Proceedings of the workshop on

Multilingual Corpora: Linguistic Requirements and Technical Perspectives

A pre-conference workshop to be held at
Corpus Linguistics 2003

Lancaster, 27 March 2003

Stella Neumann & Silvia Hansen-Schirra
Saarland University, Saarbrücken, Germany

PROGRAM COMMITTEE:

Silvia Bernardini, Bologna
Sabine Brants, Palo Alto
Andreas Eisele, Saarbrücken
Stefan Evert, Stuttgart
Tony Hartley, Leeds
Natalie Kübler, Paris
Mick O'Donnell, Madrid
Maeve Olohan, Manchester
Elke Teich, Saarbrücken
Spela Vintar, Ljubljana
Federico Zanettin, Bologna

WORKSHOP WEBSITE:

http://www.coli.uni-sb.de/muco03

FURTHER INFORMATION:

Stella Neumann
Applied Linguistics, Translation and Interpreting
Saarland University
P.O. Box 151150
66041 Saarbrücken, Germany
st.neumann@mx.uni-saarland.de

Silvia Hansen-Schirra
Computational Linguistics
Saarland University
P.O. Box 151150
66041 Saarbrücken, Germany
hansen@coli.uni-sb.de

TABLE OF CONTENTS

The challenge of working with multilingual corpora

Silvia Hansen-Schirra
Computational Linguistics

Stella Neumann
Applied Linguistics, Translation and Interpreting

Saarland University

1. Introduction

Corpus linguistics has flourished in recent years, evolving from the need to empirically investigate linguistic theories and hypotheses. In order to encourage the cross-fertilization between work being done in language engineering and linguistics, the Corpus Linguistics 2001 conference offered a forum for those concerned with the computer-assisted empirical analysis of natural language. The Corpus Linguistics 2003 conference extends the scope to deal with more fine-grained research questions such as multi-layer annotation or the exploitation of multilingual resources. On the technical side, the available solutions comprise automatic or semi-automatic annotation tools on every language level and for many different languages. Treebank-initiatives all over the world show that even more complex linguistic annotation is state of the art. The achieved advances, however, lie mainly in the area of monolingual corpora. When it comes to multilingual corpora, the picture changes.

While monolingual and multilingual corpus research share a number of problems such as multi-layer annotation and exploitation, they are also different in a number of ways. Multilingual corpus research raises numerous problems which are specific to this kind of research and which are caused by the influence of differing language systems involved as well as the more complex methodological requirements. Among these issues are the comparability of segmentation, of features to be analysed and annotated as well as the question of how to query multilingual information.

The present paper gives an overview of the methodological crux of carrying out multilingual corpus research and which decisions are involved at each stage. As it is an introductory paper to a workshop it will then present the papers dealing with different aspects of the work with multilingual corpora. The paper concludes with an outlook beyond the workshop and a publication list of related work in which analysis scenarios in multilingual language research are explained in more detail.

2. Methodological steps in exploring multilingual corpora

As has been stated, working with multilingual corpora includes taking into account the differences and commonalities between the languages under investigation. This has to be done at each methodological stage of corpus-based research, thus involving

- the appropriateness and comparability of the corpus design,
- the different kinds of segmentation,
- the alignment techniques for translation corpora,
- the diverging annotation schemes,
- the possibly different corpus representations and finally
- the again converging querying across different languages.

Mistakes or inconsistencies which happen at one stage of the multilingual corpus development have negative influences on the following steps and might result in worse mistakes or inconsistencies at later stages.

Without methodological precautions, it is questionable whether the gained results are reliable and sufficiently comparable to give answers to the initially posed research question. While a research design that is well thought-out will help avoiding these problems, evaluation of the conducted research is a decisive step in which some aspects are often neglected. Standards contribute to avoiding difficulties, but they do not replace evaluation and – as will be discussed in some of the workshop papers – are not necessarily what is needed in each specific case.

In the following, each of the above mentioned steps will be scrutinized in turn. The respective options will be presented and open questions will be identified.

2.1 Corpus design

How do researchers go about building up multilingual corpora? Consciously or unconsciously, every corpus linguist makes decisions with respect to corpus size and representativeness. Apart from practical questions that influence these decisions like how much data can be collected and processed under the given constraints, the size depends on the research question for which the corpus is compiled. That means, the expected frequency of the feature in question determines the size: e.g. for investigating extraposed relative clauses a larger corpus has to be built up than for analysing a more frequently occurring feature like the usage of time adverbials.

The underlying question with respect to corpus representativeness is, which language variety should be represented. With this in mind, it has to be determined how much data is necessary to build up a corpus representative of the language variety under investigation. Moreover, the format of the collected language instances has to be defined, e.g. whether full texts are chosen or whether text samples are sufficient, but also whether the investigated linguistic features can rather be expected in spoken or written data. The researcher has to settle on whether the corpus is meant to only answer one specific research question or whether it should serve some general purpose. In the latter case the corpus design has to follow some common design principles in order to safeguard reusability.

Furthermore, building up multilingual corpora also involves decisions on the comparability of the sub-corpora. Not only should corpus size and the degree of representativeness be comparable, but the sampled linguistic data should also belong to the same register in terms of comparable selections of field, tenor and mode of discourse.

Another important issue in the stage of corpus design is the inclusion of a tertium comparationis: At a later point in the research process, the interpretation of the gained results will require referring to a basis of comparison for identifying characteristic patterns. For the corpus design, this means that the inclusion of a reference corpus has to be considered.

Having put together the corpus keeping the discussed design questions in mind, the corpus has to be processed. Segmentation of the data is the first preparatory step.

2.2 Segmentation

Due to language typological divergences, merely comparing multilingual raw corpora raises questions concerning the comparability of the respective segmentation. For instance, French clitics have to be lemmatised in order to make the segmentation between French and German comparable and thus analysable. Other questions that cannot be solved by lemmatisation are: How can a German compound noun be compared with its English counterpart consisting of several orthographic words? How should the English NP be tagged and aligned? Which query provides complete information on these diverging yet comparable structures?

One option in this concrete case might be chunking the German compounds into their components. This would make the German corpus more comparable to the English one. Nevertheless, instead of aiming at a direct comparison of the use of compounds in the German and English sub-corpus, it would be more reasonable to include a tertium comparationis as described above. This would offer the possibility to dissociate the realisation of compound noun patterns typical of the represented language variety from typological differences between the languages. That is to say, the solution for the segmentation problem might lie in the previous methodological step, namely the corpus design.

All of the problems discussed here do not take into account units of segmentation above the word or phrase level. On higher levels similar problems arise like the diverging segmentation of non-finite clauses in German and English. In this case, the recognition of segments should be based on the comparison of the respective functions in each language.

Closely related to questions of segmentation is the problem of aligning texts when working with translation corpora.

2.3 Alignment

A translation corpus usually consists of source texts in one language and corresponding target texts in another. In order to investigate the relationship between the two sub-corpora, the matching texts have to be parallelised. Now, what exactly is to be aligned? Whole Texts, sentences, or smaller units?

For practical or technical reasons, usually the sentence is chosen as alignment unit, since sentence boundaries can easily be detected automatically. For many research questions, however, this is a compromise, as actually the relationship between smaller units is under investigation. Whereas word alignment is useful for multilingual term extraction, this procedure does not make sense for the analysis of translations, because it breaks up the flow of text The desirable unit for alignment is somewhere in between words and sentences, and should be flexible enough to cope with typological differences between source and target language. This would offer, for instance, the possibility to align a finite clause in German with a non-finite construction in English.

Other alignment problems arise from the fact that different languages show different patterns with respect to sentence length and that during the process of translating L1-texts into L2-texts sentences are split or merged. Both phenomena make the whole endeavour of aligning sentences questionable.

While the previous steps of designing and pre-processing the corpus are in close connection with each other, annotation initiates the next phase involving the linguistic enrichment of the thus prepared raw data.

2.4 Annotation

For the development of an annotation scheme that meets the requirements of a multilingual analysis there seem to be two methods taking into consideration the typological characteristics of the involved languages,: First, the multilingual corpus is split up into monolingual sub-corpora which are then annotated independently. For the second method, one language serves as the basis for building up and analysing a multilingual corpus, whereas the other has to be adapted. Both methods, however, are rather problematic. The latter forces the adapted language to fit into the system of the language used as a fundament. In the former, questions of cross-linguistic comparison are merely pushed to a later point and come into play at the stage of interpreting the data obtained from this research design. In order to avoid both problems, the annotation scheme has to abstract from the level of language-specific realisations to functional categories that are comparable across languages.

On a more technical note, a number of computer tools exist for corpus annotation which serve monolingual as well as multilingual purposes. These include some tools that are language-independent, but their high degree of flexibility results in a low degree of automation. Other tools that allow automatic or interactive annotation require language-specific training, thus again leading back to the previously discussed question of comparability across multilingual annotations. More difficulties result from the fact that many of these tools are focused on Indo-European languages. The comparison of an Indo-European language with one from another language family poses new problems for the existing tools.

While the peculiarities in the annotation process of multilingual corpora concentrate on linguistic aspects, concerning the specificities for the next two steps of corpus representation and querying attention shifts to technical concerns.

2.5 Representation

The representation of multilingual corpora requires matching formats of the monolingual annotation, as this establishes the basis for simultaneous queries into all languages involved. Therefore, if language-specific tools are employed, they have to operate on the same format. Since the format of translation corpora needs to reflect the alignment of the parallel texts, the source and target units have to be linked to each other either within one file or by matching IDs in separate files. The researcher has to decide on this according to the linguistic needs but also to the input requirements of the corpus tools.

Additionally, the differences and commonalities of the language systems involved intertwine with issues of multi-layer annotation. This means that the language-specific characteristics in syntax,

semantics, morphology, discourse etc. are to be reflected in the annotated corpora. For the complex combination of multilingual and multi-layer annotation, at present, XML seems to be the most practicable solution, because this format allows the multi-layer annotation of embedded and overlapping segments on the basis of stand-off mark-up. More and more tools operate on an XML-based representational format and more and more representational XML-based standards evolve. But is this really practicable? And do linguists get along with techniques such as stand-off mark-up? Is the use of XML maybe a technical compromise? At the moment, this may look problematic. Nevertheless, the currently available prototypes as well as the tools under development promise an improvement both with respect to manageability for the linguist and to technical standardisation.

However, in order to offer the possibility of exchanging and reusing multilingual corpora the representational format should follow some common principles. This is also a crucial issue for importing a multi-layer annotated corpus into a query tool.

2.6 Querying

Irrespective of the multilingual aspect, for the step of querying the following questions have to be answered: Do we want to query instances, tags or both? Querying instances, i.e. raw text, poses no problem for monolingual as well as translation corpora, since the existing concordance tools sufficiently support this task. Things become more difficult when it comes to comparable corpora. At present, comparable corpora can only be used in monolingual concordance tools and therefore cannot be queried simultaneously in all sub-corpora.

Monolingually, the area of querying tags is widely exploited by tools for querying part-of-speech tags, syntactic information, semantic annotation etc. and, if needed, also in combination with raw text. When working with aligned translation corpora, only part-of-speech tags and/or instances can be retrieved. Still, this has to be regarded as a monolingual query, as the search operates only on the basis of one language. The aligned parallel segment is displayed in the concordance, simply because it is linked to the correspondent segment. A query for source language and target language instances and/or tags at the same time is not yet supported. Queries for syntactic, semantic or even multi-layer annotated data are not possible to date. Comparable corpora do not allow any simultaneous querying in tags at all.

With respect to displaying the output of a query, it would be desirable to show both raw text and annotation for all languages involved. Moreover, output functionalities are not guaranteed for all languages similarly to what has been said previously for the annotation of languages other than Indo-European ones.

When trying to query the linguistically interpreted multilingual corpus difficulties may become visible. It is our belief that these can be regarded as obvious symptoms for weaknesses of one of the methodological steps. This leads us to the issue of evaluating the research design with a view to all of the steps discussed here.

2.7 Evaluation

Under the perspective of evaluation the methodological steps that have been listed here in chronological order can be rearranged into different categories: technical evaluation strategies, standards and “manual” evaluation. Elaborated evaluation strategies can be found for the scalable, technically operationalised stages in multilingual corpus work. It is obvious to employ, for instance, precision and recall calculations for querying results. Annotation schemes and tools can be contrasted to a gold standard. Although there is no technical evaluation method for alignment, this step can be categorised under the technical heading. The above discussed problems with alignment techniques necessitate manual check of the aligned corpus prior to further processing.

Standardisation comes into play in the second category which is related to common principles for corpus representation. In the process of agreeing on standards, the existing representational formats have to be examined. On the basis of the thus defined best practice, a standard can be determined that will – if properly applied – make the evaluation at the step of representation pointless. Nevertheless, a researcher may choose to check his or her representational format against the defined standards in the evaluation process.

Finally, the steps of corpus design and segmentation escape technical measurement. Possibly, this is the reason why the evaluation of these steps is frequently neglected. Assessing the corpus design rather involves intellectual decisions: The evaluation of a large scale corpus project might result in the insight that above a certain amount of analysed data only redundant information was gained. The researcher should have pondered the corpus design more thoroughly. Evaluating segmentation has to be linguistically motivated. In the case of a multilingual syntax analysis, for instance, the research can verify whether the unit chosen – usually the sentence – helped to obtain the expected information on syntactic structures. In both cases, the researcher has to question the explanatory power of statements made on the basis of the research design.

Constraints with respect to time and money as well as the resulting compromises on methodological questions are in the way of – if not altogether prevent – the evaluation of the achieved results.

The present workshop makes its contribution to further work on both the linguistic as well as technical aspects of each of the steps in multilingual corpus research discussed here.

3. The workshop

In the call for papers we had asked for contributions on the following issues:

- problems and their possible solutions in the design, segmentation, annotation, representation and querying of multilingual corpora,
- computational tools which support these steps and
- international standards which facilitate the development and exchange of multilingual corpora.

All of these points are in some way or the other dealt with in the papers contained in this volume.

The debate is opened by Diana Santos' paper that takes a stance against a multilingual design of corpora, shifting the meaning of the term *multilinguality* away from our broad understanding of 'comprising more than one language' to a more strict sense of 'comprising more than two languages'. This view may be tested in the light of the next paper, in which Sattar Izwaini introduces a concrete corpus design including Arabic, English and Swedish. Izwaini's corpus is designed for the purpose of analysing translation procedures. Belinda Maia advocates the investigation of comparable corpora, not necessarily created for general purposes and therefore not needing to conform to standards essential for corpus reuse.

The issue of standardisation is the main interest of Pernilla Danielsson and Andrius Utka's paper. It weighs the use of XML as a representational format adding a more technical note to the discussion.

In the technical perspective, Christian Bering and his colleagues present a tool for corpus enrichment and information extraction showing how shared grammars can be used for multilingual named entity recognition. As their tool supports different languages such as Chinese and Spanish, it addresses the above mentioned problem of handling non-Indo-European languages. The discussion is concluded by Valia Kordoni who reports on some general principles of building up corpus resources exemplifying her point with the VERBMOBIL corpus of spontaneous speech.

The scope of these contributions shows that the challenge of multilingual corpora incites a lively discussion reviewing the usefulness of what has been achieved so far and showing how existing shortcomings may be overcome.

4. Conclusions and outlook

Some of the issues raised in this and the following papers require a linguistic perspective, others pose requests as to the adaptation of computer tools according to the needs of multilingual research. We have seen so far that the methodological problems not only arise at each stage of corpus research. They also multiply with the growing complexity of the research design. The limitations of the present paper allowed only to touch on the options and questions at the different stages. Each of the seven points discussed here deserves a thorough analysis in its own right. The paper might remind researchers of the options at their command and help choosing individual solutions. It also became obvious that there is still a lot to do for language engineers. One aim of the workshop is to inspire language engineers to

develop tools adapted to the needs of multilingual corpus research.

Hopefully, the workshop papers will incite a discussion of a linguistic requirement catalogue in combination with a selection of specifically adapted technical solutions. It could thus serve as a starting point for the development of an annotation typology which takes into account different languages as well as different annotation layers. Another aim could be the elaboration of a decision tree on the basis of which a multilingual corpus builder should be able to cope with possible problems in each of the above explained steps in corpus development. Within such a standardised research framework, the comparability of multilingual multi-layer annotated corpora can be guaranteed.

Related work[1]

Abeille A., S. Hansen & H. Uszkoreit (eds.), to appear. Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03). Budapest.

Hansen S., 2002. The Nature of Translated Text - An Interdisciplinary Methodology for the Investigation of the Specific Properties of Translations. PhD Thesis, Saarland University, Saarbrücken.

Hansen S., to appear. Linguistic enrichment and exploitation of the Translational English Corpus. In Proceedings of the Corpus Linguistics 2003 conference. Lancaster.

Hansen, S., M. Klaumann & S. Neumann, 2002. How to Overcome Registerial Translation Problems: A Corpus-based Approach. In Revista Brasileira de Lingüística Applicada, v. 2, no. 2, 15-23.

Hansen S. & E. Teich, 2001. Multi-layer analysis of translation corpora: methodological issues and practical implications. In D. Cristea, N. Ide, D. Marcu & M. Poesio (eds.) Proceedings of EUROLAN 2001 Workshop on Multi-layer Corpus-based Analysis. Iasi: 44-55.

Hansen S. & E. Teich, 2002. The creation and exploitation of a translation reference corpus. In Proceedings of the First International Workshop on Language Resources for Translation Work and Research (Third International Conference on Language Resources and Evaluation (LREC-2002)). Las Palmas: 1-4.

Neumann, S., 2002a. Register Characteristics in Contrastive Corpora. Paper presented at the 29th International Systemic Functional Conference. Liverpool, UK, July 15th – 19th, 2002.

Neumann, S. 2002b. Die Beschreibung von Textsorten und ihre Nutzung beim Übersetzen. Eine systemisch-funktionale Korpusanalyse englischer und deutscher Reiseführer. PhD Thesis, Saarland University, Saarbrücken.

Teich E.& S.Hansen, 2001a. Towards an integrated representation of multiple layers of linguistic annotation in multilingual corpora. In Online Proceedings of Computing Arts 2001: Digital Resources for Research in the Humanities. Sydney.

Teich E. & S. Hansen, 2001b. Methods and techniques for a multi-level analysis of multilingual corpora. In P. Rayson, A. Wilson, T. McEnery, A. Hardie & S. Khoja (eds.) Proceedings of the Corpus Linguistics 2001 conference. Lancaster: 572-580.

Teich E., S. Hansen & P. Fankhauser, 2001. Representing and querying multi-layer annotated corpora. In Proceedings of the IRCS Workshop on Linguistic Databases. Philadelphia: 228-237.

[1]While this paper represents a short methodological discussion on the work with multilingual corpora, concrete applications and examples as well as more in-depth methodological considerations can be found in the publications listed here.