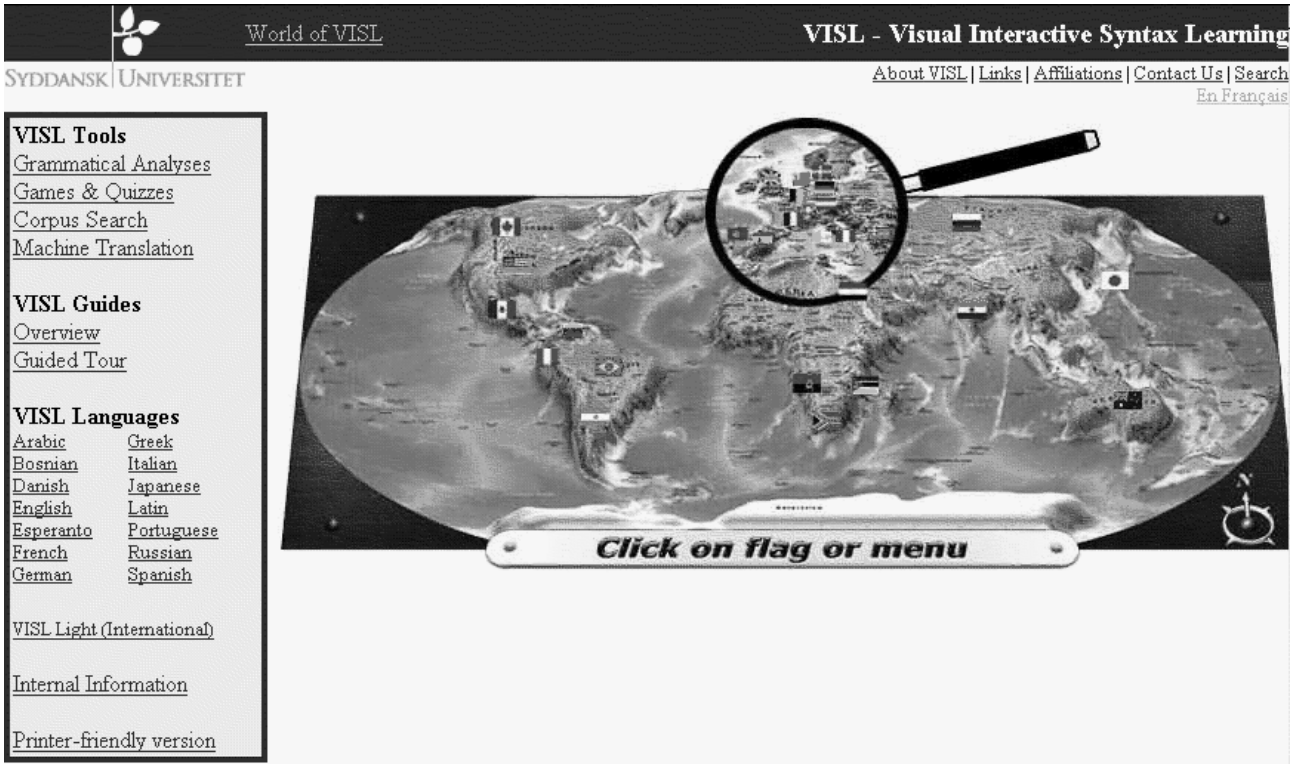
The VISL System:
Research and applicative aspects of IT-based learing

Eckhard Bick
e-mail: lineb@hum.au.dk
web: http://visl.sdu.dk

1. Abstract

*The paper presents an integrated inter active user interface for teaching grammatical analysis through the Internet medium (**V**isual **I**nteractive **S**yntax **L**earning), developed at Southern Denmark University, covering 14 different languages , half of which are supported by live grammatical analysis of running text. For reasons of robustness, efficiency and correctness, the system's internal tools are based on the Constraint Grammar formalism (Karlsson, 1990, 1995), but users are free to choose from a variety of notational filters, supporting different descriptional paradigms , with a current teaching focus on syntactic tree structures and the form-function dichotomy. The original kernel of programs was built around a multi -level parser for Portuguese (Bick, 1996, 2000) developed in a dissertation framework at Århus University and used as a point of departure for similar systems in other languages . Over the past 5 years, VISL has grown from a teaching initiative into a full blown research and development project with a wide range of secondary projects, activities and language technology products. Examples of application oriented research are NLP-based teaching games, machine translation and grammatical spell checking. The VISL group has repeatedly attracted outside funding for the development of grammar teaching tools, semantics based Constraint Grammars and the construction of annotated corpora.*



1. Background

When the VISL project started in 1996, its primary goal was to further the integration of IT tools and IT based communication routines into the university language teaching milieu at Odense University (Denmark), and more specifically, to develop tools for Visual Interactive Syntax Learning. The initiative was funded jointly by CTU (Center for Teknologi-Støttet Uddannelse) and Odense University for 3 years, and the languages involved were English, German, French and Portuguese.

Already in the early stages of the project it became clear that a distinction would have to be made as to whether the language data to be used in the teaching interface would be limited text book examples or unlimited natural language text. We decided to develop both a "closed" and an "open" system, and to design the teaching applications for maximal synergy, such that they would be able to take input from both the closed and open language data sources, - and do so in a largely language independent way.

For the closed system a notational formalism was developed that allowed the textual expression of graphical syntactic tree structures, and data bases of manually analysed sentences were built for all participating languages, the original target being 500 text book sentences and 500 running text sentences. With the help of enthusiastic students and teachers, these "closed corpus" data bases are constantly being enlarged, and today VISL covers 14 languages, among them the basic Romance and Germanic languages as well as a number of more exotic specimen, like Arabic, Japanese and Esperanto.

The open system is research based and centered around the Constraint Grammar paradigm, introduced by Fred Karlsson at Helsinki University in the early 1990ies (Karlsson, 1991, 1995). The 1996 role model for the syntactic VISL system was my Portuguese CG parsing system (Bick 1996, 2000), which featured a full dependency analysis of subclause structure and a prototype CG-to-tree syntax transformation grammar. I have since developed similar CG based systems for Danish, Spanish and Esperanto. For English and German, VISL has corrected and amplified licensed commercial CG systems from the Finnish software firm Lingsoft.

2. A unified approach to grammar

The central principle of VISL's language analysis is its focus on surface structure (expressed as either dependency relations or syntactic tree structures) and the form-function dichotomy. Following Bache et.al. (1993, 1999), function symbols start with upper case letters, form symbols with lower case letters, and both are combined in a combined colon-separated symbol (text) or function-over-form symbol (graphics). For the dependency notation, international CG conventions are followed, with upper case letters for all primary tags, using the @-symbol to introduce function tags, and arrow heads (>,<) for head oriented dependency markers.

VISL light vertical tree (non-graphical notation)	**VISL vertical tree** (non-graphical notation)
UTT:cl(fcl) S:prop VISL P:v(v-pr) er Cs:g(np) =D:art et =H:n forskningsprojekt =D:cl(fcl) ==S:pron(pron-rel) der ==P:v(v-pr) involverer ==Od:g(np) ===D:pron(pron-indef) mange ===D:adj forskellige ===H:n sprog	STA:fcl S:prop VISL P:v-fin(v-pr) er Cs:np =DN:art et =H:n forskningsprojekt =DN:fcl ==S:pron-rel der ==P:v-fin(v-pr) involverer ==Od:np ===DN:pron-indef mange ===DN:adj forskellige ===H:n sprog

VISL	[VISL]	<heur> <*>	**PROP** NOM	@SUBJ>
er	[være]	<vk>	**V** PR AKT	@FMV
et	[en]		**ART** NEU S IDF	@>N
forskningsprojekt	[forskningsprojekt]		**N** NEU S IDF NOM	@<SC
,				
der	[der]	<rel>	**INDP** nG nN NOM	@SUBJ>
involverer	[involvere]	<vt>	**V** PR AKT &MV	@FS-N<
mange	[mange]	<quant>	**DET** nG P NOM	@>N
forskellige	[forskellig]		**ADJ** nG P nD NOM	@>N
sprog	[sprog]		**N** NEU P IDF NOM	@<ACC
.				

Meeting regularly over 4 years, the VISL group of university teachers has invested considerable effort in discussing the compatibilities, incompatibilities and blind spots of different national and linguistic grammar traditions, and agreed on a common superset of symbols. Recently, a reduced symbol set for propedeutic use and schools, "VISL light", was agreed upon, and the Danish X-and-O-system adapted to match the function categories used in VISL light. At the lowe st level, 11 word classes and 14 primary functions are used.

- ● Predicator (P), Verbal (V)
  - ○ Auxiliary (Vaux), also as <> (D)
  - ✪ Main verb (V*, Vm), also as ★ (H, K)
  - ⊙ Verb chain particle (Vp), also as <> (D), simplified as ⩛(A)
  - ➲ Infinitive marker (Vi, INFM), also as <> (D)
- ✘ Subject (S)
  - (✘) Formal or provisional subject (Sf), possibly with the subclass of situative subject (Ss)
- ▲ Direct (accusative) object (Od)
  - (▲) Formal or provisional object (Of)
- ■ Indirect (dative) object (Oi)
- ◆ Prepositional object (Op), emneled, evt. forenklet som ⩛(A)
- ⊗ Subject complement (Cs), Subject predicative (Ps)
  - [⊗] Free subject predicative (fPs, fCs), simplified as ⩛(A)

⊕		Object complement (Co), Object predicative (Po)
	[⊕]	Free object predicative (fPo, fCo), simplified as ⩛(A)
⩛		Adverbial (A), with possible subdivision of free (⩛ fA) or bound (⩓ bA, bAs, bAo)
★		Head (H), Kernel (K)
<>		Dependents (D)
↘		Subordinator (SUB)
↔		Co-ordinator (CO)
#		Conjunct (CJT)
	«»	Underspecified constituent at clause level (e.g. clause body)

3. Internet based teaching tools

One lesson to be learned from the VISL project, is that it is not at all easy to introduce IT-based tools into an existing teaching environment. Ap art from hardware problems (there never being enough - compatible and updated - machines in the right room at the right time), there is the very central problem of psychological resistance against the new medium, simply because it may feel too "technical". All things technical have a very low acceptance rate in the Humanities, and teachers often resent the personal investment in time and effort necessary to acquire the necessary skills - not to mention changes in teaching material and exams. There is, of co urse, a fundamental difference in terms of "technicality" between a human teacher and a computer terminal, - the latter lacks the teacher's *naturalness*, *interactivity*, *flexibility* and *tutoring* capacities. On the other hand, computers do have evident teachin g advantages - they can integrate the senses, making use of colours, pictures and sounds in a more flexible and impressive manner than paper can. Also, a computer program can "know" more - in terms of facts and examples, and within a well -defined subject matter - than a human teacher. And last, but not least, a computer system, especially if accessible through the internet, can teach an unlimited number of students at the same time in what optimally still amounts to an individual manner.

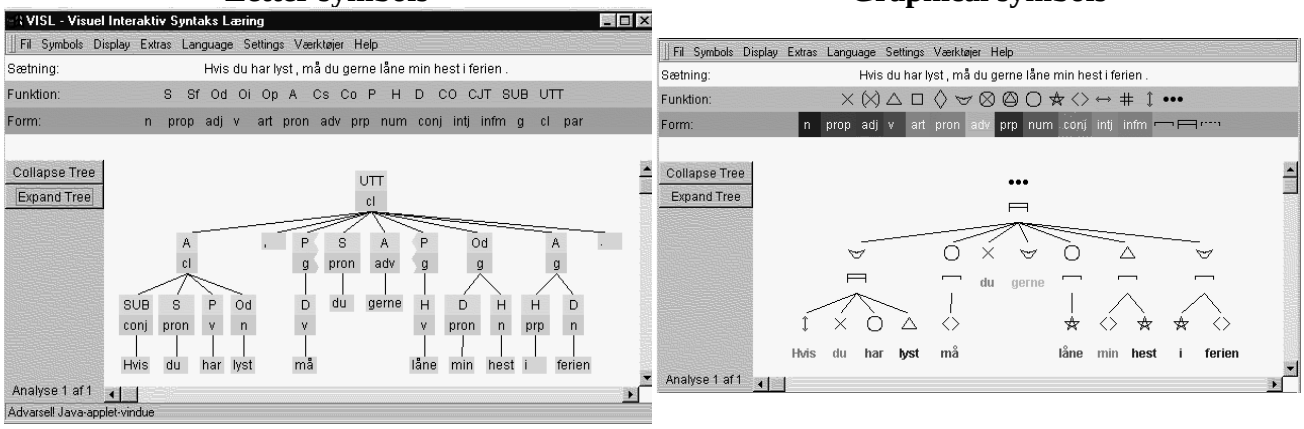
Given these advant ages, it makes sense to invest some effort in addressing the four main disadvantages, as listed above. The VISL grammar teaching interface tries to make advances with regard to the following four principles:

(i) Flexibility

The VISL interface is notationally f lexible, i.e. th e user can choose between several notational conventions (e.g. flat dependency grammar, enriched text, meta text notation, tree structures), and move back and forth between different levels of complexity. For instance, depending on the exer cise chosen, the type and number of grammatical categories used (e.g. word classes) may be changed. In order to make work more colourful, it is also possible to move between text book material, copied "live" texts, randomized test sentences and one's own creative idiolect.

In the tree structure example below, the user can switch back and forth between letter symbols and graphical symbols, more than dou ble the number of categories, or reduce the tree to a pure function tree (green only).

Letter symbols **Graphical symbols**

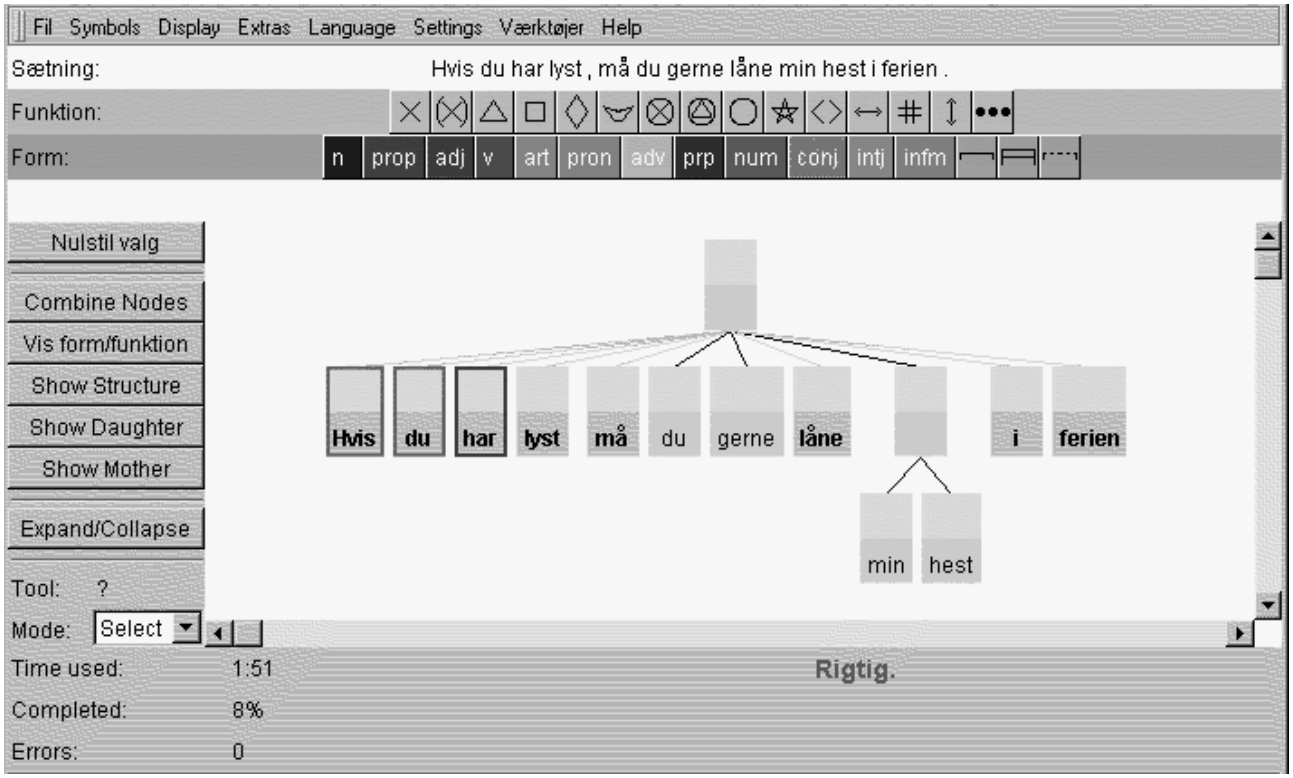


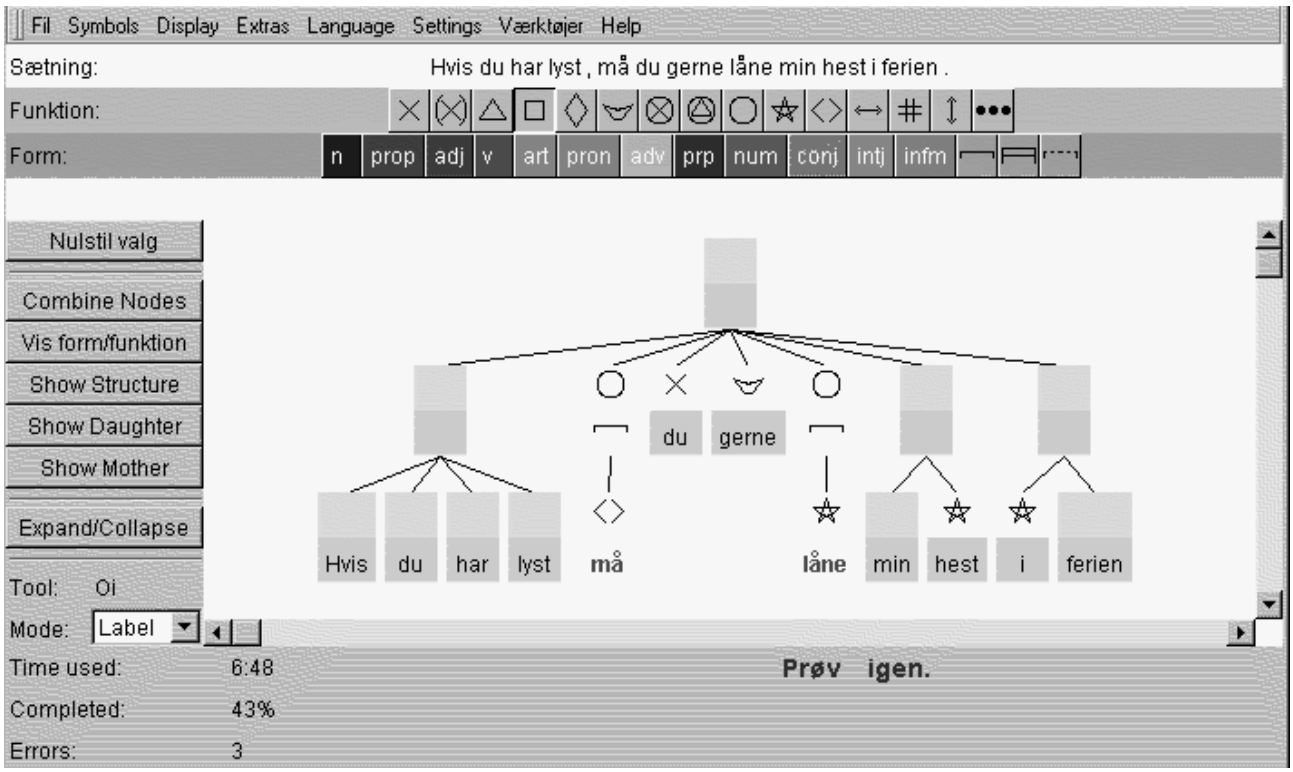
VISL's unique integration of teaching and research tools would even allow the user to experiment with different kinds of subjects or add a couple of place and time adverbials and rerun the sentence in free-text mode – with exactly the same graphical setup and paedagogical functionality.

(ii) Interactivity

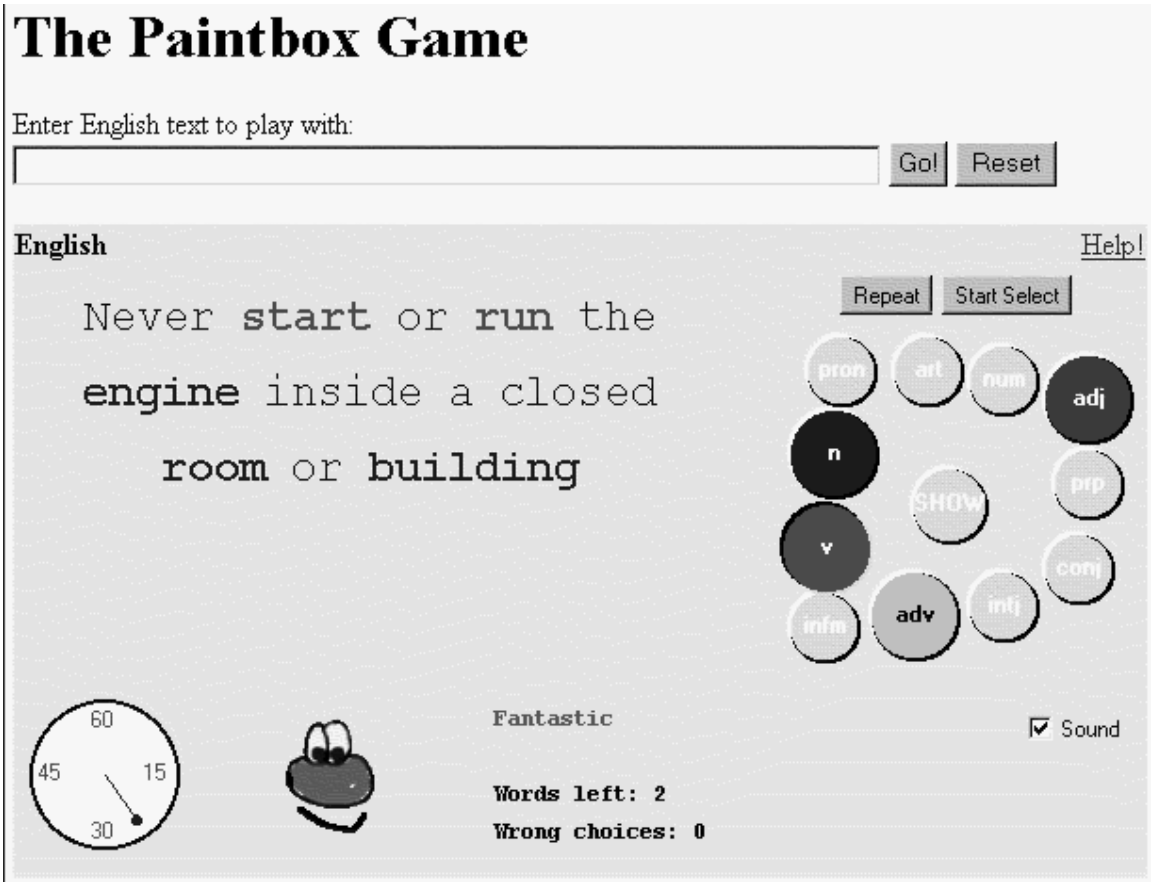
VISL's java -tree interface for grammatical analysis allows the step -by-step interactive inspection, construction and labelling of syntactic trees using menus, mouse clicks and drag -and-drop movements, all known from b asic text processor functionality.

In the first example below, a student has recognized the np *"min hest"*, but has yet to assemble *"lyst"* onto the predicator *("har")* of the adverbial subclause to the left.





When a sentence proves problematic or incomprehensible, the user can modify it, or ask for the computer 's opinion (show -me option). In grammar games like *Paintbox*, *Post Office* or *Shoot-the-Verb,* interactivity inegrates a c ertain element of competition, and is further enhanced by sound effects, timers and high-scores.



(iii) Naturalness

A major draw back of most language teaching software (or, for that matter, language analysis software) is that they do not run on free, na tural language, but only on a small set of predefined sentences or structures ("toy lexica" or "toy grammars"), that cannot be modified or replaced. In the VISL interface, for better or worse, the underlying lexica and grammars cover the whole language, supporting gradual and comparative changes in a given sentence, or confronting the user with the stimulating lexical freshness and structural unpredictability of running natural text.

The second aspect of naturalness concerns, as mentioned above, "untech nical" ergonomics, and as much keyboard -interaction as possible has therefore been replaced by graphical and mouse governed tools, like menu choices and help windows. Being internet based, the system automatically takes advantage of a browser's navigation tools, scroll bars, page memory and cut'n'paste functionality.

(iv) Tutoring

Tutoring is traditionally a human task, and difficult to simulate in a computer interface. Therefore, it has been one of the last features to be broadly implemented on the VISL site. A certain minimum of tutoring can be achieved simply by providing guided tours, help windows, clickable definitions of grammatical terms, show -me-buttons, and ready access to topic conditioned corpus examples (throug h VISL's corpus search site). Howev er, real tutoring asks for more specific and individual comments. Therefore, to help students with the tree -building and -labelling task, we have implemented so -called error -comment files , where pedagogic remarks (and suggested reading -links) are stored for all common and some rarer combinations of "correct label expected" and "wrong label chosen", as well as for different types of wrong attachment (phrase and clause grouping).

4. A methodological research paradigm

An important difference between the VISL approach and traditional schools of grammar is the fact that what unifies VISL's different strands of research is not primarily a descriptional or interpretative paradigm, but a methodological one. Constraint Grammar with its focus on corpus data , lexicography, disambiguation and word based tagging is simply a very robust method, yielding low error rates and information-rich output easy to handle and filter with relatively simple text based computer programs. In descriptional and applicative terms , Constraint Grammar is more a tool grammar than a target grammar. Thus, at the teaching level, VISL uses different represantations of the same grammatical information, for instance graphical trees with form -function nodes, word class colouring or head bas ed function indexing, and a number of different corpus annotation and corpus search schemes have been supported in collaboration with outside research partners.

Constraint Grammar can be thought of as a hierarchically organized progressive level system of lexical data bases and grammars, dynamically adaptable to different tasks and different levels or angles of grammatical description. In the

table below , a hierarchy of "pure" and "applicational" modules are shown for the present VISL languages, half of which incorporate CG modules at different levels.

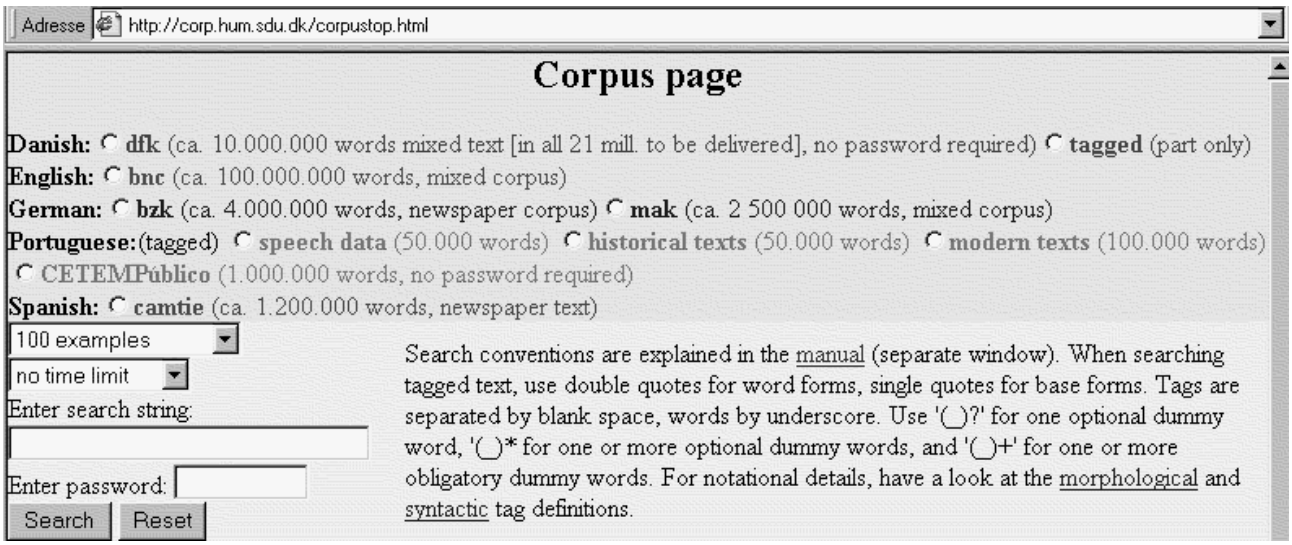
Modules **Languages**	Po	En	Da	Sp	Ge	Es	Fr	It	Ar	Ja	Gr	Ru	La	Bo
Morphological parsing lexica	+	(+)	+	+	(+)	(+)	*	*						
Valency lexica	+	(+)	+	+										
Semantic lexica	+		+											
Morphological CG	+	x	+	≈+	x	≈	*+	*						
Syntactic CG	+	x+	+	≈+		≈	≈							
CG-to-tree PSG or equivalent	+	+	+	+			?							
Polysemi CG (partial)	?													
Bilingual electronic lexica (into TL)	Da En	Da	Es En											
Machine translation to TL or from SL: (with translation mapping CG)	Da		Es Po			Da								
Spelling/grammar checker CG		?	?											
CG-to-tree compatible teaching corpora	+	+	+	+	+		+	+	+	+	+	+	+	+
CG tagged corpora	+	+	+											
CG based tree corpora	+													

+	VISL-built module
(+)	Lexicon as part of a closed CG system, licensed form Lingsoft, Helsinki
x	Closed CG, licensed from Lingsoft, Helsinki
x+	Closed commercial CG with VISL add-ons (correction module, subclause function etc.)
≈	"Cloned" from the Portuguese PALAVRAS system
*	Probabilistic Tree Decision Tagger (Helmut Schmid & Achim Stein, Stuttgart)
*+	Probabilistic Tagger with correction CG
?	Partial pilot project

5. Spin-off results

Transcending its original target area, internet based grammar teaching tools, VISL has generated a number of collateral spin-off results both technological and linguistic. Thus, a number of comprehensive bilingual lexica, valency -lexica and semantic prototype lexica are under development for several languages, and GNU -licence compilers for CG and PSG are being made available to the public. VISL's corpus site offers a search interface handling regular expressions and CG tags, and text corpora are accessible in both raw and tagged for m for VISL's core languages. S eparate sub-projects are the construction of a large freely accessi ble Danish corpus (now 10 million words, in cooperation with DSL, Denmark) and a 2 million word tree bank for Portuguese (in cooperation with the AC/DC-project, Oslo).

VISL's corpus material is partly integrated into the main site, part ly accessible through a separate search interface (http://corp.hum.sdu.dk), which allows the use of regular expressions for running text, and the combination and chaining of word forms, base forms, word class, inflexion and syntactic tags for CG-tagged text.



The table gives an overview of VISL-products within different core areas:

	Teaching	Corpus and general linguistics	Constraint Grammar
Programs	**Java-trees:** Interactive inspection, construction and labelling og syntactic trees **Paintbox:** Word class colouring game **Post office:** Syntactic function stamping game **Shooting gallery:** Selection of grammatical categories in moving sentences	**Search engine** for raw text and CG-tagged corpora **Filters** for a number of different notational conventions	flexible **CG-compiler** for Constraint Grammars **PSG-compiler** for CG-to-tree-grammars
Linguistic data	**Text book sentences:** Hand-analysed or machine analysed and proof-read "closed corpora" for 14 languages	Collection of **raw-text corpora** for 6 languages **CG-tagged corpora** for En, Po, Da New **Danish free corpus** Portuguese **treebank**	English **benchmark** text Port. **benchmark** text
Grammars	**Unified approach** to grammatical analysis and common category inventory across languages Danish **X-and-O**-symbols	Corpus driven grammar development **PSG-grammars** for CG-to-tree-conversion, for En, Da, (Po, Sp)	Port. CG (ca. 5000 rules) Dan. CG (ca. 3000 rules) Spa. CG (ca. 3000 rules) Esp. CG (port. clone) Eng. add-on CG Ger. add-on CG
Lexica	**Term bank** with definitions of grammatical categories etc. **Online dictionaries:** Po-Da, Da-Po, Da-Es, Es-Da	**Bilingual MT-lexica** for running text translation: Po-Da, Po-En, Da-En, En-Da	**Valency lexica**: Po, Da, Sp **Semantic class lexica:** Po, Da, En
Texts & documents	Online grammar **manuals**, guided tours and **tutorials** *EB: Grammy i Klostermølle-skoven (Da-En-Ge-Fr), Portuguese Syntax Manual*	Manuals, e.g. on **regular expression** in corpus searches (JMD & HK) Articles, reports and evaluations	Scientific **articles**, BA- and Ph.D.-**projects** *EB: The Parsing System "Palavras"*

Among VISL's non -teaching applications, machine translation is the most controversial one, while the tiny Danish spell-checker module is the one that even at the idea level generates most commercial interest.

A kind of "dictionary" translation service can easily be incorporated as polysemy disambiguated base form translations added onto CG tag lines, but MT proper asks for a number of additional modules, such as target language inflexion generation, syntactic transformations, instantiation of complex tenses and so on. Constraint Grammar functions here as a context sensitive mapping device fo r structural markers or special translation equivalents.

Bibliography:

Bache, Carl et. al. (1999). *English Sentence Analysis*, København: Gyldendal

Bick, Eckhard (1996). *Automatic Parsing of Portuguese.* In García, Laura Sánchez (ed.), *Anais / II Encontro para o Processamento Computacional de Português Escrito e Falado.* Curitiba: CEFET-PR.

Bick, Eckhard (1997) *Internet Based Grammar Teaching* , in: Christoffersen, Ellen & Music, Bradley (eds.), *Datalingvistisk Forenings Årsmøde 1997 i Kolding, Proceedings,* pp. 86-106. Kolding: Institut for Erhvervssprog og Sproglig Informatik, Handelshøjskole Syd

Bick, Eck hard (2000-1), *Portuguese Syntax* (Teaching Manual), http://www.portugues.mct.pt/Repositorio/Bick_Portuguese_Syntax3.doc and http://visl.sdu.dk/visl/pt

Bick, Eckhard (2000-2), *The Parsing System "Palavras" – Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework,* Aarhus: Aarhus University Press

Bick, Eckhard (2001), *Grammy i Klostermølleskoven - "VISL light": Tværsproglig sætningsanalyse for begyndere* (Teaching Manual), http://visl.sdu.dk/visl/light

Dienhart, John (2000), *VISL-projektet: Om anvendelse af IT i sprogundervisning og -forskning.* In: *At undervise med IKT,* pp. 51-70. Gylling: Narayana Press

Karlsson, Fred (1990) . *Constraint Grammar as a Framework for Parsing Running Text.* In Karlgren, Hans (ed.), *COLING-90: Papers presented to the 13th International Conference on Computational Linguistics*, Vol. 3, pp. 168-173. Helsinki: RUCL

Karlsson, Fred, et. al. (1995). *Constraint Grammar, A Language-Independent System for Parsing Unrestricted Text.* Berlin: Mouton de Gruyter.

Santos, Diana & Eckhard Bick (2000). *Providing Internet access to Portuguese corpora: the AC/DC project* , in Maria Gavrilidou et al. (eds.), *Proceedings of the Second International Conference on Language Resources and Evaluation*, LREC 2000 (Athens, 31 May-2 June 2000), pp.205-210.

Tapanainen, Pasi (1996). *The Constraint Grammar Parser CG-2.* Publication No. 27. Helsinki: Department of General Linguistics, University of Helsinki

Voutilainen, Atro & Heikkilä, Juka & Anttila, Arto (1992). *Constraint Grammar of English, A Performance-Oriented Introduction,* Publication No. 21. Helsinki: Department of General Linguistics, University of Helsinki

Voutilainen, Atro (1994). *Designing a Parsing Grammar.* Publications No. 22. Helsinki: Department of General Linguistics, Helsinki University