fortsat kun omkring 50, **som har en omsætning på over 100 mio. kr.**" #A1/1
#28 ID=[aovcphzg](#) Tivoli præsenterer "den originale amerikanske version", **som ingen til pressemødet kunne oplyse, hvornår havde haft premiere i USA**. #A1/3
#42 ID=[agrbqac](#) Russernes udbredte jødehad rammer netop en mand som Beresovskij, **som er rig, snedig og svær at stole på**. #A1/2

⁴ The search string examples are written in VISL's form & function convention, with an upper case syntactic function symbol separated from a lower case form symbol by a colon. Of course, a tree bank search can also make use of word strings, regular expressions and tgrep2 relations between constituents (mother, daughter, sister etc., cp <http://tedlab.mit.edu/~dr/Tgrep2/>).



References

- Bick, Eckhard (1997). "Internet Based Grammar Teaching", in: Christoffersen, Ellen & Music, Bradley (eds.), *Datalogvistisk Forenings Årsmøde 1997 i Kolding, Proceedings*, pp. 86-106. Kolding: Institut for Erhvervsprog og Sproglig Informatik, Handelshøjskole Syd.
- Bick, Eckhard (2002). *Grammy i Klostermølleskoven - "VISL light": Tværsproglig sætningsanalyse for begyndere*. Århus: Mnemo.
- Bick, Eckhard (2003). "A CG & PSG Hybrid Approach to Automatic Corpus Annotation", In: Kiril Simow & Petya Osenova (eds.), *Proceedings of SProLaC2003* (at Corpus Linguistics 2003, Lancaster), pp. 1-12.
- Bick, Eckhard (2004). "Grammatik for shov: IT-baseret grammatik-læring med VISL". In: Peter Juel Henriksen (ed.), *Call for the Nordic Languages: Tools and methods* (Proceedings of NorFa CALL Net Symposium Sept. 30. - Oct. 1. 2004), forthcoming
- Christ, Oli (1994). "A modular and flexible architecture for an integrated corpus query system". COMPLEX'94, Budapest.
- Dansk Sprognævn (2004). "Kommaregler". Copenhagen: Dansk Sprognævn, pp. 17-30
- Dienhart, John (2000). "VISL-projektet: Om anvendelse af IT i sprogundervisning og -forskning". In: *At undervise med IKT*, pp. 51-70. Gylling: Narayana Press.