LREC 2002. Workshop Proceedings. Customizing Knowledge in NLP Applications: Strategies, Issues, and Evaluation. Las Palmas, Spain: ELRA. 2002. 45–51.

Locating and Reusing Sundry NLP Flotsam in an e-Learning Application

Anju Saxena and Lars Borin

Department of Linguistics, Uppsala University,
Box 527, SE-751 20 Uppsala, Sweden
and
Computational Linguistics, Department of Linguistics,
Stockholm University, SE-106 91 Stockholm, Sweden

anju.saxena@ling.uu.se, lars.borin@ling.su.se

Abstract

We describe the background and motivation for an e-learning project—*IT-based Collaborative Learning in Grammar*—where NLP resource reuse has become an important issue. The resources are of several kinds: POS-tagged and syntactically annotated corpora (treebanks), parsing systems and grammar writer's workbenches, and visulization and manipulation tools for linguistically annotated corpora. Our experience thus far has been that although there are a number of such resources available e.g. on the Web, as a rule, numerous incompatibilities and lack of standardization at all levels—markup formats, linguistic annotation schemes, grammatical framework, software APIs, etc.—make the reuse of these resources into a non-trivial endeavor.

0. Preamble: the Setting

It is generally acknowledged that the goal of teaching grammar—especially at the university level—should not primarily be that students memorize definitions of concepts and grammatical constructions, but rather that they understand and learn to recognize different structural patterns. This can hardly be achieved without giving students practical training in the skill of grammatical analysis. Research has shown that hands-on problem-solving is more stimulating and thought-provoking than when the information and results are handed down to the pupils during lectures. Further, our experience has been that students learn about grammatical constructions and phenomena more actively when these constructions are discussed by comparing the system found in their native language with that of another language. An added factor contributing to an active student participation is the choice of the material forming the basis for exercises and group activities, which should preferably be as natural as possible.

With these pedagogical considerations in mind, we formulated a project for realizing a new format for teaching courses in grammar in Linguistics and Computational Linguistics (the ability to reason about grammar and to carry out grammatical analyses of language utterances being necessary prerequisites for all linguistic studies of language and thereby part of the core curriculum of these subjects). In the proposed format interactive practical training and corpus-based exercises comprise an integral part of the students' learning process, giving them the opportunity and incentive to participate more actively in their own learning process. Using IT as a tool for collaborative work allows the students to choose the problem-solving strategy which suits them best, as well as the time and place to work on the problem. A corpus of natural language material for grammatical analysis contributes to a more active participation, as it not only presents the grammatical constructions in their context, but also gives students a greater freedom to approach the material and conduct the investigation from a perspective which suits their individual learning styles. A text corpus consists of naturally occurring language in its natural physical context, since it is made up of complete texts or large text fragments, as opposed to the made-up or isolated single sentences or phrases often used to illustrate grammatical points in linguistics textbooks. This accompanying physical context makes it possible to investigate the textual, discourse-level, functions of the grammatical phenomena.

An outline of the proposed training material is presented below. It has a modular architecture, composed of four types of modules (see Figure 1, below):

1. 'Encyclopedia' module, containing descriptions of grammatical concepts and constructions. Its content will be attuned to the contents of the course and the interactive exercises (as, in their turn, the exercises will be adapted to the 'encyclopedia' contents), and at appropriate places, there will be hyperlinks to interactive exercises dealing with the current topic.

2. 'Text corpus' module, containing at least (a) POS-tagged and syntactically annotated corpora of Swedish, and (b) an annotated corpus of a foreign language. For (a), we will use the SUC and Talbanken annotated Swedish corpora (see below); for (b), we will use a corpus of Kinnauri (a Tibeto-Burman language spoken in India) narratives available on the web (`http://www.ling.uu.se/anjusaxena/corpus.html`; see figure 2), which is hyperlinked to a morpheme dictionary. Further, with the help of a graphic interface students will be able to see a 'map' of how and where one particular morpheme or a word occurs in the corpus (see Olsson and Borin (2000)), providing support in their work on the functions of grammar. The students will work with the

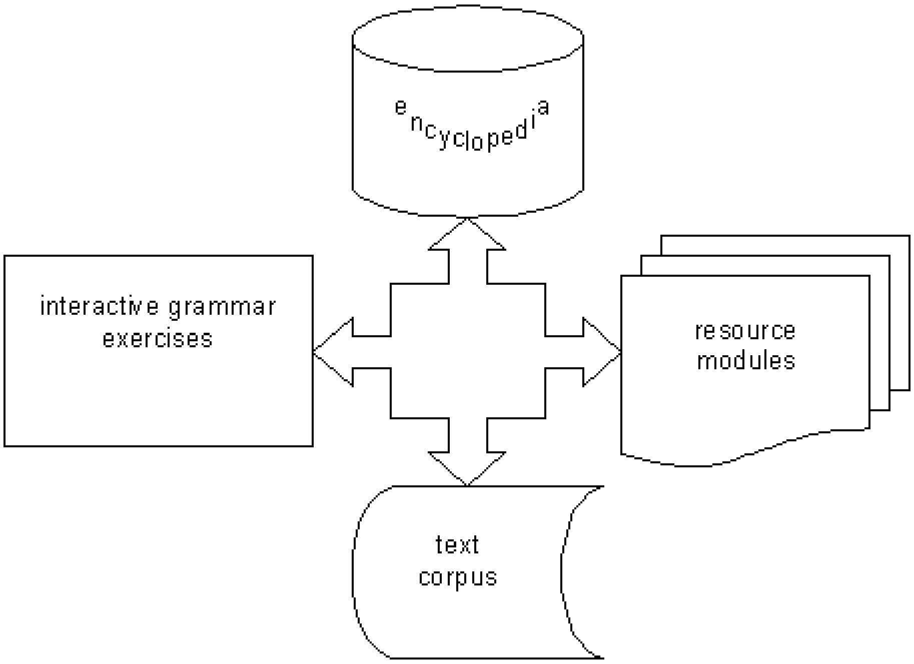

Figure 1: Organization of the proposed IT-supported grammar training application

same corpus as part of their group activities and as part of their examination.

3. 'Interactive exercise' module. Our aim here will be to provide students with a set of exercises, with basic tools for computer-mediated student cooperation in virtual work- groups (a 'spreadsheet' for problem-solving; optional 'step-by-step questions' for the grammatical topic covered; grammar rule writing exercises to be discussed in more detail below), with hyperlinks to the 'encyclopedia', to the 'resources' (see below) and to the annotated corpus of a foreign language (which, in turn, will be hyperlinked to the dictionary; see Saxena (2000)). As part of each theme, students will first discuss the construction during the lecture session, then again while examining the construction in the corpus, and finally also while comparing the results of the corpus-based analysis with the Swedish system and then discussing it in the group. This learning method where the same construction is examined from a number of mutually reinforcing practical and theoretical viewpoints will, hopefully, provide the students with support and incentive in their learning process. Further, the same corpus will be used in grammar courses in first and second semesters, providing grounds for deeper analyses in the second semester than would have been the case.

4. 'Resource' modules will provide a pool of resources for further reading and relevant links to other sites.

The architectural organization of the software proposed here has several advantages, the two most significant ones being extensibility and 'conceptual decentralization'. Extensibility means that new functions can be easily integrated in the application. 'Conceptual decentralization' is especially significant as it allows the possibility of adjusting to individual learning styles. For example, if the student prefers to start out with the 'encyclopedia' material and go from there to the appropriate exercises, when she feels the need to do so, she has that choice. At the same time, the application allows the possibility of starting out at other entry points, e.g., 'interactive exercises', with the option of calling up the relevant 'encyclopedia' material at each instant.

1. The NLP Resource Customization Problem

NLP resource customization has become an issue in this project mainly in connection with module 3 (interactive grammar exercises). It has been our aim from the conception of the project to rely mostly on standard WWW and open-source software—i.e., software which is generally free and where the source code is freely available and modifiable by the user—for implementing the modules. This design philosophy has the advantage of making the application maximally platform-independent, as well as providing a familiar interface—a standard web browser—for students and faculty.

One of the exercises that we have planned for module 3 builds upon a combination of a syntactically annotated corpus (a treebank) and a grammar writer's workbench. The basic premise of the exercises is a further refinement of the idea presented by Borin and Dahllöf (1999). We propose to use grammar rules written by students (using an existing grammar development tool) as search expressions in the

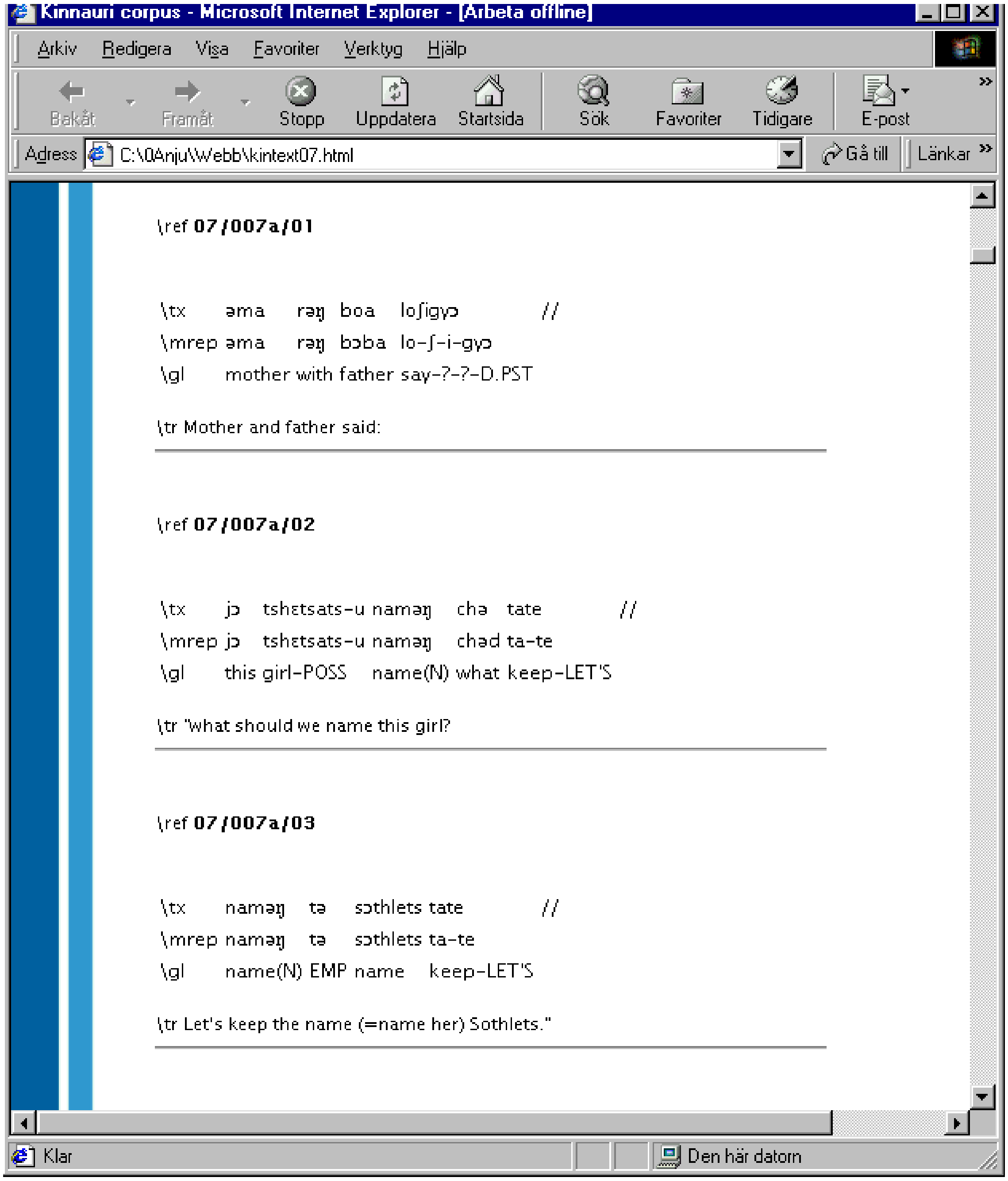

Figure 2: The Kinnauri corpus – Web format

treebank. In its simplest form, the result of the search would be expressed as precision and recall. Given an NP rule formulated by a student, we could automatically tell how many of the (maximal) treebank POS sequences matching the rule actually make up NPs, how many are not NPs, and how many NPs in the treebank are not described by the rule. There are all kinds of conceivable elaborations of this basic scheme, which could be seen as a more linguistically sophisticated parallel to the use of (unannotated) text corpora and concordancing software in so-called data-driven language learning (Flowerdew, 1996).[1] For the Computational Linguistics students, there is the additional advantage of being able to work from the very beginning of their studies with the same kind of tools and resources that they will be using 'for real' after graduating, in their professional life.

What we have found already in this beginning stage of the project, however, is that there are some serious obstacles to using available NLP resources.[2] Mostly, the issues that have arisen in this connection concern (lack of) compatibility and standardization of NLP resources. Some of

[1]The basic idea here is similar to the ICECUP FTF (Fuzzy Tree Fragment) grammatical query system for parsed corpora (Wallis and Nelson, 2000), but with a diffent use and target audience in mind.

[2]Here, we use "NLP resources" as a cover term for both *language resources* and *processing resources* in the terminology adopted by Cunningham (2002).

```
<text id=kl01>
<body>
<p>
<s id=kl01-001>
<c lem='-' msd='FI' n=1>-</c>
<w lem='vilken' msd='DH@0P@S' n=2>Vilka</w>
<w lem='djävla' msd='AQP00N0S' n=3>djävla</w>
<w lem='optimist' msd='NCUPN@IS' n=4>optimister</w>
<c lem=',' msd='FI' n=5>,</c>
<w lem='frusta' msd='V@IIAS' n=6>frustade</w>
<name type=person>
<w lem='Lasse' msd='NP00N@0S' n=7>Lasse</w>
</name>
<c lem='.' msd='FE' n=8>.</c>
</s>

<suctext id=kl01>
<p>
<s id=kl01-001>
<d n=1>-<ana><ps>MID<b>-</d>
<w n=2>Vilka<ana><ps>HD<m>UTR/NEU PLU IND<b>vilken</w>
<w n=3>djävla<ana><ps>JJ<m>POS UTR/NEU SIN/PLU IND/DEF NOM<b>djävla</w>
<w n=4>optimister<ana><ps>NN<m>UTR PLU IND NOM<b>optimist</w>
<d n=5>,<ana><ps>MID<b>,</d>
<w n=6>frustade<ana><ps>VB<m>PRT AKT<b>frusta</w>
<name type=person>
<w n=7>Lasse<ana><ps>PM<m>NOM<b>Lasse</w>
</name>
<d n=8>.<ana><ps>MAD<b>.</d>
</s>
```

Figure 3: Alternative SUC annotation formats

the issues are:

- Differences in fundamental storage and text markup formats. The three corpora that we are considering for use in the project have three different storage formats: (1) The basic format of Saxena's Kinnauri narrative corpus is as a Shoebox database (Buseman and Buseman, 1998) (see figure 4), from which a web version in HTML hyperlinked to a morpheme lexicon was semiautomatically derived (see figure 2); (2) The Stockholm Umeå Corpus (SUC; Ejerhed and Källgren (1997)) comes in an SGML corpus format as specified by the Text Encoding Initiative (TEI; `http://www.tei-c.org/`), and further, there are two different grammatical annotation formats, Parole/EAGLES format (see Monachini and Calzolari (1996)) and SUC format (see figure 3); (3) The Talbanken syntactically annotated corpus of Swedish (Einarsson, 1976a; Einarsson, 1976b; Teleman, 1974) is in an 80-column punch card format with only capital letters (see figure 5).

```
\ref 07/007a/01
\tx @ma r@N boa loshigyO //
\mrep @ma r@N bOba lo-sh-i-gyO
\gl mother with father say-?-?-D.PST
\tr Mother and father said:

\ref 07/007a/02
\tx jO tshEtsats-u nam@N ch@ tate //
\mrep jO tshEtsats-u nam@N ch@d ta-te
\gl this girl-POSS name(N) what keep-LET'S
\tr "what should we name this girl?

\ref 07/007a/03
\tx nam@N t@ sOthlets tate //
\mrep nam@N t@ sOthlets ta -te
\gl name(N) EMP name keep-LET'S
\tr Let's keep the name (=name her) Sothlets."
```

Figure 4: The Kinnauri corpus – Shoebox format

- Differences in POS tagging and syntactic annotations between corpora. The SUC and Talbanken Swedish corpora, although both are POS tagged, use different tagsets, with e.g. SUC having two and Talbanken three subclasses of nouns, and SUC, but not Talbanken, marking number in nouns, etc. Tagset incompatibilities, even within a language is a problem that has been noted in the literature (e.g. by Atwell et al. (2000)), and there has been some work on tools for automatic tagset mapping (e.g. Teufel (1995)). The problems are compounded when several languages are involved,[3] which would be desirable in our setting, where the linguistic subdisciplines of Contrastive Linguistics and Language Typology rely on explicit comparisons between languages at various linguistic levels. As stated above, we know from experience that students learn about grammatical constructions and phenomena more actively when these constructions are discussed by comparing the system found in their native language with that of another language. Preferably, the other language should be one that the students do not know already, as they then will be better able to concentrate on the analysis of 'pure' form. This is why we intend to use the Swedish and Kinnauri corpora together in our first application.
- Differences in POS categories, syntactic categories and grammatical framework between the corpora on

[3]The problem of crosslinguistic mapping of part-of-speech tags has not been extensively discussed in the computational linguistics literature (see Borin (2000); Borin (Forthcoming 2002); Borin and Prütz (2001)), but in general linguistics, there is an extensive literature on the issue of crosslinguistic properties of part-of-speech systems and the universality of proposed parts of speech, which is very relevant in this context (e.g., Anward et al. (1996); Itkonen (2001); Pawley (1993)).

```
P21803012001      0000                        <<      GM          010
P21803012002      *DET                        POOP    SS          010
P21803012003      RÖR                         VVPS    FV          010
P21803012004      SIG                         POXP  AAOO          010
P21803012005      ALLTSÅ                      ABKS    +A          010
P21803012006      OM                          PR      OAPR        010
P21803012007      FALL                        NN      OA          010
P21803012008      1000                        RC      OAET        010
P218030120091000 2DÄR                         ABRA    RA          010
P2180301201010002ORSAKEN                      NNDD    SS          010
P2180301201110002TILL                         PR      SSETPR      010
P2180301201210002PATIENTENS                   NNDDHHGGSSETDT      010
P2180301201310002SYMTOM                       NN      SSET        010
P2180301201410002INTE                         ABNA    NA          010
P2180301201510002PRIMÄRT                      AJ      AA          010
P2180301201610002ÄR                           AVPS    FV          010
P2180301201710002ÅDERFÖRKALKNING              VN  SS  SP          010
P21803012018100021100                         +F      +F          010
P2180301201911002UTAN                         ++MN    ++          010
P2180301202011002I                            ABMN    +A          010
P2180301202111002STÄLLET                      ID      +A          010
P2180301202211002BEROR                        VVPS    FV          010
P2180301202311002PÅ                           PR      OAPR        010
P2180301202411002EN                           EN      OADT        010
P2180301202511002SANNOLIK                     AJ      OAAT        010
P2180301202611002STÖRNING                     VN      OA          010
P2180301202711002I                            PR      OAETPR      010
P2180301202811002CIRKULATIONEN                VNDD    OAET        010
P2180301202911002AV                           PR      OAETETPR    010
P2180301203011002DEN                          PODP    OAETETDT    010
P2180301203111002VÄTSKA                       NN      OAETET      010
P21803012032110021110                         RC      OAETETET    010
P2180301203311106SOM                          PORP    SS          010
P2180301203411106OMGER                        VVPSSM  FV          010
P2180301203511106HJÄRNAN                      NNDD    OO          010
P21803012036      .                           IP      IP          010
```

Figure 5: The annotation format in the Talbanken treebank

the one hand and the grammar writing tools and parsers on the other. Thus, the Talbanken corpus uses a fairly traditional Swedish functional grammatical framework, where e.g. NPs are not directly recoverable, but only indirectly, through a combination of syntactic function and lexical category of the head word, while it seems that many, perhaps the majority, of the grammar writing tools freely available on the Web presuppose a phrase structure framework.

- Differences in implementation language, storage model, API, documentation and source code availability, etc. of potentially suitable software. For an excellent overview of these issues, see Olsson (2002).

Thus, we have been forced from the outset to discuss seriously how we are to integrate existing NLP resources in our application, as well as how to make the application itself extensible, so that e.g. new language corpora or new annotations can be added.[4]

2. Taking Stock and Looking Ahead

We are attempting to reuse NLP resources originally meant for NLP research—both *language resources* (notably annotated text corpora) and *processing resources* (the most important being parsers and grammar writing tools)—in an e-learning application for IT-based collaborative learning in grammar courses for Linguistics and Computational Linguistics university students. At the moment, we are locating and evaluating[5] NLP resources, mainly on the web, for the corpus-based interactive grammar exercises. As the corpora are in place already, we are now evaluating tools for the manipulation and visualization of corpus data, parsing systems, and grammar writing environments (workbenches), which raises a number of compatibility/standardization issues that need to be resolved. These compatibility/standardization issues point in two directions simultaneously, as it were:

1. backwards: How can we integrate in our application, with the least amount of effort, existing NLP resources of the kind that we need?
2. forwards: How can we ensure that we ourselves, as well as others, will be able in the future to modify the existing NLP resources, or add new ones, in the framework that we define?

The preliminary answers to these two questions are as follows.

There does not seem to be a simple answer to the first question. Generally, we think that it is more desirable to be able to reuse existing language resources—i.e., texts and corpora, lexicons, and the like—than processing

[4]Courses in Hindi and Turkish at Uppsala University will be used as testbeds during the third year of the project, based on relevant Hindi and Turkish corpus resources.

[5]The evaluation is to be mainly pedagogical, i.e. we will ask ourselves whether a particular resource will be suitable for the pedagogical framework that we have adopted for teaching grammar. However, usability—as the term is used in Human–Computer Interaction research—will also be an important evaluation criterion, as well as the the estimated effort needed to adapt the resource for our needs. See Hammarström (Forthcoming 2002) for details.

resources—in our case first and foremost grammar writing and processing environments—for the pragmatic reasons that

- constructing an annotated corpus from scratch is likely to be a much larger effort than building a grammar writing environment;

- standardization efforts have progressed further particularly in the realm of POS tagged language corpus resources than in the case of language processing resources (Monachini and Calzolari, 1996; Bird et al., 2000; Ide et al., 2000; Cotton and Bird, 2002) (and treebank formats; see Atwell et al. (2000)), although, as a rule, their use in computer-assisted language learning applications has not been considered in this connection (Borin, 2002).

Hence, we aim at being able to handle at least POS-tagged corpora using the EAGLES/Parole tag scheme and marked-up according to the TEI/CES SGML or TEI/XCES XML language corpus formats (thus recognizing, e.g., the SUC Parole format without special preprocessing).

As for the second question, it too, is easier to answer for language resources. Here, we will harmonize the underlying corpus formats with other ongoing projects in our departments,[6] while simultaneously endeavoring to conform to standards that are being worked out in the NLP community. This means that we will undertake the conversion of the Kinnauri and Talbanken corpora into this format, and that in due course we plan to make the corpora generally available in the new format.

As far as 'grammar writer's workbenches' are concerned, we have not yet been able to find a ready-made environment user-friendly enough (for our Linguistics students) and bug-free enough to be immediately useful for our purposes. Thus, it seems likely that we will have to put in some development effort in this area. If this turns out to be the case, the most likely kind of workbench that we will modify or build, will be one within the general paradigm of unification-based feature structure grammar. The evaluation of these systems is still ongoing, however (Hammarström, Forthcoming 2002).

3. Acknowledgements

The work described here forms part of the project *IT-based Collaborative Learning in Grammar*, a collaboration between the universities in Uppsala and Stockholm, funded by the Swedish Agency for Distance Education (DISTUM), for the three years 2002–2004. Anju Saxena is the principal investigator for the project. See also `http://www.ling.uu.se/anjusaxena/distum.html`.

[6]We will strive to be compatible with the corpus format developed in the CROSSCHECK (`http://www.nada.kth.se/theory/projects/xcheck/`), SVANTE (`http://www.ling.uu.se/lars/SVANTE/`) and *ASU availability* projects, in all of which formats and tools for Swedish *learner corpora* (see Granger (1998)) are being developed. The basic corpus format will adhere closely to XCES, with 'standoff' linguistic annotation (Ide et al., 2000).

4. References

Jan Anward, Edith Moravcsik, and Leon Stassen. 1996. Parts of speech: a challenge for typology. *Linguistic Typology*, 1(2):167–183.

Eric Atwell, George Demetriou, John Hughes, Amanda Schiffrin, Clive Souter, and Sean Wilcock. 2000. Comparing linguistic annotation schemes for English corpora. In Anne Abeille, Torsten Brants, and Hans Uszkoreit, editors, *Proceedings of the Workshop on Linguistically Interpreted Corpora. LINC-2000*, pages 1–10. Held at the Centre Universitaire, Luxembourg, August 6, 2000.

Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman. 2000. ATLAS: a flexible and extensible architecture for linguistic annotation. In *Proceedings of LREC 2000*, pages 1699–1706, Athens. ELRA.

Lars Borin and Mats Dahllöf. 1999. A corpus-based grammar tutor for Education in Language and Speech Technology. In *EACL'99. Computer and Internet Supported Education in Language and Speech Technology. Proceedings of a Workshop Sponsored by ELSNET and The Association for Computational Linguistics*, pages 36–43, Bergen. University of Bergen.

Lars Borin and Klas Prütz. 2001. Through a glass darkly: Part of speech distribution in original and translated text. In Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel, editors, *Computational Linguistics in the Netherlands 2000*, pages 30–44. Rodopi, Amsterdam.

Lars Borin. 2000. Enhancing tagging performance by combining knowledge sources. In Gunilla Byrman, Hans Lindquist, and Magnus Levin, editors, *Korpusar i forskning och undervisning. Corpora in Research and Teaching*, pages 19–31, Växjö Universitet, Växjö. ASLA, ASLA.

Lars Borin. 2002. Where will the standards for intelligent computer-assisted language learning come from? In *Proceedings of LREC 2002 workshop on International Standards of Terminology and Language Resource Management*. To appear.

Lars Borin. Forthcoming 2002. Alignment and tagging. In Lars Borin, editor, *Parallel Corpora, Parallel Worlds. Selected Papers from a Symposium on Parallel and Comparable Corpora at Uppsala University, Sweden, 22–23 April, 1999*. Rodopi, Amsterdam.

Alan Buseman and Karen Buseman, 1998. *The Linguist's Shoebox for Windows and Macintosh.* Summer Institute of Linguistics, Waxhaw, North Carolina:.

Scott Cotton and Steven Bird. 2002. An integrated framework for treebanks and multilayer annotations. In *Proceedings of LREC 2002*, Las Palmas. ELRA. To appear.

Hamish Cunningham. 2002. GATE, a general architecture for text engineering. *Computers and the Humanities*, 36:223–254.

Jan Einarsson. 1976a. Talbankens skriftspråkskonkordans. Corpus on CD-ROM.

Jan Einarsson. 1976b. Talbankens talspråkskonkordans. Corpus on CD-ROM.

Eva Ejerhed and Gunnel Källgren. 1997. Stockholm Umeå

Corpus version 1.0, SUC 1.0. Department of Linguistics, Umeå University.

John Flowerdew. 1996. Concordancing in language learning. In Martha C. Pennington, editor, *The Power of CALL*, pages 97–113. Athelstan, Houston, Texas.

Sylviane Granger, editor. 1998. *Learner English on Computer*. Longman, London.

Harald Hammarström. Forthcoming 2002. Overview of IT-based tools for learning and training grammar. Project report, IT-based Collaborative Learning in Grammar. Department of Linguistics, Uppsala University.

Nancy Ide, Patrice Bonhomme, and Laurent Romary. 2000. XCES: an XML-based encoding standard for linguistic corpora. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC2000)*, pages 825–830, Athens. ELRA.

Esa Itkonen. 2001. Concerning the universality of the noun vs. verb distinction. *SKY Journal of Linguistics*, 14:75–86.

Monica Monachini and Nicoletta Calzolari. 1996. Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. a common proposal and applications to European languages. EAGLES Document EAG-CLWG-MORPHOSYN/R.

Leif-Jöran Olsson and Lars Borin. 2000. A web-based tool for exploring translation equivalents on word and sentence level in multilingual parallel corpora. In *Erikoiskielet ja kännösteoria – Fackspråk och översättningsteori – LSP and Theory of Translation. 20th VAKKI Symposium*, pages 76–84, Vaasa, Finland. University of Vaasa.

Fredrik Olsson. 2002. *Requirements and Design Considerations for an Open and General Architecture for Information Refinement*. Number 35 in Reports from Uppsala University, Department of Linguistics, RUUL. Uppsala University, Department of Linguistics.

Andrew Pawley. 1993. A language which defies description by ordinary means. In W. A. Foley, editor, *The Role of Theory in Language Description*, pages 87–129. Mouton de Gruyter, Berlin.

Anju Saxena. 2000. Corpora of lesser-known languages on the internet: A pedagogical tool for the teaching of syntax. Paper presented at the workshop on IT inom språkundervisningen. Uppsala University. `http://www.ling.uu.se/anjusaxena/symposium0303.html`.

Ulf Teleman. 1974. *Manual för grammatisk beskrivning av talad och skriven svenska*. Liber, Lund.

Simone Teufel. 1995. A support tool for tagset mapping. In *Proceedings of SIGDAT 1995. Workshop in connection with EACL 95*, Dublin. Association for Computational Linguistics.

Sean Wallis and Gerry Nelson. 2000. The FTF home pages. WWW: `http://www.ucl.ac.uk/english-usage/ftfs/faqs.htm`. Accessed on 10 April 2002.