The MULINCO corpus and corpus platform

Bente Maegaard (1), Lene Offersgaard (1), Lina Henriksen (1), Hanne Jansen (2), Xavier Lepetit (2), Costanza Navarretta (1), Claus Povlsen (1)

(1) Center for Sprogteknologi, University of Copenhagen
(2) Department of English, Germanic and Romance Studies, University of Copenhagen
Njalsgade 80, DK-2300 Copenhagen S, Denmark
{bente, loff, lina, costanza, claus}@cst.dk, {hanjan, lepetit}@hum.ku.dk

Abstract

The MULINCO project (MUltiLINgual Corpus of the University of Copenhagen) started early 2005. The purpose of this cross-disciplinary project is to create a corpus platform for education and research in monolingual and translation studies. The project covers two main types of corpus texts: literary and non-literary. The platform is being developed using available tools as far as possible, and integrating them in a very open architecture. In this paper we describe the current status and future developments of both the text and tool side of the corpus platform, and we show some examples of student exercises and linguistic investigations taking advantage of tagged and aligned texts.

1. Introduction

The MULINCO project (MUltiLINgual Corpus of the University of Copenhagen) started early 2005. The purpose of this cross-disciplinary project is to create a corpus platform for education and research in monolingual and translation studies. The project itself is limited to two years, but the ultimate goal is to create a strong and long-living infrastructure for researchers and students. The infrastructure will be very open both in terms of texts and of tools, while at the same time providing a structure for extracting sub-corpora with specific characteristics, and for choosin0g tools that belong together in a package.

The project is structured into 4 themes: on the one hand the corpus platform theme (theme 1), and on the other hand translation and contrastive studies at the lexical/syntactic level (2), the text linguistic level (3) and the literary level (4). Languages in focus are: Danish, English, French, German, Italian and Spanish, where Danish texts, obviously, play a central role for the corpus platform both as source and target.

The corpus platform basically consists of the corpora collected, and of the technical solutions provided. The first six months of the project were devoted to a requirements analysis at the corpus side as well as the tool side and resulted in a requirements report. In the report and in the features of the platform chosen, the project relies on previous work done by other groups, cf. later in this paper, while at the same time extending and combining solutions.

The collaboration between computational linguists and translation researchers made it possible to provide solid input both for the corpus collection and for the technical requirements to the platform. At the same time this cross-disciplinary collaboration was rewarding for all participants.

2. Corpus

Availability of texts is often the most important criterion when it comes to corpus creation. Therefore many corpora take their texts from the Internet (where the texts are supposed to be free of copyright, even if this is not always true) or from public institutions, i.e. the European Parliament (Europarl[1]), the Canadian Parliament (Hansard[2]). Although surely useful, this kind of corpora covers only a limited range of language production. Text corpora aiming instead at representing a broader language use (BNC[3] or ANC[4]) are, on the other hand, generally constituted of text samples. This puts a series of restrictions on the kind of phenomena that can be studied within and across the texts. Consequently, we have opted for dividing the text collections into two main corpora: the literary corpus and the non-literary corpus.

The project has run into some difficulties in getting access to texts: there are no distribution agencies like ELRA for literary texts, so editors and authors have to be contacted individually. The texts are free of copyright only when the author and the translator have been dead for more than 70 years. It has been possible however, to negotiate the rights to use also a number of contemporary texts in the project. In order to make sure to provide texts for all languages and in order to provide possibilities for the comparison between literary texts and other texts, we have also made a subcorpus with other texts, in particular texts from the European Union.

Not only parallel texts are being collected, but also comparable corpora, i.e. texts of the same type written in different languages. These comprise recipes, jokes, official speeches etc.

2.1. The literary corpus

Within the framework of MULINCO, our aim has been to give the criterion of relevance higher priority than the criterion of availability. The MULINCO platform is to be used primarily in a department of foreign languages, literatures and culture studies, both for research and teaching purposes, and therefore a substantial part of the MULINCO corpus is centered on literary texts. Literature is a teaching matter in itself (text analysis and literary studies) and teaching and practicing translation is widely based on literary texts.

[1] http://www.europarl.eu.int/

[2] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC95T20

[3] http://www.natcorp.ox.ac.uk/

[4] http://americannationalcorpus.org/

Obviously, the challenges in putting together a literary corpus are not primarily technological, but they involve important financial and practical issues, as it is very difficult as well as time-consuming to get the rights to make a digital copy of a literary text (and it is even a political and philosophical issue, see the current discussion in the USA about fair use: http://fairuse.stanford.edu/).

To avoid the difficulties of obtaining copyrights many literary corpora include only parts of literary texts, i.e. samples. This procedure allows also coping with the criterion of corpus representativeness, as it is possible to choose any (extract of) text. But every text has a structure, that is, a web of lexical, grammatical, thematic, stylistic features that run through the text and constitute its textuality (i.e. the phenomena studied within the discipline of text linguistics). Within the MULINCO project, we have put much emphasis on the possibility of studying texts in their entirety. Both from a linguistic and a literary perspective, whole texts are necessary for a range of analytical purposes involving discourse and text phenomena that are not observable in samples. For instance, in order to make a study of the reference to a certain colour in a literary text (and of how this colour is being translated in different translations), you will need the whole text to be able to say anything about the variation of this specific parameter across the text.

For this reason, the kind of literary text we found most suited to be integrated in the corpus is the short story. Short stories are whole texts and thus allow making thematic and text linguistics analyses, and they are of a manageable size, which is an advantage both in educational contexts, and within specific literary and stylistic studies that work with qualitative rather than quantitative methods.

The core literary corpus in the MULINCO project will therefore be based primarily on short stories and their translation in as many of the involved languages as possible, including as a rule a Danish text, either as a source or target text.

It is an important feature of the project that the literary texts cover different periods, and that at least in some cases the same text exists for most of the languages treated, e.g. a Danish text that has been translated into all the other languages. This is the case e.g. for Hans Christian Andersen, but also some contemporary Danish authors have been widely translated. The present corpus also contains novels by Jules Verne, original French and translations into English and Danish, and short stories by E.A. Poe, original English and translations into Danish and French, as well as by Pirandello, original Italian and Danish translation.

2.2. The non-literary corpus

As mentioned above, the MULINCO project decided to acquire other text types as well. In the non-literary corpus we are collecting both parallel (translated) texts and comparable texts.

The European Union has a large amount of text which is parallel in many languages. Here the debates of the European Parliament constitute a valuable resource, with around 30 million words in 11 languages, Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish, i.e. this corpus covers all the relevant language for MULINCO.

The EU has recently given access to the Acquis Communautaire corpus (Steinberger et al 2006), a news corpus which has also been included.

Apart from these EU corpora, a collection of New Year speeches has been developed, consisting of official speeches by e.g. the Danish Queen, the German Chancelor, the French President etc. This gives a basis for comparing language, as well as language in relation to societal structure (monarchy, republic, federal republic etc.).

These new types of text collections will provide new material for the language teaching, and the platform will make them more accessible.

3. Corpus platform

The focus of the corpus platform is to make it possible to search the collected texts, which will normally be enriched by different levels of annotations. A requirements analysis (Farø et al. 2005) was carried out to determine the demands of the users concerning the corpus platform. The most general demands are:

- It should be possible to annotate the texts included in the platform with different levels of annotations, such as morphosyntactic, syntactic and semantic information.
- It should be possible to base a search on patterns of letters/words and/or annotated information both within and beyond the period boundaries in the text.
- Frequency lists for words and the search results should be available when querying.
- Parallel texts which are sentence aligned should be handled by the platform being able to display the aligned sentence for a target language together with the sentences that match a search pattern in the source language
- The interface should be an internet application giving students and researchers an easy access to the platform.

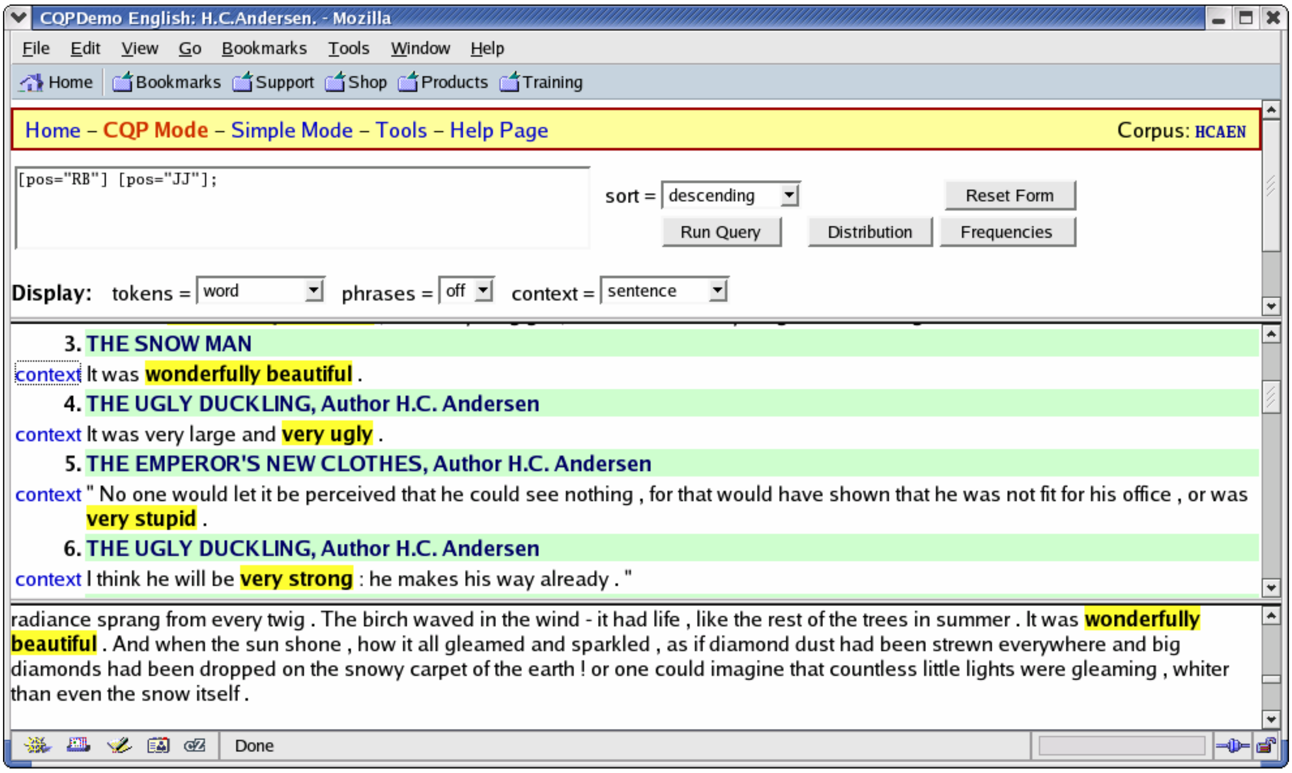
These demands has lead to choosing IMS Corpus Workbench (CWB)[5] (Christ 1994) as the core part of the corpus platform[6] and CQP as the query language. The CWB Web demo interface is used as a single language interface.

During collection all corpus texts are supplied with a header specification file, based on the XCES-header (Corpus Encoding Standard for XML) specification[7] containing administrative and bibliographic information. The XCES-header specification is extended with information about first publication date for a given text. In order to ensure that all required information is given, the collection process has been optimised by using an online corpus collection web interface, which guides the researchers involved to

[5] http://www.ims.uni-stuttgart.de/projekte/CorpusWorkbench.

[6] Two other corpus tools has been investigated: ParaConc http://www.athel.com/para.html and CorpusSearch http://corpussearch.sourceforge.net.

[7] http://www.xml-ces.org.



Corpus: HCAEN

```
[pos="RB"] [pos="JJ"];
```

sort = descending

Display: tokens = word phrases = off context = sentence

3. THE SNOW MAN

context It was **wonderfully beautiful** .

4. THE UGLY DUCKLING, Author H.C. Andersen

context It was very large and **very ugly** .

5. THE EMPEROR'S NEW CLOTHES, Author H.C. Andersen

context " No one would let it be perceived that he could see nothing , for that would have shown that he was not fit for his office , or was **very stupid** .

6. THE UGLY DUCKLING, Author H.C. Andersen

context I think he will be **very strong** : he makes his way already . "

radiance sprang from every twig . The birch waved in the wind - it had life , like the rest of the trees in summer . It was **wonderfully beautiful** . And when the sun shone , how it all gleamed and sparkled , as if diamond dust had been strewn everywhere and big diamonds had been dropped on the snowy carpet of the earth ! or one could imagine that countless little lights were gleaming , whiter than even the snow itself .

Figure 1: The web-interface for search in the English translation of fairy tales by Hans Christian Andersen. The search concerns an adverb, [pos="RB"], followed by an adjective [pos="JJ"].

specify the header information and to deliver it properly together with the corpus texts

3.1. Annotation

Annotation at different levels is done, -corresponding to the main themes mentioned above. The general approach is that the texts are annotated with part-of-speech (pos) tags[8] and lemmas (serving theme 2), but for specific studies (the other themes) the texts will also be annotated with higher level annotation. Obviously, annotation for the lower levels can be done automatically, whereas much of the higher level annotation will be manual, done through the work of the linguistic researchers and the students.

Annotation with pos-tags and lemmas is done mainly based on existing tools, but review of results has already led to adaptation of some of the tools. The TreeTagger is used for annotate German and Italian texts[9]. Italian is also analyzed by the NLP text analysis module developed by Istituto di Linguistica Computazionale, Pisa (Battista and Pirrelli, 1999; Bartolini et al. 2002). For Spanish the FreeLing-package[10] is used. For French we use Cordial[11]. For Danish CST's tools or tools trained on CST LRs are used. Furthermore we have trained TreeTagger for the old Danish literary texts as the orthography, spelling, vocabulary and punctuation in the original fairy tales of Hans Christian Andersen differ from those of contemporary texts. It has been chosen to manually validate and correct the annotated pos-tags and lemmas for a limited number of texts only, as the task is time consuming. The rest of the texts will be annotated automatically, and if project participants want to correct specific annotations on specific texts they are given this possibility and the corrected version will afterwards be used in the corpus platform.

The document structure annotation scheme covers title, filename, title page, chapters, paragraphs and periods. This annotation is done semi-automatically. If the texts are annotated with document structure, the query interface can use this information when presenting query results.

The work with syntactic and semantic annotation is not yet started. This annotation is planned to be done with XML mark-up. This work is planned to lead to new annotation types and schemes (e.g. the annotation of metaphors).

The parallel corpus web-interface of the platform includes search possibilities based on part-of-speech and lemma as well as search in parallel translation corpora. From a technical point of view this part of the MULINCO project can be seen as an already proven concept in the OPUS-project[12], but while the OPUS project focuses on providing publicly available parallel

[8] The part-of-speech tags include morphosyntactic information

[9]http://www.ims.unistuttgart.de/projekte/corplex/TreeTagger.

[10] http://garraf.epsevg.upc.es/freeling.

[11] http://www.synapse-fr.com/.

[12] http://logos.uio.no/opus

corpora, the MULINCO-platform focuses on using literary parallel corpora annotated with pos and lemma information for education purposes.

To be able to provide alignment of the parallel texts, different sentence aligners have been investigated. Our experiences so far show that the sentence splitting and alignment include several challenges for older literary texts and especially for the work of H.C. Andersen, as his use of sentence segmentation is quite different from sentence segmentation in contemporary Danish texts. Because for translation studies it is important that all sentences in the text are aligned, we have manually corrected the automatic paragraph and period mark-up in the H.C. Andersen's corpus and are now testing how different sentence aligners work on these data.

Bringing all these facilities together in the same web-interface will give the researchers a new powerful tool to enrich monolingual studies and studies in translation theory and practice.

4. Examples of the use of the MULINCO platform for education

The corpus platform may obviously be used for linguistic studies, but it will also provide the basis for literary studies, and studies of language and society (cf. New Year speeches mentioned above).

In the first year of building and using the corpus platform and the language technology tools, we have already seen beneficial side effects, e.g. the fact that there is no one single truth about grammatical function and part of speech. When the taggers are run on texts, they may assign part of speech in a way that the user did not expect. The first reaction is that the tagger is "wrong", and this can be used as input for a discussion about language description, different word classification traditions and theories etc. Given the fact that the tagging results will not always be correct, the material may also be used for the students to correct (in agreement with the tagger documentation).

As another general example, the tagger documentation or the results of tagging may be used to discuss if a specific word always belongs to the same part of speech: is the Italian *splendente* always a participle or sometimes rather an adjective? And for words for which the tagger has two or more possibilities, e.g. the Italian *quando*, how can we decide whether it is a conjunction or an adverb (and how is it possible for a tagger to discriminate between the two uses of the word?)

These are general benefits of using the tools and corpus search in the teaching. Below we present a few examples of investigations made on the corpus collected.

4.1. Fairy tales and language technology

The first fairy tales written by Hans Christian Andersen (HCA) were published in 1835. The reaction from the contemporary and established reviewers were in general quite negative. These tales were criticized for not meeting the past norms in terms of being instructive in accordance with the ruling didactic principles. The most profound element of criticism was that that not only were the tales about children and their behaviour, HCA also used various techniques in order to pretend that the content of the tales was perceived with a mind of a child.

The assessment of posterity of these literary 'tricks' - used by HCA in order to simulate a child's space of interpretation - has been that this narrative style constitutes the innovative and barrier-breaking elements of HCA's fairy tales. Recent literary analyses have focussed on this original element of the HCA writing. One concrete example of this, pointing out an HCA technique seen from a research angle, is his usage of onomatopoeic words which contribute to establishing a phantasmagoric universe (cf. Andersen, J. 2003).

This more linguistic approach to interpreting and understanding the originality of the HCA narrative style has proven useful and is used by several researches within the HCA research community.

It would be both an extension and a refinement of this approach to exploit the above mentioned HCA corpus and the search platform containing results from the NLP processes of lemmatization and PoS-tagging.

Using this refined approach it will be possible to make a mapping of all the characteristic features of HCA language usage seen from a more linguistic point of view.

4.1.1. A specific preliminary result

One example of this would be to generate a list of all the adjectival and adverbial phrases which occur in all the HCA fairy tales in order to find out whether the type of adjectival and adverbial phrases constitute a distinctive element in the fairy tales. In order to shed some light on this part of the HCA language usage, a comparative analysis was made.

The data consisted of one HCA fairy tale: *Sneemanden* (The snowman) and one contemporaneous short story, *Hosekræmmeren* (The hose merchant) by Steen Steensen Blicher. Using the corpus and search platform it was relatively straightforward to conduct the comparison of the adjectival and adverbial phrases occurring in the two literary works. The search pattern for the adjectival phrases would be [pos="ADV"] [pos="ADJ"] - similar to the example mentioned above - and for the adverbial phrases [pos="ADV"] [pos="ADV"].

The comparative corpus analysis revealed that in the HCA fairy tale the number of adjectival and adverbial phrases in which the modifying element consisted of a semantically vague adverb such as (eg *saa* (so), *ganske* (fairly), and *ret* (fairly) was significantly higher compared to *Hosekræmmeren.*

In *Sneemanden* the rate of frequency turned out to be 1 percent while in *Hosekræmmeren* the frequency was less than two per thousand running words. These preliminary results do not come as a surprise. Even though the Blicher short story contains quite many dialogues and therefore expectedly contains rather many oral features, the fairy tale in comparison embedded in the HCA narrative style has many more oral features.

4.2. Contrastive language and literary studies

The collection of literary texts and their translations are and will be used for investigating and teaching contrastive linguistic and literary phenomena, such as lexical variation, syntactic constructions, valency, discourse structure, the relation between the vocabulary and the main topics in the texts.

An example of the linguistic phenomena that are currently investigated is the translation of Danish nexus adverbials to non-Scandinavian languages. Nexus adverbials occupy a fixed position in sentences in the Actualization field in Diderichsen's (Diderichsen 1946/1957) topological language model, i.e. they occur before the main verb in subordinate clauses and after the main verb in main clauses. Nexus adverbials are very frequent and contribute to discourse in several ways, expressing phenomena such as negation, aspect, mood, the speaker's or the character's point of view, the connection between the current sentence and another sentence. Nexus adverbials can be combined in various ways and in many cases they do not have direct correspondence in the other languages the project works with. Thus they are often ignored in translations or are expressed by modal verbs or by various circumlocutions. In table 1 are shown examples of these adverbials in extracts from HCA's *Sneemanden* and their translations to English and Italian respectively.[13] The adverbs in the Danish text in the first row (in bold) are not translated at all in the English and Italian versions, while the contents of the adverbs in the text in the second row are somehow expressed by the English "certain to" and the Italian conditional "dovrebbero". The English interpretation is nearer the Danish than the Italian one.

Andersens *Sneemanden*	**English translation**	**Italian translation**
ja hun løb ***jo rigtignok*** før, da jeg saae stift paa hende, nu lister hun fra en anden Kant!	Yes, it was running itself, when I saw it a little while ago, and now it comes creeping from the other side.	Lui sì che è corso via prima, quando l' ho guardato fisso, ora sbuca fuori da un' altra parte!
det er et uskyldigt Ønske, og vore uskyldige Ønsker maae ***dog vist*** blive opfyldte	It is an innocent wish, and our innocent wishes are **certain to** be fulfilled.	È un desiderio innocente e i nostri desideri innocenti **dovrebbero** avverarsi.

Table 1: Nexus adverbials and their translations

Another research which is currently made on the MULINCO corpus material is a comparison of the discourse structure in Italian material and in its translations to Danish. In particular we have looked at some of Pirandello's short stories and their Danish translations. A first surface analysis of the stories in the two languages shows that the Danish translations contain approx 35% more sentence markers than the Italian source stories. This is mainly due to the fact that many Italian complex sentences, especially those containing gerundive clauses have been split up in series of shorter finite clauses in the Danish translation. We have now begun to annotate the rhetorical structure using the RST (Rhetorical Structure Theory) tool (Mann and Thompson, 1987; O'Donnell, 2000) on part of the data to compare the use of discourse relations in this Italian and Danish material.

4.3. Contrastive studies of motion verbs and motion expressions

The expression of motion is a well-known and well documented contrastive issue (Talmy 1975, Slobin 1996). A PoS-tagged aligned corpus such as the MULINCO corpus offers the possibility of testing and refining the hypothesis concerning significant typological differences within the coding of motion and spatial relations in general - especially differences involving Danish - on the basis of empirical studies on 1) parallel corpora, i.e. original texts and their translations, and 2) comparable texts. The infrastructure of MULINCO allows combined analyses on both kinds of sub-corpora.

Talmy distinguishes between *satellite-framed* languages (such as Germanic languages) that prototypically code the direction, or the "path", of the movement with a satellite (in English: *up, down,* etc.), and *verb-framed* languages, (such as Romance languages) that express the path of the movement within the verb itself (Fr. *monter, entrer, sortir*, vs. Da. *gå op, gå ind, gå ud*). At the same time, the Germanic languages express the manner in the verb, whereas the Romance languages typically have to add an extra phrase if it is considered necessary to express the manner explicitly (Fr. *entrer en courant*, Da. *løbe ind*).

The analytic structure of the (prototypical) expression of movements in the Germanic languages (verb + satellite) is particularly well suited to automatic retrieval methods. A query that combines a PoS-tag (verb) with a predefined list of spatial satellites (*up, down*, etc.) allows retrieving automatically, and therefore systematically and exhaustively, the entire population of a certain kind of expressions of movement from a text. Combining this search query with the alignment facility (comparing for example original Danish texts with their translated counterparts in various Romance languages), the rendering of this particular structure in translated texts can be analyzed in relation to systematicity, frequency and possible variation within the Romance languages. These results can then be tested by comparison to original, not translated texts.

In order to give some first ideas about the frequency of such translation examples, the Danish version of Hans Christian Andersen's *Sneemanden* was checked for the pattern 'verb + *op/ned*' (up/down). 9 relevant examples were found in the Danish text, out of which only one expressed the direction in the French translation, namely

Da. *Fuldmaanen stod op*
Fr. L*a pleine lune monta dans le ciel*

[13] The search string for nexus adverbials in Mulinco CQP is "([pos="V_PRES|V_PAST|V_IMP"] [pos="ADV"] [pos="ADV"]+)| ([pos="ADV"] [pos="ADV"]+)[pos="V_PRES|V_PAST|V_IMP"]", meaning search for all adverbs that follow or precede the main verb.

In all other examples, the information expressed in the direction particle was omitted in French. As an example, consider:

> Da. *Hun* [Solen] *lærer Dig nok at **løbe ned** i Voldgraven*
> Fr. *Il* [le soleil] *saura t'apprendre à **courir** dans le fossé.*

This kind of empirical results can be used both as data within scientific work and as data and/or process integrated in a teaching context. The students could be asked to use the platform interface to make a survey as the ones mentioned above on a particular text. They would thereby gain at least three kinds of insights: 1) an insight in language technology, its methodology and its benefits, included the fact that not all examples of 'verb + *op/ned*' are relevant in this context, 2) a contrastive insight in different linguistic systems and 3) an insight in translation practice/technique based on the contrastive insight.

5. Future work

The work with the platform has already been highly beneficial for all participants. Current work on collection of data, tagging and correction will continue. For the platform the next challenges are the search interface for bilingual studies, as well as experiments with word alignment. The current results are already used in teaching at the University of Copenhagen, and when more texts and facilities are available, the researchers will also be able to take even more advantage of the platform. The texts will be annotated, not only with the automatic PoS tagging, but also with manual annotations e.g. semantic and pragmatic. Collaboration with other research groups has already started, and this will be extended.

6. Acknowledgements

The authors want to thank all project participants, cf. the project home page
http://www.cst.dk/mulinco/index.html

7. References

Andersen, J. (2003): *Andersen en Biografi*, Bind I, Gyldendal, København.

Pedersen, V.H. (1999): H. C. Andersen's Fairy Tales in Translation: Prepositions and 'Small Words, In: Johan de Mylius, Aage Jørgensen and Viggo Hjørnager Pedersen (ed.): *Hans Christian Andersen. A Poet in Time. Papers from the Second International Hans Christian Andersen Conference 29 July to 2 August 1996*, Odense Universitetsforlag, Odense.

Battista, Pirrelli (1999): *Una piattaforma di morfologia computazionale per l'analisi e la generazione delle parole italiane*, ILC-CNR technical report.

Bartolini, Lenci, Montemagni, Pirrelli (2002): Grammar and Lexicon in the Robust Parsing of Italian: Towards a Non-Naïve Interplay, in *COLING 2002 Workshop on Grammar Engineering and Evaluation*, Taipei, Taiwan.

Christ (1994): A modular and flexible architecture for an integrated corpus query system. *COMPLEX'94*, Budapest.

Diderichsen, P. (1957): *Elementær Dansk Grammatik*, Gyldendal, 1946, 2. edition.

Farø, Henriksen, Jansen, Lepetit, Maegaard, Navarretta, Offersgaard, Povlsen (2005): *MULINCO Behovsanalyse*, Rapport 1, Univ. Copenhagen.

Schmid (1995): *TreeTagger---a Language Independent Part-of-speech Tagger*, Institut für Maschinelle Sprachverarbeitung (IMS) Universität Stuttgart.

Mann, W. & Thompsom, S.A (1988): Rhetorical Structure Theory: Toward a functional theory of text organization', *Text* 8 (3), 243-281.

Martin, L.E. (1990). Knowledge Extraction. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 252-262.

O'Donnell, M. (2000): RSTTool 2.4 - A Markup Tool for Rhetorical Structure Theory". *Proceedings of the INLG'2000*, 13-16 June 2000, Mitzpe Ramon, Israel. 253 -256.

Olohan, M. (2004). Introducing Corpora in Translation Studies. Routledge

Slobin, D.I. (1996). "Two ways to travel: Verbs of Motion in English and Spanish", in: Shibatani and Thompson (eds.) Grammatical Constructions: Their Forms and meanings, Clarendon Press, Oxford, 195-219.

Talmy, L. (1975). "Semantics and Syntax of Motion", in: Kimbal (ed.): Syntax and Semantics vol. 4, Academic Press, London, 181-238.

Talmy, L. (2000): *Towards a cognitive semantics*. Vol. 1 and 2. Cambridge, MA, MIT Press.