LIONEL NICOLAS, EGON STEMLE, AIVARS GLAZNIEKS, ANDREA ABEL

A Generic Data Workflow for Building Annotated Text Corpora

1. Introduction

At the beginning of the corpus building process is the selection of appropriate software tools and data formats for the acquisition and annotation of the original linguistic data. This initial phase is characterised by challenging decisions, for the software needs to be flexible (to facilitate intuitive and speedy transcription), powerful (to meet annotation demands), adaptable (to enable easy transfer from one project to another), and process data according to formal criteria (to ensure data persistence and congruency). All decisions will have substantial implications throughout the data lifecycle - its annotation, retrieval, analysis, and re-annotation and re-use later on.

At present, there are a number of carefully developed guidelines such as the ones put forward by Wynne (2005) or Garside *et al.* (1997) as well as numerous projects that can be used as drafts. However, how to transform these guidelines into ordered sets of tasks or work packages, and how the tasks interact with one another has not been fully explained. Such knowledge is usually acquired by work experience. People with no prior experience have to cope with a challenging task for which they might not have had proper training.

In this paper, we focus less on *what* should be considered or reused for building an annotated text corpus, but rather on *how* tasks can be organized together and *how* they can interact with one another to form a smooth data workflow. We describe an abstract and generic workflow that has been carefully developed through an extensive interdisciplinary collaboration between linguists, who annotate and use corpora, and computational linguists and computer scientists, who are responsible for providing technical support and adaptation or implementation of software components. This workflow has been devised as a useful blueprint providing a common mental representation of how the work can be organized to achieve the planned objectives.

The workflow has originally been designed for building learner corpora and with the linguists' research needs and technical feasibility in mind. From a technical perspective, (non-trivial) annotated text corpora (Hundt 2008) are roughly similar and, from a user perspective, manual annotations have several commonalities; therefore, we believe, such a formalised workflow can be of interest for a wider spectrum of subjects in the digital humanities domain, and that our abstract workflow addresses the general task of annotating text corpora. More generally, the workflow follows an increasingly accepted idea, of which the *DARIAH* (Abdurachman *et al.* 2008) and *CLARIN* (Váradi *et al.* 2008) initiatives are clear examples. The idea is that many tasks within the whole of the digital humanities have a similar basis with similar objectives, and that their corresponding solutions can be reused or adapted and linked to one another.

The aim of this article is to introduce the workflow and to illustrate the way we implemented it within a learner corpus project. We focus on the needs that have been considered for establishing the workflow and point out the ones specific to our annotation task. Bearing in mind that the underlying work has been carried out in the context of building learner corpora, we provide the expert reader with as much contextual and decisional information as possible to observe analogies and discrepancies with annotation objectives in other projects.

We start with the user requirements for establishing the workflow and detail which, in our understanding, may be specific to learner corpora (*Section 2*). In *Section 3* we details the abstract workflow itself, and in *Section 4* the way we implemented the workflow in practice. We finally conclude in *Section 5*.

2 Research on a Learner Corpus – the Linguists' Needs

Learner corpora are "systematic computerized collections of texts produced by language learners" (Nesselhauf 2005). Many are error-tagged, that is, orthographic, lexical, and grammatical errors in the corpus have been annotated with the help of a standardized system of error tags (Díaz-Negrillo/Fernándes-Domínguez 2006). In addition, learner corpora should provide meta-information, such as the authors' L1, age, gender, and the like. Annotations can be done automatically, which is often the case for lemma and part-of-speech (POS) information, or manually which often involves grammatical errors. Technically, the annotations are either inline (Granger 2003) or multi-layered using a stand-off format (Lüdeling *et al.* 2005; Reznicek *et al.* 2013; Zinsmeister/Breckle 2012; Hana *et al.* 2010; Hana *et al.* 2012).

Below, we describe six user requirements for building a learner corpus: we take two perspectives, one is concerned with the attributes of the corpus and the other focuses on the corpus building procedure. We also explain which requirements are specific to building learner corpora and which can be applied to a wider spectrum of tasks concerned with annotating text corpora.

2.1 Requirements Related to the Corpus

The first requirement regarding the attributes of the corpus is concerned with the *extensibility of the corpus.* For any research project that aims at annotating textual data, establishing the final set of annotations beforehand is a difficult objective that is often difficult to meet. Sometimes, particular issues that are worth being annotated become evident only after the annotation task has already started. Hence, it should be possible to add annotation layers concurrently. To ensure the extensibility of the corpus, different processing phases need to add annotation levels in a well-structured and systematic fashion.

High-quality corpus annotations are the second requirement: they are the basis for precise analyses, particularly with regard to the task of learner language annotation or any annotation task where annotation levels are interdependent. Indeed, on each annotation level, it is important to minimize the number of annotation errors in order to avoid incorrect and misleading results. A low error rate is thus crucial for subsequent interdependent annotation levels. Because errors may escalate, incorrect annotations can lead to other incorrect annotations, and differences between individuals and groups may be artificially augmented.

The last requirement concerned with the attributes of the corpus refers to its usefulness. In order to avoid time-consuming, labour-intensive and error-prone manual activities, and to exploit the advantages of corpora, *searchable corpora* are required. The interface for corpus queries should enable sophisticated queries and allow for statistics on the result sets, taking into account different annotation levels.

2.2 Requirements Concerning the Corpus Building Procedure.

Annotating textual data usually implies annotating contiguous or non-contiguous strings and linking interdependent annotations into some structures. Even though the content of the annotations can be very different from one subject to another, from a technical perspective, these annotations are all similar. Requirements concerning the corpus building procedure are thus not specific to learner corpora.

The first aspect is concerned with the *efficiency of the corpus building procedure for manual work.* Even though automatic tools carry out many tasks within the corpus building procedure, manual work remains necessary in many cases. However, human resources are usually limited. Enhancing a manual task and avoiding repetitions and unnecessary work in general can have a noticeable impact on the size of the corpus, the quality of the annotations, and thus on the validity of the argument derived from it. Therefore, within the corpus building procedure routine manual work has to be integrated in an efficient way.

The second requirement of the corpus building procedure stresses *the process to be dynamically evaluable and adaptable*. Indeed, during the corpus building process it is helpful to monitor the quality and quantity of annotated data. Thereby, problems can be identified early on, and researchers can take appropriate action to solve such problems. In the context of learner corpus research for example, if inter-annotator agreement is low for error annotation, the procedure should allow the researcher to identify the situation and amend it. Since these issues are difficult to predict in advance, the procedure should allow for a regular evaluation of the corpus and for the necessary adaptations.

Finally, the corpus building procedure should be *formalised and reproducible* where (major) decisions are highlighted in order to ensure that identical objectives and design decisions for a corpus ensure identical results. In addition, the results obtained from the workflow should be reproducible by others, provided that they have knowledge of the objectives and design decisions, and access to the intermediate data.

3 Workflow for Building Learner Corpora

We will now present the abstract workflow; first, we describe the parts with which linguists interact, and then present the entire workflow managed by computer scientists.

3.1 Abstract Workflow – the Linguists' View

In the linguists' view (see Figure), the abstract workflow is organized in an iterative, user-oriented, manner. It is designed as comprising three phases and five components that can all rely on one or several tools. The acquisition phase comprises component (1) that covers the process of obtaining a digital representation of data. The annotation phase comprises components (2) and (3); they address manual annotation tasks and automatic annotation tasks respectively, with the help of Human Language Technology (HLT) tools. The exploration phase comprises component (4) for corpus exploration that enables the linguist to search for specific elements in context, while component (5) for corpus statistics allows for the computation of general numerical values over the corpus.

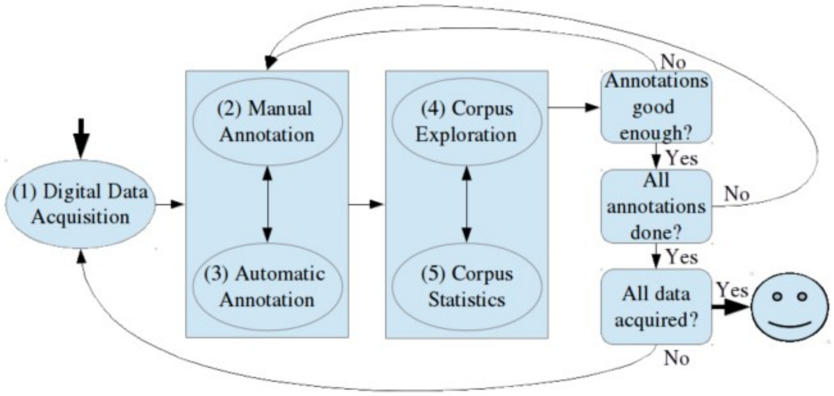

Figure 1. Linguists' view.

3.2 Abstract Workflow – the Computer Scientists' View

In the computer scientists' view (see Figure), the abstract workflow is organized in a dynamic, data-oriented manner. It is represented as comprising the previously mentioned five components along with two other components. Component (6) converts the corpus (or parts of the corpus) from and to necessary formats for data processing and exchange. Component (7) for data storage encodes the corpus in an exchangeable format that accommodates any type of annotation provided by the other components. Finally, we added an all-encompassing, non-mandatory tracking system that is of practical relevance: the change-log system (8). Its purpose is to track all relevant changes of both the corpus and the tools implementing the workflow.

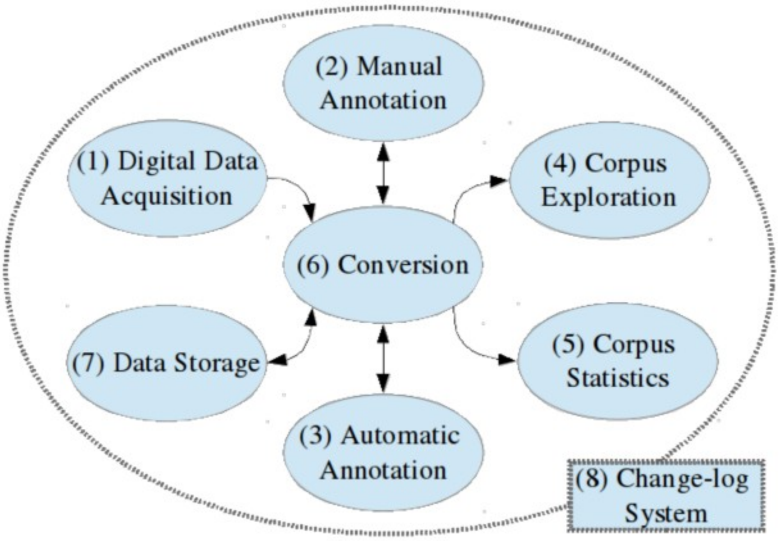

Figure 2. Computer scientists' view.

3.3 How Does the Abstract Workflow Relate to User Requirements?

In this section, we explain how the abstract workflow responds to the six user requirements defined in Section 2. We thus answer on an abstract level of reasoning[6].

First, regarding the *extensibility of the corpus*, additional annotation levels can be dynamically integrated by the manual annotation component (2) and the automatic annotation one (3). The integration of additional annotation levels is supported by the conversion component (6) and the data storage one (7). Both components (6, 7) should be able to deal with data from the process of digital data acquisition (1) and the annotation components (2, 3).

Second, the workflow refers to the requirement of *high-quality corpus annotations* by means of a well-defined data flow between all

[6] For fine-grained guidelines, we recommend the interested reader to consult Wynne (2005) or Garside *et al.* (1997).

components, and the possibility to repeat data processing steps to improve overall data quality.

Third, the aspect of creating *searchable corpora* is covered by the corpus exploration (4) and corpus statistics (5) components. Interoperability between the two components and the data from the process of digital data acquisition (1) and the annotation components (2, 3) is ensured by the conversion component (6).

The forth user requirement defined in *Section 2* refers to the *efficiency of the corpus building procedure* for manual work. This aspect is related to the implementation of the manual annotation component (2) and, when it includes human interaction, the digital data acquisition component (1). Unnecessary manual work can also be avoided by promptly detecting any issue in the performed annotations. Therefore, the corpus exploration (4) and the corpus statistics component (5) should give users the possibility to implement methods and tests to detect mistakes. In addition, semi-automatic annotation combine automatic annotations (3) with human resources needed for the digital data acquisition component (1) and the manual annotations (2) and reduce human effort. Beyond that, the optional change-log system can recover earlier versions of annotations, and thus help to retrieve and resume work quickly.

Fifth, the *dynamically evaluable and adaptable procedure* is ensured by the components for corpus exploration (4) and corpus statistics (5). Both components (4, 5) are used to perform quantitative and qualitative analyses. If predefined quality and quantity criteria are failed, annotations can be adapted with the help of the manual annotation (2) and the automatic annotation component (3). The conversion (6) and data storage component (7) facilitate this dynamic exchange process.

Finally, the description of the workflow in an abstract way is a *formalisation of the procedure* and adhering to it increases comprehensibility of the work and ensures *reproducibility* of the obtained results. In addition, the change-log system realises reproducibility on the level of tools and data versions.

4 Implementation of the Abstract Workflow

In the following section, we describe how we implemented the abstract workflow in order to build our learner corpus.

4.1 The KoKo Corpus

The KoKo project (2010-2014) is part of "Korpus Südtirol" (Abel/Anstein 2011, Anstein *et al.* 2011) – an initiative to collect, file and process South Tyrolean texts in order to make them available to the public and to document the use of written German in South Tyrol. The goal of the project is to investigate and describe the language skills of secondary-school pupils with German as L1 at the end of their school career by analysing authentic texts produced in classrooms. The corpus building process was guided by two linguistic goals, namely to describe language skills at the transition from secondary school to university, and to determine external factors that influence the distribution of language skills, such as sociolinguistic (gender, age), socio-economic, and language-related biographical factors (e.g. L1, preferred variety of German, reading and writing habits). 1,511 pupils from 85 classes and 66 schools participated in the project in May 2011 by writing a text and providing information about their background. Classes were sampled randomly using as strata the size of the cities in which the schools were located (small vs. medium vs. big) and the type of school (providing general education vs. education specific to a particular profession). 1,503 essays containing around 811,000 tokens were used for the KoKo corpus (version December 2012). The predominant part of the corpus (1319 essays with 716,405 tokens) consists of essays written by pupils with German as L1 (see Abel *et al.* 2014 for a detailed description of the corpus). We refer to this corpus as *L1 learner corpus*, since all essays were written by pupils (Abel/Glaznieks in press). All essays were manually transcribed, on-the-fly annotated, and automatically processed. In the next version of the corpus (end of 2014), lexical and grammatical annotations will be integrated.

4.2 Digital Data Acquisition and On-the-Fly Inline Annotation

For the transcription of the handwritten documents, we used the XML editor *XMLmind.*[7], a strictly validating near WYSIWYG editor, which can be used to create documents conforming to a custom schema. We combined the digital data acquisition (1) with the manual annotation component (2) by using *XMLmind* as a tool for transcription and on-the-fly inline annotations. During the transcription, the corpus was manually annotated with surface features of the text, such as graphical arrangement (header, paragraphs, emphasis, etc.) and self-corrections (insertions, deletions). For specific deviations of the standard written variety of German (orthographical errors, uncommon abbreviations), the correct versions were added on a separate level as target hypotheses (cf. Lüdeling *et al.* 2005). This ensures an annotation level that provides an error-free version of the corpus, which can be used to search for canonical word forms and improves the accuracy rate of automatic processing such as POS-tagging (cf. Glaznieks *et al.* 2014 for the evaluation of this process). All these annotations were done on-the-fly.

The main shortcomings in using *XMLmind* are inherent limitations of XML related to the cumbersome or impossible annotation of crossing hierarchies and of discontinued constituents, as well as problematic handling of multiple annotations of the same layer. For this reason, we chose to rely o a stand-off annotation format for further linguistic analyses.

4.3 Manual Non-Inline Annotation

To perform linguistic analyses, elaborated annotations are added in sequentially dependent and independent phases concerning new lexical and grammatical annotations as well as annotations for phenomena on the text level. These types of multi-layered annotations demand a stand-off annotation tool such as *Mmax2* (Müller/Strube 2006). *Mmax2* is well suited for annotating linguistic elements at the

[7] <http://www.xmlmind.com/xmleditor>

level of the text or discourse. It allows the definition of customized annotation schemes, and provides useful means to customize displays and user interaction.

4.4 Automatic Annotation

Regarding the automatic annotation component (3), as we are interested in tokenisation, sentence splitting, POS-tagging and lemmatisation, we chose to rely on the *TreeTagger* (Schmid 1994): a tool for annotating text with POS and lemma information that includes tokenisation and sentence splitting as pre-processing steps.

4.5 Corpus Exploration and Corpus Statistics

The corpus exploration (4) and the corpus statistics component (5) is covered by *ANNIS* (Zeldes *et al.* 2009). As explained on its website [8], *ANNIS* is an open source, versatile web-browser-based search and visualization architecture for complex multi-level linguistic corpora with diverse types of annotation. *ANNIS* addresses the need to visualise annotations covering various linguistic levels, such as syntax, semantics, morphology, prosody, referentiality, lexis and more. It also provides means to build highly elaborated queries.

4.6 Conversion

For the conversion component (6), we have commited ourself to the *SaltNPepper* framework (Zipser/Romary 2010). *SaltNPepper* is an open source project[9] developed to tackle an important issue in HLT research: there is a range of formats and no unified way of processing them. This issue derives from the fact that many expert tools for annotating and interpreting linguistic data have been developed for

[8] <http://www.sfb632.uni-potsdam.de/annis/>

[9] <https://korpling.german.hu-berlin.de/p/projects/saltnpepper/wiki/>

specific purposes. In order to fill that gap, a metamodel called *Salt*, which abstracts over linguistic data, and a pluggable universal converter framework called Pepper have been designed and implemented. It currently handles *PAULA*, *Mmax2* and a large variety of other formats.

4.7 Data Storage

The KoKo corpus is currently stored in *Mmax2* format. However, our Data Storage component (7) will be migrated to *PAULA* XML format (Dipper *et al.* 2007), which, just like the *Mmax2* format, is a stand-off one. Nevertheless, *PAULA* has originally been designed to be an exchange format for linguistic content and takes into account more recent technical developments. As such, it is able to represent a wider range of annotations more efficiently.

4.8 Change-Log

Many versioning systems could be used to implement the change-log system. Being widely adopted, *SVN* is a good mean for managing changes to documents and tools. Indeed, several clients are available on major operating systems; the HTTP transport layer can use well-established proxies and thus be integrated into corporate security configurations. Finally, some clients enable point-and-click interaction, which is an important feature for computer laymen.

5 Conclusion

In this paper we presented an abstract and generic workflow and detailed how we implemented it to build and annotate learner corpora. This workflow has been established while taking into account the needs of the users and is the result of an extensive collaboration between linguists, computational linguists, and computer scientists. For this reason, the workflow also refers to corpus exploration and corpus statistics as an integral part, and considers the way all parts can interact with each other. We also explained why its usefulness goes beyond learner corpora and can thus be of interest for a wider spectrum of subjects aiming at annotating textual data. We explained our reasoning by providing expert readers with contextual and decisional information so as to observe analogies and discrepancies with their own annotation objectives.

References

Abel, Andrea / Anstein, Stefanie 2011. Korpus Südtirol – Varietätenlinguistische Untersuchungen. In Abel, Andrea / Zanin, Renata (eds) *Korpora in Lehre und Forschung.* Bozen-Bolzano: University Press, 29-54.

Abel, Andrea / Glaznieks, Aivars in press. Wo Sprachkompetenzforschung auf Varietätenlinguistik trifft: Empirische Befunde aus dem Varietäten-Lernerkorpus “KoKo”. In Lenz, Alexandra / Glauninger, Manfred (eds) *Variation und Varietäten des Deutschen in Österreich – Theoretische und Empirische Perspektiven.* Frankfurt: Peter Lang.

Abel, Andrea / Glaznieks, Aivars / Nicolas, Lionel / Stemle, Egon 2014. KoKo: An L1 Learner Corpus for German. In Calzolari, Nicoletta *et al.* (eds) *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*. Reykjavik: ELRA.

Abdurachman, Adimihardja / Dariah, Ai / Mulyani, Anny 2008. Strategi dan teknologi pengelolaan lahan kering mendukung pengadaan pangan nasional. *Jurnal Litbang Pertanian* 27/2, 43-49.

Anstein, Stefanie / Oberhammer, Margit / Petrakis, Stefanos 2011. Korpus Südtirol – Aufbau und Abfrage. In Abel, Andrea / Zanin, Renata (eds) *Korpora in Lehre und Forschung.* Bozen-Bolzano: University Press, 15-28.

Díaz-Negrillo, Ana / Fernándes-Domínguez, Jesús 2006. Error Tagging Systems for Learner Corpora. *Spanish Journal of Applied Linguistics (RESLA)* 19, 83-102.

Dipper, Stefanie / Götze, Michael / Küssner, Uwe / Stede, Manfred 2007. Representing and Querying Standoff Xml. In Rehm, Georg *et al.* (eds) *Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007*. Tübingen: Narr, 337–346.

Garside, Roger / Leech, Geoffrey N. / McEnery, Tony 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London/New York: Longman.

Glaznieks, Aivars / Nicolas, Lionel / Stemle, Egon / Abel, Andrea / Lyding, Verena 2014. Establishing a Standardised Procedure for Building Learner Corpora. In Jantunen, Jarmo *et al.* (eds) *Proceedings of Learner Language, Learner Corpora - LLLC 2012*. Center for Applied Language Studies: University of Jyväskylä.

Granger, Sylviane 2003. Error-tagged Learner Corpora and Call: A Promising Synergy. *CALICO Journal* 20/3, 465–480.

Hana, Jirka / Rosen, Alexandr / Škodová, Svatava / Štindlová, Barbora 2010. Error-tagged Learner Corpus of Czech. In Ide, Nancy *et al.* (eds) *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*. Uppsala: The Association for Computational Linguistics, 11-19.

Hana, Jirka / Rosen, Alexandr / Stindlová, Barbora / Jäger, Petr 2012. Building a Learner Corpus. In Calzolari, Nicoletta *et al.* (eds) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*. Istanbul: ELRA, 3228-3232.

Hundt, Marianne 2008. Text Corpora. In Lüdeling, Anke / Kytö, Merja (eds) *Corpus Linguistics: An International Handbook*. Berlin: de Gruyter, 168-187.

Lüdeling, Anke / Walter, Maik / Kroymann, Emil / Adolphs, Peter 2005. Multi-level Error Annotation in Learner Corpora. In *Proceedings from the Corpus Linguistics Conference Series* 1 (1). Birmingham: University of Birmingham.

Müller, Christoph / Strube, Michael 2006. Multi-level Annotation of Linguistic Data with MMAX2. In Braun, Sabine *et al.* (eds) *Corpus Technology and Language Pedagogy. New Resources, New Tools, New Methods*. Frankfurt: Peter Lang, 197-214.

Nesselhauf, Nadja 2005. *Collocations in a Learner Corpus*. Amsterdam: John Benjamins.

Reznicek, Mark / Lüdeling, Anke / Hirschmann, Hagen (2013). Competing Target Hypotheses in the Falko Corpus: A Flexible Multi-layer Corpus Architecture. In Díaz-Negrillo, Ana et al. (eds) *Automatic Treatment and Analysis of Learner Corpus Data.* Amsterdam: John Benjamins, 101-124.

Schmid, Helmuth 1994. Probabilistic Part-of-speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*. Manchester.

Váradi, Tamás / Krauwer, Steven / Wittenburg, Peter / Wynne, Martin / Koskenniemi, Kimmo 2008. Clarin: Common Language Resources and Technology Infrastructure. In Calzolari, Nicoletta *et al.* (eds) *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech: ELRA, 1244-1248.

Wynne, Martin (ed) 2005. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books.

Zeldes, Amir / Ritz, Julia / Lüdeling, Anke, / Chiarcos, Christian 2009. Annis: A Search Tool for Multi-layer Annotated Corpora. In Mahlberg, Michaela *et al.* (eds) *Proceedings of Corpus Linguistics 2009.* Liverpool.

Zinsmeister, Heike / Breckle, Margit 2012. The Alesko Learner Corpus: Design–annotation–quantitative Analyses. In Schmidt, Thomas / Wörner, Kai (eds) *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam: John Benjamins, 71-96.

Zipser, Florian / Romary, Laurent 2010. A Model Oriented Approach to the Mapping of Annotation Formats Using Standards. In Calzolari, Nicoletta *et al.* (eds) *Proceedings of the Workshop on Language Resource and Language Technology Standards (LREC 2010)*. Valetta: ELRA, 7-18.