

LREC10-W4

**Language Resource and Language Technology
Standards – state of the art, emerging needs, and future
developments**

18th of May, 2010

WORKSHOP PROGRAMME

- 9 a.m. Introduction and overview: *Gerhard Budin, Thierry Declerck, Laurent Romary, Peter Wittenburg*
- 9.15 a.m. Setting data free – on two Open Content, data-sharing, TEI-related projects: *Piotr Banski*
- 9.40 a.m. A model oriented approach to the mapping of annotation formats using standards: *Florian S. Zipser, Laurent Romary*
- 10.05 a.m. Multilingual Lexical Support for the SEMbySEM Project: *Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem, Tarik Osswald, Vincent Herrmann*
- 10.30 a.m. Coffee break
- 11 a.m. ISOcat Definition of the National Corpus of the Polish Tagset: *Agnieszka Patejuk, Adm Przepiórkowski*
- 11.25 a.m. Referencing ISOcat Data Categories; The OWL and the ISOcat: Modelling Relations in and around the DCR: *Menzo Windhouwer, Marc Kemps-Snijders, Sue Ellen Wright*
- 11.50 a.m. annSem: Interoperable Application of ISO Standards in the Annotation and Interpretation of Multilingual Dialogue: *Kiyong Lee, Alex C. Fang, Jonathan J. Webster*
- 12.15 a.m. Lexicon standards: from de facto standard Toolbox MDF to ISO standard LMF: *Jacqueline Ringersma, Sebastian Drude, Marc Kemps-Snijders*
- 12.40 a.m. Grounding an Ontology of Linguistic Annotations in the Data Category Registry: *Christian Chiarcos*
- 1.05 p.m. Lunch break
- 2.30 p.m. Creating & Testing CLARIN Metadata Components: *Folkert de Vriend, Daan Broeder, Griet Depoorter, Laura van Eerten, Dieter van Uytvanck*
- 2.55 p.m. Climbing onto the shoulders of standards: TEI as annotation glue and metadata wrapper: *Piotr Banski, Adam Przepiórkowski*
- 3.20 p.m. Implementing a Variety of Linguistic Annotations through a Common Web-Service Interface: *Adam Funk, Ian Roberts, Wim Peters*
- 3.45 p.m. Preliminary proposal for a metadata scheme for the description of LRs: *Maria Gavrilidou, Penny Labropoulou, Elina Desipri, Stelios Piperidis*
- 4.10 p.m. Coffee break
- 4.30 p.m. Survey on the Use of XLIFF in Localisation Industry and Academia: *Dimitra Anastasiou*
- 4.55 p.m. Towards a Standard for Annotating Abstract Anaphora: *Stefanie Dipper, Heike Zinsmeister*
- 5.20 p.m. Towards a Standardized Linguistic Annotation of Fairy Tales: *Thierry Declerck, Kerstin Eckart, Piroska Lendvai, Laurent Romary, Thomas Zastrow*
- 5.45 p.m. Annotating a Historical Corpus of German: A Case Study: *Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt*
- 6.10 p.m. Discourse Structure Annotation: Creating Reference Corpora: *Alexandre Labadié, Patrice Enjalbert, Yann Mathet*
- 6.35 p.m. Linguistic Tool Development between Community Practices and Technology Standards: *Thomas Schmidt*
- 7.00 p.m. Are Chunking Standards Needed for Language Technology? Case Study From Hindi And Bangla: *Kalika Bali, Monojit Choudhury*
- 7.25 p.m. Wrap-up and Closure: *Gerhard Budin*

Workshop Organizers

Gerhard Budin, University of Vienna; Austrian Academy of Sciences
Thierry Declerck, DFKI
Laurent Romary, INRIA & HUB-IDSL
Peter Wittenburg, Max Planck Institute for Psycholinguistics, Nijmegen

Workshop Programme Committee

Nuria Bel, UPF, Barcelona
Harry Bunt, Tilburg University
Stelios Piperidis, ILSP, Athens
Maria Gavrilidou, ILSP, Athens
James Pustejovsky, Brandeis University
Dan Tufis, RACAI, Bucharest
Gerhard Budin, University of Vienna; Austrian Academy of Sciences
Lou Burnard, TGE-ADONIS (CNRS) and TEI
Nicoletta Calzolari, Istituto di Linguistica Computazionale del CNR, Pisa
Eric de la Clergerie, Team Alpage at INRIA
Thierry Declerck, DFKI GmbH
Gil Francopoulo, Tagmatica
Erhard Hinrichs, University of Tübingen
Nancy Ide, Vassar College
Marc Kems-Snijders, Max Planck Institute for Psycholinguistics, Nijmegen
Stelios Piperidis, ILSP, Athens
Laurent Romary, INRIA & HUB-IDSL
Florian Schiel, BAS, Munich

Table of Contents

Setting Data Free – On Two Open Content, Data-sharing, TEI-related Projects	1
<i>Piotr Banski, Institute of English Studies, University of Warsaw</i>	
A model oriented approach to the mapping of annotation formats using standards	7
<i>Florian S. Zipser, HUB-IDSL, Laurent Romary, INRIA-Gemo & HUB-IDSL</i>	
Multilingual Lexical Support for the SEMbySEM Project	19
<i>Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem, Tarik Osswald, Vincent Herrmann, Centre de Recherche INRIA Nancy Grand-Est, Nancy Université, LORI</i>	
ISocat Definition of the National Corpus of Polish Tagset	23
<i>Agnieszka Patejuk, Adam Przepiórkowski, Jagiellonian University, Cracow, University of Warsaw, Institute of Computer Science, Polish Academy of Sciences, Warsaw</i>	
Referencing ISocat Data Categories	27
<i>Menzo Windhouwer¹, Marc Kemps-Snijders¹, Sue Ellen Wright², ¹Max Planck Institute for Psycholinguistics, ²Kent State University Institute for Applied Linguistics</i>	
annSem: Interoperable Application of ISO Standards in the Annotation and Interpretation of Multilingual Dialogue	30
<i>Kiyong Lee¹, Alex C. Fang², Jonathan J. Webster², ¹Korea University, Seoul, South Korea, ²City University of Hong Kong, Hong Kong SAR, China</i>	
Lexicon standards: from de facto standard Toolbox MDF to ISO standard LMF	34
<i>Jacqueline Ringersma¹, Sebastian Drude² and Marc Kemps-Snijders¹, ¹Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, ²Goethe-Universität, Frankfurt, Germany</i>	
Grounding an Ontology of Linguistic Annotations in the Data Category Registry	37
<i>Christian Chiarcos, Universit`at Potsdam</i>	
Creating & Testing CLARIN Metadata Components	41
<i>Folkert de Vriend (1), Daan Broeder (2), Griet Depoorter (3), Laura van Eerten (3), Dieter van Uytvanck (2), 1) Meertens Institute, Amsterdam, 2) Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands, 3) Institute for Dutch Lexicology, Leiden, The Netherlands</i>	
Climbing onto the shoulders of standards: TEI as annotation glue and metadata wrapper	44
<i>Piotr Bański, Institute of English Studies Institute of Computer Science, Adam Przepiórkowski, University of Warsaw Polish Academy of Sciences</i>	
Implementing a Variety of Linguistic Annotations through a Common Web-Service Interface	46
<i>Adam Funk, Ian Roberts, Wim Peters, Department of Computer Science, University of Sheffield</i>	
Survey on the Use of XLIFF in Localisation Industry and Academia	50
<i>Dimitra Anastasiou, Centre for Next Generation Localisation, Localisation Research Centre, Department of Computer Science and Information Systems, University of Limerick</i>	
Towards a standard for annotating abstract anaphora	54

Stefanie Dipper, Institute of Linguistics, Bochum University, Heike Zinsmeister, Institute of Linguistics, Konstanz University

Towards a standardized linguistic annotation Standardised Linguistic Annotation of Fairy Tales 60

Thierry Declerck¹, Kerstin Eckart², Piroska Lendvai³, Laurent Romary⁴, Thomas Zastrow⁵, 1 DFKI GmbH, Language Technology Lab, 2 Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, Germany, 3 Research Institute for Linguistics, Hungarian Academy of Science, Budapest, Hungary, 4 INRIA, France & HUB-ISDL, Berlin, Germany, 5 Seminar für Sprachwissenschaft, Universität Tübingen

Annotating a historical corpus of German: A case study 64

Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt, University of Manchester

Linguistic tool development between community practices and technology standards 69

Thomas Schmidt, SFB 538 'Multilingualism' University of Hamburg

Setting data free – on two open-content, data-sharing, TEI-related projects

Piotr Bański

Institute of English Studies
University of Warsaw
Nowy Świat 4
00-497
Warszawa, Poland
pkbanski@uw.edu.pl

Abstract

The paper presents two related open-content projects hosting multilingual data and designed in such a way as to make it possible for the component resources to interact. These projects are FreeDict, hosting bilingual dictionaries, and the Open-Content Text Corpus – a multilingual corpus with a parallel component. Both are located at SourceForge.net, and make use of many of the advantages of this dissemination and collaboration platform. Both use the TEI Guidelines as the encoding format. The paper shows how the design of both projects facilitates standards-related research.

1. Introduction

The most recent ISO meeting at the City University in Hong Kong held a panel on data sharing and the notorious problems connected with it. One way to cope with what Doug Cooper – one of the conveners of the panel – called shy data (something you can “meet in public places, but can't take it home with you”, cf. (Cooper, 2010)) is to respect its shyness and expose only specific fragments via web APIs.

The two open-content projects reviewed here approach the problem of shy data from a different angle: by providing a platform that makes it possible for at least some data to come out in the light, with a life-long guarantee of freedom. The projects are cross-platform and multilingual; they are also designed in such a way as to make it possible for their component resources to interact. Both are part of SourceForge.net, with all the advantages of this web distribution platform, and both make use of different pieces of a single set of Text Encoding Initiative (TEI) Guidelines (TEI Consortium, 2010). The first project is FreeDict, hosting free bilingual dictionaries (<https://sourceforge.net/projects/freedict/>), and the other is the Open-Content Text Corpus (OCTC, <https://sourceforge.net/projects/octc/>).

Both projects offer a wide variety of data across numerous languages, provided in a common format across-the-board, thus forming a useful test-bed for research on standards and interoperability. FreeDict has 73 dictionaries at the moment (with new ones being encoded), while the OCTC currently has seeds (minimal subcorpora) for 55 languages. Apart the monolingual component, the OCTC also has an alignment part, in essence a project-wide parallel corpus component – this component is one of the planned points of synergy with FreeDict, because it has the potential for yielding bilingual material for the purpose of facilitating lexicographic study.

FreeDict has been part of GNU/Linux package repositories for years and has gone from TEI P3 in 2000 through TEI P4 and on towards TEI P5 (the migration is not yet complete), while the OCTC is a very fresh development, using the newest version of TEI P5. These projects have the potential to become a testing ground for various standards, practically as far away from “the armchair” as it gets – where virtually every design decision has consequences for the entire system, and thus it becomes important what exactly standards do to constrain the possible choices.

Below, we look in turn at the history of FreeDict and the Open-Content Text Corpus and then at what their points of synergy are and at what research opportunities they offer.

2. FreeDict

The FreeDict project started in the year 2000 to fulfil its founder's, Horst Eyermann's, vision of creating a repository of free bilingual dictionaries available on every platform via clients using the DICT protocol (cf. Goltzsch, 2000).¹ The timing was perfect: the DICT protocol (Faith and Martin, 1997) had been launched a couple of years before, SourceForge.net had already been a year old and a stable platform, and the Ergane project (<http://download.travlang.com/Ergane/>) had released under an open license its small but numerous dictionaries, the products of crossing several basic bilingual dictionaries of Esperanto, which was used as the bridge language for the creation of derivative lexicons. These lexicons were what Bański and Wójtowicz (2010) would later, in the context of the OCTC, call *seeds*: often, they

¹ The earliest 27 FreeDict databases (in the DICT format) that can be found at <ftp.dict.org> date from January 2000 (by September 2000, there were already 40 of them). At the beginning, FreeDict operated from freedict.de, only after a while moving to SourceForge as the primary site.

had only a few hundred headwords, but they *existed* and that was what mattered: it is much easier to extend and enhance an existing resource than to create one from scratch – as we shall see below, the same principle lies at the foundation of the OCTC.

There was one more fitting piece for the puzzle: the TEI had already gained recognition in the field of the Humanities and was the SGML application of choice to use for all kinds of encoding, from prose through verse and drama to entire collections of texts, culminating with its major flagship back then, the British National Corpus (<http://www.natcorp.ox.ac.uk/>). Very soon after its inception, FreeDict became a TEI project,² using the freshly revised final version of TEI P3 SGML (Sperberg-McQueen and Burnard, 1999). A couple of years later, the dictionaries were transduced into the then brand-new TEI P4 XML format (Sperberg-McQueen and Burnard, 2004) by the project's second administrator, Michael Bunk. Transduction into TEI P5 was initiated in 2008 by the present author, in the context of version 0.3 of the Swahili-English xFried/FreeDict dictionary, compiled by Beata Wójtowicz (cf. Bański and Wójtowicz, 2009a).

As is usual in projects which are fruit of passion and at the same time characterised by a high “bus factor”,³ a sudden break came after Michael Bunk's efforts at re-importing all Ergane dictionaries on the basis of fresh Travlang databases went down the drain when it was belatedly discovered that Travlang had changed the licensing of its databases (announced solely in Esperanto). All the affected dictionary sources had to be withdrawn from distribution, which was the beginning of a 3-year stagnation period, during which only Debian package maintenance began to function. The work on the Swahili-English dictionary and enabling FreeDict tools for TEI P5 marked the revival of the project.

While the present author gives himself a large part of the credit for stirring FreeDict from sleep, it has to be stressed that the little FreeDict community was easy to awake after some care was given to it, and some actual results demonstrated. This involved converting the FreeDict HOWTO to the MediaWiki format and creating every other part of the project wiki as well as not letting the community forget about the project, by getting the mailing archives free of spam and making sure that news

² Sadly, without the TEI knowing much about it. Had FreeDict been recognized within the TEI, the dictionary module of the Guidelines would probably not have to wait until mid-2007 for conversion into a format fully suited to encoding *electronic* as opposed to print dictionaries. It seems that what failed was one of the components of successful standards creation: community expectations and pressure. The FreeDict community became passive end-users of the TEI, with no attempt at becoming part of the TEI community. On the other hand, it has to be pointed out that most of the dictionaries were very simple glossaries that did not put many demands on the encoding format.

³ A developer's high bus factor means a high risk that the given project stalls after that developer gets “hit by a bus”, i.e. leaves or suspends their activity, for whatever reason, cf. (Collins-Sussman and Fitzpatrick, 2007).

of how things develop are published to the mailing list. The author's modification of the XSLT part of the FreeDict build system (which is one of Michael Bunk's invaluable contributions to the project) to support dynamic conversion of TEI P5 sources into DICT databases made it possible to continue the process of upgrading the sources, and getting the DICT project wiki⁴ installed by Rickard Faith and co-maintaining it helped secure FreeDict's position as a sister project to DICT rather than giving it the status of a distant satellite. This does not mean that DICT is treated instrumentally: the author co-maintains the vOoCabulum project (currently in the alpha phase), offering the first DICT client for Open Office, coded by Oleg Tsygany.⁵

Numerous open-content databases have been added to FreeDict over the years. Currently, the project has dictionaries for 73 language pairs, with several new ones in the works (some of them, as is the case of Welsh dictionaries, as complete replacements for the existing ones). Some of the existing databases are in the process of being further developed and enhanced – the sub-projects include John Derrington supplying gender information for the French-English dictionary or Kevin Donnelly supplying Arabic script spelling variants for the Swahili-English dictionary, cf. (Omar and Frankl, 1997) – in both cases, the new additions are planned to be automatically carried over to other relevant existing FreeDict resources. A change of the versioning system from CVS to SVN eliminated a long-standing awkwardness concerning the architecture of the repository, and it is now easier to ensure communication between the repository and the static pages located at freedict.org. Currently, there is hardly a week without an SVN commit; thanks to Kęstutis Biliūnas, Debian GNU/Linux packages are released regularly.⁶ The FreeDict dictionaries begin to be part of scholarly research – for example, De Pauw *et al.* (2009) included the Swahili-English dictionary as one of the four resources that they used when evaluating bilingual coverage on a parallel Swahili-English corpus.

3. Open-Content Text Corpus

The OCTC started as a generalization of an idea to build a parallel Polish-Swahili corpus, coupled by a reflection on the current state of affairs in African language technology and similar areas, where data are hard to come by not only because they haven't been produced, but because they are closed by various more or less reasonable licensing restrictions. The project is a continuation of its founders' attempts to increase the degree of collaboration among

⁴ The DICT project (<http://dict.org/>), led by Aleksey Cheusov and Rickard Faith, also distributes its deliverables via SourceForge. Its wiki documentation is located at <http://dict.org/w/>.

⁵ The prototype of vOoCabulum can be downloaded from <http://extensions.services.openoffice.org/en/project/voocabulum>

⁶ See <http://qa.debian.org/popcon.php?package=freedict> for some statistics regarding the use of FreeDict packages in Debian GNU/Linux.

African language technology projects as well as to make sure that even small language resources that would otherwise be discarded as not worthy of dissemination can be presented to others for extension and enhancement, as seeds that, produced by one, may be tended by others in subsequent projects. This is not only an attempt to rescue linguistic data in areas where data is scarce, but also an attempt to avoid wasting the effort and expertise of those who produce it, by creating a platform where everyone can donate as much time as they can afford to and still get credit and satisfaction for it.

And that does not need to mean “satisfaction but no money”. Firstly, many small resources would not ever be turned into a monetary gain because they would be deemed too small or not valuable enough, or selling them would entail too many bureaucratic problems. Secondly, many resources are produced under various forms of licenses with a non-commercial restriction on their use, even though their commercial variants are not offered. This is believed to be “the right thing” to do for academics, and is a plague of academic projects that, by not making their deliverables truly free (as in “free to do whatever you wish with it”), close them to further re-use in the much more popular open-content/open-source applications. Thirdly, as Koster and Gradmann (2004) show, it is open-content and open-source strategies that benefit scholarship in the long run, also in the sense of potential financial gains. Fourthly, in the academic world, it is sometimes more important and profitable to be able to claim credit for a job well done than to sell to a few specialist centers or libraries.

The OCTC consists of two major parts, monolingual and parallel, with the former grouping opportunistic corpora for individual languages and the latter holding documents that remotely point to selected parts of the former in order to create aligned texts (for some details of implementation see (Bański and Wójtowicz, 2010)). Note that this maximizes the gain from storing language resources in the monolingual part: the OCTC is not just the sum of its individual monolingual subcorpora – these subcorpora form a potential basis for the parallel part. Furthermore, single texts from the monolingual part are not meant to be enriched with linguistic annotations directly – they are stored in separate documents, which can be accompanied by annotation documents of various kinds, arranged in layers, the first of which is the layer of segmentation that separates running text into tokens and gives each token an identifier, which in turn can be referenced by other layers of annotation that contain e.g. morphosyntactic, syntactic, semantic or discourse information. There can be more than a single instance of the given layer of annotation, which makes it possible to e.g. compare the layers containing POS tagging, etc. There is a separate component for corpus tools, currently containing some general-purpose XSLT scripts.

A system like that is called a stand-off system (Ide and Romary, 2007) and the architecture of the OCTC is an extension of the architecture of the National Corpus of

Polish that the present author proposed and that was further refined with the participation of the NCP team, cf. Przepiórkowski and Bański (forthcoming).

The OCTC contains seeds (minimal corpora) for 55 languages at the time of writing, as well as a demonstration of a Polish-Swahili aligned document. It is placed under version control in the SourceForge Subversion repository and is accompanied by a wiki, a bulletin board, a mailing list for Subversion commits and a general mailing list (there can be as many mailing lists as there are subprojects). It also has a bug/patch/issue tracker and access to the SourceForge file release system.

As can be gleaned from the above description, the research possibilities that the OCTC offers, both for data collection and manipulation, are plentiful, and the entire well-tested infrastructure is in place. Individual researchers or teams can co-maintain individual monolingual subcorpora or concentrate on the parallel component. Tools can be tested and produced for handling individual subcorpora as well as the entire corpus (the existing XSLT scripts for indexing and whitespace normalization are of the latter kind). Research advancement can be traced in the public mailing archives and the version control system. The system inherently enhances the possibilities for peer code- and content-review as well as for cross-project collaboration. Research on the parallel part of the OCTC can produce lexical resources for FreeDict.

4. FreeDict and OCTC vis-à-vis standards and interoperability

Both FreeDict and the OCTC use the TEI Guidelines for XML encoding. FreeDict has been through three subsequent versions of the TEI, and the OCTC is created according to the latest version, TEI P5. The adoption of a single standard for multiple multilingual resources both puts this standard to the test of versatility and interoperability, and testifies to its strength. This does not mean, however, adopting a single rigid schema for all dictionaries and all parts of the OCTC. Quite on the contrary: FreeDict aims towards three-level conformance, and the OCTC has separate schemas for the source text, the individual annotation layers, and for the aligned part.

The idea of multiple stages of conformance derives from the Corpus Encoding Standard (CES and later XCES, cf. Ide *et al.*, 2000), which defined the lowest level of conformance for source texts produced by automatic encoding, and two other levels for subsequent refinements of markup. A similar approach was suggested by Bański and Wójtowicz (2009b) for FreeDict: a fairly loose lowest conformance level for mass-derived glossaries, with two more refined levels: for semi-automatic encoding and for hand-crafted dictionaries that border on lexical databases. The OCTC defines a loose and a strict schema for source texts and time will tell whether this duality is useful. In each case, it is possible to define the given level as a subset of the general TEI schema thanks to the ODD (TEI “literate config file”, written in the TEI itself, cf. (Burnard

and Rahtz, 2004)). Maintaining three ODDs, each of which defines a subset of the previous one, raises maintenance problems that could be overcome with the idea of ODD inheritance, recently brought up by Laurent Romary (Romary, 2009). Especially FreeDict may be useful for testing this new development.⁷

Both projects should expect to reach “critical mass”, both in the number of lexicons/subcorpora and in their content: for FreeDict, this means the ability to concatenate (cross) dictionaries, forming new ones; for OCTC, this means attaining a stage at which useful aligned subcorpora may begin to be created. To achieve this, mass conversion of data may in some cases need to be performed, which raises problems of its own, concerning the quality of the result given the lack of human supervision. Again, representation standards become crucial here, and again, implementing standard-conformance levels may help a lot.

FreeDict has already shown the need for standardization not only in terms of the encoding schema, but also in terms of data content describing grammatical or lexical properties – it is not a sensible approach to attempt to impose a single set of data categories (parts of speech, agreement features or even usage-note categories and values) on all the current and future dictionary maintainers, the more so that some resources are third-party donations. It is more likely to expect that each dictionary may declare the equivalence of the categories it uses with a set of standard categories. This is where the ISO Data Category Registry (Kemps-Snijders *et al.*, 2008) or a linguistic ontology such as GOLD (Farrar and Langendoen, 2003) may come to rescue (it is actually tempting to subject both systems to the test of applicability and scalability).

Another, related point of interaction with standards is the issue of interoperability of tools. It is to be expected that e.g. a memory-based tagger used for one subcorpus of the OCTC will be tried on another, in order to create an annotation layer that can later be used for comparisons with annotations created by other tools. Similarly with sentencers, aligners or any other kind of tools designed for large-scale applications – a resource such as the OCTC may turn out to be valuable for creators of such tools, who will be presented with a unified format of the input data from multiple languages and multiple text types. A similar challenge that FreeDict presents is in the area of dictionary concatenation, where entries from two bilingual dictionaries have to be satisfactorily aligned in order to produce a third one. A tool that is planned to be deployed for both projects and whose scalability and flexibility may thus be put to test is eXist, a native XML database (<http://www.exist-db.org/>).

⁷ Sharing parts of XML across ODDs can be partially achieved by XML Inclusions, but these can only help with maintenance to a certain extent and would create an appearance of complexity where what is needed is simplicity, also in the general outlook.

The adoption of the TEI for both projects raises questions concerning other standards. One of the long-term goals of FreeDict is creation of transducers into three popular dictionary interchange formats: LIFT (Lexicon Interchange Format, <http://code.google.com/p/lift-standard/>), OLIF (Open Lexicon Interchange Format, <http://www.olif.net/>) and ISO LMF (Lexical Markup Framework, <http://www.lexicalmarkupframework.org/>, ISO:24613). One of the questions that arise in this context is whether to prepare three separate transducers, in effect using the FreeDict TEI schema as the pivot, or whether to nominate the LMF as the hub, which, given that the three-level FreeDict conformance is still to be instituted, may be a much more sensible solution. A similar question concerns interoperability of the TEI against other corpus-encoding formats, such as the ISO LAF family of standards (Ide and Romary, 2007) or PAULA (Chiaros *et al.*, 2008).

Some of the problems arising in the context of implementing the TEI for complex corpus encoding are mentioned in Bański and Przepiórkowski (2009). An attempt to implement the TEI Guidelines for complex corpus encoding, especially with running, untokenized text at the bottom of the annotation hierarchy, reveals the still insufficient level of stand-off support in the current version of the standard (cf. (Bański (2010)) for more detailed discussion), from something that is independent of the specification, namely the lack of tools to support the TEI-defined XPointer extensions, to what can be called incomplete re-absorption of XCES innovations into the modern TEI.⁸ On the other hand, Przepiórkowski and Bański (2010) show why the TEI is nevertheless a sensible choice of an encoding standard for large, multi-layer linguistic resources.

If the TIGER-XML format for treebank annotation (Mengel and Lezius, 2000) becomes re-cast as part of the TEI Guidelines, as Laurent Romary (p.c.) suggests, the OCTC will become an ideal alpha-testing environment for it as well.

5. Conclusion

The OCTC is designed to be to a large degree synergistic with FreeDict in terms of the data and the encoding standard used by both projects. The corpus may feed the dictionary project, which in turn may support various annotation tools for the corpus.

Both are either gradually approaching full TEI P5 conformance (FreeDict) or have been designed to be TEI-P5-conformant from the beginning (OCTC), which increases their potential for interoperability with various tools and with other formats.

⁸ For a sketch of the problem and a suggestion of a solution, see e.g. <http://listserv.brown.edu/archives/cgi-bin/wa?A2=ind1003&L=TEI-L&T=0&F=&S=&P=35185> or <http://listserv.brown.edu/archives/cgi-bin/wa?A2=ind1003&L=TEI-L&T=0&F=&S=&P=36329>.

The fact that both projects potentially involve numerous research teams who may have the same aims for different languages or whose aims may be complementary while targeting the same set of data, creates the potential for a positive feedback loop, leading to improved co-operation and better results. The existence of many research groups within each project (this is especially true for the OCTC) may reduce the currently high “bus factor” and thus increase the resilience of individual subprojects.

FreeDict and the Open-Content Text Corpus offer a wealth of opportunities for research and for testing interoperability among standards on many levels. Both have the potential to become serious players in the standardization game in their respective categories.

6. Acknowledgements

I would like to take this opportunity to thank Michael Bunk for his trust, and for being an absolutely benevolent non-tyrant and a far-sighted project leader.

Neither my participation in FreeDict nor the creation of the OCTC would have taken place without Beata Wójtowicz. I would like to thank Beata for encouragement at every step and the incentive to first get involved with FreeDict and then, in a way by extension, to implement the idea of a Swahili corpus in a much broader and much more sensible context than we initially mused.

I would also like to thank Adam Przepiórkowski for turning my interest back to corpus architecture in the late 2008, after a long break from the IPI PAN corpus (where my stand-off fantasies did not pass the test of reality), by hiring me to clean up a certain mess and suggest something sensible in its place. Without my work on the NCP architecture, the OCTC would not be created in the shape it is in now.

Lastly, I am grateful to the SourceForge admin teams past and present for creating, maintaining, and constantly improving the site that means so much to all of us in open-source/open-content movement.

7. References

Bański, P. (2010). Why TEI stand-off annotation doesn't quite work. Manuscript, University of Warsaw.

Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009, Singapore.

Bański, P., Wójtowicz, B. (2009a). A Repository of Free Lexical Resources for African Languages: The Project and the Method. In De Pauw, G., de Schryver, G-M., Levin, L. (Eds). *Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages (AfLaT)*, 31 March 2009, Athens, Greece. Greece: Association for Computational Linguistics., pp. 89–95.

Bański, P., Wójtowicz, B. (2009b). Freedict: an Open

Source repository of TEI-encoded bilingual dictionaries. Presentation given at the Conference and Members' Meeting of the TEI Consortium, Nov. 9–15, 2009, University of Michigan, Ann Arbor. Available from <http://www.lib.umich.edu/spo/teimeeting09/files/Banski+Wojtowicz-TEIMM-presentation.pdf>.

Bański, P., Wójtowicz, B. (2010). Open-Content Text Corpus for African languages. In the proceedings of the LREC 2010 Workshop on Language Technologies for African Languages (AfLaT).

Burnard, L., Rahtz, S. (2004). RelaxNG with Son of ODD. Presented at Extreme Markup Languages 2004, Montréal, Québec. Available from <http://conferences.idealliance.org/extreme/html/2004/Burnard01/EML2004Burnard01.html>

Chiarcos, Ch., Dipper, S., Götze, M., Leser, U., Lüdeling, A., Ritz, J., Stede, M. (2008). A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In *Traitement Automatique des Langues* 49(2), pp. 271–293.

Collins-Sussman, B., Fitzpatrick, B. (2007). How to protect your Open-Source project from poisonous people. Google Tech Talks January 25, 2007. Available from <http://video.google.com/videoplay?docid=-4216011961522818645>

Cooper, D. (2010). When nice people don't share: 'Shy' data, web APIs, and beyond. Position presentation for panel discussion by the same title, organized during the 2nd International Conference on Global Interoperability for Language Resources (ICGL) at the City University of Hong Kong in January 2010. Available at <http://sealang.net/archives/ICGL2010.pdf>.

De Pauw, G., de Schryver, G-M., Wagacha, P.W. (2009) A Corpus-based Survey of Four Electronic Swahili–English Bilingual Dictionaries. *Lexikos* 19, 340–352.

Faith, R., Martin, B. (1997). A Dictionary Server Protocol, Network Working Group Request for Comments #2229. <http://www.ietf.org/rfc/rfc2229.txt>. Accessed 29 January 2010.

Farrar, S., Langendoen, D.T. (2003). A linguistic ontology for the Semantic Web. *GLOT International*. 7 (3), pp. 97–100.

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Standard for Linguistic Corpora. Proceedings of the Second Language Resources and Evaluation Conference (LREC), Athens, Greece, pp. 825–830.

Goltzsch, P. (2000). Internet Babylon. Available at <http://www.heise.de/tp/r4/artikel/5/5927/1.html> (accessed on April 4, 2010)

Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hensen, H., Minker, W. (eds.), *Evaluation of Text and Speech Systems*, Springer, pp. 263–84.

Kemps-Snijders, M., M.A. Windhouwer, P. Wittenburg and S.E. Wright. (2008). ISOcat: Corraling Data Categories in the Wild. In Calzolari, N., K. Choukri, B. Maegaard, J. Mariani, J. Odjik, S. Piperidis, D. Tapias

- (Eds.). *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*. Marrakech, Morocco: European Language Resources Association (ELRA), pp. 887--891.
- Koster, C.H.A., Gradmann, S. (2004). The language belongs to the People!. In *Proceedings of LREC'04*. Lisbon, Portugal.
- Mengel, A., Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 121--126.
- Omar, Y.A., Frankl, P. J. L. (1997). An Historical Review of the Arabic Rendering of Swahili Together with Proposals for the Development of a Swahili Writing System in Arabic Script (Based on the Swahili of Mombasa). *Journal of the Royal Asiatic Society of Great Britain & Ireland (Third Series)*, 7 , pp. 55--71, doi:10.1017/S1356186300008312
- Przepiórkowski, A., Bański, P. (2010). TEI P5 as a text encoding standard for multilevel corpus annotation. In Fang, A.C., Ide, N. and J. Webster (eds). *Language Resources and Global Interoperability. The Second International Conference on Global Interoperability for Language Resources (ICGL2010)*. Hong Kong: City University of Hong Kong, pp. 133--142.
- Przepiórkowski, A., Bański, P. (forthcoming). XML Text Interchange Format in the National Corpus of Polish. In S. Goźdz-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.
- Romary, L. (2009). ODD as a generic specification platform. Presentation given at the Conference and Members' Meeting of the TEI Consortium, Nov. 9--15, 2009, University of Michigan, Ann Arbor. Available from <http://www.lib.umich.edu/spo/teimeeting09/files/FutureOfODD2.pptx>
- Sperberg-McQueen, C.M., Burnard, L. (Eds.) (1999). TEI P3: Guidelines for Electronic Text Encoding and Interchange. Revised Reprint, Oxford, May 1999. Available at <http://www.tei-c.org/Vault/GL/P3/index.htm> (accessed on April 4, 2010).
- Sperberg-McQueen, C.M., Burnard, L. (Eds.) (2004). TEI P4: Guidelines for Electronic Text Encoding and Interchange. XML-compatible edition. TEI Consortium. Available at <http://www.tei-c.org/release/doc/tei-p4-doc/html/> (accessed on April 4, 2010).
- TEI Consortium (Eds.) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.6.0. Last updated on February 12th 2010. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>

A model oriented approach to the mapping of annotation formats using standards

Florian S. Zipser, HUB-IDSL

Laurent Romary, INRIA-Gemo & HUB-IDSL

Abstract: In this paper, we present, Salt, a framework for mapping heterogeneous linguistic annotation formats into each other using a model-based approach, i.e. independently of the actual formats in which the corresponding linguistic data is being expressed. As we describe the underlying concept of this framework, we identify how it echoes ongoing standardisation activities within ISO committee TC 37/SC 4, and in particular, the possible conceptual equivalences with ISO CD 24612 (LAF) combined with ISO 24610-1 (FSR), as well as the possible role of the central data category registry (ISOCat), currently under deployment. We thus show the adequacy of our methodology and its capacity to integrate a wide range of possible linguistic annotation models.

1 The issue of mapping and the current standardization landscape

1.1 The importance of mapping when managing heterogeneous language resources

Over the years, the linguistic research community has seen the development of a wide variety of tools ([schmidt02], [lezius02] and [zeldes09] specifically targeted at the extraction, representation and analysis of many different phenomena. For example, a tool such as the search tool Tiger Search [lezius02] was primarily developed for syntactic analysis, whereas a tool like the annotation tool EXMARaLDA [schmidt02] covers discourse analysis. Most of these tools are built around the use of one specific format, which was developed specifically for this tool and for a certain type of analysis. The focus of such formats has in general been to supply all necessary information for the tool to proceed in an efficient manner (limited coverage, optimized representation). Because of their specialization, these formats are difficult to reuse in other contexts for which they were not intended.

Providing standardized formats is one of the possible answers to this issue. One of the benefits of a standardized format can be the interoperability between tools or the keeping of existing data for some years and being assured these will also be legible in the future. At present, however, there is very few linguistic data that is represented in standardized formats. As long as the tools do not have a direct import or export for standardized formats, it would be necessary to map the used formats from or to standardized formats. As a consequence, defining mappings between existing formats and more standardized representations represents an important component of any further development relying on the use of external data.

1.2 Difficulties related to mapping formats

Existing standards such as LAF [iso24612], MAF [iso24611] or SynAF [iso24615] mainly focus on the provision of persistent models and formats to provide a stable descriptive framework for linguistic information. In particular, they do not address the mapping between themselves and the already used formats, with the exception of ISO 16642 (TMF), which provide an explicit mapping framework across terminological data formats. It is thus necessary to define appropriate solutions to get existing data into standard formats by 1) defining a conceptual mapping between them and 2) having a concrete implementation which realizes the mapping thus defined.

Most standards, because they basically aim at providing an interchange format, include a strong

technical part to specify, for instance, how they can be implemented in a given XML representation or a relational database structure. In this context, it is quite often the case that the very existence of such format definitions, with the associated technical constraints, impact on the actual expressive power of the corresponding model. For example, an attribute value of an XML element cannot contain additional mark-up. To create a mapping, one therefore has to consider both the conceptual mapping and the technical realizations. This requires the implementer to have a good level of understanding of the underlying format description, for instance expressed by means of a schema language (DTD, RelaxNG or W3C schema) in the case of XML. Covering both aspects makes the mapping generation extremely complex, for anyone who just wants to focus on the underlying linguistic concepts or constraints.

A conceptual mapping has to cover two aspects. First, there has to be a mapping for each structural object like the representation of tokens or representations of primary data. Second, the mapping has to regard semantic mappings for data categories. In this paper we want to propose an approach to structural mappings via a model like Salt (introduced in section 2) and a semantic mapping using the ISOCat [kemps09] system (shown in section 3).

1.3 A model based approach to mapping

A solution for clarifying the actual interdependence between conceptual and technical levels is to adopt a model-based approach as for instance in MDA ([miller03]). The idea is to separate the meaning of data (the model layer) from their representation (the format layer, cf. figure 1) especially in the case of persistence constraints. When a separation between a conceptual model and a persistent format is made, one can avoid taking care of persistence issues and focus on processing data through the elicitation of a mapping between models. For example, a specialist in the linguistic domain, can create or describe a mapping between two morphosyntactic tagsets, leaving it for a further stage, and a more technical expertise, to implement a mapping for the underlying formats.

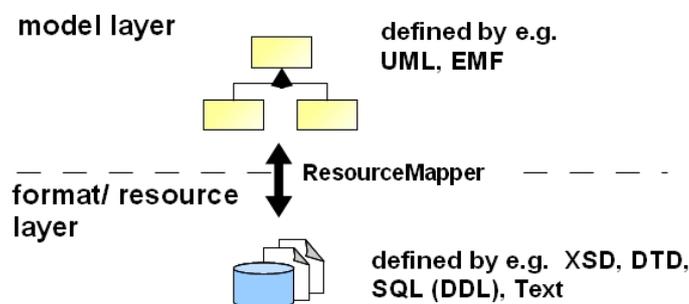


figure 1: correlation between the model and the resource or format layer

Model-based development frameworks such as MDA [miller03] or EMF [steinberg09] support 1) a graphical representation for models and 2) a generation of processable object models for further work (in terms of an API for instance). The graphical representation of a model can be used as a communication base between linguists and technical experts. The generated API can be used for implementing tools working with the model, such as an annotation tool or, in our case, a converter. The EMF framework that we use also generates a persistent format based on XML. This generated format is called a resource and can be exchanged with other formats, by re-implementing the “ResourceMapper” in figure 1.

Figure 2 shows an example of a resource mapping between the format description of Tiger XML [menge00] and the corresponding model.

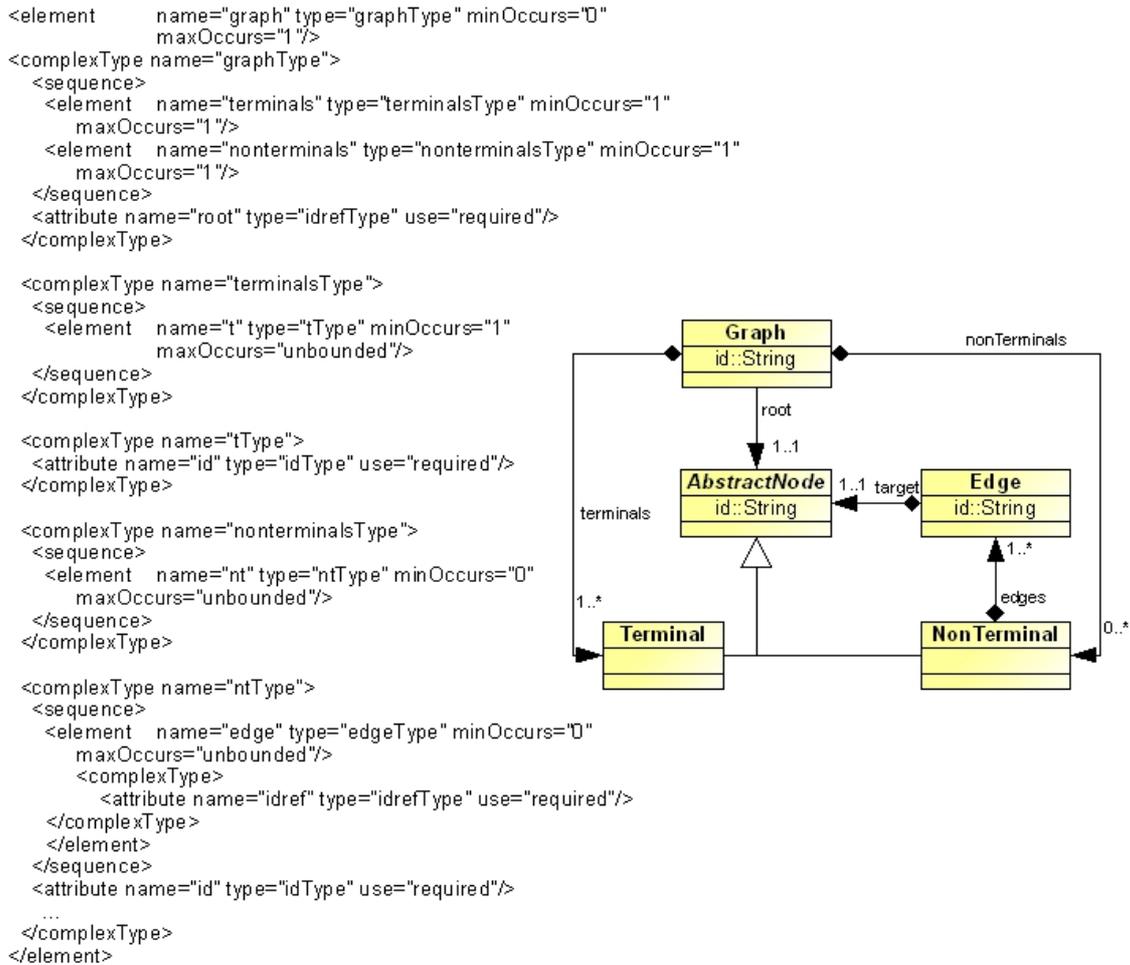


figure 2: on the left side: an excerpt of the xsd description of Tiger XML [menge100]; on the right side: the correlated model for this excerpt in UML-like notation

1.4 Same but different – shared advantages with a format based approach

As pointed out in [ide07], the number of mappings can be reduced by mapping data over a common format, or in this case a common model. Instead of creating n^2-n mappings to map n models to each other in the case of 1:1 mappings, the number of mappings via a common model decreases to $2n$ mappings. In this paper we want to follow this approach. Figure 3 shows this approach using a common model for mappings simultaneously to the mapping of data via a pivot format defined by LAF/GrAF [ide07].

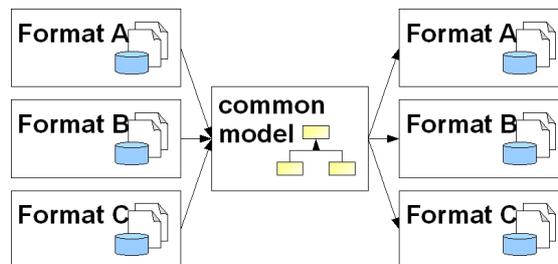


figure 3: common model as middleware between formats to import and to export

In the remaining sections of this paper, we present the main characteristics of the framework that we have developed to implement such a perspective by the comparisons of models.

2 An overview of Salt and its relation to LAF

2.1 Basic principles

Salt is a common model for linguistic annotated data. This model defines a conceptual abstraction of data, independent of persistence techniques. This means that one can use Salt as an object representation of data. This allows us to process data with respect to the object model, with no prejudice with respect to the actual storage (or linearisation) format, be it XML or a relational database, in which the data will be represented.

Salt was influenced by several existing linguistic formats such as EXMARaLDA [schmidt02] TigerXML [mengel00] and above all PAULA [dipper05]. Salt unifies the concepts of these formats e.g. common timeline, multiple layers of annotation etc. and represents them in a common model. Salt is a model for representing the underlying organization of linguistic data, and as such, does not take into consideration their underlying semantics. Furthermore, Salt is independent of specific linguistic theories or analyse.

2.2 The underlying graph structure of Salt

Salt is based upon a directed, labeled and layerable graph structure model. The model contains a graph structure component, which contains 1) a set of nodes or vertices, 2) a set of directed edges, 3) a set of layers, which embraces a set of nodes and edges and 4) a set of labels, used to label a node, an edge, a layer or a label. This means that a label can be used as a recursive structure and therefore enables the possibility to annotate an annotation.

The Salt model is a refinement of the general graph structure model, in effort to apply Salt to linguistic needs e.g. primary data, tokens, relations, annotations and so on. But every element in Salt is still an element of a general graph structure model and can be processed with general graph structure methods e.g. traversing. Figure 4 shows this refinement on the basis of some elements of Salt. Here one can see, for example that a textual representation of primary data (STextualDS) is still a node. Although nodes get a more linguistic meaning, nodes and relations are just placeholders for annotations.

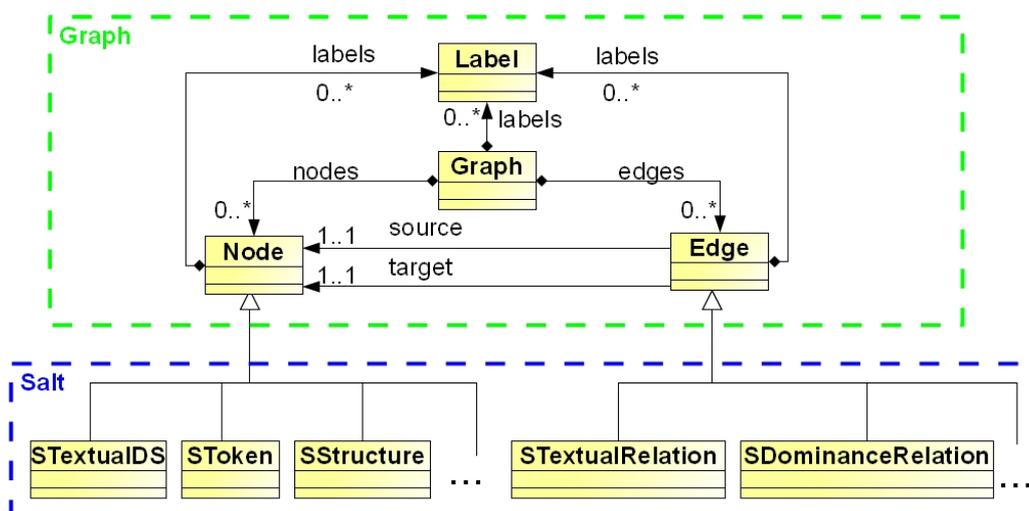


figure 4: excerpt of the refinement between the graph structure model and the common model Salt. The elements *STextualDS*, *SToken* and *SStructure* are still nodes and the elements *STextualRelation* and *SDominanceRelation* are still edges.

We used the element `STextualDS` as a model representation of the primary data. Therefore this element contains a `String` representation of the primary data. Continuous spans of the primary data can be addressed by using the node type `SToken` and the edge type `STextualRelation`. A node of type `SToken` represents the tokenization of the primary data and is the basis for further structural objects and annotation. To relate such a token node with the primary data node, an edge of type `STextualRelation` can be created. This edge contains the start and end position of the referred span. To create hierarchical annotation graphs for example in case of syntactic analysis one can use nodes of type `SStructure` and relate them via edges of type `SDominanceRelation` to one or more nodes of type `SToken` or `SStructure`. Figure 5 shows an example of data represented in the Salt model. Salt offers further types of nodes and edges to create annotation graphs which are not shown in figure 4 and not mentioned here. For example it contains further edge types to realize different relations between nodes.

2.3 Salt and LAF

The graph-based approach is very similar to the one taken in the linguistic annotation framework (LAF, [iso24612]). Our objective is indeed to let Salt and LAF be identified as complementary tools on their specific abstraction level. LAF can be used as a persistence and exchange format for data whereas Salt can be used 1) as a conceptual abstraction which can be easily understood by non technical experts 2) as basis for a processable API. To do so we need a mapping between the Salt object model and the XML-representation of LAF (the GrAF format [ide07]). Although both GrAF and Salt are very similar, there are some core differences between them. One is the way they deal with edges: as opposed to GrAF, Salt allows edges to be annotated. A second difference lies in the referencing to primary text: In Salt there is a relation (`STextualRelation`) between a token node (`SToken`) and the primary data node (`STextualDS`), whereas in GrAF there is just one span concept for both. A third difference is that in Salt a copy of primary data is part of the model in terms of a node (see `SText1` in figure 5). The first two differences can be handled as shown in figure 5. The figure shows a Salt model representation and an XML representation according to GrAF. The third difference can be handled by storing primary data in a separate document or by loading primary data from a text file into the Salt model.

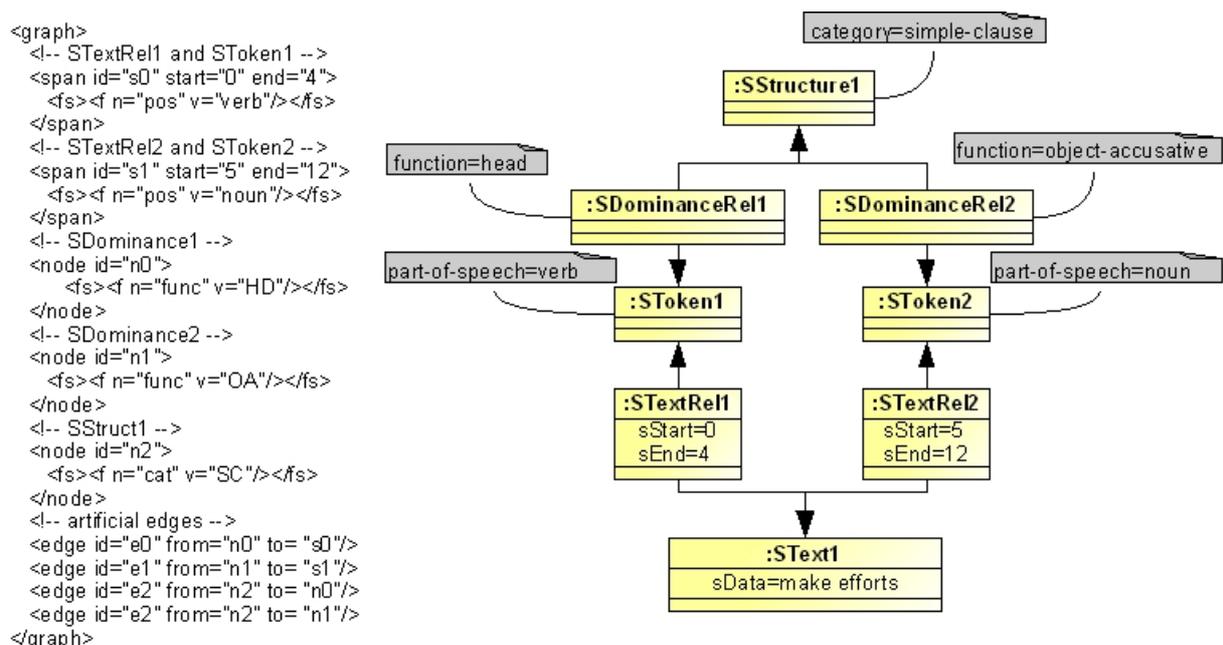


figure 5: on the left side: an example corpus represented in the format GrAF (the primary data "make efforts" can be stored in a external file); on the right side: the same example represented in a Salt model

Moreover, we developed Salt to be able to take into account some important phenomena that LAF would not handle in its current state:

- The representation of a common timeline (e.g. for audio-video and dialog data such as those produced by EXMERaLDA)
- The management of higher level structures, in particular for the implementation of the notion of corpus (in particular, embedded corpus or sub-corpus relations)
- The typing of annotations e.g. as textual, numeric or more complex values.

3 The relation of Salt to ISOCat and FSR

3.1 The need to consider the meaning of annotations

As already mentioned, Salt does not deal with the semantics of annotations. Similarly to GrAF [ide07] annotations are understood as an attribute-value pair, the entries of which do not have an interpretable meaning for the system. In the case of converting data, the meaning could be important. For example some formats like TreeTagger [schmid94] need to have part-of-speech or lemma annotations. If these data were mapped in a format or a model which handles annotations as attribute-value pair the meaning of the annotations would get lost. For example a problem occurs if one tries to map to a format which needs specific annotations, because the data for a part-of-speech annotation appear in different forms: pos=verb, POS=verb, PartOfSpeech=verb. Because of different surface representations of the attribute name for part-of-speech, annotations cannot be unified by the system. The system does not know that all these names actually have the same meaning.

It is therefore essential to have a possibility for unifying syntactical representations, or rather to make clear the meaning of such a representation. In this respect, ISOCat [kemps09] supplies the possibility of a central reference for elementary descriptors (data points) to which data model can refer. The meaning of a data point can be defined by the experts of the domain, whereas a system just has to check equality of references to the data points. In the case of part-of-speech annotations in format data, we can for instance use the reference <http://www.isocat.org/datcat/DC-396>, which in turn provides the actual definition of this data point as stored in ISOCat (“A category assigned to a word based on its grammatical and semantic properties”).

Indeed, many formats which support attribute-value pairs for representing annotations only support String values e.g. TigerXML [menge100], PAULA [dipper05] etc. . This means that a reference can be stored, but not necessarily interpreted as a reference. Thus we have to mark the data type of an attribute as well as of a value as references. In Salt there is a possibility for marking this, therefore we now take a closer look at an annotation. In figure 5 annotations are shown as simple attribute-value pairs beside the nodes and edges. Annotations are slightly more complex than what figure 5 shows. The annotation shown in figure 6 is the same as in figure 5 beside the node “SToken1” first as a String representation and second as a representation using ISOCat references.

:anno1		:anno1	
name	=part-of-speech	name	= http://www.isocat.org/datcat/DC-396
nameType	=String	nameType	=URI
value	=verb	value	= http://www.isocat.org/datcat/DC-1424
valueType	=String	valueType	=URI

figure 6: on the left side: an annotation using simple string values as an attribute-value pair; on the right side: an annotation using references to ISOCat

3.2 Salt and FSR

As in GrAF, Salt nodes can be multiply annotated. For example, one can attach a part-of-speech and a lemma annotation to one node. But actually in Salt, there is no grouping function for annotations. Every annotation stands alone for itself. GrAF uses feature-structures (FSR) defined by ISO [iso24610-1] and used in the TEI P5 guidelines [burnard08]. For example some features can be grouped to a “morpho-syntactic annotation”. GrAF does not yet support naming or typing of a feature structure as TEI describes (@type attribute in the <fs> element). Figure 7 shows an example taken from the TEI P5 guidelines for representing a grouping of annotations via feature structures.

```

<fs type="morpho-syntax">
  <f name="case">
    <symbol value="accusative"/>
  </f>
  <f name="gender">
    <symbol value="feminine"/>
  </f>
  <f name="number">
    <symbol value="plural"/>
  </f>
</fs>

```

figure 7: sample from the TEI P5 guidelines of grouping features by using feature structures

In Salt you can either represent the given three annotations as independent annotations, or you can represent them by using recursive annotations (means creating annotations on annotations). The second way simulates such a grouping as feature structures achieve. Both ways are shown in figure 8.

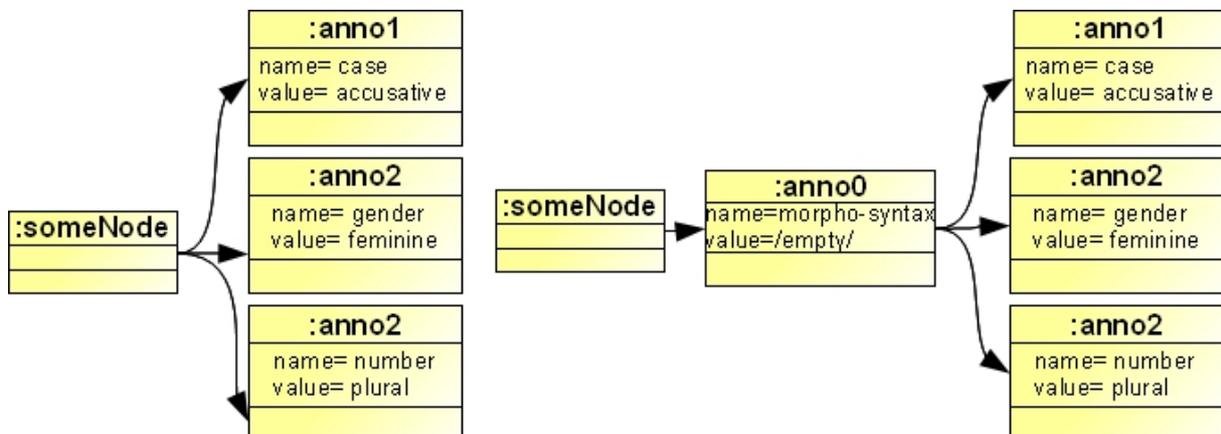


figure 8: on the left side: the sample from figure 7 without grouping; on the right side the same sample with grouping via the recursive structure of annotations in Salt

In addition to the types URI and String, we introduce additional types for annotation names and annotation values. On the one hand, there are additional simple types such as numeric (for numeric data), float, and boolean. On the other hand, there is a complex type called object. This complex type is defined in a flexible way, so that a value of this type can be any kind of object. As a consequence, it is possible to define a complex structure as a collection with conditions on their elements in terms of alternations or negations as mentioned in TEI [burnard08] chapter 18.

The main element of Salt is a SaltProject. This element contains the corpus structure. The corpus

structure is a tree, which defines super- and sub-corpus relations between corpora. A corpus contains one or more documents in which the primary data, tokens, hierarchical structures annotations and so on can be found. Additionally to the corpus structure a SaltProject can also contain a library graph structure. This graph structure consists of nodes, which define data points as well as ISOCat do. These nodes can be referenced by URI's using the scheme *salt*. A library structure can therefore be modeled as a graph structure. For example the STTS tagset [schiller95] for German part-of-speech can be described as shown in figure 9.

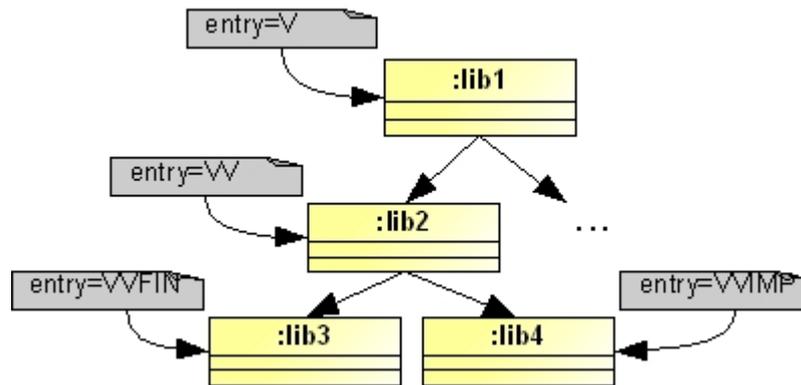


figure 9: an excerpt of the STTS tagset represented in the library graph structure of Salt. This example shows how refinements between entries can be handled.

Figure 9 contains the nodes „lib1“, „lib2“, „lib3“ and „lib4“ as data points. These nodes can be annotated with annotations like entry, for the tagset name, a description, which explains the usage of this tag and an example, which shows the usage in a specific case. The relations between the nodes “lib1”, “lib2”, “lib3” and “lib4” can be interpreted as a refinement. This means, that the node “lib3” which defines the entry “VVFİN”¹ is also of type “V”². Further we propose a grouping relation to group the represented entries of several nodes under one node. This way of grouping is similar to the grouping function of the “fvLib” element of the FSR. Figure 10 shows the grouping mechanism by using a grouping relation.

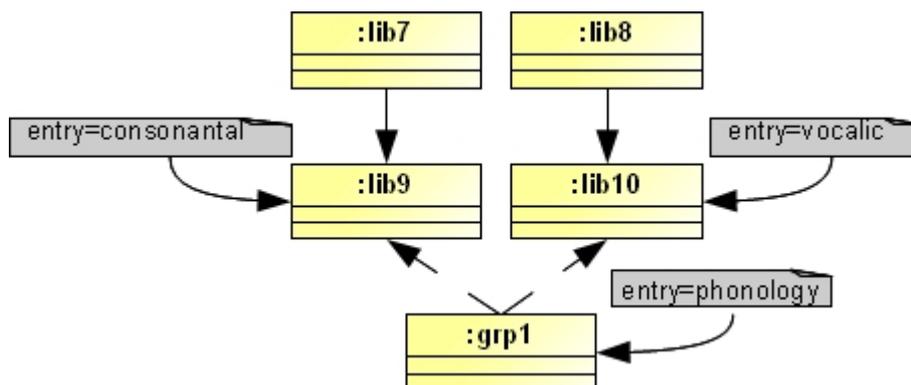


figure 10: grouping mechanism to group several data points e.g. consonantal and vocalic to one data point. This example is an excerpt from the TEI P5 guidelines (chapter 18).

The dashed arrow of figure 10 shows such a grouping relation, whereas the continuous arrow shows a refinement. The node “grp1” groups the nodes “lib3” and “lib4”, and also stands for the entry “consonantal” as well as for the entry “vocalic”.

1 tag for a finite full verb in the STTS

2 general tag prefix for a verb in the STTS

To use a data point such as a document structure, one can use the attribute value of an annotation typed as URI. The value then contains a URI entry. This URI starts with the scheme name *salt*, followed by the path which is the identifier for the library structure and the fragment which is the identifier of a node of the library structure graph. This node either can be a node standing for such an entry as “lib3” for example, or a grouping node as “grp1”. Figure 11 shows the referencing mechanism for annotations using a URI value for a reference to the library graph structure.

:anno1		:anno2	
name=	pos	name=	phonology
nameType=	URI	nameType=	URI
value=	salt://featLib#lib3	value=	salt://featLib#grp1
valueType=	URI	valueType=	URI

figure 11: on the left side: an annotation which references a library entry; on the right side: an annotation which references a grouping.

4 Validation (using Salt in Pepper)

4.1 What is Pepper?

To validate the Salt model, we define Pepper, a Salt based converter framework. This framework was developed to convert data from x formats into y different formats, with a constant number of mapping steps. As shown in figure 3 Salt and Pepper makes it possible to convert several formats via a common model into each other with a minimal number of needed mappings and just two steps.

Pepper thus forms a use case for Salt with which we can check whether Salt can represent data from several formats. Furthermore, it is possible to trace information losses during conversion operations. For example one can convert a corpus from format A into Salt and then export the data back to format A. The import and export can then be compared for losses.

4.2 How does Pepper work?

Pepper can be separated into three components: 1) the framework, 2) a common instance of the Salt model and 3) mappers to several formats. Figure 12 shows the general architecture of Pepper and the relations of the components.

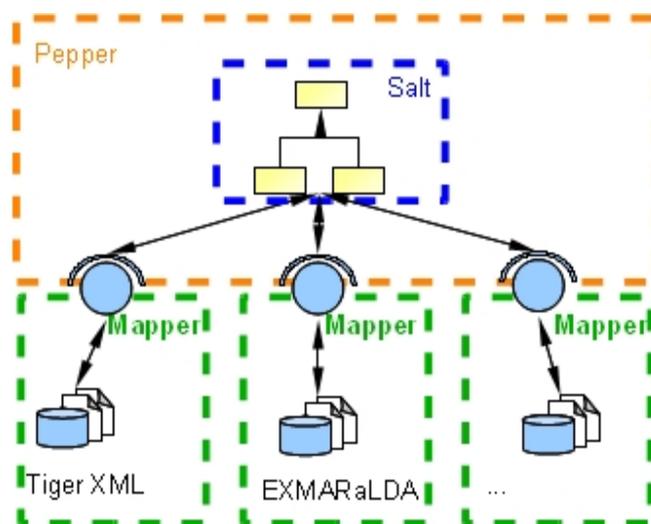


figure 12: architecture of the converter framework Pepper and the relation between the components of Pepper

The framework controls the given workflow, for example importing a corpus from TigerXML [menge100] and exporting it to the EXMARaLDA format [schmidt02] via Salt. It creates a common instance of the Salt model, which can be used by mappers to import, or export their data. A mapper has to realize a mapping from an external format to the Salt instance, a mapping from the Salt instance to an external format, or both. A mapper is implemented in terms of a module, which can be plugged into the framework. Such a module can either be 1) an import module, 2) a manipulation module or 3) an export module.

- 1) An import module maps data from external formats to a Salt instance.
- 2) A manipulation module can manipulate a Salt instance, for example by changing the names of an annotation to upper case or to ISOCat data points.
- 3) An export module maps data from a Salt instance to an external format.

The example in figure 13 describes a mapping for an import module between TigerXML [menge100] and Salt, with respect to the persistence and the model layer. The mapping can be described as

map: TigerXML \rightarrow Salt

and can be done in two ways.

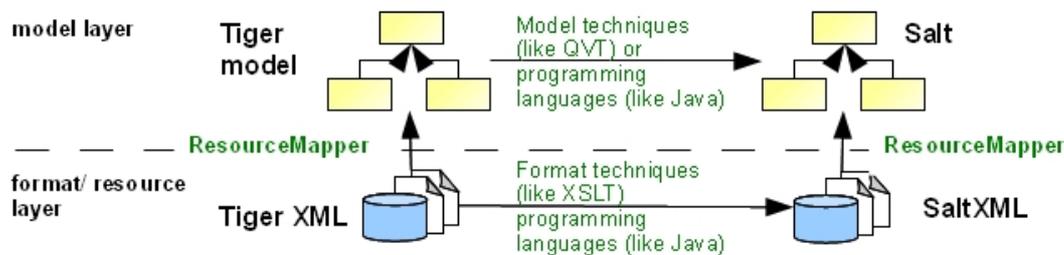


figure 13: two different mechanisms to map data from the format Tiger XML to a Salt model (the first way via Tiger XML \rightarrow Tiger model \rightarrow Salt, the second way via Tiger XML \rightarrow SaltXML \rightarrow Salt).

Both ways address different technical mechanisms, the first one handles the mapping via format techniques with no abstraction between persistence layer and conceptual layer and the second one handles a conceptual mapping on the conceptual layer. For the second way we need to have a mapping between model and format for example to the format developer and in creating a mapping, which can be done by another person or team. Figure 14 shows the representations of the three stages of the first way: 1) the data in the origin format Tiger XML, 2) the data in a Tiger model representation and 3) the data in a Salt model representation.

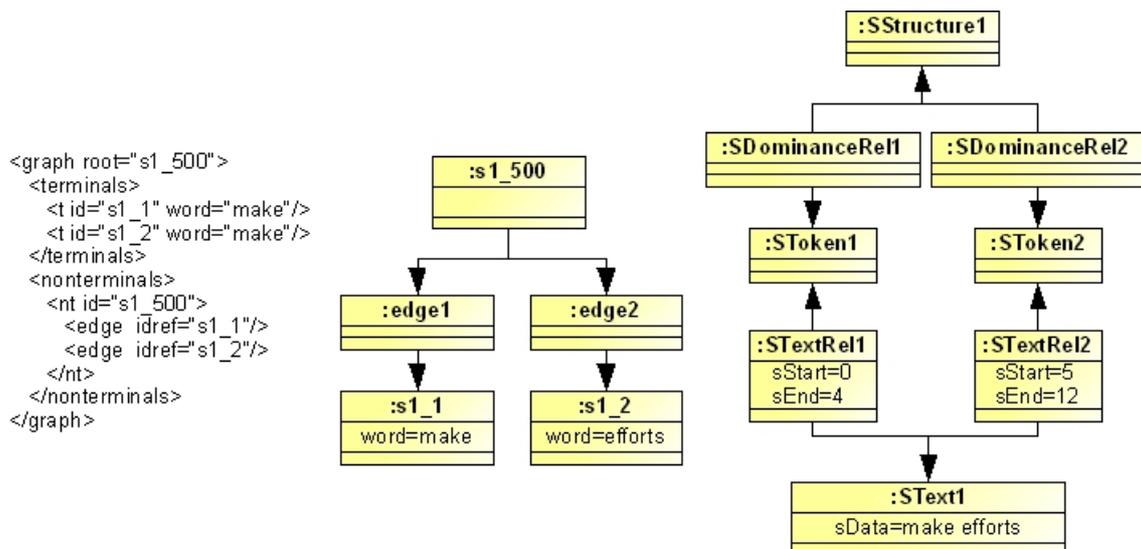


figure 14: on the left side: an example of data in the Tiger XML format; in the middle: the same example in the model of Tiger XML; on the right side: also the same data in a Salt model

Model based developing of mappings on a conceptual layer becomes much easier especially if a usable API also exists. In the case of using programming languages, one has a well-defined, context specific object model to map with, instead of working with a general model, e.g. a DOM model.

4.3 Evaluation

There are two ways To attach GrAF to Salt: 1) GrAF can be treated as an actual format, therefore a mapper can be implemented and plugged into the Pepper framework or 2) GrAF can be used as a native resource of Salt. GrAF then gains the same status as the automatically generated format Salt-XML³. The second approach makes Salt and GrAF become closer and will melt them as a unit consisting of a format and a model. This would be helpful for both, Salt gets a standardized format for persisting data and GrAF gets a processable API with a defined model.

Both ways need an isomorphic mapping, the general way of mapping was shown in section 2, but some losses remain in terms of the element types of Salt. As shown above, Salt elements such as edges have types: for example they can define a dominance, a coverage relation and further more between nodes. GrAF includes a type attribute for nodes, but no defined value domain, so the mapping from Salt to LAF/GrAF can be made, but the way back would be difficult, if the attribute does not contain Salt-types.

Another loss also occurs for the recursive structure of annotations in Salt. As long as features in GrAF [ide07] cannot contain feature structures, an annotation of an annotation is not possible.

The current implementation of Pepper covers modules for the mapping between Salt and the formats EXMARaLDA [schmidt02], TigerXML [menge100], TreeTagger [schmid94], PAULA [dipper05] and relANNIS (the relational format of the search and visualization system for multilevel linguistic corpora: ANNIS [zeides09]). These data can be represented in Salt. To support other formats it must be discovered if the structure of Salt is powerful enough to cover them, or if Salt has to be expanded.

³ automatically generated by the modeling framework used, EMF [steinberg09], as mentioned in section 1

5 References

- [burnard08] Burnard, L. and Bauman, S., editors (2008). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Oxford. <http://www.tei-c.org/Guidelines/P5/>.
- [dipper05] Stefanie Dipper (2005) XML-based Stand-off Representation and Exploitation of Multi-Level Linguistic Annotation. In: Eckstein R, Tolksdorf R (Hrsg.) Berliner XML Tage.
- [ide07] Nancy Ide, Keith Suderman (2007) GrAF: A Graph-based Format for Linguistic Annotations. In: Proceedings of the Linguistic Annotation Workshop, Prague, Czech Republic.
- [iso24610-1] ISO:24610-1 (2005). Language resource management – feature structures – part 1: Feature structure representation. ISO/DIS 24610-1, 2005-10-20.
- [iso24611] ISO:24611 (2005). Language resource management – Morphosyntactic annotation framework (MAF). ISO/CD 24611, ISO TC 37/SC 4 document N225 of 2005-10-15.
- [iso24612] ISO:24612 (2008). Language resource management – Linguistic annotation framework. ISO/WD 24612[2], ISO TC 37/SC 4 document N463 rev00 of 2008-05-12.
- [iso24615] ISO:24615 (2009). Language resource management – Syntactic annotation framework (SynAF). ISO/CD 24615, ISO TC 37/SC 4 document N421 of 2009-01-30.
- [kemps09] Kemps-Snijders, M., Windhouwer, M., Wittenburg, P., & Wright, S. E. (2009). ISocat: Remodeling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies (IJMSO)*, 4(4), 261-276.
- [lezius02] Wolfgang Lezius (2002) Ein Suchwerkzeug für syntaktisch annotierte Textkorpora. [http://www.ims.uni-](http://www.ims.uni-stuttgart.de/projekte/corplex/paper/lezius/diss/)
- [mengel00] Andreas Mengel, Wolfgang Lezius (2000) An XML-based encoding format for syntactically annotated corpora. In: Proceedings of the Second International Conference on Language Resources and Engineering (LREC 2000), Athen. [stuttgart.de/projekte/corplex/paper/lezius/diss/](http://www.stuttgart.de/projekte/corplex/paper/lezius/diss/).
- [miller03] J. Miller, J. Mukerji (2003) MDA Guide Version 1.0.1. Object Management Group (OMG).
- [schiller95] A. Schiller, S. Teufel, and C. Thielen (1995). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Universität Stuttgart and Universität Tübingen.
- [schmid94] Helmut Schmid (1994) Probabilistic Part-of-Speech Tagging Using Decision Trees. In: Proceedings of International Conference on New Methods in Language Processing.
- [schmidt02] Thomas Schmidt (2002) EXMARaLDA - ein System zur Diskurstranskription auf dem Computer. Arbeiten zur Mehrsprachigkeit, Folge B 34:1 ff. <http://www.exmaralda.org/files/AZM.pdf>.
- [steinberg09] David Steinberg, Frank Budinsky, Marcelo Paternostro and Ed Merks (2009) EMF: Eclipse Modeling Framework 2.0. Addison-Wesley Professional.
- [zeldes09] Amir Zeldes, Julia Ritz, Anke Lüdeling, Christian Chiarcos (2009) ANNIS: A Search Tool for Multi-Layer Annotated Corpora. In: Proceedings of Corpus Linguistics 2009, Liverpool, July 20-23, 2009.

Multilingual Lexical Support for the SEMbySEM Project

Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem, Tarik Osswald, Vincent Herrmann

Centre de Recherche INRIA Nancy Grand-Est, Nancy Université, LORIA
{ingrid.falk, samuel.cruz-lara, nadia.bellalem, tarik.osswald, vincent.herrman}@loria.fr

Abstract

In this paper we describe how multilingual linguistic and lexical information is stored and accessed within the framework of the SEMbySEM project. The SEMbySEM project is dedicated to defining tools and standards for the supervision and management of complex and dynamic systems by using a semantic abstract representation. To provide the project with multilingual linguistic and lexical information and in order to achieve an appropriate, flexible, reusable and accurate representation of this information we chose the Linguistic Information Repository representation (Peters et al., 2009) model and adapted it to our needs. In this paper we discuss the rationale for this choice, describe its implementation and also the employment of other linguistic standards.

1. The SEMbySEM project.

1.1. Description

The SEMbySEM project¹ aims at providing a framework for universal sensors management using semantic representations. A detailed description can be found in (Brunner et al., 2009b), here we give a brief overview and concentrate on the aspects related to language and linguistic information.

A sensor system supervises and manages the data coming from various sensors with varying technical specifications and placed on various objects. The sensors collect and transmit data and a sensor management system must make sense of and visualise this data.

To achieve this the SEMbySEM system will be organised in a three layered architecture (Fig. 1). The interaction with the sensors (registering and processing events from the sensors) is done in the basic layer, the *Façade Layer*. The information from the sensors is unified and processed and may then trigger an update of the semantic model of the system. The semantic model together with a rule system make up the middle layer, the *Core Layer*. End-users connect to the system through the top layer, the *Visualisation Layer*. They have access to tailored view points designed by expert users and HMI experts through which the data from the semantic model is displayed.

From the linguistic point of view the relevant modules are the *Core* and the *Visualisation Layer*.

The semantic representation is based on a business-oriented model, the **MicroConcept** model (Brunner et al., 2009a). It was decided against OWL and Description Logic which are habitually employed to represent semantic information in this setting (Brunner et al., 2009b) because of its being difficult to handle by business users and its deficiencies in expressing some specific business needs. However, the **MicroConcept** model also uses existing standards and it is therefore possible to leverage standards and methods developed for OWL as for *eg.* the lexicalisation tools to be discussed later in this paper.

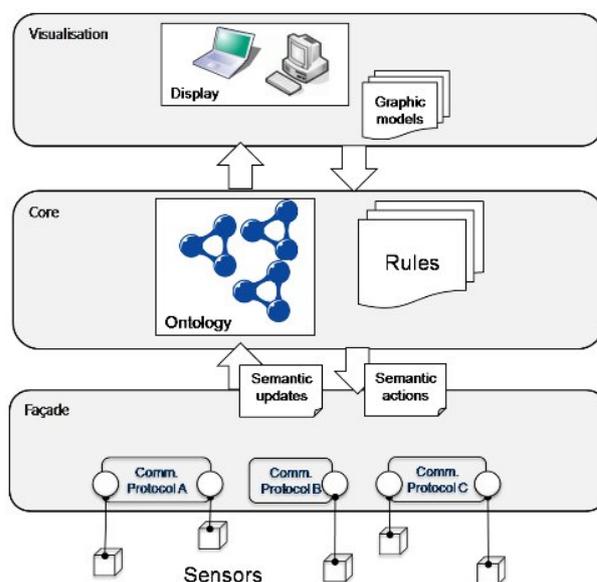


Figure 1: SEMbySEM system architecture.

1.2. Linguistic needs in SEMbySEM

SEMbySEM needs (multilingual) linguistic information:

- on the conceptual level, the *Core Layer* (cf. 2.),
- on the GUI or visualisation level (cf. 3.) .

2. Linguistic information on the conceptual level.

The most common way to provide linguistic and lexical information to a conceptualisation is by using the `rdfs:label` and `rdfs:comment` tags with `xml:lang` attributes. However, this approach, albeit presumably sufficiently expressive for SEMbySEM needs

- is only suitable when there are one to one equivalents for the ontology elements in each language and can not account for any conceptualisation mismatches,
- is not user friendly,
- is hardly reusable.

¹SEMbySEM (<http://www.sembysem.org>) is a research project within the European ITEA2 programme (<http://www.itea2.org/>). It started June 2008 and will end December 2010.

We identified two recent models for representing linguistic information for ontologies: LIR, (Peters et al., 2009) and LexInfo, (Buitelaar et al., 2009). In both models the linguistic information is stored in a lexical ontology and elements of the domain conceptual representation are linked via an ontology relation (or property) to concepts of the lexical ontology. Both lexical ontologies use LMF (the Lexical Markup Framework, (The LMF Working Group, 2008)) as building blocks. However the resulting ontological structures differ not only from a syntactic point of view but also semantically: LexInfo rather emphasises the representation of properties (relations) and in particular the syntax ↔ semantics interface whereas LIR adopts a more traditional lexicographic position, describing translation (partial) equivalents and linguistic phenomena as synonymy. We finally opted for LIR as representation model for SEMbySEM for the following reasons:

- LIR’s lexicographic point of view seemed to fit the SEMbySEM needs better,
- the project seemed more advanced and tested than LexInfo,
- LIR’s alignment with other linguistic and lexicographic standards in addition to LMF: TMX, MLIF and XLIFF.

However, due to time constraints and also to LIR’s complexity, the model finally integrated into SEMbySEM had to be further simplified.

2.1. Structure of the lexical ontology.

The structure of the simplified ontology is shown in Figure 2. We (re)used the following ontology classes from LIR:

LexicalEntry is a language-related group of lexicalizations. This is the entry point for the whole data base, for ontology concepts which are linked to the database via lexical entry ids.

Lexicalization is a way to write a specific lexical entry. One lexical entry may have several lexicalizations, which are mainly distinguished by their variance type (basic form, acronym, abbreviation, etc.). A lexical entry is linked to its language of origin.

Language is a table used for representing the several languages managed by the database. This table is necessary for a good database maintenance.

Sense represents the sense of a lexical entry, given by its definition. The sense is not linked to any language, so that several lexical entries may have the same sense.

Definition is a textual description representing the sense of the lexical entry. One sense may have several definitions.

Source contains information about the source of a definition.

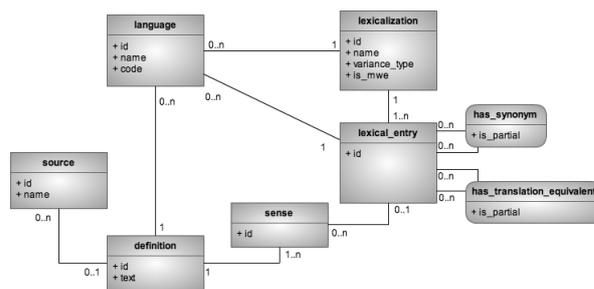


Figure 2: Merise diagram for a simplified LIR-like database. Cardinalities have to be read as Merise cardinalities (unlike UML cardinalities for example)

We also represented (with relations between the tables) the following relations (properties): belongsToLanguage, hasSynonym, hasTranslation, hasLexicalization, hasSense, hasDefinition, hasSource. It is possible to express that the synonymy or translation relations hold only partly via the *is_partial* attribute (datatype property). We only reduced the classes and properties of LIR in number, we did not change their semantics.

The classes Lexicalization, Sense, Definition and Source and the relations hasSynonym, hasTranslation and belongsToLanguage are equivalent to elements of the LMF model, whereas the LMF LexicalEntry is more general than the LIR LexicalEntry. The LIR LexicalEntry and hasTranslation are also equivalent to MLIF components.

We will illustrate the model and some of its benefits and limitations in a few examples. First consider the concept *wagon* as it appears in the following snippet:

```
<smc:Concept rdf:about="&sembysem;#AssetTracking/Wagon"/>
```

This concept is linked to the LIR lexical ontology as shown in the following:

```
<smc:Concept
  rdf:about="&sembysem;#AssetTracking/Wagon">
  <lir:hasLexicalEntry rdf:resource="&lexo;#LE-1-En"
    xml:lang="eng"/>
  <lir:hasLexicalEntry rdf:resource="&lexo;#LE-1-Fr"
    xml:lang="fr"/>
</smc:Concept>
```

Here the *hasLexicalEntry* elements point to the elements with identifier *LE-1-En* and *LE-1-Fr* in the lexical ontology. These could be represented as follows in the lexical ontology:

```
<lir:LexicalEntry rdf:about="&lexo;#LE-1-En">
  <lir:partOfSpeech>noun</lir:partOfSpeech>
  <lir:belongsToLanguage rdf:resource="&lexo;#English"/>
  <lir:hasLexicalization rdf:resource="&lexo;#Lex-1-En"/>
  <lir:hasSense rdf:resource="&lexo;#Sense-1-En"/>
  <lir:hasTranslation rdf:resource="&lexonto;#LE-1-Fr"/>
</lir:LexicalEntry>
```

This lexical entry describes the word *wagon*, it states that it is an English noun. It’s sense is given in a *Sense* instance of the lexical ontology by a definition. The actual lexicalization (the word string *wagon*) together with possibly other linguistic and terminologic properties is given in the Lexicalization instances of the lexical ontology. In addition, a translation is given through the *hasTranslation* relation, in this case it is the lexical entry *LE-1-Fr*.

In this simple example, the mapping between ontology elements and lexical entries in several languages is straight forward. However, in cases where the conceptual mapping

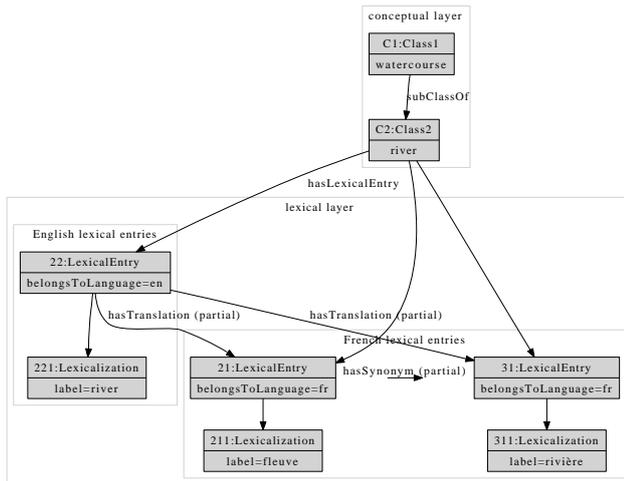


Figure 3: Example of a localisation in case of conceptual mismatches between English and French.

is different across different languages, the model allows to account for certain discrepancies, as shown in the following (fictitious) example, where one would like to localise to French the concept labeled by the English word *river* (Fig. 3).

The localisation choices made explicit here are the following: The English label *river* is lexicalised in English by the lexical entry *river* and in French by the two lexical entries *fleuve* and *rivière*. However, *fleuve* and *rivière* don't have exactly the same meaning in French, they are both more specific than *river*. This is expressed through the partial synonymy relation and by the fact that the translation relation between *river* and *rivière* and *fleuve* is marked as partial. Note that both synonymy and translation are relations in the lexical ontology. These localisation choices can be easily adapted, refined or reverted. The next example shows the lexicalisation of a concept where the label consists of several words:

```
<smc:Concept rdf:about=
"&sembysem;#AssetTracking/WagonMovement_Notification"/>
```

In our simplified model this concept would be associated to one lexical entry corresponding to the entire expression and would also be marked as *mwe* (multi-word expression). In contrast, the latest version of LIR represents a multi-word expression and its components using the LMF ListOfComponents constructs. It is thus possible to link the components to the corresponding lexical entries. However, this multi-word expression also contains relational information reflected in the syntactic realisation of the noun phrase: it represents the action of issuing a notification about the movement of a wagon. Within LIR it is currently not possible to capture and represent accurately the corresponding interactions between the lexical units forming the multi-word expression. On the other hand this is possible within the LexInfo model, it would therefore be profitable if the two models could be made compatible and merged. Such efforts are currently under way in the Monnet project².

²http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.

2.2. Implementation

The NeOn project also proposes an API which allows to automatically generate a skeleton of the lexical ontology from the domain ontology labels and then to enhance and maintain the lexical ontology. Unfortunately it was not possible to reuse this API due to its complexity and our tight schedule. Therefore, the lexical ontology is currently developed and maintained at the LORIA as a database which can be exported to the OWL or MLIF format. The designer of the conceptual SEMbySEM model is in most cases located elsewhere and uses a web service to require lexical information from the lexical ontology. More specifically, the designer enters a word in natural language and is returned, via the web service the identifiers of the *LexicalEntry* in the lexical ontology for the corresponding word. This information is returned in the MLIF format.

2.2.1. Managing the database

In order to use the database representation described by Figure 2 for our web service, we converted it into a MySQL database according to the Boyce-Codd normal form rules. We can see that many links between tables are represented in this database, that is why we could not maintain the database and add content without a dedicated application. Therefore, we created web formulars in order to be able to add, modify and delete entries without making the whole database inconsistent. These formulars are accessible to anyone who would like to add manually new entries. In the future, we plan to implement the import of MLIF or OWL files.

Figure 4 shows the web page for adding and modifying existing lexical entries and lexicalizations.

Figure 4: Web GUI for managing the database content

2.2.2. Web service for retrieving information

In order to be easily compatible with any amount of applications, we implemented a web service (written in PHP), which provides a set of information according to the parameters of the request. For example, the link <http://sembysem.loria.fr/webService.php?action=getDefinitions&word=Wagon> html

&langFrom=en returns the following MLIF data (slightly simplified):

```
<?xml version="1.0" encoding="utf-8"?>
<MLDC>
  <GroupC>
    <MultiC class="definitions">
      <MonoC xml:lang="en" class="definedWord">
        <SegC class="Basic form">
          <lir:hasLexicalEntry rdf:resource="7"/>
          <SegC>Wagon</SegC>
        </SegC>
      </MonoC>
      <MonoC xml:lang="en" class="definition">
        <SegC class="definition">
          <lir:lexicalEntry rdf:resource="7"/>
          <SegC class="paragraph">
            A wagon is a heavy four-wheeled vehicle.
          </SegC>
        </SegC>
      </MonoC>
    </MultiC>
  </GroupC>
</MLDC>
```

This data represents "the definition of the English lexical entry having 'wagon' as lexicalization (basic form), which resource id is 7".

3. Linguistic information on the visualisation level.

While on the conceptual level the linguistic and lexical information provides multilingual support, on the visualisation level lexicalisation and translation activities pertain to a more traditional localisation task. SEMbySEM's visualisation layer consists of end-user interfaces displaying and giving access to elements of the core semantic representation. The end-user interfaces are designed by HMI experts in a language independent way. Currently the data format used is XUL, the XML User Interface Language developed by the Mozilla project. It has no formal specification and does not inter-operate with non-Gecko implementations. However, it uses an open source implementation of Gecko and relies on multiple *de facto* web-standards and web-technologies. It was chosen because there was no other suitable standard or norm available.

Language dependant data (ie. the strings labeling and describing the elements of the visual user-interface) are provided in a file in the MLIF format. The Multi Lingual Information Framework (MLIF) is a standard under development with the ISO/TC37/SC4 group. Its objective is to provide a generic platform for modelling and managing multilingual information in various domains while also providing strategies for the inter-operability and/or linking of other formats of interest for localisation and translation including for example TMX and XLIFF.

Finally, at run time the XUL description containing links to the corresponding MLIF components and the MLIF information are combined to render the user-interface in the end-user's language.

4. Conclusion

In this paper we report about efforts to provide linguistic and lexical information to the SEMbySEM project, whose aim it is to implement a sensor supervision and management framework based on an semantic representation. Linguistic and lexical information intervenes at two levels:

First it is attached to the conceptual representation through a lexical ontology based on LMF and aligned with other linguistic and lexical standards. Thus conceptual and lexical representations can be developed and maintained separately while allowing for a flexible and accurate coupling. Second, language support is necessary at the visualisation level for the localisation of the end-user interfaces. Here the user-interface itself is specified in a language independent manner using XUL and linguistic information is provided through the MLIF format. We describe when and where it was possible to use existing or emerging standards or best practices and discussed arising issues.

5. References

- J. S. Brunner, J. Beck, P. Gatellier, J. F. Goudou, I. Falk, S. Cruz-Lara, and N. Bellalem. 2009a. Micro-concept: Model reference. Technical report, D2.3v1.4 SEMbySEM working draft.
- Jean-Sébastien Brunner, Jean-François Goudou, Patrick Gatellier, Jérôme Beck, and Charles-Eric Laporte. 2009b. SEMbySEM: a Framework for Sensors Management. In *1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009)*, June 1st.
- Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards Linguistically Grounded Ontologies. In *The 6th Annual European Semantic Web Conference (ESWC2009)*, Heraklion, Greece.
- W. Peters, M. Espinoza, E. Montiel-Ponsoda, and M. Sini. 2009. Multilingual and localization support for ontologies. Technical report, D2.4.3 NeOn Project Deliverable.
- The LMF Working Group. 2008. Language Resource Management - Lexical Markup Framework (LMF). Technical report, ISO/TC 37/SC 4 N453 (N330 Rev. 16).

ISOCat Definition of the National Corpus of Polish Tagset

Agnieszka Patejuk^{1,2} and Adam Przepiórkowski^{2,3}

¹Jagiellonian University, Cracow

²University of Warsaw

³Institute of Computer Science, Polish Academy of Sciences, Warsaw

agnieszka.patejuk@gmail.com adamp@ipipan.waw.pl

Abstract

This paper describes the first definition of a complete morphosyntactic tagset, The National Corpus of Polish Tagset, in the ISOCat Data Category Registry. Although the task of implementing such a sophisticated tagset in ISOCat turned out to be significantly more challenging than expected, it was successfully completed. The result of this work, the *nkjp* Data Category Selection containing 85 carefully defined Data Categories owned by the NKJP group, is publicly available at <http://www.isocat.org/interface/index.html>. Discussing various solutions considered during this implementation, this paper presents certain limitations of ISOCat and offers some suggestions for its further development.

1. Introduction

The aim of this paper is to report on the process of defining the NKJP Tagset in the ISOCat Data Category Registry, commenting on the experience of using this system and suggesting ways in which it could be improved. First sections provide background information about the National Corpus of Polish, its tagset and ISOCat. Next, the implementation of the tagset is presented, discussing the limitations from which particular alternatives suffered and explaining how the tagset was eventually defined. A section highlighting various technical aspects of ISOCat and how these influenced the implementation of the tagset follows, offering some directions for further development of ISOCat. Finally, a succinct summary of the results achieved is provided.

2. NKJP and its tagset

The National Corpus of Polish (Pol. *Narodowy Korpus Języka Polskiego*; NKJP; <http://nkjp.pl/>) is a 3-year project terminating in December 2010, carried out at 4 Polish institutions. Background description of the project can be found in Przepiórkowski *et al.* 2008, 2010. One of the annotation levels in NKJP is morphosyntax.

The tagset of the National Corpus of Polish (henceforth, the NKJP Tagset; Przepiórkowski 2009) is a slightly modified version of the IPI PAN Tagset (Przepiórkowski and Woliński, 2003), a *de facto* standard tagset for Polish. There are 36 grammatical classes approximately corresponding to parts of speech, 13 grammatical categories and their possible values (36 in total). Each grammatical class has an associated list of appropriate grammatical categories, which may be specified as obligatory or optional for the particular class. Furthermore, there are a number of constraints on the possible values of categories appropriate for some classes, which will be discussed in more detail in § 4.

Most grammatical classes have a list of categories for which they inflect or whose value is specified lexically. Gerunds (*ger*) inflect for number, case and negation

while their aspect and gender (always neuter) are lexical. The complete morphosyntactic tag for ‘piciem’, *ger:sg:inst:n:imperf:aff*, provides the following information: it is a gerund whose values of the categories of number, case, gender, aspect and negation are singular, instrumental, neuter, imperfective and affirmative respectively. Certain categories may be optional – while some prepositions, like ‘do’ which is always *prep:gen*, have only one form, many others take different forms depending on the context: ‘pod’ *prep:inst:nwok* as opposed to ‘pode’ *prep:inst:wok* which have different values of vocalicity. Some classes such as conjunctions (*conj*) or predicatives (*pred*) are non-inflecting and, having no associated categories, their complete tags consist only of grammatical class tags: ‘i’ (*conj*), ‘to’ (*pred*). Finally, there are classes such as abbreviation (*brev*), bound word (*burk*) and unknown form (*ign*) which, rather than being traditionally understood parts of speech, serve technical purposes.

Alongside widely known traditional grammatical categories such as case (7 values), number, person, gender (5 values), degree, aspect, there are categories such as negation and accentability as well as some classes rather specific to Polish. These include accommodability which determines the syntactic behaviour of numerals, post-prepositionality describing the behaviour of certain pronouns in relation to prepositions, agglutination optionally applicable to one class of verbs and vocalicity which regulates the distribution of agglutinates. The remaining category, fullstoppedness, is a technical category taking one of two values depending on whether the abbreviation segment has to be followed by a full stop.

3. ISOCat and its architecture

The ISOCat project is an implementation of the ISO 12620 standard which is described as follows in the abstract available on the ISO website (<http://www.iso.org/>):

ISO 12620:2009 provides guidelines concerning constraints related to the implementation of a

Data Category Registry (DCR) applicable to all types of language resources, for example, terminological, lexicographical, corpus-based, machine translation, etc. It specifies mechanisms for creating, selecting and maintaining data categories, as well as an interchange format for representing them.

The architecture of ISOcat is therefore determined by the requirements set by the ISO 12620 standard: the DCR model consists of administrative information, descriptive information and linguistic information. These components of the DCR are reflected in the organisation of the data contained in individual Data Categories (DCs), all of which have an Administration Information Section, Description Section and potentially Conceptual Domains whose values are specified independently of the content of each other for different languages.

The Administration Information Section contains the Administration Record which provides information about the mnemonic identifier (Identifier, as opposed to PID, the unique Persistent Identifier), version, registration status, justification together with its origin as well as the dates of creation and the last change made to the DC.

The Description Section (DS) contains the Language Section which organises information about a given DC according to language. It provides details such as the definition, its source and, optionally, some notes in the Definition Section. There is also the Name Section which specifies the DC names together with their status in the given language and the Example Section where relevant examples together with their sources can be provided. The DS also contains the Data Element Section which is intended as the place for storing language-independent names of the DC.

The Conceptual Domain (CD), an inherent feature of *complex* DCs, contains the information about its possible values in a given language, which in turn depend on the type of the particular DC. There are three types of *complex* DCs: *closed*, *open* and *constrained*. The CDs of a *complex/closed* DC are represented as finite sets of *simple* DCs which, being the only non-complex DC type, do not have any associated CDs and therefore have no associated values themselves. The two remaining types of *complex* DCs, *complex/open* and *complex/constrained*, are characterised by the fact that the sets of their values cannot be enumerated exhaustively: the *open* DC is a ‘complex data category whose conceptual domain is not restricted to an enumerated set of values’ while the *constrained* DC is a ‘complex data category whose conceptual domain is non-enumerated, but is restricted to a constraint specified in a schema-specific language or languages’ (ISOcat Glossary, 2010).

More information about ISOcat can be found on the website of the project (<http://www.isocat.org/>).

4. Defining the NKJP Tagset in ISOcat

The original idea was to enter into the ISOcat DCR the grammatical classes, grammatical categories and their corresponding values. After having completed this stage, the values would be attached to appropriate grammatical categories and these, subsequently, would be related to appropriate grammatical classes as their attributes. Ideally,

the relations between a given grammatical class or category and its possible values or attributes would be expressed in its CD for the particular language. Adopting such a strategy would be preferable not only because of being the most economic one in terms of the amount of time necessary to define the NKJP Tagset as ISOcat DCs but also because it would closely reflect the design and structure of the tagset.

4.1. Elegant but impossible

Regrettably, such a solution, even though it would certainly be the most elegant one, could not be implemented because of the architecture of the ISOcat DC types. The DC type which matches best the requirements set by this task is, in most cases, the *complex* one as every complete morphosyntactic NKJP tag consists of a tag signalling the grammatical class followed by tags corresponding to values of appropriate grammatical categories, if there are any. Using the range of DC types offered by ISOcat, the values of grammatical categories were classified as *simple* DCs since, being simple atoms, they have no values themselves. They were subsequently related to corresponding grammatical categories whose DC type was set to *complex/closed* because they have well-defined repertoires of enumerable values. Ideally, it would be possible to list in an analogous way all the corresponding grammatical categories in the CD of a given grammatical class as its attributes. This way, both grammatical classes and categories would be classified as *complex/closed* DCs while values of grammatical categories would be *simple* DCs – it is here that a serious problem is encountered. While it is possible to provide *simple* DCs as values in the CD of a *complex/closed* DC, it is not possible to specify such a *complex/closed* DC as one of the attributes of another *complex/closed* DC, which would be the case here. The reasons are manifold: non-simple DC types cannot be linked to the CD of a *complex* DC, only *complex/closed* DCs can have CDs with enumerated content and, more importantly, there is no support for representing any relations other than that of *Value* in the CD at the moment.

4.2. Clever but impossible

Another approach at defining the tagset in ISOcat was based on the idea of entering complete NKJP tags directly into the DCR as *complex/open* DC types. A complete morphosyntactic NKJP tag, for instance `subst:sg:nom:m2`, the template for which would be `class:number:case:gender` has the following structure: the first element represents the grammatical class, followed by values of appropriate grammatical categories, if there are any. If the complete tag consists of more elements than the obligatory grammatical class, every segment is separated from the following one with a colon. With 36 grammatical classes, 13 grammatical categories and 36 values in total, there are more than 1500 possible complete tags, which makes the task of creating and entering them manually unfeasible. Due to this fact, the complete tags need to be either generated or extracted from the corpus. The latter solution is given preference because it avoids problems encountered in the case of baseline tag generation which include accounting for restrictions on the

values of grammatical categories as well as the optionality of some categories. On the other hand, choosing to extract the complete tags from the corpus, there is the risk of some not being represented due to their absence from the data. Having obtained the complete tags by either method, the original tag separators must be replaced by some other character due to the fact that the use of colons in the DC Identifier is restricted and the system will not accept such DCs. Subsequently, automatic descriptions of DCs would be generated (`subst:sg:nom:m2 = noun, singular, nominative, animate masculine`) and, together with corresponding complete morphosyntactic tags as identifiers, fitted into the frame provided by the Data Category Interchange Format (DCIF), the XML export format for DCs grouped into Data Category Selections (DCSs). Finally, the modified DCIF file would be fed into the DCR.

Unfortunately, this solution could not be implemented as DC import is not supported at all at the moment. According to the obtained information, although the DCIF DC import is given priority, there is no set date of its introduction and, more importantly, it is either not going to be publicly available or it is going to be subject to certain restrictions on the allowed data import limit.

4.3. Successful but time-consuming

Due to the fact that the implementation of the previous solution was not possible because of technical limitations, another strategy had to be adopted. 36 DCs defining grammatical classes which roughly correspond to parts of speech were created manually. Due to the uniqueness of many solutions adopted in the tagset which include, for instance, a separate class for depreciative forms (`depr`) and two distinct classes of pronouns, it was not possible to use already existing DCs and new ones tailored to the needs of NKJP were created. Definitions of NKJP DCs were written, with minor modifications, on the basis of extracts from publications about the IPI PAN Corpus (mainly, Przepiórkowski 2004) and NKJP with appropriate bibliographic source provided, following ISOcat guidelines.

Since the appropriate grammatical categories could not be represented in the CD of grammatical classes as their attributes, the definition is followed by a line containing detailed information about the grammatical categories associated with the particular grammatical class. In order to make it easier to trace associated grammatical categories, the list is accompanied by corresponding PIDs in plain text since the use of hyperlinks in definitions is not supported. Furthermore, if a category happens to be optional for some class, information about its optionality is also provided in brackets after the corresponding PID. Since grammatical classes are sets of lexemes which are not defined in ISOcat themselves, the DC type of defined grammatical classes was set to *complex/open*.

4.4. Another alternative

There is an alternative solution which, although it has not been implemented, is worth mentioning as it could improve the results achieved through the application of the previous one. This would, however, come at a considerable cost – grammatical classes would need to be reclassified as *com-*

plex/closed, which would distort the ontology modelled in this implementation where grammatical classes are consistently *complex/open* DCs whose values are appropriate lexemes. The values of grammatical categories could be related directly, bypassing the level of categories, to appropriate grammatical classes in their CDs. Though the intermediate level of grammatical categories would not be represented in the CD of the given grammatical class, this information would be still available in its justification as well as definition. In this way, the information provided in the description of the Data Category (DC) would complement the specification of grammatical category values in its CD. Furthermore, such a solution would make it possible to account in a straightforward way for most constraints on the values of categories appropriate for classes, with the exception of more complex ones as in the case of imperative (`impt`) where the range of appropriate values of the category of person is restricted by the value of the category of number.

However, there are some serious drawbacks which have to be taken into consideration. These include a great deal of manual work due to the complete lack of support for templates or multiple changes to the DC or even the entire DCS. Moreover, since values in the CD are listed in an alphabetical order, the values of corresponding grammatical categories would not be grouped. Finally, there is no means to account for the optionality of values of certain categories directly in the CD of the given grammatical class – such information could only be retrieved in the definition and justification of the DC.

5. Technical issues

As it is openly acknowledged on the project website, ISOcat is constantly under development – it is emphasised that the Web Interface (WI) available at the moment is a beta version. As a result of this implementation which required a considerable amount of time spent using ISOcat WI, a few bugs were reported and many more features were requested. Most of the identified bugs were fixed while only some of the suggested functionalities have been introduced. This section recaps the main points concerning the technical side of the defining the NKJP Tagset in ISOcat and brings into focus some issues which require particular attention.

5.1. What has been done

So far, only two of the requested features have been implemented, the first one being the update of the ISOcat DC search following earlier requests from other users. It is now possible to refine the search results using a variety of parameters such as the matching method, the language of keywords as well as fields, profiles and scopes to be considered. On the one hand the update introduces many more features than requested, but on the other it does not include an important functionality that was suggested – the possibility to use DC search when specifying the values in the CD of a DC. The second update brought the possibility to delete a DCS but not a DC which, being persistent, which is a part of ISOcat policy, cannot be removed once created. Since there is no way to delete a DC, the only solution at the moment is to recycle it – the only element of the DC

which cannot be changed is its PID. In the future, however, an alternative in the form of having the possibility to deprecate a DC is going to be made available while at the moment it is only possible to set the name status of the DC to deprecated.

5.2. What needs to be done

There are many vital functionalities which could make the work with ISOcat significantly easier and more efficient but, unfortunately, are not supported at the moment. These include the introduction of basic tools such as DC templates which could be created from scratch or on the basis of a particular DC chosen by the user. A multiple change tool, possibly with regular expression support, making it possible to apply multiple changes at the same time instead of doing it manually would certainly be in place. It could easily be applied to editing the definitions (to change some key term shared by a number of DCs), their source, but also to changing features such as DC name status, DC type or even CD values (if some of them are shared). Multiple change tool would also be particularly useful when changing the scope of chosen DCs or even the entire DCS as currently changing the scope of a DCS does not result in an automatic change of the scope of DCs it contains. At the moment all of the above must be done manually, which is extremely time consuming when handling a greater number of DCs. The next feature request, whose importance is supported by ample evidence presented earlier, is the implementation of the DCIF import which would not only enable automating the management of the DCS to a large extent but it would also provide the first basic alternative to managing DCs via the WI which is currently the only means of accessing ISOcat.

Finally, there are some minor issues which could still have a considerable positive impact on the experience of using ISOcat. The introduction of password change would certainly be appreciated by many, not only for security reasons. It might be a good idea to resign from the obligatory comment required by the WI when saving a new DC, which would make working with ISOcat even more smooth and reduce the number of whitespace comments. Last but not least, the ISOcat WI provides a brilliant platform for work supported by state-of-the-art technology but a faster, more lightweight alternative would certainly be welcome.

6. Conclusion

In spite of a number of problems encountered during this implementation, the main objectives were achieved – the NKJP Tagset has been successfully defined in the ISOcat DCR and it is now available as a public DCS, `nk_jp`, owned by the NKJP group which is the first and used to be the only group with a public DCS in the ISOcat DCR. The `nk_jp` DCS contains 36 *complex/open* DCs corresponding to grammatical classes, 13 *complex/closed* DCs defining grammatical categories and 36 *simple* DCs specifying values of these categories. In total, 85 DCs were created, all of which have an associated definition describing its function in the tagset, accompanied by relevant examples from Polish. Due to the technical limitations discussed at length,

it was not possible to reproduce the original design of the tagset and alternative solutions had to be adopted.

At a glance, the NKJP Tagset was modelled in the ISOcat DCR in the following way: grammatical classes are *complex/open* DCs with appropriate lexemes as their values, grammatical categories are *complex/closed* DCs whose CDs are populated with their possible values modelled as *simple* DCs. Due to the fact that, using currently implemented ISOcat solutions, it was not possible to express formally the relation of *Attribute* between grammatical classes and categories, definitions of grammatical classes provide additional information about appropriate categories together with their PIDs, details about their optionality and, if applicable, constraints on their values.

The idea of standardising linguistic concepts is undeniably appealing and ISOcat provides a convenient platform to assist this process. It offers functionalities such as DC checking which support the creation of DCs in accordance with ISO standards and gives the unique possibility to submit DCs for standardisation. Though the experience of defining the NKJP Tagset in the ISOcat DCR suggests that there is still some room for improvement, the current implementation of ISOcat was flexible enough to allow a successful realisation of this task – this is the first public ISOcat definition of any complete tagset, for any language.

References

- ISOcat Glossary (2010). <http://www.isocat.org/interface/JSXAPPS/ISOcat/help/ISOcatGlossary.html>.
- Przepiórkowski, A. (2004). *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.
- Przepiórkowski, A. (2009). A comparison of two morphosyntactic tagsets of Polish. In V. Koseska-Toszewa, L. Dimitrova, and R. Roszko, editors, *Representing Semantics in Digital Lexicography: Proceedings of MONDILEX Fourth Open Workshop*, pages 138–144, Warsaw.
- Przepiórkowski, A. and Woliński, M. (2003). The unbearable lightness of tagging: A case study in morphosyntactic tagging of Polish. In *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, pages 109–116.
- Przepiórkowski, A., Górski, R. L., Lewandowska-Tomaszczyk, B., and Łaziński, M. (2008). Towards the National Corpus of Polish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech. ELRA.
- Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (2010). Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Referencing ISOcat Data Categories

Menzo Windhouwer¹, Marc Kemps-Snijders¹, Sue Ellen Wright²

¹Max Planck Institute for Psycholinguistics

Wundtlaan 1, 6525 XD Nijmegen, The Netherlands

²Kent State University Institute for Applied Linguistics

109 Satterfield Hall, Kent, Ohio 44242 USA

E-mail: Menzo.Windhouwer@mpi.nl, Marc.Kemps-Snijders@mpi.nl, sellenwright@gmail.com

1. The ISO 12620 Data Category Registry

After 10 years in development, a revision of ISO 12620 has been published by ISO in December 2009. Where the previous version of ISO 12620 contained a list of standardized data categories in hardcopy format, this revision describes a web-based electronic Data Category Registry (DCR).

To meet its potential as a fundamental tool for semantic interoperability, the link between data categories as available in the DCR and instances of their use in linguistic resources has to be made explicit. For this purpose the DCR assigns a Persistent Identifier (PID) to each data category in compliance with ISO DIS 24619. For example, the data category */part of speech/* from the Terminology profile has the following PID: <http://www.isocat.org/datcat/DC-396>. ISO 12620:2009 also proposes a small DC Reference XML vocabulary¹ for embedding these data category PIDs in XML documents. For example, a Relax NG declaration of a POS element could be associated with this */part of speech/* data category as follows²:

```
<rng:element name="POS"
  dcr:datcat="http://.../DC-396" />
```

In addition to this generic vocabulary (markup) languages can also offer specific constructs to refer to data categories.

2. Standards and Data Category references

This section investigates the current support for embedding data category references in some standards.

2.1 XML schema languages

Document Type Definitions (DTDs) are part of the XML standard, and are still used widely to specify XML vocabularies. However, the DTD language does not provide any construct which can be used to associate an element, attribute or value with a data category.

Nor do Relax NG and W3C XML Schema provide such a construct, but they provide constructs for using other XML vocabularies with different namespaces, like the DC Reference vocabulary, in order to embed this information. The Relax NG example in section 1 shows this procedure for an XML element. The same can be done for attribute and value declarations. W3C XML Schema even provides

the `xs:appinfo` element as a specific location to embed this kind of annotations.

The TEI ODD³ (One Document Does all) specification can be used as a schema specification which can generate a DTD, a Relax NG or W3C XML Schema upon demand. ODD provides the `equiv` element which can take an `uri` attribute to refer to an equivalent external structure or value.

```
<elementSpec ident="pos">
  <equiv name="partOfSpeech"
    uri="http://.../DC-396"
  />
</elementSpec>
```

These XML schema languages are very generic but existing standards for linguistic resources have also dealt with the embedding of data category references.

2.2 Markup frameworks

Both the Terminological Markup Framework (TMF; ISO 16642) and the Lexical Markup Framework (LMF; ISO 24613) describe abstract metamodels. For specific uses these metamodels are complemented by a Data Category Selection (DCS).

The Generic Mapping Tool (GMT) is a canonical representation of the TMF model. GMT refers to ISO 12620:1999 data categories using their name in a `type` attribute of the `feat` or `annot` element:

```
<feat type="definition">
  <annot type="broader concept generic">
    pencil
  </annot>
  whose
  <annot type="characteristic">
    casing
  </annot>
  is fixed around a central
  <annot type="characteristic">
    graphite
  </annot>
  medium which is
  <annot type="characteristic">
    used for writing or making marks
  </annot>
</feat>
```

Both the DTD and XML schema given for GMT would allow the use of the new data category PIDs.

LMF mentions several times that data categories should

¹ See <http://www.isocat.org/12620/>

² Due to space limitations in the examples part of the data category PID has been replaced by ellipses.

³ See <http://www.tei-c.org/Support/Learn/odds.xml>

be taken from ISOcat. However, in contrast to TMF, LMF does not define a canonical representation of the model. The foreword to Annex R, which gives an example representation, states that a user can use any schema to implement LMF. From the viewpoint of ISO 12620:2009 these schemas should preferably use the proposed DC Reference vocabulary.

2.3 Terminological markup languages

TermBase eXchange (TBX; ISO 30042) is an XML-based framework for representing structured terminological data. With TBX, various terminological markup languages (TMLs) can be defined. These TMLs may differ in respect to which data categories may be allowed and where they are allowed. These constraints are formally expressed in an XCS (eXtensible Constraint Specification) file. The data category references currently found in ISO 30042 are based on sub-clause numbers from ISO 12620:1999:

```
<termNoteSpec
  name="partOfSpeech"
  datcatId="ISO12620A-020201">
  <contents datatype="plainText"
    forTermComp="yes"
  />
</termNoteSpec>
```

However, the use of the `datcatId` attribute is limited to complex data categories only. The XCS file does allow the specification of possible values, but this is done using a space separated list and there are no provisions to associate simple data category references with these values. The `datcatId` attribute is defined by the XCS DTD as CDATA, which would allow the use of the new data category PIDs.

Geneter⁴ is a TMF compliant TML. The schema for Geneter XML documents is provided by a modular collection of DTDs. Element names and values are taken from the data category specifications in ISO 12620:1999. However, there are no language constructs to embed references to these data categories, so instead these are placed inside comments:

```
<!--[part_of_speech] A category assigned to
a word based on its grammatical and semantic
properties. [ISO 12620 - A.2.2.1]-->
<!ELEMENT PartOfSpeech %Inline;>
```

As described in Section 2.1 an XML-based schema vocabulary, like W3C XML Schema or Relax NG, would allow the use of the DC Reference vocabulary.

2.4 Annotation frameworks

The Morpho-syntactic Annotation Framework (MAF; ISO 24611) specifies the relationship between a feature, a value or a feature type and a complex or simple data category using the `dcs` element:

```
<dcs local="genre"
  registered="dcs:morphosyntax:gender:fr"
  rel="eq"
/>
```

Here the registered attribute is defined in the schema as containing a URI, so this construction is well suited for embedding the new data category PIDs. In addition the `dcs` element also allows refinement of the relationship, e.g., equals, subset or generic, between the tag and the data category. This information will have to be considered by any DCR-based interoperability mechanisms.

The Linguistic Annotation Framework (LAF; ISO 24612) documents are associated with a standalone header file. This file can contain one or more `typeDescription` or `featureDescription` elements, each of which describes a content category or feature. This description can be external, in which case the URI, e.g., a PID, of this external description is put in a `loc` attribute. Although the description of this setup is rather short and no actual example or schema is given, the fact that a `featureDescription` element can take a list of possible values suggests that it is only suitable for complex data categories.

3. Conclusion and future work

The standards considered all have some support for referring to data categories. Even for the older standards like TBX and TMF it seems that the existing constructs could be reused for the new data category PIDs. But the impact on relevant tool chains needs more careful inspection. Especially as this reuse is only possible due to under specification of the DTDs involved, i.e., the attributes containing the references are specified as CDATA. Some of the newer standards make use of Relax NG or XML Schema and properly type the references as any URI. Revisions of the older standards should consider the same approach. In some cases, e.g., TBX and LAF, values cannot be associated with simple data categories. This would have to be resolved in future revisions of these standards as trying to inference the proper data category based on a complex data category and a value may not be possible due to ambiguity.

4. References

- ISO 12620 (1999) Computer applications in Terminology — Data categories
- ISO 12620 (2009) Terminology and other language and content resources — Specification of data categories and management of a Data Category Registry for language resources
- ISO DIS 16642 (2001) Computer applications in terminology - Terminological markup framework
- ISO DIS 24611 (2008) Language resource management — Morpho-syntactic annotation framework
- ISO DIS 24619 (2009) Language Resource Management — Persistent Identification and Access in Language Technology Applications
- ISO DIS 24612 (2009) Language resource management — Linguistic annotation framework (LAF)

⁴ See <http://www.geneter.org/>

ISO 24613 (2008) Language resource management —
Lexical markup framework (LMF)

ISO DIS 30042 (2008) Terminology and other language
and content resources — Computer applications in
terminology — TermBase eXchange Format
Specification (TBX)

OASIS Relax NG Technical Committee. Relax NG,
<http://relaxng.org/>

W3C. XML (2008) Extensible Markup Language (XML)
1.0 (Fifth Edition),
<http://www.w3.org/TR/2008/REC-xml-20081126/>

W3C. XML Schema: Tools, Usage, Resources,
Specifications and Development,
<http://www.w3.org/XML/Schema>

***annSem*: Interoperable Application of ISO Standards in the Annotation and Interpretation of Multilingual Dialogue**

Kiyong Lee¹, Alex C. Fang², Jonathan J. Webster²

¹Korea University, Seoul, South Korea

²City University of Hong Kong, Hong Kong SAR, China

E-mail: ¹klee@korea.ac.kr, ²{acfang,ctjjw}@cityu.edu.hk

Abstract

This paper proposes graph-based annotation semantics, named *annSem*, for the annotation and interpretation of multilingual dialogue. *annSem* is constructed in a graph-theoretic pivotal format (GrAF; Ide & Suderman, 2007) by interoperable use of ISO annotation schemes. GrAF first converts XML-based various standoff annotations into graphs and then merges them into a single directed coherent network. Within such a merged graph, *annSem* selects and converts semantically relevant parts of annotated data to logical forms for appropriate interpretation. This process is claimed to be effective and robust especially for capturing the semantic content of multilingual dialogue simply because it does not totally rely on the syntactic well-formedness of input dialogue data and also because it does not require language-to-language translation.

1 Introduction

This paper proposes a graph-theoretic annotation-based informal semantics (*annSem*) for multilingual dialogue. By interoperably using some of the ISO annotation standards for language resource management in a graph-theoretic pivotal format, developed in Ide and Suderman (2006, 2007) and LAF (Ide and Romary, 2004; ISO, 2009b), the proposed semantic scheme is designed (1) to convert separately annotated parts of a multilingual dialogue to graphs, (2) to merge them into a coherent graph, and (3) transduce semantic representation in very rudimentary logical forms for coherent interpretation. It captures the semantic content of multilingual dialogue with its background information without the otherwise required process of language-to-language translation.

Consider the following dialogue, in Korean and English, between two elderly married people in a small Korean village:¹

- (1) Sample Multilingual Dialogue (Dia1)
 - a. *ye.bo, sa.lang.hay.yo.*
 - b. Me, too.

While (b) is uttered in English by the shy husband who was embarrassed at his affectionate wife, the first line (a) is spoken in Korean by his wife. It roughly translates to English as in (2) below:

- (2) Honey, (I) love (you).

(1a) is a well-formed sentence in Korean, although it only consists of a verb. Here both the Subject and the Object are understood only contextually, although they are not expressed in the utterance as common in spoken Korean. The understanding of such a dialogue, especially a type of dialogue that is carried in more than one language, thus

requires adequate addition of notes to a text that is called *annotation*.

In compliance with the current conventions, annotations are represented in XML-based markup languages in this paper. For the annotation of a dialogue, *annSem* particularly makes use of seven main ISO annotation standards: FSR (ISO, 2006), MAF (ISO, 2009a), LAF (ISO, 2009b), SynAF (ISO, 2009c), SemAF-Time (ISO, 2009d), SemAF-DActs (ISO, 2009e) and SemAF-SRL (ISO, 2009f). FSR and LAF provide basic descriptors and mechanisms for language resource management, while the other five standards treat different levels of linguistic description in annotating language resources.

The rest of the paper is organized as follows. Section 2 will describe FSR-based standoff annotation framework. Section 3 will present the conversion of annotations into graphs and Section 4 graph-based annotation semantics before concluding remarks in Section 5.

2 FSR-based Standoff Annotation

2.1 Basic Requirements

The proposed *annSem* follows at least three of the major requirements of LAF (ISO, 2009b). First, each annotation is marked standoff from a primary data and stored in a different file, as is in ANC. Second, each data structure consists of a referencing structure and a content structure. Third, each content structure is represented by a list of feature specifications that conform to FSR (ISO, 2006). This is illustrated with the sample multilingual dialogue (dia1) given in (1).

2.2 Annotation of Multilingual Dialogue

Dialogue is understood as a type of communications by means of a language that involves more than one participant who exchange their roles through turn-taking. The use of a language is primarily spoken, but may have multimodal aspects such as gestures or facial expressions. Multilingual dialogue is typically characterized by the use of more than one language as a medium of communications.

¹ This sample was obtained from a KBS television broadcast on 2010-02-04, 6 pm evening village tour program.

Each multilingual dialogue is then annotated with respect to its background, structure and function, and content as well as the use of a language for each utterance.

2.2.1 Background of a Dialogue

The background provides a list of the participants, senders (speakers) and addressees, for the whole dialogue with any other relevant information related to these participants. This is annotated in XML as below:

```
<sText xml:id="dial">
<diaML>
  <background>
    <fs type="participants">
      <fs xml:id="p1"/>
      <f name="sex" value="female"/>
      <f name="age" value="65+"/>
    </fs>
    <fs xml:id="p2"/>
    <f name="sex" value="male"/>
    <f name="age" value="70+"/>
    </fs>
    <fs type="married">
      <f name="wife" target="#p1"/>
      <f name="husband" target="#p2"/>
    </fs>
  </background>
</diaML>
</sText>
```

2.2.2 Structure and Function of a Dialogue

The structural part of the annotation specifies how the dialogue is segmented into utterances as well as functional segments:²

```
<sText xml:id="dial"
target="#string-range(dial,0,28)>
<diaML>
  <diaStruc>
    <seg type="utterance" xml:lang="kr" xml:id="u1"
target="#string-range(dial,0,9)>
      <fs type="roles">
        <fs type="sender" target="#p1"/>
        <fs type="addressee" target="#p2"/>
      </fs>
      <fs type="funcSeg">
        <fs xml:id="u1fs1" target="#wd1">
          <f name="comFunc" value="calling"/>
        </fs>
        <fs xml:id="u1fs2" target="#wd2">
          <f name="comFunc" value="statement"/>
        </fs>
      </fs>
    </seg>
    <seg type="utterance" xml:lang="en" xml:id="u2"
target="#string-range(dial,20,8)">
      <fs type="roles">
        <fs type="sender" target="#p2"/>
        <fs type="addressee" target="#p1"/>
      </fs>
      <fs xml:id="u2fs1" target="#u2">
        <f name="comFunction" value="reply"/>
      </fs>
    </seg>
  </diaStruc>
</diaML>
</sText>
```

Dialogue 1 is segmented into utterances, *u1* and *u2*, spoken in Korean and in English, respectively. Each utterance is further segmented into functional segments: *u1* into two functional segments, *u1fs1* and *u1fs2*, respectively serving the communicative functions of *calling* and *statement*, and *u2* into one functional segment, *u2fs1*, only which carries the communicative function of *reply*. In addition to the

²The turn-taking is specified indirectly with the sender and the addressee(s) specified for each utterance.

primary data, this annotation, as marked with the attribute *@target*, references several other annotations: the part of *<diaML>* that provides background information and the word form segmentation of *<MAF>*.

2.2.3 Content of a Dialogue

The content of a dialogue or its parts basically consists of a set of atomic propositions. Each atomic proposition is a well-formed formula in elementary logic, consisting of a predicate and a list of arguments. SemAF-SRL (ISO, 2009f), for instance, annotates the argument structure of *u1fs2*, the second functional segment of the first utterance that consists of a verb only, as below:

```
<sText xml:id="dial">
  <semAF-SRL>
    <argStruc xml:id="argSt1" target="#u1fs2">
      <fs type="predicate" xml:id="pred1"
target="#e1"/>
      <fs type="arguments">
        <fs type="agent" xml:id="arg1"
target="#subj"/>
        <fs type="patient" xml:id="arg2"
target="#obj"/>
      </fs>
    </argStruc>
  </semAF-SRL>
</sText>
```

Besides referencing to the functional structure of the sample dialogue *dial* annotated by SemAF-DActs (ISO, 2009e), *<semAF-SRL>* references *<isoTimeML>* for the event *#e1* LOVE that is targeted at by the predicate and also *<synAF>* for the two arguments *#subj* and *#obj* that have the roles of being an agent and a patient, respectively. These two arguments are then linked to the sender *#p1* and the addressee *#p2* by referencing the dialogue structure and also the background of the dialogue annotated by *<diaML>*. Hence, the interpretation of *<SemAF-SRL>* references practically all of the seven ISO standards that have been developed for linguistic annotation.

3 Converting Annotations into Graphs

Originally based on GrAF (Nancy and Suderman, 2007), LAF (ISO, 2009b) provides a theoretical framework for converting various standoff annotations into a coherent pivotal format in a graphic structure. *annSem* adopts this format and converts each of the XML-based annotations into a graph.

3.1 Implementing GrAF for annSem

For the graph representation of feature structures, FSR (ISO, 2006) introduces the notion of a directed graph with a unique root. Formally, this graph is defined as a quadruple $G = \langle N, r, R, L \rangle$ such that N is a nonempty set of nodes, r a unique member of N , called *root*, R a partially ordered binary relation over N , and L a set of labels. Each node is labeled by a nonempty set of type or feature specifications. For referencing it is also labeled with a tag name, an id or a target. Edges themselves are not labeled, while branching edges are understood as collection or alternation. This conforms to Ide and Suderman (2007, 2.2) that suggests that the labeling of edges be accommodated into feature value nodes so that only nodes are labeled uniformly. As an illustration, consider the following *<synAF>* XML-based annotation:

```

<synAF>
  <synStruct xml:lang="kr" xml:id="sent1"
    target="#u1fs2">
    <fs type="verb" xml:id="v1" target="#wd2" />
    <fs type="NP" xml:id="subj"
      target="#p1" />
    <fs type="NP" xml:id="obj"
      target="#p2" />
  </synStruct>
</synAF>

```

This annotation represents a flat syntactic structure for a sentence in Korean, consisting of a verb *wd2* with two arguments: the Subject NP and an Object NP. But on the surface the sentence consists of a verb only, while its Subject and Object are referenced contextually in a dialogue. The Subject refers to the sender (speaker) *#p1* and the Object to the addressee (hearer) *#p2* of the utterance *#u1*, or the functional segment *#u1fs2*, as annotated here. This annotation is then systematically converted into the following graph.

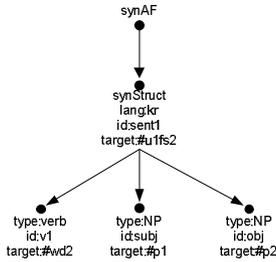


Figure 1: Syntactic annotation with branching edges

3.2 Merging Various Segmentations

A variety of segmentations can be represented on different paths, sequences of branches or edges. Each path represents a list of possible alternatives or collections. These paths can also be linked to show their coreferentiality. The following figure shows three different ways of segmenting the utterance *u1* to: (1) tokens, (2) word forms, and (3) utterances. Each utterance is also segmented to functional segments.

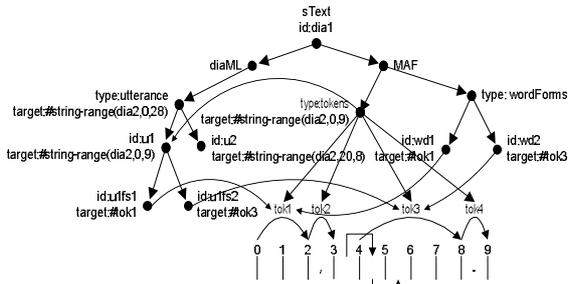


Figure 2: Interlinked segmentations

4 Graph-based Annotation Semantics

4.1 Focus

Just as annotation itself is selective, *annSem* is selective in choosing what to be interpreted among various possible items that carry a variety of information on an annotation graph. Given the first utterance *u1* in our sample data *dia1*, it may focus either on the first functional segment *u1fs1* that carries the function of calling the addressee or on the

second functional segment *u1fs2*, *sa.lang.hay.yo*, for its propositional content, for it carries the function of statement. For this segment, the following frame of a propositional form is immediately proposed in the form of elementary logic without quantification:

$$\sigma_{u1fs2} := [\sigma_{pred} \ \& \ \sigma_{arguments}]$$

4.2 Logical Forms for Propositional Content

Now relying on a merged graph like figure 3, *annSem* elaborates relevant parts of the graph. We first focus on two nodes, predicate and arguments, on *<SemAF-SRL>* on the leftmost edge. Here we navigate through the graph to see how they are linked by each occurrence of the attribute *@target* to relevant attribute *@id*'s in other parts of the graph, namely to *<isoTimeML>* and *<synAF>*. The node *{predicate, id: pred1, target:#e1}* is linked to the node *{EVENT, id:e1}* on *<isoTimeML>* that also leads to the node *{pred:LOVE}*. The nodes *{id:arg1, type:agent, target:#subj}*, and *{id:arg2, type:patient, target:#obj}* on *<SemAF-SRL>* are linked to the nodes *{type:NP, id:subj, target:#p1}* and *{type:NP, id:obj, target:#p2}*, respectively. We thus obtain the following logical forms:

$$\begin{aligned} \sigma_{pred} &:= \sigma_{pred1} \ \& \ pred1=e1 && \langle \text{SemAF-SRL} \rangle \\ &:= LOVE(e1) \ \& \ pred1=e1 && \langle \text{isoTimeML} \rangle \\ \sigma_{arguments} &:= [[AGENT(pred1,x) \ \& \ PATIENT(pred1,y)] \\ &\ \& \ [SUBJECT(pred1,x) \ \& \ OBJECT(pred1,x)]] && \langle \text{SemAF-SRL} \rangle \\ &\ \& \ [x=p1 \ \& \ y=p2]] && \langle \text{synAF} \rangle \end{aligned}$$

Second, by combining these two with some substitutions, we obtain the following logical form:

$$\begin{aligned} \sigma_{u1fs2} &:= [\sigma_{pred} \ \& \ \sigma_{arguments}] \\ \sigma_{u1fs2} &:= [[LOVE(e1) \ \& \ [AGENT(e1,x) \ \& \ PATIENT(e1,y)] \\ &\ \& \ [SUBJECT(e1,x) \ \& \ OBJECT(e1,y)]] \\ &\ \& \ [x=p1 \ \& \ y=p2]] \end{aligned}$$

This logical form is understood as representing the propositional content of the second functional segment of the first utterance, *u1fs2*. This is also understood as a statement by referencing the node *{id:u1fs2, target:#wd2, comFunc:statement}* on *<diaML>*, where *target:#wd2* needs to reference *<MAF>* for *sa.lang.hay.yo* in Korean.

4.3 Background for the Utterance

In order to anchor the parameters *x* and *y* properly, we need to reference the background node of *<diaML>*. Looking at Figure 3 again, we see that the nodes *{NP, id:subj, target:#p1}* and *{NP, id:obj, target:#p2}* on *<synAF>* are linked to the nodes *{id:p1, sex:female, age:65+}* and *{id:p2, sex:male, age:70+}* on *<diaML>*, respectively. Furthermore, these two are referenced by the other two nodes, *{type:married, wife:#p1, husband:#p2}* and *{type:roles}* with two edges *{type:sender, target:#p1}* and *{type:addressee, target:#p2}* on *<diaML>*. With such information, we can drive the following logical form for background:

$$\begin{aligned} \beta_{u1fs2} &:= [[SENDER(x,u1) \ \& \ ADDRESSEE(y,u1)] \\ &\ \& \ [FEMALE(x) \ \& \ AGE(x,65+)] \\ &\ \& \ [MALE(y) \ \& \ AGE(y,70+)] \ \& \\ &\ \& \ [MARRIED(x,y) \ \& \ WIFE(x,y) \ \& \ HUSBAND(y,x)] \\ &\ \& \ [x=p1 \ \& \ y=p2]] \end{aligned}$$

Lexicon standards: from de facto standard Toolbox MDF to ISO standard LMF

Jacqueline Ringersma¹, Sebastian Drude² and Marc Kemps-Snijders¹
¹Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
²Goethe-Universität, Frankfurt, Germany

Abstract

This paper discusses possible solutions for the apparent incompatibility between two standards for lexicon structure and concept naming: the de facto standard MDF, which is part of the widely used lexicon application Toolbox [1] and the newly accepted ISO standard LMF, ISO FDIS 24613:2008 [2], implemented in the online lexicon tool LEXUS [3]. The basic difference between the two standards is that in MDF, the form-related and meaning-related parts of lexical entries are embedded in each other, while in LMF there is a strict separation of the two parts. The difference might be related to the final medium for which the standards have been created; although Toolbox is a tool for digital lexicon creation, the MDF format was created for printed dictionaries, whereas LMF is created for digital presentation of lexicon resources. At first sight the difference seems to be fundamental and impossible to overcome. However, in this paper we would like to show possible solutions, and would like to probe them in the LREC2010 workshop on Language Resource and Language Technology Standards, and thoroughly discuss them amongst a wide linguistic public, before implementing a conversion procedure in the Toolbox import module of the LEXUS tool.

Multi-Dictionary Formatter (MDF)

In linguistic field work on minor languages, Toolbox is a widely used data management and analysis tool, designed for maintaining lexicons and for parsing and interlinearizing of text. Toolbox is text-oriented [1]. A lexicon structure is defined as a set of rules which declares the lexicon structure elements (markers), their value domains and their hierarchy. Toolbox delivers a default structure definition file for dictionary formatting: the Multi-Dictionary Formatter (MDF). Lexical entries content can be built following the MDF structure. The hierarchy, however, is not explicitly represented in Toolbox databases (which are in the 'Standard Format', a flat list of feature-value pairs). MDF structures facilitate not only the creation of digital lexicons, but also structured and formatted output of the Toolbox lexicon in a rich text format, which can be imported into Microsoft Word. MDF has become a de facto standard for lexicon structures in field linguistics.

In the MDF hierarchy, there are three main primary markers: `lexeme (\lx)`, `part of speech (\ps)` and `sense number (\sn)`. Lexicon structures are either 'part of speech'-oriented or 'sense'-oriented, and users of the tool are free to choose (in recent versions of Toolbox, the 'part of speech'-orientation is the default). In 'part of speech'-oriented structures, `\lx` is superordinate to `\ps` and `\ps` to `\sn`. One `\lx` can have multiple `\ps` markers and likewise one `\ps` can have multiple `\sn` markers. In 'sense'-oriented structures the hierarchy is `\lx>\sn>\ps`. In both orientations, `\lx` is also superordinate to a set of markers which apply to the lexical entry as a whole, e.g. `homonym number (\hm)` or `variant form (\va)`. In addition, `\sn` can be followed by a flat set of markers, like `english gloss (\ge)`, `vernacular gloss (\gv)`, `english definition (\de)` etc. Sense number can also be followed by structured sets of markers, for instance those for example sentences (`\xv\xe\xn`). MDF accommodates sub-entries (`\se`); these are integrated elements of lexical entries, subordinate to `\lx`, and the same hierarchy that applies to a full lexeme entry can also apply to a sub-entry.

Lexical Markup Framework (LMF)

The Lexical Markup Framework model (LMF) [2] was recently accepted as ISO standard for Natural Language Processing lexicons and Machine Readable Dictionaries (ISO-FDIS-24613:2008). LMF prescribes a basic model for lexicon structure elements ('data categories'), and a registry for data category naming and value domains. LMF also defines the constraints on the relations between the data categories. The main goal of LMF is to enhance true content interoperability between all aspects of lexical resources; in specific data exchange between resources, searching across and merging of the resources.

In LMF, the structure of a Lexical Entry consists of three basis components: `Lemma`, `Form` and `Sense`. `Lemma` is the conventional form chosen to represent a lexeme. `Form` manages the

orthographical variants of a lexical entry, as well as any other data category that represents the attributes of the word form (e.g. `writtenForm`, `inflections`). `Sense` represents one meaning of a lexical entry, with attributes like `definition` or `gloss`. `Part of speech` is considered to be neither form nor sense; in LMF, `part of speech` is an attribute of the Lexical Entry.

LEXUS

LEXUS [3] implements an instantiation of LMF. It is the online lexicon tool of the Language Archiving Technology suite (LAT) developed by the Max Planck Institute for Psycholinguistics (MPI) in The Netherlands. With LEXUS, users may create, manipulate and visualize lexicons and enrich the lexical entries with multimedia fragments. The default lexicon structure in LEXUS for new lexica is based on LMF. LEXUS offers the ISOcat data category registry for data category naming and value domain specifications (ISOcat is the ISO implementation of the ISO 12620:2009 standard and offers standard linguistic concepts to be used in linguistic resources). LEXUS has been operational since 2007 and currently has about 450 registered users, of which some 20 are active. The active users have developed around 60 lexica. Most of these lexica were imported in LEXUS, initially created in XML or Toolbox. For both formats LEXUS provides an import and export facility. However, in LEXUS it is possible to avoid the LMF structure; this means that when Toolbox lexica are imported into LEXUS it is possible to maintain the structure defined in the Toolbox typ file in the LEXUS lexicon.

From Toolbox MDF to LEXUS LMF

A first difference between MDF and LMF is in the naming of the concepts. However, in a recently created working group of the RELISH project [4] on lexicon standard interoperability, it was proposed to add the MDF markers to the ISOcat data category registry. MDF is thereby acknowledged as an important de facto standard in lexicography, and its data elements can be related to data elements used elsewhere.

One principal difference between MDF and LMF structures is that MDF does allow for sub-entries, whereas LMF does not. Since LMF is created for digital formatting of lexicons, this gap seems to be easy to overcome: for every sub-entry within a Toolbox lexeme, create a new Lexical Entry in LMF and attribute it with a cross reference and pointer to the Lexical Entry of the lexeme (and vice versa). Lexicographers might argue that the status of the two Lexical Entries is not equal, but also this difference can be covered with an attribute at the Lexical Entry level.

For 'part of speech' oriented MDF structures, the conversion from MDF to LMF is not too problematic. Lexical Entries in LMF can have multiple senses, so for each group of markers under `\ps\sn`, there will be a separate Sense container in the LMF structure. In case `\lx` contains multiple `\ps`, the option is again the creation of multiple Lexical Entries for each `\ps` block, possibly with several sense blocks within, with cross reference attributes.

For 'sense number' oriented MDF structures the situation is more complicated, since in this orientation, one `\sn` can have more than one `\ps`. But again the gap can be bridged by splitting the Toolbox lexical entries in multiple LMF Lexical Entry's. An algorithm for this will not be too hard to define, but it is not trivial to define the multiple cross referencing attributes which indicate the relations among the different entries.

In our paper we will describe the possible conversion from MDF to LMF, on the basis of examples taken from the Marquesan lexicon *Dico général - tekao tapapatia* [5] and the Iwaidja lexicon [6]. These lexica were initially created in Toolbox, with an MDF structure. We will discuss lexical entries with and without sub-entries and we will discuss both the part of speech and the sense number orientation. We will make a qualitative description of the required algorithms. On the basis of the examples we show how the conversion from MDF to LMF could be realized in the future import module of LEXUS.

The MDF format is not a suitable format for interoperability because it is based on a textual database system and because it is very prone to inconsistencies. Since the trend is that more resources will become digital, the need for interoperability will increase. However, when LMF is to become one new major standard for lexicon structures, it is important to suggest to the research community which concerns exist when converting MDF to LMF. Our paper is meant to initiate the discussion.

References

[1] Toolbox, <http://www.sil.org/computing/toolbox/> Visited on January 13, 2010

- [2] ISO technical committee ISO/TC37 (2008) Language resource management - Lexical Markup Framework (LMF) <http://www.lexicalmarkupframework.org/> Visited on January 13, 2010
- [3] Ringersma, J., & Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme, & R. van Son (Eds.), Proceedings of Interspeech 2007 (pp. 65-68). Baixas, France: ISCA-Int.Speech Communication Assoc.
- [4] RELISH project: <http://www.neh.gov/ODH/Default.aspx?tabid=111&id=117> Visited on February 1, 2010
- [5] Cablitz, G. (2007-2009) Marquesan lexicon Dico général - tekao tapapatia, published in the online lexicon tool LEXUS (<http://corpus1.mpi.nl/mpi/lexusDojo>). Accessible after permission from the owner.
- [6] Birch, B. and others (2006-2009) Iwaidja lexicon, published in the online lexicon tool LEXUS (<http://corpus1.mpi.nl/mpi/lexusDojo>). Accessible after permission from the owner.

Grounding an Ontology of Linguistic Annotations in the Data Category Registry

Christian Chiarcos

Universität Potsdam
chiarcos@uni-potsdam.de

Abstract

This paper presents an ontology-based approach to link linguistic annotations to repositories of linguistic annotation terminology. It describes the experimental integration of the morphosyntactic profile of the Data Category Registry with this architecture.

1. Background

In the last 15 years, the heterogeneity of linguistic annotations has been identified as a key problem limiting the interoperability and reusability of NLP tools and linguistic data collections. The multitude of linguistic tagsets complicates the combination of NLP modules within a single pipeline. The Rosana coreference resolution system (Stuckardt, 2001), for example, requires Connexor (Tapanainen and Järvinen, 1997) parses. Similar problems exist in language documentation, typology and corpus linguistics, where researchers are interested to access and to query data collections on a homogeneous terminological basis.

In order to enhance the consistency of linguistic metadata and annotations, several repositories of linguistic annotation terminology have been developed by the NLP/computational linguistics community (Leech and Wilson, 1996; Aguado de Cea et al., 2004) as well as in the field of language documentation/typology (Bickel and Nichols, 2002; Saulwick et al., 2005). The General Ontology of Linguistic Description (Farrar and Langendoen, 2003, GOLD) and the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al., 2009, DCR) address both communities.

With a terminological reference repository, it is possible to abstract from the heterogeneity of annotation schemes: Reference definitions provide an interlingua that allows to map linguistic annotations from annotation scheme *A* to annotations in accordance with scheme *B*. This application requires a linking of annotation schemes with the terminological repository, and here, I propose a formalization of this linking in OWL/DL.

2. Linking annotations with terminology repositories

2.1. Annotation mapping

The classic approach to link annotations with reference concepts is to specify rules that define a direct mapping (Zeman, 2008). It is, however, not always possible to find a 1:1 mapping.

One problem is **conceptual overlap**: A common noun may occur as a part of a proper name, e.g., German *Palais* ‘baroque-style palace’ in *Neues Palais* lit. ‘new palace’, a Prussian royal palace in Potsdam/Germany. *Palais* is thus both a proper noun (in its *function*), and a common noun

(in its *form*). Such conceptual overlap is sometimes represented with a specialized tag, e.g., in the TIGER scheme (Brants and Hansen, 2002). The DCR (and other terminological repositories) do currently not provide the corresponding hybrid category, so that *Palais* is to be linked to both `properNoun/DC-1371` and `commonNoun/DC-1256` if the information carried by the original annotation is to be preserved. **Contractions** pose similar problems: English *gonna* combines *going* (PTB tag `VBG`, (Marcus et al., 1994)) and *to* (`TO`). If whitespace tokenization is applied, both tags are to be assigned to the same form.¹

A related problem is the representation of **ambiguity**: The SUSANNE (Sampson, 1995) tag `ICSt` applies to English *after* both as a preposition and as a subordinating conjunction. The corresponding DCR category is thus *either* `preposition/DC-1366` or `subordinatingConjunction/DC-1393`. Without additional disambiguation, `ICSt` is to be linked to both data categories.

Technically, such problems can be solved with a 1:*n* mapping between annotations and reference concepts. Yet, overlap/contraction and ambiguity differ in their underlying meaning: While overlapping/contracted categories are in the intersection (\cap) of reference categories, ambiguous categories are in their join (\cup). This difference is relevant for subsequent processing, e.g., to decide whether disambiguation is necessary. A standard mapping approach, however, fails to distinguish \cap or \cup .

The linking between reference categories and annotations requires a formalism that can distinguish intersection and join operators. A less expressive linking formalism that makes use of a 1:1 (or 1:*n*) mapping between annotation concepts and reference concepts can lead to inconsistencies when mapping annotation concepts from an annotation scheme *A* to an annotation scheme *B* that make use of the same terms, although with slightly deviating definitions, as noted, for example, by Dimitrova et al. (2009) for MULTEXT/East. Further, a formalism to express negation is desirable for the linking of annotation categories that have a narrower definition than the corresponding reference cate-

¹The TnT Tagger (Brants, 2000) that uses whitespace tokenization circumvents this problem by assigning `TO` and suppressing `VBG`. Then, however, `TO` is ambiguous between the original PTB tags `TO` and `VBG+TO`.

gories.²

2.2. Annotation linking with OWL/DL

OWL/DL represents a formalism that supports the necessary operators and flexibility: With reference concepts and annotation concepts are formalized as OWL classes, the linking between them can be represented by `rdfs:subClassOf` (\sqsubseteq). OWL/DL provides `owl:intersectionOf`, `owl:unionOf` and `owl:complementOf` operators, and it allows to define properties and restrictions on the respective concepts.

An OWL/DL-based formalization has the additional advantage that it can employ existing terminological repositories, e.g., GOLD (native OWL/DL) and the DCR (with an OWL/DL conversion as described below). GOLD and the DCR are, however, under development. The efforts to maintain the linking between annotations and the terminological repository can be reduced if another ontology is introduced that mediates between the terminological repository and the annotation models: If a major revision of the repository occurs, only the linking between the intermediate ontology and the repository is to be revised, but the linking with not every single tagset.

Moreover, this intermediate ontology allows us to link annotations to multiple terminological repositories at the same time. This may be necessary, as in some case, we do observe considerable disagreement between the current developmental stages of GOLD and the DCR.³

The idea of a modular ontological architecture with an ontology mediating between terminological repositories and annotation schemes is formalized the OLiA architecture.

3. The OLiA ontologies

The Ontologies of Linguistic Annotations – briefly, OLiA ontologies (Chiarcos, 2008) – represent a modular architecture of OWL/DL ontologies that formalize several intermediate steps of the mapping between annotations, a ‘Reference Model’ and existing terminology repositories (‘External Reference Models’).

The OLiA ontologies were developed as part of an infrastructure for the sustainable maintenance of linguistic resources (Schmidt et al., 2006), and their primary fields of application include the formalization of annotation schemes and concept-based querying over heteroge-

²In current praxis, the STTS (Skut et al., 1998) tag for auxiliary verbs (`VAFIN`) that applies to all forms of German *haben* ‘to have; to own’, *sein* ‘to be; to exist’ independently of their syntactic function is mapped to the reference concept `auxiliary`. If the DCR category `auxiliary/DC-1244` is redefined as pertaining to *potential* auxiliaries, other tagsets with a function-oriented definition `auxiliary`, e.g., the Connexor (Tapanainen and Järvinen, 1997) tag `@AUX`, are to be correctly represented as instances of `auxiliary/DC-1244` but not `mainVerb/DC-1400`.

³As one example, a GOLD Numeral is a Determiner (Numeral \sqsubseteq Quantifier \sqsubseteq Determiner, <http://linguistics-ontology.org/gold/2008/Numeral>), whereas a DCR Numeral (DC-1334) is defined on the basis of its semantic function, without any references to syntactic categories. Thus, *two* in *two of them* may be a DCR Numeral but not a GOLD Numeral.

neously annotated corpora (Rehm et al., 2007; Chiarcos et al., 2009).

In the OLiA architecture, four different classes of ontologies are distinguished:

- The OLiA REFERENCE MODEL specifies the common terminology that different annotation schemes can refer to. It is derived from existing repositories of annotation terminology and extended in accordance with the annotation schemes that it was applied to.
- Multiple OLiA ANNOTATION MODELS formalize annotation schemes and tagsets. Annotation Models are based on the original documentation, so that they provide an interpretation-independent representation of the annotation scheme.
- For every Annotation Model, a LINKING MODEL defines `rdfs:subClassOf` (\sqsubseteq) relationships between concepts/properties in the respective Annotation Model and the Reference Model. Linking Models are interpretations of Annotation Model concepts and properties in terms of the Reference Model.
- Existing terminology repositories can be integrated as EXTERNAL REFERENCE MODELS, if they are represented in OWL/DL. Then, Linking Models specify \sqsubseteq relationships between Reference Model concepts and External Reference Model concepts.

The OLiA Reference Model specifies classes for linguistic categories (e.g., `olia:Determiner`) and grammatical features (e.g., `olia:Accusative`), as well as properties that define relations between these (e.g., `olia:hasCase`). Far from being yet another annotation terminology ontology, the OLiA Reference Model does not introduce its own view on the linguistic world, but rather, it is a derivative of EAGLES (Leech and Wilson, 1996), MULTTEXT/East (Erjavec, 2004), and GOLD (Farrar and Langendoen, 2003) that was introduced as a technical means to allow to interpret linguistic annotations with respect to these terminological repositories.

With respect to morphosyntactic annotations, the OLiA annotation models comprise 16 annotation schemes applied to 42 languages.⁴ Conceptually, Annotation Models differ from the Reference Model in that they include not only concepts and properties, but also individuals: Individuals represent concrete tags, while classes represent abstract concepts similar to those of the Reference Model.

Fig. 1 illustrates the linking between the STTS Annotation Model and the OLiA Reference Model for two STTS tags.

4. Linking with the Data Category Registry

4.1. The morphosyntactic profile of the DCR

The morphosyntactic profile of the Data Category Registry can be accessed through ISOcat (Kemps-Snijders et

⁴There are currently 5 annotation models for English, 5 annotation models for German, 2 annotation models for Russian, one annotation model for Tibetan, one for Old High German, the Connexor annotation model (10 European languages), and one annotation model for a typologically-oriented annotation scheme (29 languages).

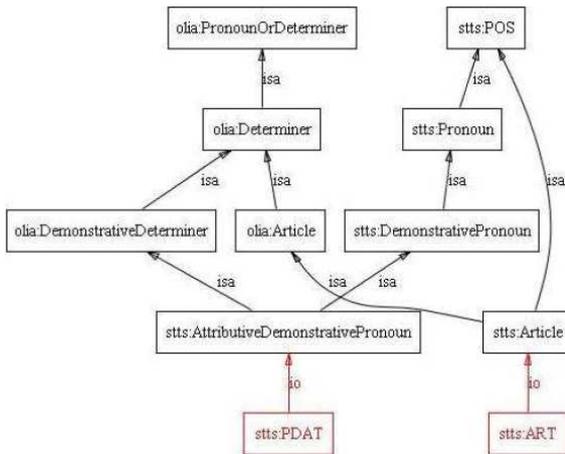


Figure 1: The STTS tags PDAT and ART, their representation in the STTS Annotation Model (stts) and linking with the OLIA Reference Model (olia).

al., 2009, <http://www.isocat.org>). It provides 383 data categories that represent either an attribute name like `partOfSpeech` or a value assigned to this attribute, e.g., `noun` (Francopoulo et al., 2006). Data categories are organized in 7 directories. The directories `PartOfSpeech` (115 categories), `Cases` (34 categories), and `MorphologicalFeaturesExcludingCases` (78 categories) contain word classes and grammatical features that correspond to conceptions in the OLIA Reference Model.⁵

Each data category is assigned a unique id (e.g., DC-1243), a textual identifier (e.g., `attributiveAdjective`), and a definition. Data categories can be hierarchically structured by `dcif:isA` ('has a broader data category') or `dcif:conceptualDomain` ('has one of these values').

4.2. Building an OWL representation of the DCR

The DCIF representation (Kemps-Snijders et al., 2009) of the morphosyntactic profile was transformed to OWL with XSL/T: Data categories are represented as `owl:Class`, definitions as `rdfs:comment`, references to the DCR by `dcr:datcat` (also used in ISOcat's RDF export); both hierarchical relations were mapped onto `rdfs:subClassOf`. Fig. 2 shows the resulting concept hierarchy and the properties of DC-1596.

It should be noted that this OWL representation of the DCR is not an ontology in a strict sense. The data category registry is a *collection* of data categories; hierarchical relations between categories are *optional*, and numerous data categories stand in no hierarchical relationship with other data categories, e.g., `attributiveAdjective/DC-1243` (as evident from its absence in the `adjective` hierarchy in Fig. 2).

⁵The other directories pertain to morphological and annotation processes (FormRelated, 35 categories; Operations, 29 categories) or contain unclassified categories (Basics, 60 categories), and usage-related metadata (RegisterDatingFrequency, 19 categories).

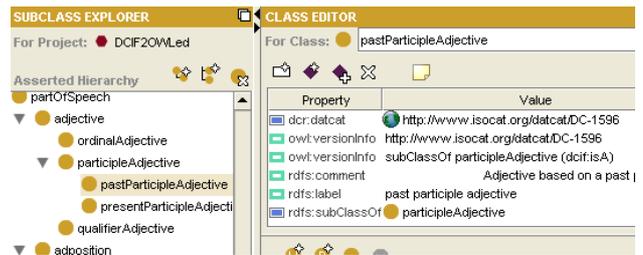


Figure 2: Past participle adjective in the OWL version of the morphosyntactic profile of the DCR in the Protégé OWL editor (Knublauch et al., 2004)

4.3. Semiautomatic linking with the OLIA reference model

For every OLIA Reference Model concept o , the following linking procedure was applied:

- (i) If the name of o matches the name of an DCR concept d , then set $o \sqsubseteq d$.
- (ii) If no matching DCR concept is found, retrieve all DCR concepts whose names contain at least one word that also occurs in o . The correct match d is manually selected (if there is any), and $o \sqsubseteq d$ is added to the linking model.
- (iii) If no candidate is found, leave a comment in the linking model.

This algorithm generates a preliminary linking, and under consideration of the comments generated during the procedure, the linking was then manually validated. Step (i) was applicable to 76 OLIA Reference Model concepts (e.g., `olia:AttributiveAdjective` and `attributiveAdjective/DC-1243`); for 155 concepts, the corresponding match could be established in step (ii) (e.g., for `olia:ExclamatoryPronoun` and `exclamativePronoun/DC-1285`). For 48 concepts, no candidate could be confirmed in step (ii), and for 79 concepts, no candidate was found.

5. Discussion

This paper described the application of modular OWL/DL ontologies to link annotations with terminological repositories: Annotation schemes and reference terminology are formalized as OWL/DL ontologies, and the linking is specified by `rdfs:subClassOf` descriptions. Currently, multiple repositories of linguistic annotation terminology are applied in the fields of NLP and language documentation/typology, e.g., GOLD and the DCR; both differ in their conceptualizations (fn. 3), and both are still under development.

Therefore, another ontology of linguistic annotations was applied, the OLIA Reference Model, that provides a stable intermediate representation between terminology repositories and ontological models of annotation schemes. This paper described how the morphosyntactic profile of the DCR can be integrated in this architecture as an External Reference Model: The morphosyntactic profile of the DCR was transformed into OWL/DL; then, OLIA Reference Model concepts were linked to DCR data categories. In this experiment, the majority of OLIA Reference

Model concepts (64.5%, 231/358) could be linked with the DCR. The remaining 35.5% (127/358) OLiA Reference Model concepts are partly concerned with issues of syntax and semantics, but to some degree, also different conceptions of morphosyntactic features are involved. Reflexivity, for example, is represented in the DCR only in combination with word classes, e.g., by the concepts `reflexivePronoun/DC-1378` and `reflexiveDeterminer/DC-1377`, reflexivity as a feature of verbs (e.g., in Russian) is missing in the DCR. For some phenomena, the OLiA Reference Model provides a richer feature set (e.g., with respect to `voice/DC-1413`), for other phenomena, the DCR provides a more granular concept repository (e.g., the subclassification of `particle/DC-1342`).

Unlinked concepts indicate where extensions or revisions of the DCR or the OLiA Reference Model may be necessary. As such, the OLiA Reference Model does currently not cover operations, processes and register/usage-related meta data. The DCR, on the other hand, is only partially hierarchically structured (only 210 of 383 data categories in the morphosyntactic profile are assigned a superclass), and the concept hierarchy of the OLiA Reference Model may be exploited to augment the DCR with a more exhaustive hierarchical organization.

A number of technical applications of the OLiA ontologies has been proposed, including corpus querying (Rehm et al., 2007), the specification of tag-set independent interface representation in NLP pipelines (Buyko et al., 2008), and information retrieval (Hellmann, 2010). With the DCR linked to the OLiA ontologies as an External Reference Model, the ontological representations of linguistic annotations applied in these contexts can be interpreted as DCR data categories, thereby enhancing their interoperability with other DCR-conformant annotation schemes.

6. References

- G. Aguado de Cea, A. Gomez-Perez, I. Alvarez de Mon, and A. Pareja-Lora. 2004. OntoTag's linguistic ontologies. In *Proc. Information Technology: Coding and Computing (ITCC'04)*, Washington, DC, USA.
- B. Bickel and J. Nichols. 2002. Autotypologizing databases and their use in fieldwork. In *Proc. LREC 2002 Workshop on Resources and Tools in Field Linguistics*, Las Palmas, Spain.
- S. Brants and S. Hansen. 2002. Developments in the TIGER annotation scheme. In *Proc. LREC 2002*, pages 1643–1649, Las Palmas, Spain.
- T. Brants. 2000. TnT—a statistical part-of-speech tagger. In *Proc. ANLP 2000*.
- E. Buyko, C. Chiarcos, and A. Pareja-Lora. 2008. Ontology-based interface specifications for a NLP pipeline architecture. In *Proc. LREC 2008*, Marrakech, Morocco.
- C. Chiarcos, S. Dipper, M. Götze, U. Leser, A. Lüdeling, J. Ritz, and M. Stede. 2009. A Flexible Framework for Integrating Annotations from Different Tools and Tag Sets. *Traitement Automatique des Langues*, 49(2).
- C. Chiarcos. 2008. An ontology of linguistic annotations. *LDV Forum*, 23(1):1–16.
- L. Dimitrova, R. Garabík, and D. Majchráková. 2009. Comparing Bulgarian and Slovak Multext-East morphology tagset. In *Organization and Development of Digital Lexical Resources*, pages 38–46. Kyiv.
- T. Erjavec. 2004. MULTEXT-East version 3. In *Proc. LREC 2004*, pages 1535–1538, Lisboa, Portugal.
- S. Farrar and D.T. Langendoen. 2003. A linguistic ontology for the semantic web. *Glott International*, 7(3):97–100.
- G. Francopoulo, M. Monachini, T. Declerck, and L. Romary. 2006. Morpho-syntactic Profile in the ISO-TC37/SC4 Data Category Registry. In *Proc. LREC 2006*, Genova, Italy.
- S. Hellmann. 2010. The semantic gap of formalized meaning. accepted at the 7th Extended Semantic Web Conference (ESWC 2010), May 30th – June 3rd 2010, Heraklion, Greece.
- M. Kemps-Snijders, M. Windhouwer, P. Wittenburg, and S.E. Wright. 2009. ISOcat: Remodelling metadata for language resources. *International Journal of Metadata, Semantics and Ontologies*, 4(4):261–276.
- H. Knublauch, R.W. Ferguson, N.F. Noy, and M.A. Musen. 2004. The Protégé OWL plugin. *The Semantic Web—ISWC 2004*, pages 229–243.
- G. Leech and A. Wilson. 1996. EAGLES recommendations for the morphosyntactic annotation of corpora. Version of March 1996.
- M.P. Marcus, B. Santorini, and M.A. Marcinkiewicz. 1994. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
- G. Rehm, R. Eckart, and C. Chiarcos. 2007. An OWL-and XQuery-based mechanism for the retrieval of linguistic patterns from XML-corpora. In *Proc. RANLP 2007*, Borovets, Bulgaria.
- G. Sampson. 1995. *English for the computer: The SUSANNE corpus and analytic scheme*. Oxford University Press.
- A. Saulwick, M. Windhouwer, A. Dimitriadis, and R. Goedemans. 2005. Distributed tasking in ontology mediated integration of typological databases for linguistic research. In *Proc. 17th Conf. on Advanced Information Systems Engineering (CAiSE'05)*, Porto.
- T. Schmidt, C. Chiarcos, T. Lehmborg, G. Rehm, A. Witt, and E. Hinrichs. 2006. Avoiding data graveyards. In *Proc. E-MELD Workshop 2006*, Ypsilanti.
- W. Skut, T. Brants, B. Krenn, and H. Uszkoreit. 1998. A linguistically interpreted corpus of German newspaper text. In *Proc. ESSLLI Workshop on Recent Advances in Corpus Annotation*, Saarbrücken, Germany.
- R. Stuckardt. 2001. Design and enhanced evaluation of a robust anaphor resolution algorithm. *Computational Linguistics*, 27(4):479–506.
- P. Tapanainen and T. Järvinen. 1997. A nonprojective dependency parser. In *Proc. 5th Conference on Applied NLP*, pages 64–71, Washington, DC.
- D. Zeman. 2008. Reusable Tagset Conversion Using Tagset Drivers. In *Proc. LREC 2008*, Marrakech, Morocco.

Creating & Testing CLARIN Metadata Components

Folkert de Vriend (1), Daan Broeder (2), Griet Depoorter (3), Laura van Eerten (3), Dieter van Uytvanck (2)

1) Meertens Institute

Joan Muyskenweg 25, Amsterdam, The Netherlands

2) Max Planck Institute for Psycholinguistics

Wundtlaan 1, Nijmegen, The Netherlands

3) Institute for Dutch Lexicology

Matthias de Vrieshof 2-3, Leiden, The Netherlands

Folkert.de.Vriend@meertens.knaw.nl, {Daan.Broeder, Dieter.vanUytvanck}@mpi.nl,
{Laura.vanEerten, Griet.Depoorter}@inl.nl

1. Introduction

The metadata infrastructure developed in CLARIN (Common Language Resources and Technology Infrastructure) [3] will make use of a computer-supported framework that combines the use of controlled vocabularies with a component-based approach. This framework is called CMDI and is described in [1].

The goal of the project “Creating & Testing CLARIN Metadata Components” is to create metadata components and profiles for existing resources housed at two CLARIN-NL data centres according to the CLARIN Metadata Infrastructure (CMDI) specifications. In doing so the principles supporting the framework are tested. The results of the project will be of benefit to other CLARIN-projects who are expected to adhere to the framework and its accompanying tools. The Max-Planck Institute for Psycholinguistics carries out the coordination and management of the project. The Institute for Dutch Lexicology and the Meertens Institute are the two CLARIN-NL data centres that house the resources for which CMDI metadata profiles are created and tested.

2. Data centres and resources

The Meertens Institute (MI) studies diversity in language and culture in the Netherlands. Its focus is on contemporary research into factors that play a role in determining social identities in Dutch society. Its main fields of research are:

- Ethnological study of the function, meaning and coherence of cultural expressions.
- Structural, dialectological and sociolinguistic study of language variation within Dutch in the Netherlands, emphasising on grammatical and onomastic variation.

The Institute for Dutch Lexicology (Instituut voor Nederlandse Lexicologie; INL) collects and studies Dutch words, stores them in databases – along with various additional linguistic data – and uses them to make scholarly dictionaries. The INL also manages and preserves external (third party) digital language resources, of which availability is facilitated by the HLT Agency department (Human Language Technology Agency).

In the project metadata components were created only for a sub selection of all the resources housed at the two data centres. The most important selection criterion for the resources at MI and INL was that the resources were non multi-media and multi-modal type of resources. For such resources it is expected that the default existing CMDI set that was derived from IMDI [2] (ISLE Meta Data Initiative <http://www.mpi.nl/IMDI/>) is already sufficient. Non-IMDI types of resources are lexica, dictionaries, text corpora and also metadata components for describing collections.

The resources that were selected at the two data centres

vary greatly:

- MI: Lexica, corpora and collections with ethnological data (folktales, songs, probate inventories and pilgrimages).
- INL: Lexica (monolingual and bilingual), corpora (spoken and written), bible texts and dictionaries.

Although CMDI can also be used for creating metadata for tools and web services, in the project these were not taken into account.

3. Creating CMDI metadata

For the selected resources metadata profiles were created. CMDI should be flexible enough for any researcher to decide what metadata fits his or her needs best. The framework offers ready-made metadata components that were derived from existing metadata sets like IMDI and OLAC but it also allows for creating new metadata components if necessary.

Newly introduced metadata elements had to be properly linked to existing concepts in the DCR (Data Category Registry: <http://www.isocat.org/>). Only if existing concepts in the DCR were not accurate enough new concepts were added to the DCR.

4. Testing of CMDI and tool kit

In the project two aspects of the CMDI framework were tested while creating the CMDI metadata:

- Suitability of CMDI for describing the non-IMDI resources at the two data centres.
- Workflow and usability issues in using the tool kit.

Currently an XML editor has to be used for creating the components and XSLT style sheets with the tool kit. Eventually the tool kit will be replaced by a component registry and editor web application.

5. Problems and Challenges

During the project we identified some interesting issues for which no standard solution was provided by the CMDI yet:

- The issue of granularity. Do we provide one big metadata profile for a complete corpus or do we neatly provide profiles at the collection, sub collection and the resource level. Here we tried to offer guidance by developing some special profiles and components.
- When trying to apply the CMDI approach to existing databases, it is often a matter of intuition where to lay the boundary between data and metadata since the data base design makes no difference between them. In the deliverable of the project we hope to give some guidance for this.

6. Final remarks

In our full paper we will discuss our findings in creating and testing of the CLARIN metadata components in detail.

7. References

- [1] D. Broeder, T. Declerck, E. Hinrichs, S. Piperidis, L. Romary, N. Calzolari, and P. Wittenburg. Foundation of a component-based flexible registry for language resources and technology. In 6th International Conference on Language Resources and Evaluation (LREC 2008), 2008.
- [2] D. Broeder and P. Wittenburg. The IMDI metadata framework, its current application and future direction. *International Journal of Metadata, Semantics and Ontologies*, 1(2):119–132, 2006.
- [3] Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*.

Climbing onto the shoulders of standards: TEI as annotation glue and metadata wrapper

Piotr Bański

Institute of English Studies
University of Warsaw
pkbanski@uw.edu.pl

Adam Przepiórkowski

Institute of Computer Science
Polish Academy of Sciences
adam@ipipan.waw.pl

Abstract

The talk reviews and elaborates on our findings concerning standards and best practices regulating various aspects of multi-level corpus annotation: XCES, TIGER-XML, PAULA, as well as the publicly available versions of the TC 37 SC 4 proposed standards. We conclude that the Text Encoding Initiative Guidelines offer mechanisms that straightforwardly encompass these aspects within a single system of interconnected XML schemas. Additionally, these schemas embed documentation that can be easily updated and shared among the distributed centres that participate in corpus creation. Our test case is the 1-billion-word National Corpus of Polish, a TEI-encoded text corpus featuring a hierarchy of stand-off annotation levels, from lightly-tagged source text, through tokenization and sentence boundary layers, a disambiguated morphosyntactic layer and up towards syntactic annotation and into semantics (NE and WSD levels).

1. Contents of the talk

It is by now obvious that standards do not get past their youth if they are not created with care concerning the needs and even quirks of the community that they are meant for. This is why language-resource-related standards are usually built with an eye towards the established best practices and various *de facto* guidelines recommending representation formats and ways to manipulate them.

A challenge presents itself when a single resource requires seamless convergence of several standards, some of them still in the process of being refined, as is the case with ISO SynAF (ISO:24615) or MAF (ISO:24611), and some of them still only in the planning stage (as is true of the Named Entities representation in ISO). Such a challenge is posed by corpora containing multiple layers of linguistic description, from tokenization (not an entirely straightforward task, as the long path of ISO WordSeg-2 demonstrates, ISO:24614-2) through morphosyntactic and syntactic, to semantic levels of annotation (involving the identification of Named Entities and word senses). An additional challenge emerges when the task of creation of such a resource is distributed across several centres and the need for unambiguous guidelines arises.

There are two logical ways to proceed: either to follow the development of standards step by step, with their ups and downs, and temporary misalignments, or to aim at convergence with the best practices of the field that inform the evolving standards. For resources nearing the mature phase, the first option is risky – that was the route followed by the KYOTO project, a route that led up a garden path and away from conformance with the evolving ISO standards (Aliprandi *et al.*,

2009).¹ Following the second option requires a careful survey of the strengths and weaknesses of the current community-adopted practices and a glue or a wrapper that is sufficiently expressive to be able to combine various standards into a coherent whole, sometimes also providing a data model that may be expected to be easily mappable to standards still in the planning phase.

The talk reviews and elaborates on our findings presented earlier, in (Bański and Przepiórkowski, 2009; Przepiórkowski and Bański, 2010; Przepiórkowski and Bański, forthcoming), concerning standards and best practices regulating various aspects of multi-level corpus annotation: XCES (Ide *et al.*, 2000), TIGER-XML (Mengel and Lezius, 2000), PAULA (Dipper, 2005), as well as the publicly available versions of the TC 37 SC 4 proposed standards. We conclude that the Text Encoding Initiative Guidelines offer mechanisms that straightforwardly encompass these aspects within a single system of interconnected XML schemas. Additionally, these schemas embed documentation that can be easily updated and shared among the distributed centres that participate in corpus creation.

Our test case is the 1-billion-word National Corpus of Polish, a TEI-encoded text corpus featuring a hierarchy of stand-off annotation levels, from lightly-tagged source text, through tokenization and sentence

¹ This is not meant as criticism, merely as an example of the danger in following *evolving* standards too closely. On the other hand, it is a truism that not every standard matches exactly the needs or best practices in the given field – compare the ISO/TEI feature structure representation (FSR) standard, ISO:24610-1, which can be criticized for embedding type information with data – something that is stigmatized in both database- and programming language design (cf. Ide, 2010). This has the unfortunate consequence that the ISO LAF standard (ISO:24612) does not follow the FSR standard defined by the same Committee, ISO/TC 37/SC 4.

boundary layers, a disambiguated morphosyntactic layer and up towards syntactic annotation and into semantics (NE and WSD levels). We present these layers and discuss the standards and community practices followed at each step, together with the way in which all of them can be interconnected thanks to the features of the TEI.

We show that the TEI – thanks to its richness that can easily be constrained and tailored to the particular task – provides glue mechanisms for annotations obeying different standards, while wrapping them into a single metadata envelope. The fact that the TEI ODD driver files provide a way to document project-wide decisions and that they embed documentation and markup examples is an additional bonus for distributed projects.

2. References

- Aliprandi, C., Neri, F., Marchetti, A., Ronzano, F., Tesconi, M., Soria, C., Monachini, M., Vossen, P., Bosma, W., Agirre, E., Artola, X., de Ilarraza, A. D., Rigau, G., and Soroa, A. (2009). Database models and data formats. KYOTO Deliverable NR. 1/WP NR. 2, Version 3.1, 2009-01-31.
- Bański, P. and Przepiórkowski, A. (2009). Stand-off TEI annotation: the case of the National Corpus of Polish. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III) at ACL-IJCNLP 2009*, Singapore.
- Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005* (BXML 2005), pages 39--50, Berlin.
- Ide, N. (2010). What does “interoperability” mean, anyway?. Keynote presentation delivered at ICGL-2010, City University of Hong Kong.
- Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based Standard for Linguistic Corpora. *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 825--830.
- Mengel, A., Lezius, W. (2000). An XML-based encoding format for syntactically annotated corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 121--126.
- Przepiórkowski, A., Bański, P. (2010). TEI P5 as a text encoding standard for multilevel corpus annotation. In Fang, A.C., Ide, N. and J. Webster (eds). *Language Resources and Global Interoperability. The Second International Conference on Global Interoperability for Language Resources (ICGL2010)*. Hong Kong: City University of Hong Kong, pp. 133--142.
- Przepiórkowski, A., Bański, P. (forthcoming). XML Text Interchange Format in the National Corpus of Polish. In S. Goźdz-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.
- TEI Consortium (Eds.) (2010). TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 1.6.0. Last updated on February 12th 2010. TEI Consortium. <http://www.tei-c.org/Guidelines/P5/>

Implementing a Variety of Linguistic Annotations through a Common Web-Service Interface

Adam Funk, Ian Roberts, Wim Peters

Department of Computer Science
University of Sheffield
Regent Court, Sheffield, S1 4DP, UK
{a.funk,i.roberts,w.peters}@dcs.shef.ac.uk

Abstract

We present a web service toolkit and common client for a series of natural language processing (NLP) services as a contribution to CLARIN'S European Demonstrator. We have also deployed and tested several natural language processing and information extraction services for English and propose to develop further compatible services using resources for other languages.

1. Introduction

An important goal of CLARIN is to make language resources and technology available to humanities researchers and other end users, especially through web services. The European Demonstrator (Kemp-Snijders et al., 2009) prototype system will integrate a number of resources and services available from the participating institutions. We present here a web service toolkit and client, along with the implementation of a series of NLP and IE services, as a contribution to that system.

2. Web service implementation

Our CLARIN services are standard SOAP web services. They take their input as binary data (as a MIME attachment to the SOAP message, according to the MTOM specification)—though the services are intended to process text they can handle input in many formats including XML, HTML and PDF, extracting the text from the source data using the format handling mechanism provided by GATE. The service loads the input data into a GATE *Document* object, then processes that Document using a GATE *DocumentProcessor* (typically a wrapper around a saved GATE application), and returns its output as XML data (any valid XML element is allowed for versatility). (Please refer to the GATE manual (Cunningham et al., 2010) and API documentation¹ for details of the GATE library.)

All the services share a common WSDL interface as their inputs and outputs are the same; only the underlying GATE application needs to vary between services. The standard interface simply specifies the output as any XML in any namespace, and the implementation does not restrict the XML that the underlying application can produce. However if the output types for a specific service are known and there is a suitable W3C XML Schema available then there is the option to use a custom WSDL for that service with the more

constraining schema included, which may be beneficial for certain types of client.

The various components making up the service implementation are configured using the Spring framework, making it simple to slot in alternative *DocumentProcessor* implementations for different services without changes to the code. The aspect-oriented programming tools provided by Spring are used to allow pooling of several identical DocumentProcessors, to support multiple concurrent web service clients. The web service layer is provided by the Apache CXF toolkit, which itself uses Spring extensively and thus was a good fit with the Spring-driven architecture adopted for the business logic.

This toolkit will work easily for any GATE application, typically a *SerialAnalyserController* or *ConditionalSerialAnalyserController* (corpus pipeline); furthermore, with suitable modification in the Spring beans configuration, it can use any class that *implements DocumentProcessor*—in effect, any class that can analyse a single GATE Document (or a GATE *Corpus* containing one Document) and produce any valid XML document (the root element of which which we treat as the result and embed in the SOAP response). Each web service provides a WSDL file available from the server and offers three methods:

- *process* send only the document content as a `byte[]`;
- *processWithURL* also sends the document URL—the GATE *Factory* will take the filename into consideration when instantiating the Document (to distinguish PDFs properly, for example);
- *processWithParams* sends a parameter list, which allows the client to specify the original URL, encoding, and mime type (this method allows the greatest flexibility).

3. Services currently available

We have implemented the following services so far, making use of standards which we have worked with in

¹<http://gate.ac.uk/documentation.html>

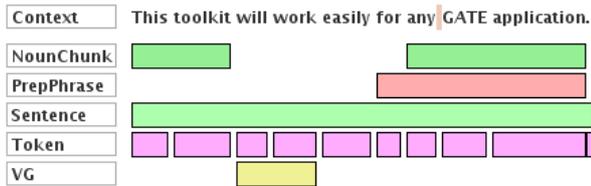


Figure 1: Spans of syntactic elements produced within the chunking service

previous projects (particularly LIRICS², SEKT³, and MUSING⁴).

- The *annie-alpha* service runs the ANNIE (Cunningham et al., 2002) named-entity recognition and orthographic co-reference pipeline and returns the fully annotated document in GATE XML format. The file saved by the client contains ANNIE’s output in the default AnnotationSet and the input document’s HTML or XML mark-up in the “Original markups” AnnotationSet.
- The *maf-en* service runs GATE’s sentence-splitter, tokenizer, POS-tagger and morphological analyser (lemmatizer) for English, and returns an XML document containing the morphosyntactic information according to the MAF (ISO, 2008) standard.
- The *chunking-synaf-en* service runs GATE’s sentence-splitter, tokenizer, POS-tagger, and NP and VP chunkers (Cunningham et al., 2010, §17) for English, as well as a simple PP chunker, to produce annotations as shown in Figure 1 (where *VG* means verb group). The application then constructs a simple syntactic tree for each sentence based on simple containment (each phrase or token annotation is a constituent of the smallest sentence or phrase annotation containing it), as shown in Figure 2, and returns an XML document according to the SYNAF (ISO, 2010) standard.

The syntactic detail is not complete but chunking and constructing a tree this way is reasonably accurate and reliable for many purposes and much faster (especially for large documents over a web service) than full parsing. (The verb chunker’s annotations also contain features indicating tense, voice, etc., which will be incorporated into the SynAF output in the improved version of this service.)

- The *annie-rdf* service runs ANNIE, then analyses ANNIE’s annotations by type and features and generates RDF representing the recognized entities as instances according to the PROTON⁵

²<http://lirics.loria.fr/>

³<http://www.sekt-project.com/>

⁴<http://www.musing.eu/>

⁵<http://proton.semanticweb.org/>

(Terziev et al., 2005) ontology, and returns an RDF-XML document.

4. Reference client

We also provide a GUI Java client, supplied as a ZIP file with the necessary libraries, so the user needs only a Java 5 runtime environment (JRE). This client uses the *processWithParams* method and sends the `file://` URL, and user-selected encoding along with the content of the selected local file. The user selects the service from the list of endpoint URLs included with the client, but can also type in a URL if he is aware of a service that has been added since the client software was issued. Figure 3 shows this client’s main panel used for sending files to the services, and Figure 4 shows the output panel, which allows the user to inspect the output and save it to a local file.

Of course, developers can also use the services’ WSDL files to produce their own clients for users’ direct use or embedment in other software.

5. Conclusion and future work

The services described here have been proposed as contributions to CLARIN’s European Demonstrator. We also plan to deploy services with MAF output for some other European languages (probably a selection from Bulgarian, Dutch, German, and Spanish) in the near future, based on the resources we have available, and are open to suggestions for others, especially if suitable language resources and processing tools are available to be shared with us and suitable for integration with GATE.

Acknowledgements

This research is partially supported by the European Union’s Seventh Framework Program project CLARIN (FP7-212230).

6. References

- H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL’02)*.
- H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, M. Dimitrov, M. Dowman, N. Aswani, I. Roberts, Y. Li, A. Funk, G. Gorrell, J. Petrak, H. Saggion, D. Damjanovic, and A. Roberts. 2010. *Developing Language Processing Components with GATE Version 5.0 (a User Guide)*. The University of Sheffield.
- ISO. 2008. Language resource management—morphosyntactic annotation framework (MAF). Standard ISO/DIS 24611, ISO TC37/SC4, December.
- ISO. 2010. Language resource management—syntactic annotation framework (SynAF). Standard ISO/DIS 24615, ISO TC37/SC4.

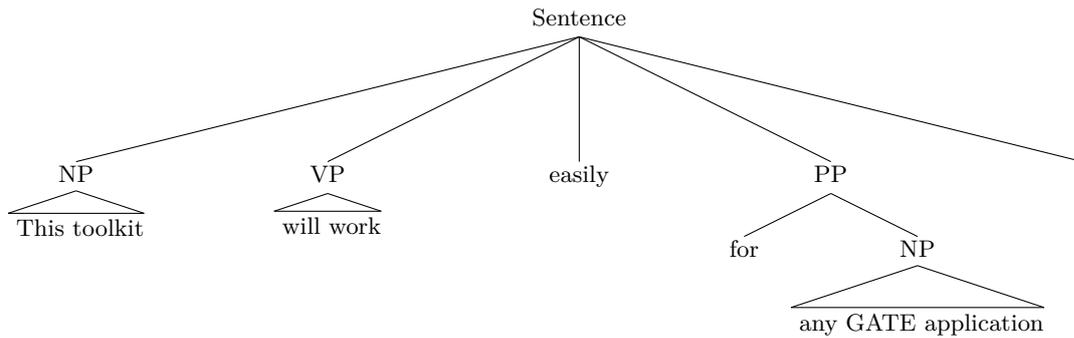


Figure 2: Approximate syntactic tree produced from the annotations in Figure 1

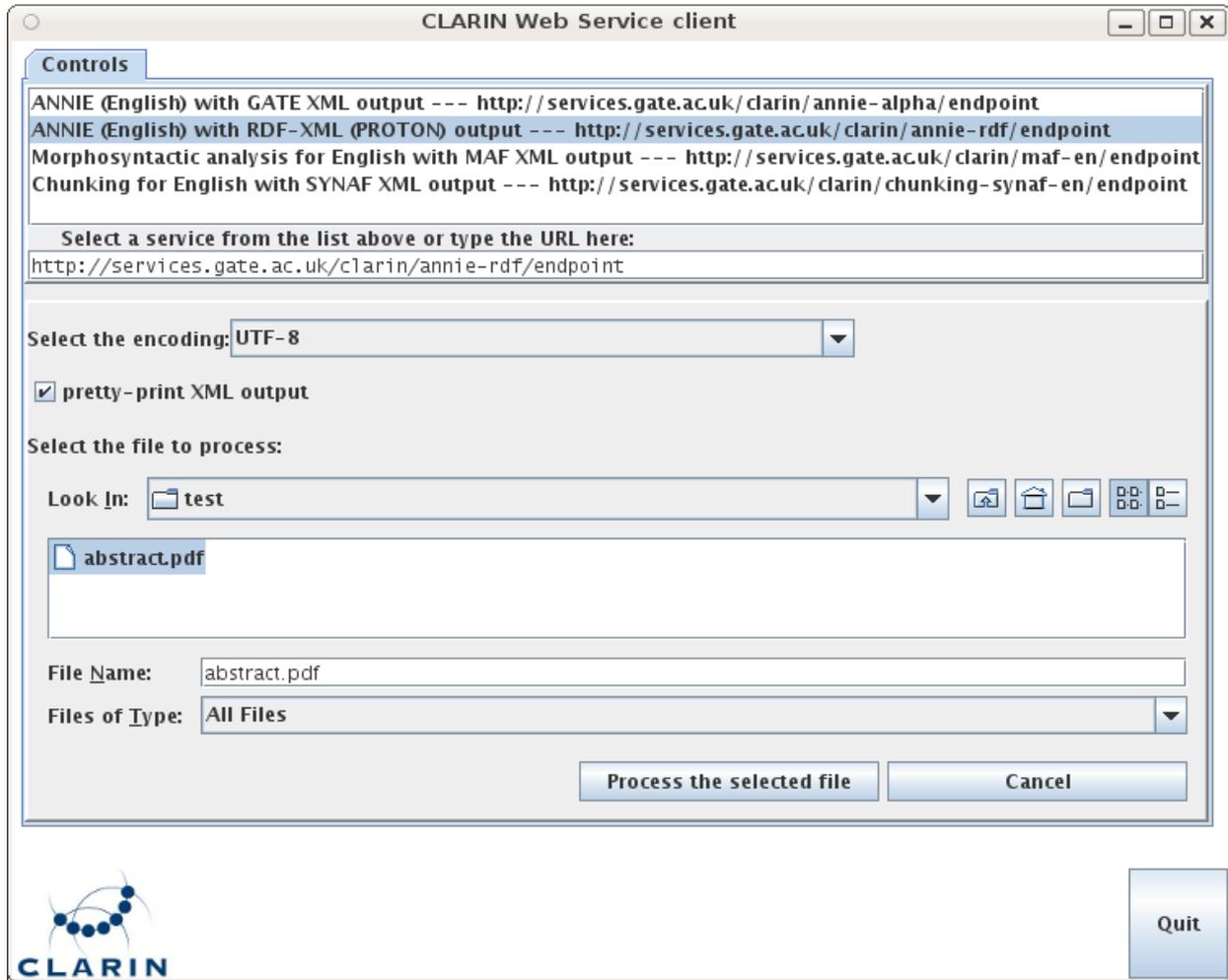


Figure 3: Main panel of the GUI client

Marc Kemps-Snijders, Núria Bel, and Peter Wittenburg. 2009. Proposal for a CLARIN European demonstrator. Technical report, CLARIN Consortium, September.

I. Terziev, A. Kiryakov, and D. Manov. 2005. Base upper-level ontology (BULO) guidance. Deliverable D1.8.1, SEKT Consortium, July.

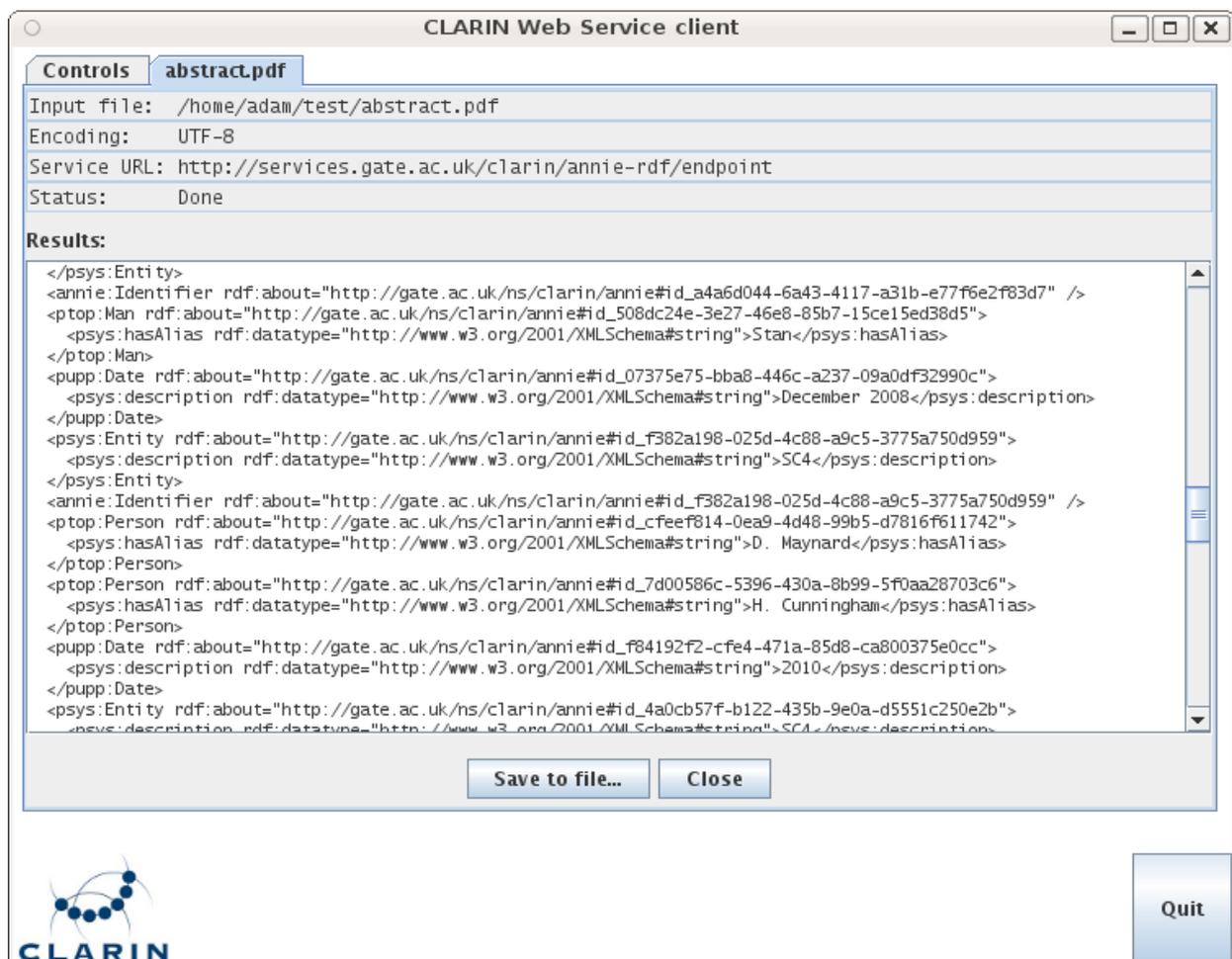


Figure 4: Output panel of the GUI client

Survey on the Use of XLIFF in Localisation Industry and Academia

Dimitra Anastasiou

Centre for Next Generation Localisation,

Localisation Research Centre,

Department of Computer Science and Information Systems

University of Limerick,

E-mail: Dimitra.Anastasiou@ul.ie

Abstract

The XML Localisation Interchange File Format (XLIFF), managed by Organization for the Advancement of Structured Information Standards (OASIS), first released in 2002, is an open standard for translation and localisation which allows for the exchange of data between software publishers, localisation vendors, or localisation tools. The OASIS XLIFF Technical Committee defines and promotes the adoption of a specification for the interchange of localisable software-based objects and metadata, so that the real content (data) and the data about data (metadata) can be transmitted smoothly through several localisation phases, from digital content creation, through to internationalisation, localisation, and actual content generation. This paper presents the results of an XLIFF survey that was conducted in order to evaluate the adoption of XLIFF by different stakeholders, both from industry and academia, be they tool developers, customers, or translators. The general structure of XLIFF was rated as good by more than half of the respondents, the synergy with other standards, such as ITS (W3C) and TMX (LISA) was rated more than desirable, while some of the changes recommended for XLIFF's structure are increased simplicity, modularisation, clarification of necessary metadata, and better workflow control.

1. Introduction

According to the Localisation Industry Standards Association¹ (LISA), localisation is defined as follows:

“Localisation involves the adaptation of any aspect of a product or service that is needed for a product to be sold or used in another market.”²

Localisation is distinct from translation, because it is not only text being transferred from one source language (SL) to a target language (TL), but also, icons/images, audio, video, colours, layout, and other “aspects of products or services”.

In terms of localisation, while *data* is the actual content to be localised, *metadata* is the data about data, i.e. describing, explaining, and processing other data. Metadata is undoubtedly as important as data, since the former provides structure and order to the latter, and generally defines a clear workflow.

Localisation metadata not only defines and supports a clear workflow, but also connects the data present at different localisation workflow stages. Metadata is very useful during the digital content creation, linguistic annotation, content maintenance, translation (Translation Memory (TM) and Machine Translation (MT) usage), proofreading/postediting, content generation of the localised content, and process management in general.

In section 2 of this paper we examine some localisation standards, and focus on the XML Localisation Interchange File Format (XLIFF) standard. Section 3, the main body of the article, focuses on a survey that we conducted pertaining to the adoption of XLIFF, the difficulties of supporting it, recommended changes, etc.

In section 4 we introduce the Centre for Next Generation Localisation (CNGL) project and we describe the tasks of the metadata group set up within CNGL. A conclusion and future prospects of our research combining our membership of the XLIFF Technical Committee (TC) and chairing the metadata group in CNGL are found in the last section of the paper 5.

2. Standards – Related work

There are many standards bodies, such as ISO, W3C, LISA, OASIS, etc. Each of these organizations manages standards on different aspects of information management and technology topics; we focus though more on the localisation-related standards.

In 2006, an initiative called MultiLingual Information Framework (MLIF)³ was standardised (ISO/TC37/SC4) providing a common platform for all existing tools and promoting the use of a common framework for the future development of several different formats: TMX (LISA), XLIFF (OASIS), etc. MLIF introduces a metamodel for multilingual content in combination with data categories as a means to ensure interoperability between several multilingual applications and corpora. MLIF examines at morphological description, syntactical annotation, or terminological description. An important point about MLIF is that it does not propose a closed list of description features, but rather provides a list of data categories, which are much easier to update and extend.

In addition, the Open Architecture for XML Authoring and Localization⁴ (OAXAL) is another framework which takes advantage of the Darwin Information Typing Architecture (DITA) standard from OASIS and also

³ <http://mlif.loria.fr/> and

http://www.tc37sc4.org/new_doc/ISO_TC37-4_N266_WD_Multilingual_resource_management.pdf, 29.04.2010

⁴ <http://www.xml.com/pub/a/2007/02/21/oaxal-open-architecture-for-xml-authoring-and-localization.html>, 29.04.2010

¹ <http://www.lisa.org/>, 29.04.2010

² <http://www.lisa.org/Localization.61.0.html>, 29.04.2010

XML-based text memory (xml:tm) standard from LISA's standards committee called Open Standards for Container/content Allowing Reuse (OSCAR).

A localisation standard released by W3C, the Internationalization Tag Set (ITS), is designed to be used by schema developers, content authors, and localisation engineers to support the internationalisation and localisation of schemas and documents. It includes data categories and their implementation as a set of elements and attributes. In ITS there is a global and local approach; one of the benefits of the global approach is that the content authors make changes in a single location, rather than by searching and modifying the markup throughout a document.

To give an example related to localisation in ITS global approach, one can look at the translateRule element. It includes a translate attribute with boolean value (in this case "no") and a selector. The selector contains an XPath expression which selects the nodes. Rules apply to these nodes (see Table 1).

```
<its:rules
xmlns:its="http://www.w3.org/2005/11/its"
  version="1.0">
<its:translateRule translate="no" selector
="//code"/>
</its:rules>
```

Table 1: ITS example of global approach
Source: <http://www.w3.org/TR/its/#translatability-implementation>

The conversion tool from ITS2XLIFF⁵ (v 0.6) developed by Felix Sasaki⁶ is worth mentioning here. His tool allows users to generate up to date XLIFF files (v 1.2) from XML files for which W3C ITS rules are available. LISA/OSCAR standards⁷ include:

- Translation Memory eXchange (TMX);
- Segmentation Rules eXchange (SRX);
- Term-Base eXchange (TBX);
- XML Text Memory (xml:tm);
- Global Information Management Metrics eXchange - Volume (GMX-V).

TMX is probably the most well known LISA/OSCAR standard as it exchanges TM data between applications, being commercial or open-source.

We now turn our focus to XLIFF. For information about a brief history see Reynolds and Jewtushenko (2007). As previously mentioned, XLIFF is managed by OASIS, a not-for-profit consortium which also produces many Web service procedures. According to OASIS, XLIFF is defined as follows:

"XLIFF is [...] designed by a group of software providers, localisation service providers, and localisation tools providers. It is intended to give any software provider

*a single interchange file format that can be understood by any localisation provider."*⁸

An example⁹ of an XLIFF translation unit <trans-unit> element is visible in Table 2:

```
<trans-unit id="#1" datatype="plaintext">
<source xml:lang="en-us">file</source>
<target state="needs-translation"
xml:lang="de-DE" resname="String"
coord="-0;-0;-0;-0">Datei</target>
</trans-unit>
```

Table 2: XLIFF example of trans-unit element

The actual data in this example shows that *file* in the source language (SL) English (US) English means *Datei* in German. XLIFF is a standard that can carry a large amount of metadata, as we can see from the example in Table 2: data type (datatype="plaintext"), resource name (resname="String"), and coordinates (coord="- 0;-0;-0;-0").

Another XLIFF element important to localisation is the alttrans element (see Table 3). This element contains possible alternative translations, e.g. in <target> elements along with optional context, notes, etc. (see Table 3):

```
<alt-trans match-quality="80%" tool="XYZ">
<source>file type</source>
<target xml:lang="de-DE"
phase-name="pre-trans#1">Dateityp
</target>
</alt-trans>
```

Table 3: XLIFF example of alt-trans element

Here we have a matching percentage of 80%, because in the TM of the tool XYZ we had the entry *file type* translated as *Dateityp*.

There is a relationship between TMX and XLIFF in that XLIFF 1.2 borrows from the TMX 1.2 specification, but they are different standards, each having their own format. Inline markup XLIFF support in TMX 2.0 is currently in progress.

Based on Rodolfo Raja's (2007) article "*XML in localisation: Reuse translations with TM and TMX*"¹⁰, we created the following table which distinguishes between TMX and XLIFF and also brings them into symbiotic relationship, see Table 4.

	Definition	Synergy	Conversion
TMX	Standard for the exchange of TM data created by CAT and localisation tools	Should be used as a complement to XLIFF	XSL transformation to convert an XLIFF file to TMX format
XLIFF	Format for	Possible	

⁸ XLIFF Specification 1.2: <http://docs.oasis-open.org/xliff/v1.2/os/xliff-core.html>, 29.04.2010

⁵ <http://fabday.fh-potsdam.de/~sasaki/its/>, 29.04.2010

⁶ Also, Christian Lieske contributed to the development of this tool.

⁷ <http://www.lisa.org/OSCAR-LISA-s-Standa.79.0.html>, 29.04.2010

⁹ This is not a full valid XLIFF file, but only an element and its contents.

	exchanging localisation data	translations contained in the <alt-trans> elements of an XLIFF file are extracted from a TM database	
--	------------------------------	--	--

Table 4: XLIFF-TMX

It is noteworthy that XLIFF has a working relationship with LISA/OSCAR standards and is a requirement for the standards TMX, GMX-V, and xml:tm.

1. XLIFF survey

As aforementioned, XLIFF was first released in 2002 and its latest version is 1.2 (approved as a Specification in February 2008). Currently, the XLIFF TC is working on the specifications of the next XLIFF release 2.0¹¹.

Recently both commercial and open-source tools have supported XLIFF. Examples of commercial tools supporting XLIFF are Swordfish, XTM, SDL TRADOS, Alchemy Catalyst, memoQ, and some open-source examples are OmegaT, Virtaal, etc. There are more tools which support the format of XLIFF, but mentioning these tools and their diverse support is outside the scope of this particular paper.

In fact, the interest in XLIFF is not total and it is only in recent years that more and more tools have begun to support it. But is this XLIFF support full and proper support or does it only cover the basic features? Also, how often are there cross-tool operations and when are they successful?

These reasons motivated us to conduct a survey about XLIFF in terms of primary research. One questionnaire consists of eight questions, five of which are multiple choice questions and three ask for general feedback. The survey was created by the author with the help of Reinhard Schäler, director of the Localisation Research Centre (LRC). The questionnaire is available online¹² and copies of the questions can be found in the Appendix (section 8).

70 respondents completed the questionnaire. Half of the responses were received after distributing the questionnaire at the tc world conference in Wiesbaden, 2009; the other half was collected by sending the questionnaire to mailing lists.

The 70 respondents of the survey were either from industry (tool providers (33%) and localisation service providers (17%)) or from Academia (CNGI¹³ researchers (22%)). The remainder were translators (11%), content publishers (6%), and others, e.g. consultants, students (11%). The distribution of the survey's respondents (first question of the questionnaire) can be seen in Diagram 1:

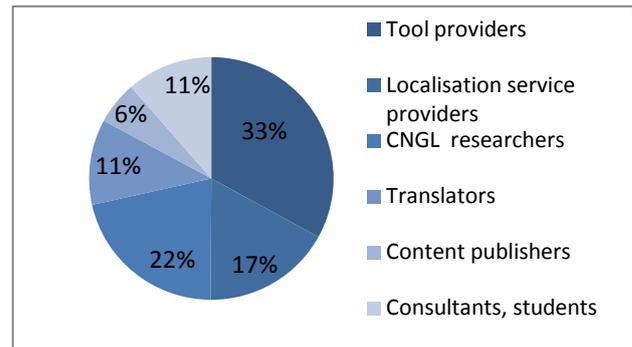


Diagram 1: Respondents

The second question posed is whether the technologies and tools the respondents use are XLIFF-compliant. This question received a positive response from 33% and a negative response from 20%. The remaining 47% was split to four categories. The first category is by those where some technologies are XLIFF-compliant while some others are not (20%). The second category featured 17% who heard of but were not exactly aware of what XLIFF is about, and the third category concerns 3% who gave other answers, such as "not yet, as tools are now compliant". In the fourth category, 6% stated that they had never heard of XLIFF before. The distribution of the percentages regarding this question is shown in Diagram 2:

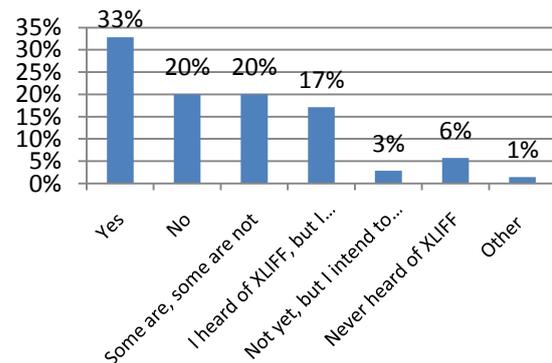


Diagram 2: XLIFF-compliance

The third question concerned the use of XML in XLIFF and whether it should be supported by namespaces. This question is included, because, in our opinion, tool providers often customise the XLIFF document at an extreme level. The responses were diverse; some people would prefer for more extensibility, while some others argue that XLIFF is already too flexible. What most respondents answered (and we agree with) is that if XLIFF is extremely user-defined and there are custom namespaces for every different CAT tool, then the cross-tool operations will lead to data loss. According to some respondents' answers, the solutions to that would be "stronger standards and not just guidelines", "tools that comply to proper XLIFF coding", and "starting with a simple base and expanding".

Moving towards the fourth question "Should there be more synergy between XLIFF and other standards?", the predominant answer is yes. The feedback received was that XLIFF should be in synergy with TMX, ITS, and GMX-V. According to some respondents, the

¹¹<http://wiki.oasis-open.org/xliff/XLIFF2.0/FeatureTracking>, 29.04.2010

¹²http://ai.cs.uni-sb.de/~stahl/d-anastasiou/Survey/XLIFF_questionsnaire.pdf, 29.04.2010

¹³<http://www.cngl.ie/>, 29.04.2010

LISA/OSCAR process of developing specification should be more open, so that LISA and OASIS work better together. One noteworthy answer was that strict synchronisation is not necessary, as every standard focuses on a particular part of the localisation workflow. The fifth question asked about problems that users of XLIFF face. If we categorise the answers, we come up with the following:

- Difficulties converting end client formats;
- Tools unable to handle and support XLIFF (and also the same way);
- Infrequent use by publishers/clients and lack of acceptance by professional translators;
- Lack of filters.

Based on these problems, we asked the respondents what they would change in XLIFF. Again, we tried to categorise the feedback which is as follows:

- Simplification of the inline markup;
- Stronger compliance requirements ;
- Clarification of necessary metadata;
- Better workflow control;
- Support for multilingual content.
- Tools which create and use rather than just import XLIFF;

We ended the questionnaire by asking how people would evaluate the general structure of XLIFF. Most regarded it as good (59%), followed by very good (24%), and average (17%); nobody chose the “not good” option.

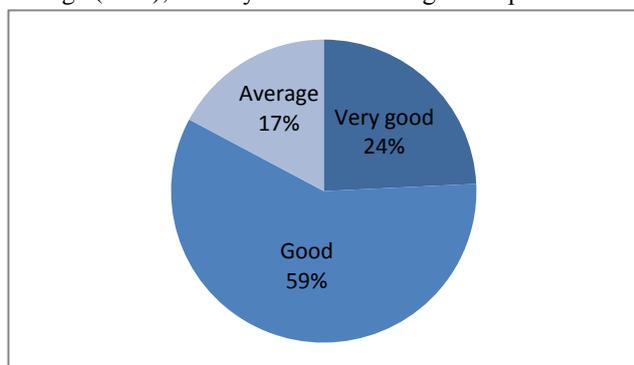


Diagram 3: General structure of XLIFF

2. CNGL

The Centre for Next Generation Localisation (CNGL) project is an Academia-Industry partnership with 100 researchers working on MT and Speech, Digital Content Management, Next Generation Localisation, and System Framework.

As the chair of the metadata group set up in June 2009 in the terms of the CNGL project, the author investigates which metadata is currently used in this project and makes recommendations for future metadata requirements. The goal of the metadata group is to develop a framework which subsumes all of the metadata which must ensure the integrity and interoperability of data as it passes through the areas of content production, localisation, and consumption, as well as asset and process management.

As a member of the XLIFF Technical Committee, the author also examines the use of XLIFF within CNGL;

more precisely, we see whether XLIFF’s specifications suffice for the CNGL’s needs and if not, we collect XLIFF’s limitations and make recommendations for the next XLIFF releases.

3. Conclusion and Future prospects

A clear distinction between data and metadata is necessary, particularly in the process of developing specification for standards. Our survey has shown that XLIFF’s structure is generally regarded as good, although more simplification, modularisation, as well as more and better adoption by both tools and customers is required. All of the feedback was useful and certainly the XLIFF TC takes that on board and will go towards direction that suffices the needs of the users.

Our future work will be divided in two directions. Firstly, we intend to provide a common metadata framework within CNGL which subsumes all meta-information needed at the different localisation stages. Secondly, we started collecting and arranging, in an hierarchical order, the metadata that exists in XLIFF v1.2 and make recommendations for the next release. To sum up, as chair of the metadata group in CNGL and a member of the XLIFF TC, the author intends to take the outcomes of CNGL research and implement it into the XLIFF standard.

4. Acknowledgements

This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at the University of Limerick.

5. References

- Raja, R. (2007). XML in localisation: Reuse translations with TM and TMX.
<http://www.maxprograms.com/articles/tmx.html> and
<http://www.ibm.com/developerworks/library/x-localizations3/>.
- Reynolds, P., Jewtushenko, T. (2005). What Is XLIFF and Why Should I Use It?. *XML journal*:
<http://xml.sys-con.com/node/121957?page=0,3>.

6. Appendix

Questions of the XLIFF questionnaire

1. You are a: Tools provider/Content Publisher/LSP/Translator/Consultant/Other
2. Are the technologies and tools you use XLIFF-compliant?
3. If you have not implemented XLIFF, why not?
4. What is your opinion about the XML implementation in XLIFF, e.g. namespaces?
5. Should there be more synergy between XLIFF and other standards (Internationalization Tag Set – w3c or LISA’ standards). In which way?
6. Which problems do you face using XLIFF?
7. What changes would you recommend in XLIFF?
8. What is your opinion about the general structure of XLIFF?

Towards a standard for annotating abstract anaphora

Stefanie Dipper, Heike Zinsmeister

Institute of Linguistics, Institute of Linguistics
Bochum University, Konstanz University
dipper@linguistics.rub.de, Heike.Zinsmeister@uni-konstanz.de

Abstract

This paper presents a survey on the annotation of abstract (= discourse-deictic) anaphora, i.e. anaphora that involves reference to abstract entities such as events and propositions. The survey identifies features that are common to the majority of relevant annotation efforts. Based on these, we propose a small set of recommendations, which can be viewed as a first, small step towards a standard for the annotation of abstract anaphora. As the overview shows, English is the language that most of the resources have been created for. However, many of them contain only few instances of abstract anaphora. Hence, the currently available evidence only supports preliminary conclusions.

1. Introduction

The paper presents a survey on the annotation of abstract (= discourse-deictic) anaphora, i.e. anaphora that involves reference to abstract entities such as events and states (Asher, 1993). The survey tries to identify the features that are common to the majority of relevant annotation efforts, and can be viewed as a first, small step towards a standard for the annotation of abstract anaphora.

In the last couple of years, several related but nevertheless different approaches have been proposed for both the inventory of annotation tags as well as the coding schemes of the relations and the markables in the texts. The differences can be traced back to the following reasons:

- the “theory” behind abstract entities: e.g. whether abstract entities are defined by reference to syntactic, semantic, and/or pragmatic properties;
- the kind of data that is analyzed: e.g. dialogues/spoken language or written text;
- the language under consideration: e.g. languages with zero pronouns or clitics require annotation schemes different from schemes for English data.

The contribution of this paper is a survey of the state of the art of abstract anaphora annotation which highlights categories that are generally agreed on and takes different points of view on anaphoric encoding into account. The paper first addresses proposals that have been made with regard to representational issues (Sec. 2.). In Sec. 3. we describe relevant annotation efforts, followed by our recommendations in Sec. 4. The Appendix contains a synoptic table of the studies considered in this paper.

2. Standards

Annotation of abstract anaphora is not an easy task. Hence, just like all precious resources, corpora annotated with abstract anaphora relations should be maximally reusable and exploitable for further applications. Maximal reusability can be achieved by adherence to standards, which regulate both *content* and *form* of annotation. This paper deals with the first aspect.

Content is standardized by means of tagsets, specifying obligatory and/or optional tags (“data categories”), along

with annotation guidelines. We do not know of any proposal to standardize content of abstract anaphora annotation, and propose a small set of recommendations in Sec. 4. *Form* is standardized by reference to data models and physical data structures, which are used, e.g., for data interchange. Data structures are specified, e.g., in the form of DTDs or XML schemata, which define an XML representation format. We would like to point out two proposals, MATE/GNOME (Poesio 2000a/b)¹ and RAF (Reference Annotation Framework, Salmon-Alt and Romary (2004)). Both proposals agree in that they do *not* encode anaphoric relations by pointers that are attached to the anaphor and point to the antecedent. Instead, they define extra, autonomous elements that represent the anaphoric relation. This opens up the possibility of easily annotating a discourse entity with multiple anaphoric links, as well as recursively defining complex markables, or annotating empty strings, such as zero pronouns.

A MATE-style XML example of an anaphoric “identity” link would look as follows (the outer element refers to the anaphor, the embedded element to its “anchor”, i.e. its antecedent):

```
<coref:link type="ident" href="...">  
  <coref:anchor href="..." />  
</coref:link>
```

3. State of the art

In this section, we present a series of relevant work on abstract anaphora.² The main features are summarized in Table 1 in the Appendix. This overview will lead us to a comparative assessment of the features used in the annotations. Column 1 of Table 1 lists the authors of the study. Columns “Data” (2–4) inform about the data used in the research:

¹The original GNOME scheme is restricted to concrete anaphora annotation.

²See Recasens (2008, ch.2) for a similar overview of coreference annotation in general (MUC, ACE, MATE, and AnCora schemes), including a comparison of annotated English and non-English corpora. Müller (2008, ch.2 and ch.5) contains a discussion of different projects of abstract anaphora annotation.

column 2 displays the codes of the language(s) that the papers deal with; columns 3 and 4 presents general and statistical information about the corpus.³ Columns “Anaphor” and “Antecedent” focus on the syntactic and semantic properties that are taken into account by these studies.⁴ “Reliability” columns report whether inter-annotator agreement has been computed. For anaphors, agreement is computed for semantic annotation; for antecedents, agreement usually concern the marking of segment boundaries. Column “Criteria” indicates whether the study provides tests (e.g., in the form of annotation guidelines) that can be applied by the annotators.^{5, 6}

3.1. The anaphor

The majority of research considered here restricts their investigations to pronominal anaphors. Exceptions are Vieira et al. (2002), Poesio and Modjeska (2005), and Botley (2006), who consider *this*- and *that*-NPs, i.e. “full” NPs which start with the respective (translated) demonstrative determiner, and Recasens and Martí (2010), who take all kinds of NPs and pronouns into account.

Identifying pronouns in general is considered a trivial task. However, identifying *abstract* (also called: discourse-deictic, indirect) anaphors and distinguishing them from *concrete* (also called: individual) anaphors is a relevant issue. Hence, reliability studies for this task provide important information. However, not all studies distinguish between abstract and concrete anaphora but define other basic classes. In addition, more fine-grained labels are sometimes introduced. In this paper, we only consider labels that apply to abstract anaphora. Labels for concrete anaphora and pronouns are subsumed under “others” (see Table 1), and reported along with the total number of such labels.

Eckert and Strube (2000), Navarretta and Olsen (2008), and Dipper and Zinsmeister (2009a) define *vague* anaphora, which refers to some general discourse topic which is not overtly expressed.⁷ Müller (2008) uses the label *vague* in a more general sense, to mark pronouns with no clearly-defined textual antecedent.

³Abbreviations used: *T*: total number of tokens; *C*: anaphora candidates (e.g., number of NPs); *AA*: number of abstract anaphors.

⁴Abbreviations used: *Dem*: demonstrative pronouns, *Pers*: personal pronouns, *Poss*: possessive pronouns, *Rel*: relative pronouns, *Zero*: zero pronouns, *Cl*: clitics, *Expl*: expletives/idioms; *Dem-NP*: NP with a demonstrative determiner; *AA*: abstract anaphors, *concr*: concrete, *abstr*: abstract, *non-ref*: non-referring, *indir*: indirect.

Clauses means that antecedents are syntactically defined, e.g. as sentences, infinitives, gerunds; *V-head* means that only the verbal head is marked.

⁵Several annotation guidelines make use of GNOME, e.g., Poesio and Modjeska (2005) and Navarretta and Olsen (2008), but only for the annotation of concrete anaphora.

⁶Goecke et al. (2008) present an annotation scheme for anaphoric relations in German, which includes specifications for abstract entities. Abstract types are defined syntactically (propositions and projective propositions) or semantically (events, event-types, states). In their project, however, only concrete anaphora has been annotated.

⁷This label is called *deict* in Dipper and Zinsmeister (2009a).

Botley (2006), who considers *this*-NPs, investigates the semantics of the (abstract) anaphoric head nouns in detail. He distinguishes three main types of abstract anaphora: (i) “Label” anaphora, which serves to encapsulate (or to label) stretches of text (following Francis (1994)). Label anaphora is further classified as *general* or as *metalinguistic*, with subtypes *illocutionary*, *language activity*, *mental process*, *text*. (ii) “Situation” anaphora, with subtypes *eventuality* (e.g. events, processes, states) and *factuality* (e.g. fact, proposition) (following Fraurud (1992)). (iii) “Text deixis”.

Distinctions similar to Botley’s “situation” anaphora subtypes are made by Hedberg et al. (2007), Navarretta and Olsen (2008), and Dipper and Zinsmeister (2009a). Recasens and Martí (2010) define subtypes *token*, *type*, *proposition*. In contrast to most other work, Dipper and Zinsmeister (2009a) annotate these subtypes both to the abstract anaphors and their antecedents (see Sec. 3.2.).

Poesio and Artstein (2008) annotate the reference status of NPs and pronouns: *anaphoric*, *discourse-new*, *non-referring*. In addition, they classify them semantically, e.g. as *person*, *animate*, *concrete*, *space*, *time* etc.

Sometimes, abstract anaphora is subsumed under the more general label *indirect*, see Botley (2006), Vieira et al. (2002), and, with a slightly different classification, Hedberg et al. (2007).⁸ Other members of these classes are bridging relations, occurring with concrete anaphora. Bridging relations are akin to abstract anaphora in that antecedents are not readily available but require additional interpretational efforts.

Kučová and Hajičová (2004) define the label *text* for inter-sentential general coreference relations, and the label *segm* which is used for anaphors with multi-node/multi-rooted antecedents (in the dependency framework).

Usually, *all* referring pronouns are annotated, and reliability results are reported that measure inter-annotator agreement on the entire set of referring pronouns (and their antecedents). Whenever appropriate information is available, we distinguish between agreement on personal and demonstrative pronouns. Personal pronouns (at least in English) predominantly refer to concrete entities, demonstrative pronouns often refer to abstract entities. The results listed in Table 1 indicate that—as is expected—anaphora resolution is considerably easier with concrete entities than abstract entities.

If no identification criteria and/or reliability results are listed in Table 1, this means that none are mentioned in the respective papers.

3.2. The antecedent

Antecedents of abstract anaphora are abstract objects, such as actions and events. Accordingly, they correspond to linguistic entities which include at least a verb: partial clauses, clauses, sequences of sentences, or even discontinuous strings, as illustrated by the following example (the antecedent of the anaphor *it* is underlined):⁹

⁸This label is called *other* in Vieira et al. (2002).

⁹Example taken from file `ep-04-03-31.txt` of the Europarl corpus.

- (1) I would like to draw particular attention to the fact that people who have made their lives here in the European Union still do not have the right to vote, even though the European Parliament has called for **it** on many occasions.

The approaches differ as to whether they restrict the marking of the antecedent to the verbal head, as in Müller (2007) or Pradhan et al. (2007), or approximate it by predefined constituents, e.g. clauses (Byron, 2003), or whether the annotators are allowed to mark free spans of text, e.g. Vieira et al. (2002), or Dipper and Zinsmeister (2009a). Dipper and Zinsmeister aim at determining the exact scope of the anaphor, i.e., the exact extension of the antecedent's string, including examples as (1). For this task, they propose a paraphrase test. Other annotation efforts deliberately do not aim at identifying exact boundaries, some even do not mark antecedents at all (Poesio and Modjeska, 2005).

The identification of antecedents is easier in monologue or written texts than in dialogues (Poesio, 2004). In general, identifying antecedents is easier in monologues or written texts than in dialogues (Poesio, 2004). For instance, different speakers may have different assumptions about the situation. In addition, incomplete or ungrammatical sentences often occur in spoken language, due to hesitations or disfluencies. Therefore, annotations of dialogues often recur to independently-defined units, such as dialogue acts (e.g., Eckert and Strube (2000)).

Kučová and Hajičová (2004), just like other approaches, mark the verbal head of the antecedent. However, in their dependency framework, the verbal head is the root node of the clause, and, hence, the marking specifies the extension of the antecedent.

Further properties of the antecedent are investigated only in a subset of the studies. Hedberg et al. (2007) consider the saliency of the antecedent to specify the cognitive status of the anaphor. Dipper and Zinsmeister (2009a) determine the semantic subtype of the antecedent by a replacement test, deliberately ignoring the anaphor.

3.3. Summary

As can be seen from Table 1, semantic annotation is considered more relevant for anaphors than for antecedents. Annotation efforts that consider both concrete and abstract anaphors often annotate the distinction concrete–abstract for anaphors. With NP anaphors, the head noun determines its class, e.g. *this situation*. With pronominal anaphors, people apply two strategies: (i) The clausal context of the pronoun, e.g., its governing verb, impose selectional constraints on the semantic type of the anaphor. (ii) The semantics of the antecedent is used to determine the anaphor's type.

However, it is often assumed that interpreting abstract anaphora involves an additional interpretational step (Weber, 1988), and the resolution process can involve a kind of type-raising operation (coercion, Hegarty et al. (2001), Consten et al. (2007)). This has to be taken into account in the design of the annotation process.

In general, annotation efforts tend to focus on anaphors rather than antecedents, partly because it is anaphors that

have to be resolved and partly because antecedents are sometimes difficult to determine. Antecedents that are made up by arbitrary sequences are usually restricted to written texts. In contrast, dialogue annotation tends to recur to “syntactically”-defined antecedents, i.e., antecedents that correspond to segments of dialog acts (or the respective verbal heads).

4. Towards a standard

Based on the observations made in the previous sections, we propose that “reference” corpora with abstract anaphora, which aim at sustainability and reusability, should adhere to the following principles.

The anaphor

Form: Many languages distinguish between pronouns that are prototypical realizations of abstract anaphora (e.g. demonstrative pronouns), and non-typical ones (e.g. personal pronouns). In addition, NPs, depending on the semantics of their head noun, can refer to abstract entities.

Proposal: We propose that reference corpora minimally should annotate prototypical pronominal realizations.

Semantics: The distinction between concrete and abstract can be made rather easily and reliably. For finer-grained labels, the situation is more complex: no commonly-used set of labels has been yet proposed, and people annotate considerably different types of information, such as speech acts, eventualities and factualities, or type-token distinction.

Proposal: Minimal annotation should include the distinction concrete–abstract.

The antecedent

Form: In most cases, marking the verbal head vs. the entire clause are equivalent solutions. It only makes a difference if the antecedent does not contain a verb, or if it consists of multiple clauses (in this case, a necessarily discontinuous string of multiple verbal heads would have to be marked). In both cases, clause marking seems more suitable than verbal-head marking. However, marking of verbal heads does not require any preprocessing.

Proposal: Antecedents are to be marked. Minimally, (sequences of) clauses or verbal heads should be annotated.

Semantics: Only very few projects have annotated semantic properties of abstract antecedents so far, and the issue still waits further investigations to be better understood.

Proposal: Currently none.

As we have seen, many annotation studies deal with anaphora in general, and—since concrete anaphors occur considerably more frequently than abstract anaphors—are restricted by an extremely low number of abstract anaphors. Hence, current results achieved so far are of limited significance, and considerably more data has to be produced to allow for serious investigations. We therefore call for annotation efforts focusing on the annotation of abstract anaphora.

5. References

- Ron Artstein and Massimo Poesio. 2006. Identifying reference to abstract objects in dialogue. In *Proceedings of Brandial 2006*.

- Nicholas Asher. 1993. *Reference to Abstract Objects in Discourse*. Kluwer Academic Publishers, Boston MA.
- Simon Botley and Tony McEnery. 2001. Demonstratives in English: A corpus-based study. *Journal of English Linguistics*, 29:7–33.
- Simon Botley. 2006. Indirect anaphora: Testing the limits of corpus-based linguistics. *International Journal of Corpus Linguistics*, 11(1):73–112.
- Donna K. Byron. 2002. Resolving pronominal reference to abstract entities. In *Proceedings of the ACL-02 conference*, pages 80–87.
- Donna K. Byron. 2003. Annotation of pronouns and their antecedents: A comparison of two domains. Technical Report, University of Rochester.
- Manfred Consten, Mareile Knees, and Monika Schwarz-Friesel. 2007. The function of complex anaphors in texts: Evidence from corpus studies and ontological considerations. In *Anaphors in Text*, pages 81–102. John Benjamins, Amsterdam/Philadelphia.
- Stefanie Dipper and Heike Zinsmeister. 2009a. Annotating discourse anaphora. In *Proceedings of LAW III*, pages 166–169.
- Stefanie Dipper and Heike Zinsmeister. 2009b. Annotation guidelines “Discourse-Deictic Anaphora”. Draft. Universities of Bochum and Konstanz.
- Miriam Eckert and Michael Strube. 2000. Dialogue acts, synchronising units and anaphora resolution. *Journal of Semantics*, 17(1):51–89.
- Gill Francis. 1994. Labelling discourse: an aspect of nominal group lexical cohesion. In Malcolm Coulthard, editor, *Advances in Written Text Analysis*, pages 83–101. London: Routledge.
- Kari Fraurud. 1992. Situation reference: What does ‘it’ refer to? GAP Working Paper No 24, Fachbereich Informatik, Universität Hamburg.
- Daniela Goecke, Anke Holler, and Maik Stührenberg. 2008. Coreference, cospecification and bridging: Annotation scheme. Technical Report, University of Bielefeld.
- Nancy Hedberg, Jeanette K. Gundel, and Ron Zacharski. 2007. Directly and indirectly anaphoric demonstrative and personal pronouns in newspaper articles. In *Proceedings of DAARC-2007*, pages 31–36.
- Michael Hegarty, Jeanette K. Gundel, and Kaja Borthen. 2001. Information structure and the accessibility of clausally introduced referents. *Theoretical Linguistics*, 27(2-3):163–186.
- Lucie Kučová and Eva Hajičová. 2004. Coreferential Relations in the Prague Dependency Treebank. In *Proceedings of DAARC-2004*, pages 97–102.
- Marie Mikulová et al. 2005. Annotation on the tectogrammatical level in the Prague Dependency Treebank. reference book. Technical report, ÚFAL MFF UK, Prague, Czech Republic.
- Christoph Müller. 2007. Resolving *it*, *this*, and *that* in unrestricted multi-party dialog. In *Proceedings of ACL-07 conference*, pages 816–823.
- Christoph Müller. 2008. *Fully Automatic Resolution of ‘it’, ‘this’, and ‘that’ in Unrestricted Multi-Party Dialog*. Ph.D. thesis, University of Tübingen.
- Costanza Navarretta and Sussi Olsen. 2008. Annotating abstract pronominal anaphora in the DAD project. In *Proceedings of LREC-08*.
- Costanza Navarretta and Sussi Olsen. 2009. The annotation of pronominal abstract anaphora in Danish texts and dialogues. DAD report 1. Centre for Language Technology, University of Copenhagen.
- Costanza Navarretta. 2008. Pronominal types and abstract reference in the Danish and Italian DAD corpora. In *Proceedings of the Second Workshop on Anaphora Resolution*, pages 63–71.
- Massimo Poesio and Ron Artstein. 2008. Anaphoric annotation in the ARRAU corpus. In *Proceedings of LREC-08*.
- Massimo Poesio and Natalia N. Modjeska. 2005. Focus, activation, and *this*-noun phrases: An empirical study. In António Branco, Tony McEnery, and Ruslan Mitkov, editors, *Anaphora Processing*, volume 263, pages 429–442. John Benjamins.
- Massimo Poesio. 2000a. Coreference. In Andreas Mengel et al., editor, *MATE Dialogue Annotation Guidelines*. <http://www.ims.uni-stuttgart.de/projekte/mate/mdag>.
- Massimo Poesio. 2000b. The GNOME annotation scheme manual. 4th version. http://cswwww.essex.ac.uk/Research/nle/corpora/GNOME/anno_manual_4.htm.
- Massimo Poesio. 2004. The MATE/GNOME proposals for anaphoric annotation, revisited. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue*, pages 154–162.
- Sameer Pradhan, Lance Ramshaw, Ralph Weischedel, Jessica MacBride, and Linnea Micciulla. 2007. Unrestricted coreference: Identifying entities and events in OntoNotes. In *Proceedings of the IEEE-ICSC*.
- Marta Recasens and M. Antònia Martí. 2010. AnCorACO: Coreferentially annotated corpora for Spanish and Catalan. *Language Resources & Evaluation*. DOI 10.1007/s10579-009-9108-x.
- Marta Recasens, M. Antònia Martí, and Mariona Taulé. 2007. Where anaphora and coreference meet. Annotation in the Spanish CESS-ECE Corpus. In *Proceedings of RANLP-07*, pages 504–509.
- Marta Recasens. 2008. Towards coreference resolution for Catalan and Spanish. Master’s thesis, University of Barcelona.
- Susanne Salmon-Alt and Laurent Romary. 2004. RAF: Towards a reference annotation framework. In *Proceedings of LREC-04*.
- Renata Vieira, Susanne Salmon-Alt, Caroline Gasperin, Emmanuel Schang, and Gabriel Othero. 2002. Coreference and anaphoric relations of demonstrative noun phrases in a multilingual corpus. In *Proceedings of DAARC-2002*.
- Bonnie L. Webber. 1988. Discourse deixis: Reference to discourse segments. In *Proceedings of the ACL-88 conference*, pages 113–122.

Appendix

Table 1: Survey of studies involving abstract anaphora

Study	Data		Anaphor			Antecedent			Criteria	Misc	
	Lang	Corpus	Statistics	Form	Semantics	Reliability (Semantics)	Form (AA)	Semantics (AA)			Reliability (Form)
Eckert and Strube (2000)	EN	Switchboard (telephone conversations)	T: ?, C: 678, AA: 154	Dem, Pers (no Expl)	<i>concr, abstr, vague, others</i> (1)	$\kappa = .80$ (Dem), $\kappa = .81$ (Pers)	Clauses	–	96.1–98.4% (<i>concr</i>), 85.7–94.3% (<i>abstr</i>)	Compatibility constraints, in algorithm	Algorithm (not implemented)
Vieira et al. (2002)	FR, PT	MLCC Corpus (written EU inquiries)	FR: T: 50K, C: 291, AA: 136 (<i>other</i>); PT: T: 50K, C: 243, AA: 116 (<i>other</i>)	Dem-NPs	<i>concr, abstr, other</i> (incl. AA), others (2)	FR: $\kappa = .79$ PT: $\kappa = .65$	Arbitrary sequences	–	Total/partial: FR: 69.8/79.2% PT: 51/55.1%	Some form- and meaning-based criteria	Annotation tool MMAX
Byron (2003)	EN	TRAINS93 (problem-solving dialogues), BUR (radio news stories)	TR/BUR: T: 10K/13K, C: 346/380, AA: 22/47	Dem, Pers (no Expl)	<i>concr, abstr</i>	Replacement test (TR/BUR): $\kappa = .56/.53$ (Dem), $\kappa = .71/.82$ (Pers)	Clauses	–	TR/BUR: $\kappa = .37/.62$ (Dem), $\kappa = .77/.95$ (Pers)	Replacement test; compatibility constraints, in algorithm	Implementation (Byron, 2002)
Kučová and Hajčová (2004)	CZ	PDT (mostly newspaper texts)	T: ? (35K sentences), C: 15K, AA: 274	Pers, Poss, Dem, Zero	<i>text, segm, others</i> (3)	–	V-head (dependency structure)	–	–	Guidelines (Marie Mikulová et al., 2005)	Annotation tool TRED
Poesio and Modjeska (2005)	EN	GNOME (museum descriptions)	T: ? (500 sentences), C: 112, AA: 19	Dem (<i>this, these, this-NPs</i>)	<i>abstr, type</i> (generic), others (5)	$\kappa = .82$	–	–	na	Decision tree	–
Botley (2006), Botley and McEnery (2001)	EN	AP (newswire), Hansard (parliament proc.), APHB (literary texts)	T: 300K, C: 648, AA: 10 (<i>indir</i>)	Dem, <i>this/that-NPs</i>	<i>indir</i> (incl. AA, with multiple subtypes; Sec. 3.1.), others (3)	–	–	–	na	Guidelines (cf. Botley and McEnery (2001))	–

Table 1 (cont'd)

Study	Data		Anaphor			Antecedent			Misc		
	Lang	Corpus	Statistics	Form	Semantics	Reliability (Semantics)	Form (AA)	Semantics (AA)		Reliability (Form)	Criteria
Müller (2007), Müller (2008)	EN	ICSI Meeting Corpus (dialogues)	T: ? (5 dialogues), C: 343 (anaphoric chains), AA: 59 chains	<i>it, this, that</i>	<i>referring, vague, non-ref</i>	–	V-head	–	$\alpha = .43-.52$	“Simple instructions”	Annotation tool MMAX; implementation
Pradhan et al. (2007)	EN	Wall Street Journal (news wire)	T: 300K, C: 11,400 (coref chains), AA: ?	Dem, Pers (no Expl), Poss, NPs (specific)	–	na	V-head	(Only events are annotated)	–	Annotated examples	Implementation
Navarretta and Olsen (2008), Navarretta (2008)	DA, IT	DAD Corpus (mixed)	DA: T: 125K, C: 1,612, AA: 455 ; IT: T: 135K, C: 890, AA: 114	Dem, Pers, Zero	Multiple labels (Sec. 3.1.)	AA: DA: $\kappa = .71-.89$ IT: $\kappa = .78-.89$	Clauses and larger seq.	–	Segments: DA: $\kappa = .79-.81$ IT: $\kappa = .87-.89$	Guidelines (Navarretta and Olsen, 2009)	Annotation tool PALinkA
Poesio and Artstein (2008), Artstein and Poesio (2006, Experiment 1)	EN	ARRAU (mixed: TRAINS dialogue, narrative, WSI, ...)	T: 95K, C: 24,321, AA: 455	all NPs/ Prons	Multiple labels (Sec. 3.1.)	–	Arbitrary sequences/clauses	–	Exp. 1, TRAINS (T: 1,421, C: 181, AA: 35); $\alpha = .55$ (for best 16 AA)	MATE/ GNOME manuals (modified)	Annotation tool MMAX
Dipper and Zinsmeister (2009a)	DE	Europarl (parliament proc.)	T: ? (32 texts), C: 48, AA: 48	Dem (<i>dies</i> ‘this’)	Multiple labels (Sec. 3.1.)	$\alpha = .37-.66$	Arbitrary sequences	Multiple labels (Sec. 3.1.)	Segments: 85%, labels: $\alpha = .52-.60$	Guidelines (Dipper and Zinsmeister, 2009b)	
Recasens and Martí (2010), Recasens (2008)	CA, ES	AnCorà (news paper texts)	CA: T: 385K, C: 120K, AA: 643 ; ES: T: 420K, C: 135K, AA: 748	all NPs/ Prons (incl. Cl, Zero; no Expl)	AA with subtypes (Sec. 3.1.), others (4)	ES: <i>coref</i> vs. <i>non-coref</i> ; $\alpha = .85-.90$	Verbs, clauses and larger seq.	–	ES: $\alpha = .85-.89$	Guidelines (Recasens et al., 2007)	PALinkA and AnCoràPipe annotation tools

Towards a standardized linguistic annotation Standardised Linguistic Annotation of Fairy Tales

Thierry Declerck¹, Kerstin Eckart², Piroska Lendvai³, Laurent Romary⁴, Thomas Zastrow⁵

¹DFKI GmbH, Language Technology Lab

Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

²Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart

Azenbergstraße 12, D-70174 Stuttgart, Germany

³Research Institute for Linguistics, Hungarian Academy of Science

Benczúr u. 33, H-1068 Budapest, Hungary

⁴INRIA, France & HUB-ISDL, Berlin, Germany

⁵Seminar für Sprachwissenschaft, Universität Tübingen

Wilhelmstr. 19-23, D-72074 Tübingen, Germany

E-mail: declerck@dfki.de, eckartkn@ims.uni-stuttgart.de, piroska@nytud.hu, laurent.romary@loria.fr,

thomas.zastrow@uni-tuebingen.de

Abstract

In our contribution to this workshop we propose incorporating standardized linguistic annotation in semantic resources of the cultural heritage domain, more specifically in the field of fairy tales. Although there are computational resources relevant for research in this area, these currently do not include linguistic annotation. We think here in particular to the The Proppian fairy tale Markup Language (PftML, see Malec, 2001), which is an annotation scheme that enables narrative function segmentation, based on hierarchically ordered textual content objects, but lacking linguistic information. We propose an approach to enrich PftML with standardized linguistic annotation, and so to support interoperability of linguistic information when it comes to combine it with annotation structures used in the eHumanities studies.

1. Introduction

In the context of both the CLARIN¹ and the D-SPIN² projects (<http://www.sfs.uni-tuebingen.de/dspin>), we are working towards the goal of making available language resources and technologies that could be supporting research in the field of eHumanities. As a specific case of this endeavour we present a strategy (that is by now partially implemented) for the integration of linguistic annotation and annotation of character roles and typed action descriptors in the literary genre of fairytales. For the latter, our departure point is the work by Vladimir Propp (Propp, 1968) and a XML schema, called PftML, for the annotation of fairy tales suggested by (Malec, 2004) We give here just some examples of Proppian functions³:

- Hero: a character that seeks something
- Villain: who opposes or actively blocks the hero's quest
- Donor: who provides an object with magical properties
- Dispatcher: who sends the hero on his/her quest via a message
- False Hero: who disrupts the hero's success by making false claims
- Helper: who aids the hero
- Princess: acts as the reward for the hero and the object of the villain's plots
- Her Father: who acts to reward the hero for his effort

Table 1: Some examples of Proppian functions

¹ <http://www.clarin.eu>

² <http://www.sfs.uni-tuebingen.de/dspin>

³ <http://www.adamranson.plus.com/Propp.htm>

Looking at the concrete XML representation proposed by Scott Alexander Malec of Vladimir Propp's Morphology of the Folk Tale, one can notice that the text of the tale itself is annotated in a coarse-grained manner and following an inline annotation strategy. Below we can see an example:

```
<Folktale Title="The Swan-Geese" AT="480"  
NewAfanasievEditionNumber="113" ProppConformity="Yes"> ....
```

```
<CommandExecution>
```

```
<Command subtype="Interdiction">
```

```
"Dearest daughter," said the mother, "we are going to work. Look after  
your brother! Don't go out of the yard, be a good girl, and we'll buy you a  
handkerchief."
```

```
</Command>
```

```
<Execution subtype="Violated">
```

```
The father and mother went off to work, and the daughter soon enough  
forgot what they had told her. She put her little brother on the grass  
under a window and ran into the yard, where she played and got  
completely carried away having fun.
```

```
</Execution>
```

```
</CommandExecution>
```

Figure 1. A part of a tale annotated with Propp's functions

While in a closely related paper (Lendvai et al., 2010), we describe the whole integration chain, also introducing ontological resources modelling character roles and action descriptors in the fairy tale domain, we could not address the issue of the standardization of linguistic annotation we integrate with PftML or the ontologies. At the actual stage of work we use a configuration of natural language processing tools supported by the WebLicht⁴ web services, as they are implemented in the D-SPIN project. WebLicht makes use of (but is not restricted to) of TextCorpus format (TCF), which has been chosen for efficiency reasons for the internal process of the various levels of linguistic annotation that can be supported by WebLicht. Our aim is to map this format to the family of standards developed within TEI (Text Encoding Initiative)⁵ and ISO TC 37/SC4⁶, also in order to verify the potential of those standards for serving as pivot format in the representation of textual and linguistic information. In the following we just present examples of the actual mapping of the TCF format, when applied to the text “Rotkäppchen” (*Little Red Riding Hood*), as it is stored in the Gutenberg project⁷.

2. TEI Annotation

Figure 4: List of possible linguistic annotation for an ontology label

As a first step we apply the TEI encoding standard, so that we get clearly marked textual content objects. We distinguish here between the TEI header and the text properly speaking:

```
<?xml version="1.0" encoding="UTF-8"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0"
  xmlns:ht="http://www.w3.org/1999/xhtml">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>Das Rotk ppchen</title>
        <author>Charles Perrault</author>
        <respStmt>
          <resp>translator</resp>
          <persName>nacherz hlt von Moritz
Hartmann</persName>
        </respStmt>
        <respStmt>
          <resp>sender</resp>
          <persName>reuters@abc.de</persName>
        </respStmt>
      </titleStmt>
      <publicationStmt>
        <p>http://projekt.gutenberg.de; created in
20040916</p>
      </publicationStmt>
      <sourceDesc>
      <biblStruct>
```

```
<monogr>
  <imprint>
    <publisher>Der Kinderbuchverlag
Berlin</publisher>
    <pubPlace>Berlin</pubPlace>
    <date when="1987"/>
  </imprint>
</monogr>
<idno type="isbn">3-358-00163-6</idno>
</biblStruct>
</sourceDesc>
</fileDesc>
<revisionDesc>
  <change when="2010-3-18">Tokenised</change>
</revisionDesc>
</teiHeader>
```

In the following TEI example, the text properly speaking is encoded in markup that describes embedded textual content objects (<p> for paragraphs, <w> for words etc.:

```
<text>
  <front>
    <docAuthor>Charles Perrault</docAuthor>
    <docTitle>
      <titlePart>Das Rotk ppchen</titlePart>
    </docTitle>
  </front>
  <body>
    <p>
      <w xml:id="t0">Es</w>
      <w xml:id="t1">war</w>
      <w xml:id="t2">einmal</w>
      <w xml:id="t3">ein</w>
      <w xml:id="t4">kleines</w>
      <w xml:id="t5">M dchen</w>
      <c xml:id="c0">,</c>
      <w xml:id="t6">ein</w>
      <w xml:id="t7">herziges</w>
      <w xml:id="t8">Ding</w>
      <c xml:id="c1">,</c>
      <w xml:id="t9">das</w>
      <w xml:id="t10">alle</w>
      <w xml:id="t11">Welt</w>
      <w xml:id="t12">liebhatte</w>
      <c xml:id="c2">.</c>
      <w xml:id="t13">Am</w>
      <w xml:id="t14">liebsten</w>
      <w xml:id="t15">hatte</w>
      <w xml:id="t16">es</w>
      <w xml:id="t17">die</w>
      <w xml:id="t18">Gro mutter</w>
      <c xml:id="c3">,</c>
      <w xml:id="t19">die</w>
      <w xml:id="t20">kaufte</w>
      <w xml:id="t21">ihm</w>
      <w xml:id="t22">ein</w>
      <w xml:id="t23">M ntelchen</w>
      <w xml:id="t24">mit</w>
      <w xml:id="t25">einer</w>
      <w xml:id="t26">roten</w>
      <w xml:id="t27">Kapuze</w>
      <w xml:id="t28">daran</w>
      <c xml:id="c4">,</c>
```

⁴ Details on the implementation of WebLicht, is given in <http://weblicht.sfs.uni-tuebingen.de/englisch/weblicht.shtml>

⁵ <http://www.tei-c.org/index.xml>

⁶ <http://www.tc37sc4.org/>

⁷ http://www.gutenberg.org/wiki/Main_Page

```

<w xml:id="t29">und</w>
<w xml:id="t30">danach</w>
<w xml:id="t31">hieÄ</w>
<w xml:id="t32">es</w>
<w xml:id="t33">RotkÄ€ppchen</w>
<c xml:id="c5">.</c>

```

...

```

<p>
  ....
  <w xml:id="t135">sieh</w>
  <w xml:id="t136">nicht</w>
  <w xml:id="t137">rechts</w>
  <c xml:id="c31">,</c>
  <w xml:id="t138">nicht</w>
  <w xml:id="t139">links</w>
  <c xml:id="c32">,</c>
  <w xml:id="t140">und</w>
  <w xml:id="t141">lasse</w>
  <w xml:id="t142">dich</w>
  <w xml:id="t143">durch</w>
  <w xml:id="t144">niemanden</w>
  <w xml:id="t145">vom</w>
  <w xml:id="t146">geraden</w>
  <w xml:id="t147">Weg</w>
  <w xml:id="t148">ablocken</w>
  <c xml:id="c33">!</c>
  <c xml:id="c34">Ä«</c>
</p>

```

3. Morpho-Syntactic Annotation

On the top of TEI we are the MAF standard for morpho-syntactic annotation (http://pauillac.inria.fr/~clerger/MAF/html/body_1_div.5.html), and link those to the words as they are marked by the TEI annotation (whereas still some alignment work is to be done, and some incertitudes in the mapping are still to be solved): The MAF notation refers to the “tokens” identified as <w> elements in the TEI annotation-

```

<?xml version="1.0" encoding="UTF-8"?>
<maf:MAF xmlns:maf="_">
  <maf:tagset>
    <dcs local="KON" registered=
      "http://www.isocat.org/datcat/DC-1262" rel="eq"/>
    <!-- _ -->
  </maf:tagset>
  <maf:wordForm tokens="t135">
    <fs>
      <f name="lemma"><symbol value="sehen"/></f>
      <f name="partOfSpeech"><symbol value="VVIMP"/></f>
      <f name="grammaticalNumber"><symbol value="singular"/></f>
    </fs>
  </maf:wordForm>
  <maf:wordForm tokens="t136">
    <fs>
      <f name="lemma"><symbol value="nicht"/></f>
      <f name="partOfSpeech"><symbol value="PTKNEG"/></f>
    </fs>
  </maf:wordForm>

```

```

<maf:wordForm tokens="t137">
  <fs>
    <f name="lemma"><symbol value="rechts"/></f>
    <f name="partOfSpeech"><symbol value="ADV"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t138">
  <fs>
    <f name="lemma"><symbol value="nicht"/></f>
    <f name="partOfSpeech"><symbol value="PTKNEG"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t139">
  <fs>
    <f name="lemma"><symbol value="links"/></f>
    <f name="partOfSpeech"><symbol value="ADV"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t140">
  <fs>
    <f name="lemma"><symbol value="und"/></f>
    <f name="partOfSpeech"><symbol value="KON"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t141">
  <fs>
    <f name="lemma"><symbol value="lassen"/></f>
    <f name="partOfSpeech"><symbol value="VVIMP"/></f>
    <f name="grammaticalNumber"><symbol value="singular"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t142">
  <fs>
    <f name="lemma"><symbol value="_"></f>
    <f name="partOfSpeech"><symbol value="_"></f>
    <f name="grammaticalNumber"><symbol value="singular"/></f>
    <f name="case"><symbol value="accusativeCase"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t143">
  <fs>
    <f name="lemma"><symbol value="durch"/></f>
    <f name="partOfSpeech"><symbol value="PREP"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t144">
  <fs>
    <f name="lemma"><symbol value="niemand"/></f>
    <f name="partOfSpeech"><symbol value="PIS"/></f>
    <f name="case"><symbol value="accusativeCase"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t145">
  <fs>
    <f name="lemma"><symbol value="vom"/></f>
    <f name="partOfSpeech"><symbol value="APPRART"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t146">
  <fs>
    <f name="lemma"><symbol value="gerade"/></f>
    <f name="partOfSpeech"><symbol value="ADJA"/></f>
    <f name="grammaticalNumber"><symbol value="singular"/></f>
    <f name="case"><symbol value="dativeCase"/></f>
    <f name="grammaticalGender"><symbol value="masculine"/></f>
  </fs>

```

```

</maf:wordForm>
<maf:wordForm tokens="t147">
  <fs>
    <f name="lemma"><symbol value="Weg"/></f>
    <f name="partOfSpeech"><symbol value="NN"/></f>
    <f name="grammaticalNumber"><symbol value="singular"/></f>
    <f name="case"><symbol value="dativeCase"/></f>
    <f name="grammaticalGender"><symbol value="masculine"/></f>
  </fs>
</maf:wordForm>
<maf:wordForm tokens="t148">
  <fs>
    <f name="lemma"><symbol value="ablocken"/></f>
    <f name="partOfSpeech"><symbol value="VVINF"/></f>
  </fs>
</maf:wordForm>

</maf:MAF>

```

We can not go into the details of the annotation here, but just to stress that in this way we have all the morpho-syntactic annotation attached to the TEI <w> elements.

We are currently working on mapping the syntactic annotation provided by the used configuration of WebLicht to the ISO SynAF model (http://www.iso.org/iso/catalogue_detail.htm?csnumber=37329).

The reader can see how the linguistic objects are pointing to the tokenized terms, and how the terms point then to the classes. On the basis of this model, we can obtain a matrix of linguistic objects, terms, and classes (including attributes and relations). This matrix can then deliver interesting insights on the use of natural language in knowledge representation systems. In the longer term, this can lead to proposal for a normalization of natural language expressions that fit best for building a terminology representing most adequately a formal representation of a domain.

4. Integration with the PftML annotation scheme

This step is straightforward: we take the functional annotation proposed by Scott A. Malec out of the document and include as an attribute the span of words that is in fact concerned by the Propp's function. This can look like:

```

<semantic_propp>
  <Command subtype="Interdiction" id="Command1"
  inv_id="Violated1" from="t135" to="t148">
</semantic_propp>

```

T135 and t148 are used here as defining a region of the text for which the Propp function holds. Navigating through the different types of IDs included in the multilayered annotation, the user can extract all kind of (possibly) relevant information.

We can also add to the functional annotation an additional ID which refers to a related detected function (here we point to the violation of the command that happens later in the text).

We plan also to use the ISO data category registry for entering the "labels" of Proppian functions (as for example shown in Table 1), with an adequate definition of those.

5. Acknowledgements

The research presented in this paper is partially funded by the European Commission in the context of the FP7 project CLARIN MONNET - Common Language Resources and Technology Infrastructure, with grant agreement number Grant Agreement Number 212230, and by the AMICUS network, which is sponsored by a grant from the Netherlands Organization for Scientific Research, NWO Humanities, as part of the Internationalization in the Humanities programme.

6. References

- Afanas'ev, A. 1945. Russian fairy tales. Pantheon Books: New York.
- Boas, H. 2005. From Theory to Practice: Frame Semantics and the Design of FrameNet. In: Semantisches Wissen im Lexikon, pp.129-160. Tübingen: Narr.
- Jason, H. 1977. Precursors of Propp: Formalist Theories of Narrative in Early Russian Ethnopoetics. Poetics and Theory of Literature, 3, pp. 477-485.
- Levi-Strauss, S. 1955. The structural study of myth. Journal of American Folklore, 68, pp. 428-444.
- Lendvai, P, Declerck ,T., Darányi, S., Hervás R., Malec, S. and Peinado, F. 2010. Integration of Linguistic Markup into Semantic Models of Folk Narratives: The Fairy Tale Use Case. Proceedings of LREC 2010.
- Malec, Scott A, 2004. Proppian structural analysis and XML modeling. <http://clover.slavic.pitt.edu/sam/propp/theory/propp.html>.
- Peinado, F., Gervás, P., Diaz-Agudo, B. 2004. A Description Logic Ontology for Fairy Tale Generation. Proceedings of LREC.
- Propp, V.J. 1968. Morphology of the folktale. University of Texas Press: Austin.
- Takahashi, N, Ramamonjisoa, D., and Takashi, O. 2007. A tool for supporting an animated movie making based on writing stories in XML. IADIS International Conference Applied Computing
- Tuffield, M. M., Millard, D. E. and Shadbolt, N. R. 2006. Ontological Approaches to Modelling Narrative. In: 2nd AKT DTA Symposium, January 2006, AKT, Aberdeen University.)

Annotating a historical corpus of German: A case study

Paul Bennett, Martin Durrell, Silke Scheible, Richard J. Whitt

University of Manchester
Oxford Road, Manchester, M13 9PL
{Paul.Bennett, Martin.Durrell, Silke.Scheible, Richard.Whitt}@manchester.ac.uk

Abstract

We report on our experiences in annotating a historical corpus of German with structural and linguistic information, providing an example of the needs and challenges encountered by smaller humanities-based corpus projects. Our approach attempts to follow current standardisation efforts to allow for future comparative studies between projects and the potential extension of our annotation scheme. Structural information is encoded according to TEI (P5) guidelines, and the corpus is further being annotated with linguistic information in terms of word tokens, sentence boundaries, normalised word forms, lemmas, POS tags, and morphological tags. The major problem encountered to date has been how to merge the linguistic mark-up with the TEI-annotated version of the corpus. In the interest of interoperability and comparative studies between corpora we would welcome the development of clearer procedures whereby structural and linguistic annotations might be merged.

1. Introduction

GerManC is an ongoing project based at the University of Manchester and funded jointly by the ESRC and AHRC. Its goal is to develop a representative corpus of Early Modern German covering the years 1650-1800. The corpus is modelled on the ARCHER corpus for English, which aims to be a representative corpus of historical English registers and consists of samples of continuous texts for a number of genres/registers. GerManC includes nine different genres and is subdivided into three 50-year periods and the five major dialectal regions of the then German Empire. Like ARCHER, it consists of sample texts of 2,000 words (yielding 900,000 words altogether), and two-thirds of the digitisation is now complete.

We shall report on our experiences in annotating the GerManC corpus with structural and linguistic information, providing an example of the needs and challenges encountered by smaller humanities-based corpus projects. As we are collaborating with various other historical projects that are currently in progress (for example, addressing other stages of German¹, or other languages, Biber et al. (1994)), it is of major importance to choose a standardised annotation format to enable interoperability and comparison. Our approach therefore attempts to follow current standardisation efforts to allow for future comparative studies between projects and the potential extension of our annotation scheme². Our results will be of particular interest to related projects which still use their own specialised annotation formats.

2. Corpus compilation and design

As GerManC is a historical corpus which will primarily be used in corpus linguistic studies, its design and annotation needs differ significantly from current large-scale corpus compilation projects. First of all, digitised historical data from the Early Modern German period is scarce, which

means that the majority of texts included in the corpus have to be digitised first. A manual approach to digitisation was chosen as texts from this period are usually printed in black letter fonts of variable sizes (Fraktur), as illustrated in Figure 1. Initial tests showed that scanning Fraktur with OCR technology is impractical and prone to error, especially because text samples are taken from a variety of genres and printed in different locations. Further problems are the arbitrary variation in font size, the denseness of the print on the page in many texts, and frequent variation between black letter and Roman fonts, even within words. The most reliable method for the digitisation of such older texts is by means of double-keying, i.e. each text is keyed in by two individuals and the results compared electronically to eliminate errors. This technique was adopted for the project and found to be wholly satisfactory.

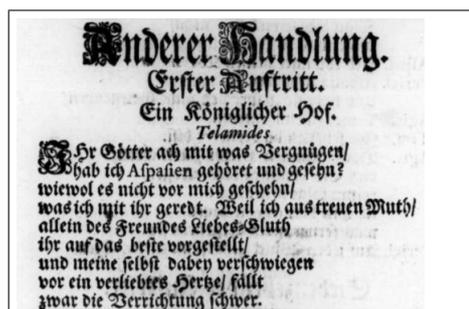


Figure 1: Drama excerpt

Although representativeness is difficult to achieve, GerManC aims to provide a broad picture of Early Modern German and takes three different levels into account. First of all, the corpus includes a range of registers or text types and, as far as possible, each register is represented by a sample of equal size. This means that the corpus does not consist of complete texts (which could mean that one text type, for example long novels, would be overrepresented), but of relatively short samples. The sample size of the Brown and ARCHER corpora, with extracts of some 2000

¹<http://www.linguistics.ruhr-uni-bochum.de/dipper/projectddd.html>

²For example, to include syntactic mark-up.

words (Meyer, 2002), has proved its viability over time, and we decided to follow this model. GerManC thus includes nine different genres, which are modelled on the ones used in ARCHER: four orally-oriented genres (dramas, newspapers, letters, and sermons), and five print-oriented ones (journals, narrative prose, scholarly writing in the humanities, scientific texts, and legal texts).

Secondly, in order to enable historical developments to be traced, the period is divided into fifty year sections (in this case 1650-1700, 1701-1750 and 1751-1800), and the corpus includes an equal number of texts from each register for each of these periods. The periodisation follows the model used in the Bonn corpus (Hoffmann and Wetter, 1987), which proved to be adequate to capture chronological variation at this time. The combination of historical and text-type coverage should enable research on the evolution of style in different genres, along the lines of previous work for English (Atkinson, 1992; Biber and Finegan, 1989).

Finally, the sample texts also aim to be representative with respect to region. This dimension has not been seen as essential for English corpora. ARCHER, for instance, only considers the two varieties British English and American English, but no further regional variation among these areas. The reason why different speech areas are taken into account in GerManC is that regional variation remained significant much longer in the development of standard German than it did in English (Durrell, 1999). However, this variation diminished over the period in question as the standard originating in the Central German area was gradually adopted in the South. Enabling this development to be traced systematically is one of the crucial desiderata for this corpus.

Altogether, per genre, period, and region, around three extracts of at least 2000 words are selected, yielding a corpus size of around 900,000 words altogether. Although this only a relatively small size, the unusual structure of the corpus represents a significant challenge for annotation.

3. Structural annotation

Annotation of historical texts needs to be very detailed with regard to document structure, glossing, damaged or illegible passages, foreign language material and special characters such as diacritics and ligatures. For this purpose, the raw input texts are annotated according to the guidelines of the Text Encoding Initiative (TEI) during manual transcription. The Text Encoding Initiative has published a set of XML-based encoding conventions recommended for meta-textual markup in corpus projects around the world and across different computer systems³. The principal aim of TEI is to minimise inconsistencies across projects and to maximise mutual usability and data interchange.

For the purpose of annotating these issues in our texts we use the TEI P5 Lite tagset as it offers a wealth of strategies for encoding structural details, and serves as standard for many humanities-based projects. Transcription and structural annotation are carried out using the OXygen XML editor⁴, which provides special support for inline TEI annotation. Only the most relevant tags are selected from the set to

keep the document structure as straightforward as possible. Due to the great variability of our corpus with respect to different genres, regions, and time periods, the full subset of TEI tags used will only be known at the end of the digitisation stage. Figure 2 shows the structural annotation of the above drama excerpt, including headers, stage directions, speakers (including a “who” attribute for co-reference), as well as lines.

```
<div type="act" n="2"><head>Anderer Handlung.</head>
<div type="scene" n="1"><head>Erster Auftritt.</head>
<head>Ein Ko&#868;niglicher Hof.</head>
<stage>Telamides.</stage>
<sp who="Telemides">
<l>Ihr Go&#868;tter ach mit was Vergnu&#868;gen/</l>
<l>hab ich <hi rend="antiqua">Aspasien</hi>
  geho&#868;ret und gesehn?</l>
<l>wiewol es nicht vor mich geschehn/</l>
<l>was ich mit ihr geredt. Weil ich aus treuen Muth/</l>
<l>allein des Freundes Liebes-Bluth/</l>
<l>ihr auf das beste vorgestellt/</l>
<l>und meine selbst dabey verschwiegen/</l>
<l>vor ein verliebtes Hertze/ fa&#868;llt/</l>
<l>zwar die Berrichtung schwer.</l>
```

Figure 2: Structural annotation

4. Linguistic annotation

The corpus is further being annotated with linguistic information in terms of word tokens, sentence boundaries, normalised word forms, lemmas, POS tags, and morphological tags. To reduce manual labour, a semi-automatic approach was chosen whose output is manually corrected. More detail about the annotation procedure can be found in Section 5.

Each annotation type requires careful consideration and adaptation as German orthography was not yet codified in the Early Modern period. Decisions on the level of tokenisation are especially important, as (with the exception of sentence boundaries) all other annotation types are token-based. Word boundaries are at times hard to determine as printers often vary in the amount of whitespace they leave between two words. For instance, sometimes they attempt to squeeze in an extra word at the end of a line, and as a result it is not straightforward to determine if one or two words were intended. Clitics and multi-word tokens are particularly difficult issues: lack of standardisation means that clitics can occur in various different forms, some of which are difficult to tokenise (e.g. *wirstu* instead of *wirst du*). Multi-word tokens, on the other hand, represent a problem as the same expression may be sometimes treated as compound (e.g. *obgleich*), but written separately at other times (*ob gleich*). While our initial tokenisation scheme takes clitics into account, it does not yet deal with the issue of multi-word tokens. This means that whitespace characters act as token boundaries, and multi-word expressions will be identified in a later step.

Annotation of sentence boundaries is also affected by the non-standard nature of the data. Punctuation is not standardised in Early Modern German and varies not only over different genres but also over time, and even within a single text. For example, the virgule symbol “/” survived longer in German than in English, and was used to separate textual segments of varying length and grammatical status. It is often used in place of both comma and full-stop, which

³<http://www.tei-c.org>

⁴<http://www.oxygenxml.com>

makes it difficult to identify sentence boundaries. This is particularly relevant for dramas and academic texts, where virgules are used alongside commas and full stops, and it is not always apparent which punctuation mark serves which function.

The tokenised text is further annotated with normalised word forms, lemmas, POS tags, and morphological tags. While the latter two tasks use standard tagsets (STTS for POS tagging⁵, and extended STTS for morphological information), we have defined special guidelines for annotating normalised word forms and lemmas. The normalisation stage aims to address the great amount of spelling variation that occurs in written historical documents, which proves problematic for automated annotation tools. Before final codification words often appear in a variety of spellings, sometimes even within the same paragraph or text. As most current corpus processing tools (such as POS-taggers or lemmatisers) are tuned to perform well on modern language data which follows codified orthographic norms, they are not usually able to account for variable spelling, resulting in lower overall performance (Rayson et al., 2007). Our goal is to develop a tool similar to Baron and Rayson’s variant detector tool (VARD) for English (Baron and Rayson, 2008), and complementing the work of Ernst-Gerlach and Fuhr (2006) and Pilz and Luther (2009) on historic search term variant generation in German, which will help to improve the output of the POS tagger and lemmatiser and will thus reduce manual labour.

In addition to creating a gold-standard annotation of our corpus which can be used to carry out reliable corpus-linguistic studies of Early Modern German, we also plan to make the following contributions:

- Provide detailed annotation guidelines for all proposed annotations
- Test and evaluate current corpus annotation tools on gold standard data
- Identify techniques for improving the performance of current tools
- Create historical text processing pipeline

To allow for future comparative studies between projects, we provide detailed annotation guidelines for both structural and linguistic annotation. Furthermore, as GerManC displays a wealth of variation in terms of different genres, time periods and regions, it lends itself as an ideal test bed for evaluating current corpus annotation tools (POS taggers etc.). This will be of particular interest to future corpus compilation projects faced with the difficult decision of which tools are most suitable for processing their data, and are likely to require the least manual correction. The second goal then utilises the findings of this evaluation study to improve the performance of existing tools, with the goal of creating a historical text processing pipeline which will contain a tokeniser, a sentence boundary detector, a lemmatiser, a POS tagger, and a morphological analyser. It will

also have options to specify the type of input data present, i.e. the components of the pipeline will be tuned to deal with genre variation, and possibly also temporal and spatial variation in Early Modern German.

5. Annotation procedure and challenges

In order to create the gold-standard annotation of our corpus and achieve the goals outlined in the previous section, our team was faced with a number of challenges. With no historical text processing platform yet available, we had to identify a suitable framework which would satisfy the following requirements:

1. Automate linguistic annotation (for subsequent manual correction)
2. Provide facilities for manual correction (annotation tool)
3. Produce standardised annotation format (suitable for further processing and comparison with other projects)
4. Merge structural (TEI) annotation with linguistic annotation

We identified GATE (Cunningham et al., 2002) as the most suitable framework for the tasks described above. GATE (“General Architecture for Text Engineering”) is open source software “capable of solving almost any text processing problem”⁶. To address point 1.), we used GATE’s German Language plugin⁷ and the TreeTagger (Schmid, 1994) to obtain annotations in terms of word tokens, sentence boundaries, lemmas, and POS tags. As GATE also offers facilities for manual annotation, we simultaneously use it as an annotation tool, correcting the errors produced by the automated tools to produce a gold standard annotation (point 2.).

The major problem encountered to date has been how to incorporate linguistic information in the TEI-annotated version of the corpus without invalidating the existing XML structure or ending up with two separate versions of the corpus. Structural and linguistic annotations cannot be merged into an inline XML format, as conflicts arise on a number of levels. For example, the inline XML structure of the drama excerpt in Figure 2 would be invalidated if sentence mark-up was added. Here, the sentence “wiewol es nicht vor mich geschehn/ was ich mit ihr geredt” (“although we did not speak at my behest”) stretches across a line boundary, leading to a crossed (and consequently invalid) XML structure. Furthermore, in words perceived as ‘foreign’, the typeface is frequently changed in the middle of a token, with Roman type used for the ‘foreign’ root and black letter (‘Gothic’) for the inflectional ending, as for example in the last word of the line in Figure 3 (marked as `<hi>repetir</hi>et`). Adding word token mark-up would only be possible if the typeface mark-up `<hi>` was nested within the token mark-up, which creates a conflict with the requirement that token tags should occupy the lowest level in the hierarchy.

⁵<http://www.ims.uni-stuttgart.de/projekte/corplex/TagSets/ststable.html>

⁶<http://gate.ac.uk>

⁷<http://gate.ac.uk/sale/tao/splitch19.htmlx24-45900019.1.2>

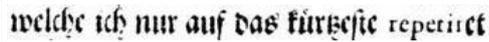


Figure 3: Typeface changes

The TEI guidelines offer some guidance to annotating cases where there is a conflict between the XML hierarchy established by the physical structure of a text (e.g., paragraphs, lines) and its linguistic structure (e.g., sentences, tokens)⁸. One suggested solution is to mark boundaries with empty elements, and including information about the start and end points of non-nesting material. This prevents the document from becoming invalid, and furthermore preserves the beginnings and end points for further processing. The TEI guidelines further discuss the use of stand-off format, in which text and mark-up are separated (for example by using XML elements which contain links to other nodes in another XML document).

Although TEI offers some solutions for merging structural and linguistic annotation, no information is provided on how the required annotations could be added automatically. Crucially, most automatic processing tools for German do not yet support TEI and require plain text as input (e.g. Lemnitzer’s Perl tokenizer⁹, TreeTagger). Given that the structural annotation is added first (in our case, by using inline XML tags during first inputting), a framework is required whereby automated linguistic annotation tools can be run on TEI-encoded texts and merge the newly created linguistic annotations with the existing structural mark-up. It seems that to date no such framework is available, and little documentation is available on how structural TEI annotations can be merged with linguistic annotations. We further found that some annotations proposed in the TEI P5 manual are unsuitable for further linguistic processing, as they allow manipulation of the original document by adding information on the text level. For example, TEI’s treatment of abbreviations suggests the use of a “choice” element to record both the abbreviation and its expansion, as illustrated in Figure 4.

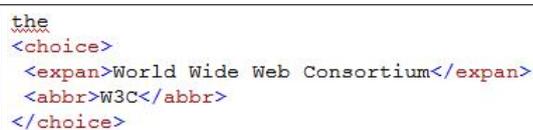


Figure 4: TEI’s choice tag

Another example is the inclusion of descriptive elements, which can be used to provide short textual descriptions of omitted figures or graph. The implications of such additions on the text level are twofold. First of all, the original text flow is interrupted, which represents a problem for further processing tools such as POS taggers (which would treat the added material as part of the original text). Secondly, the fact that TEI allows manipulation of the original text means that an approach where the structural and lin-

guistic annotations of a text are merged at a later stage is not straightforward as the underlying documents (and character offsets) may differ.

Although GATE does not specifically support TEI, it allows XML-encoded text as input, which means that the above-mentioned German Language plugin can be applied to the TEI-annotated version of our corpus, and both annotation layers can be saved within the same document. This can be achieved in two different ways: by using a GATE-particular stand-off format, or by saving the annotated text as inline XML using the ‘Save Preserving Format’ option (which attempts to preserve the original XML mark-up alongside the new annotations). However, from our point of view both formats are problematic: the stand-off architecture is GATE-specific and needs to be transformed into other formats for external processing (e.g. by using XSLT stylesheets). The inline format, on the other hand, has to deal with overlapping elements as the ones described above (crossed line boundaries in the drama corpus and changes in typeface). Ideally, the ‘Save Preserving Format’ option should address such issues by extending the spans of the original structural mark-up to wrap around the newly created linguistic annotations, and add information about the start and end points of the non-nesting material (as suggested in the TEI manual, see above). Instead, the original mark-up is manipulated to accommodate the new annotations in a way which can lead to inaccuracies on the structural level. For example, the token `<hi>repetir</hi>et` shown in Figure 3 is wrapped as `<hi><w>repetir</w></hi><w>et</w>` by the tokeniser module, incorrectly splitting the token into two parts. To deal with these issues we created scripts which “repair” such cases by using fragmentation techniques similar to the ones described in the TEI manual.

From the point of view of a smaller humanities-based project it would be desirable if text processing platforms such as GATE provided explicit support for texts encoded according to the TEI P5 guidelines, as a great deal of time has to be spent on writing scripts to deal with formatting issues. Additionally, we would welcome clearer guidelines from the Text Encoding Initiative on how structural and linguistic mark-up should be merged in practice.

6. Conclusion

With no single processing platform suitable for our needs and no clear set of guidelines to follow, identifying an adequate annotation format for our corpus has turned up a range of problems which had not been anticipated. Given the amount of investigation and pre- and postprocessing work necessary to create a standardised annotation, it comes as no surprise that projects on a limited budget still prefer to use their own specialised formats. In the interest of interoperability and comparative studies between corpora we would welcome the development of clearer procedures whereby structural and linguistic annotations might be merged, and would wish to contribute actively to this process by sharing our experiences.

⁸<http://www.tei-c.org/release/doc/tei-p5-doc/en/html/NH.html>

⁹<http://www.lemnitzer.de/lothar/KoLi/>

7. References

- Dwight Atkinson. 1992. The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal. *Applied Linguistics*, 13:337–374.
- Alistair Baron and Paul Rayson. 2008. VARD 2: A tool for dealing with spelling variation in historical corpora. In *Proceedings of the Postgraduate Conference in Corpus Linguistics, Birmingham, UK*.
- Douglas Biber and Edward Finegan. 1989. Drift and the evolution of English style: a history of three genres. *Language*, 65:487–517.
- Douglas Biber, Edward Finegan, and Dwight Atkinson. 1994. ARCHER and its challenges: compiling and exploring A Representative Corpus of Historical English Registers. In Udo Fries, Peter Schneider, and Gunnell Tottie, editors, *Creating and using English language corpora. Papers from the 14th International Conference on English Language Research on Computerized Corpora*, pages 1–13.
- Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
- Martin Durrell. 1999. Standardsprache in England und Deutschland. *Zeitschrift für germanistische Linguistik*, 27:285–308.
- Andrea Ernst-Gerlach and Norbert Fuhr. 2006. Generating search term variants for text collections with historic spellings. In *Proceedings of the 28th European Conference on Information Retrieval Research (ECIR 2006), London, UK, London, UK*.
- Walter Hoffmann and Friedrich Wetter. 1987. *Bibliographie frühneuhochdeutscher Quellen*. Frankfurt am Main: Lang.
- Charles F. Meyer. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.
- Thomas Pilz and Wolfram Luther. 2009. Automated support for evidence retrieval in documents with nonstandard orthography. In Sam Featherston and Susanne Winkler, editors, *The Fruits of Empirical Linguistics*, pages 211–228.
- Paul Rayson, Dawn Archer, Alistair Baron, Jonathan Culpeper, and Nicholas Smith. 2007. Tagging the Bard: Evaluating the accuracy of a modern POS tagger on Early Modern English corpora. In Matthew Davies, Paul Rayson, Susan Hunston, and Pernilla Danielsson, editors, *Proceedings of the Corpus Linguistics Conference (CL2007), University of Birmingham, UK*.
- Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK.

Linguistic tool development between community practices and technology standards

Thomas Schmidt

SFB 538 'Multilingualism'

Max Brauer-Allee 60

D-22765 Hamburg

E-mail: thomas.schmidt@uni-hamburg.de

Abstract

This contribution addresses the workshop topic of “standardising policies within eHumanities infrastructures”. It relates 10 years of experience with language resource standards, gained in the development of EXMARaLDA, a system for the construction and exploitation of spoken language corpora. Section 2 gives an overview of the EXMARaLDA system focussing on its relationship with existing and evolving standards for language resources. Section 3 presents the HIAT system as an example of an established community practice. Section 4 then addresses several issues that were encountered when trying to bring together HIAT, EXMARaLDA and the wider standard world.

1. Introduction

This contribution addresses the workshop topic of “standardising policies within eHumanities infra-structures”. It relates 10 years of experience with language resource standards, gained in the development of EXMARaLDA, a system for the construction and exploitation of spoken language corpora.

EXMARaLDA is targeted mainly at an audience of non-technologically oriented linguists who study, for instance, pragmatic aspects of natural interaction, language acquisition in children and adults, dialectal variation, or special forms of multi-lingual interaction like interpreting. While awareness in these different communities about the importance of standards for data exchange and sustainability is growing, there is still a large gap between their own established practices of data processing and high-level standardisation efforts in currently evolving e-infrastructures such as CLARIN. We as tool developers have therefore come to accept a role as a mediator between established community practices on the one hand, and “true” technological standards on the other hand, and it is from this perspective that I will look at language resource standards in this contribution.

The paper is structured as follows: section 2 gives an overview of the EXMARaLDA system focussing on its relationship with existing and evolving standards for language resources. Section 3 presents the HIAT system as an example of an established community practice. Section 4 then addresses several issues that were encountered when trying to bring together HIAT, EXMARaLDA and the wider standard world.

2. Standards in EXMARaLDA

EXMARaLDA, under development since 2000 at the Research Centre on Multilingualism at the University of Hamburg, is a system of data models, formats and tools for the construction and exploitation of spoken language corpora. Its main areas of application are conversation and

discourse analysis, research on learner language and dialectology (see Schmidt/Wörner 2009).

2.1 EXMARaLDA data model

EXMARaLDA's data model¹ is an application of the Annotation Graph Formalism (AG, Bird/Lieberman 2001). It is represented in two XML-based data formats of different structural complexity:

1. An EXMARaLDA Basic-Transcription is an annotation graph with a single, fully ordered timeline and a partition of annotation labels into a set of tiers (aka the “Single timeline multiple tiers” data model: STMT). It is suitable to represent the temporal structure of transcribed events, as well as their assignment to speakers and to different levels of description (e.g. verbal vs. non-verbal).
2. An EXMARaLDA Segmented-Transcription is an annotation graph with a potentially bifurcating timeline in which the temporal order of some nodes may remain unspecified. It is derived automatically from a Basic-Transcription and adds to it an explicit representation of the linguistic structure of annotations, i.e. it segments temporally motivated annotation labels into units like utterances, words, pauses etc.

A more detailed description of EXMARaLDA's data model can be found in Schmidt 2005.

2.2 Interoperability with ELAN, ANVIL, etc.

Annotation tools like ELAN, ANVIL, Praat etc. work with data models which are very similar to that of an EXMARaLDA Basic-Transcription. Schmidt et al. (2009) discusses the different variants of the STMT data model used by these tools and formulates a suggestion for an XML exchange format based on the Atlas Interchange Format (Laprun et al. 2002) which ensures that the common denominator information of their data models can be

¹ EXMARaLDA also caters for metadata descriptions, but I will restrict myself in this paper to data models and formats for representing spoken language transcriptions.

exchanged. In practice, EXMARaLDA users can profit from this interoperability by employing different tools for different tasks in their annotation workflows.

2.3 Compatibility with TEI

The principal challenge in establishing compatibility between time-based data models like AG or its different STMT derivatives and more hierarchy-oriented approaches like the TEI's is to find suitable structural units inside a directed acyclic graph (DAG) which can be ordered sequentially and underneath which other structural units of that graph nest in an ordered fashion, thus giving rise to an ordered hierarchy of content objects (OHCO) This problem can probably not be solved generically (i.e. for every conceivable type of data representable in a DAG), but, as argued in Schmidt 2005, mechanisms can be found which are at least applicable across a wider range of data types. EXMARaLDA uses one such mechanism – the combination of temporally contiguous annotation labels assigned to the same speaker – to derive a list-like representation of an annotation document from a Segmented-Transcription. This list can then be represented in an XML document following the TEI guidelines for transcriptions of speech. In terms of interoperability and data exchange, this is especially important because it creates a link between the most common way of representing time-series data (i.e. DAG) and the “natural” way of representing written language (i.e. OHCO).

The same mechanism is also used to establish interoperability between EXMARaLDA and transcription tools built on a more hierarchy-oriented conception of data – most importantly the CLAN editor of the CHILDES system.

2.4 Compatibility with LAF and GENAU

In the practice of spoken corpus construction, the Linguistic Annotation Framework (LAF) has so far not played any important role, if for no other reason than the fact that there is no transcription or annotation tool that uses or directly supports the LAF data model. Work on PAULA and the ANNIS database (Zeldes et al. 2009), however, shows at least that EXMARaLDA data can be integrated into LAF-based frameworks and thus be made accessible for analysis together with other data whose annotation follows the same principle.

Similarly, GENAU and the SPLICR platform (Rehm et al. 2008) have shown – as a proof of concept at least – that EXMARaLDA data can be transformed into data models based on the idea of multiple annotation of identical primary data (Witt 2002).

3. HIAT

HIAT is an acronym of *Halbinterpretative Arbeitstranskriptionen* (“semi-interpretative working transcriptions”). It is a transcription convention originally developed in the 1970s for the transcription of classroom interaction. The first versions of the system (Ehlich/Rehbein 1976) were designed for transcription with pencil or typewriter and paper. HIAT's main characteristic is the use of so-called Partitur (musical score) notation, i.e. a two-dimensional

transcript layout in which speaker overlap and other simultaneous actions can be represented in a natural and intuitive manner.

Not least because editing such musical scores is technically challenging, HIAT was computerized relatively early in the 1990s in the form of two computer programs – HIAT-DOS for DOS (and later Windows) computers, and syncWriter for Macintoshes. Large corpora of classroom discourse, doctor-patient communication and similar interaction types were constructed with the help of these tools. However, standardization and data exchange being a minor concern at the time, these data turned out to be less sustainable than their non-digital predecessors: The data format produced by HIAT-DOS is purely presentation-oriented and thus does not allow any structural transformations based on the actual semantics of the data. Even more problematically, syncWriter uses a largely undocumented binary format, readable and writable by no other application than syncWriter itself. The realisation that data produced by two functionally almost identical tools on two different operating systems could not be exchanged and, moreover, the prospect that large existing bodies of such data might become completely unusable on future technology, raised awareness in the HIAT community for the need for standards and was one of the major motivations for initiating the development of EXMARaLDA.

4. EXMARaLDA and HIAT

As discussed in the previous sections, EXMARaLDA as a system based on and actively supportive of different existing and developing standards for language resources, has increased the potential of transcription data to be exchanged between different applications and to be integrated into more generic frameworks for linguistic data processing. From the point of view of the HIAT community, the major challenge was to adapt the existing data processing practices in such a way that they could be realized inside the EXMARaLDA system. And, conversely, EXMARaLDA's development had to be sensitive to the needs of that community. The following sections therefore discuss how various types of standards and other – more or less conventionalized – practices continue to interact and compete with each other in this assimilation of HIAT and EXMARaLDA.

4.1 Legacy data

One non-negotiable condition for the acceptance of EXMARaLDA by the HIAT community was that it must be able to accommodate the existing bodies of data created with HIAT-DOS and syncWriter. This condition translates into three more specific requirements:

- 1) The data model and formats must contain the model(s) underlying the legacy data, i.e. every structural relation represented in the legacy data must also be representable in EXMARaLDA. Since musical score transcripts are based on a similar logic as annotation graphs, this requirement was relatively straightforward to fulfil.
- 2) Wherever the data model or formats stipulate con-

structs that go beyond the legacy data structure, they must still tolerate data that does not (yet) contain (or worse: that deals inconsistently with) such constructs. As an example, take the assignment of stretches of transcription to absolute times in the recording. While it is certainly desirable for EXMARaLDA's data model to contain a construct for this information, neither syncWriter nor HIAT-DOS provide a place for it. In order to be able to efficiently transform legacy data into and inside EXMARaLDA, the system must therefore also be able to process transcriptions *without* temporal alignment², and it must also provide the means of adding this information *ex post*. Yet, when new data is produced with the system, it should allow the user to record this kind of information at the same time the actual annotation is entered. Legacy data and new data thus pose competing requirements to the tools.

- 3) There must be efficient methods for systematically transforming legacy data into the new data model and formats. As the legacy data are known to be deficient in terms of structure and consistency, the expectation is not a fully automatic conversion procedure, but rather a workflow in which manual and automatic processing steps are combined in a maximally efficient manner. For the HIAT legacy data, this workflow consisted in a method for reading out data from the older tools, followed by a couple of semi-automatic methods for correcting structural inconsistencies, followed by several manual steps in which additional information lacking in the original data (like the above-mentioned media alignment) was added.

Of course, on top of these requirements to *enable* legacy data conversion, a further prerequisite was to find the resources to actually *carry it out* – a non-trivial requirement given that legacy data conversion (even if supported by adequate tools) is very demanding in terms of man-hours. After several years of work, a number of HIAT legacy corpora have now been fully transformed to EXMARaLDA³, and further data are in the waiting line. Experience with the data converted so far will hopefully help to speed up future transformations (see Schmidt/Bennöhr 2008 for a more detailed discussion of this aspect).

4.2 Community practices

The HIAT transcription convention is a documented community practice. It gives instructions on what phenomena to describe in an interaction, and on how to describe them. The latter type of instruction is, in principle, a formal one – it picks out certain symbols from the alphabet, assigns them certain semantics inside the transcription, and formulates rules about which combinations of such symbols are permissible and which are not. For instance, one such rule states that descriptions of pauses should have the

² Note that, for instance, Praat or ANVIL cannot deal with such data – they expect the nodes in their DAGs to correspond to some location in a recording.

³ These corpora are available through <http://corpora.exmaralda.org>

form “((1,2s))”, i.e. a decimal number followed by an ‘s’ between a pair of double round brackets. In EXMARaLDA, the transformation of Basic-Transcriptions into Segmented-Transcriptions relies on these formal regularities as the basis for a finite state parsing of annotation strings (see Schmidt 2005).

However, in times of pencil and paper transcription and also during the early computerized days of HIAT, no mechanism was available (nor was one needed) to actually *check* the “formal correctness” of a given HIAT transcription. Consequently, the formal rules were followed only loosely in practice and different dialects of HIAT developed over the years to accommodate annotation needs not covered by the “official” conventions. When the first legacy corpora had been converted and the formal regularities of HIAT were to be exploited in automatic processing of the data, it therefore soon became apparent that the conventions were in need of a revision. In Rehbein et al. (2004), the formal transcription rules were thus formulated in a more rigid manner (e.g. by providing Unicode codepoints for all symbols), and additional regulations were introduced to ensure a firm basis for automatic processing of the data. Not surprisingly (HIAT being a community practice with a tradition) this change of practice met with some opposition. In the long run, however, the additional processing methods enabled through EXMARaLDA seem to work in favour of an acceptance of the changes. In any case, the modification of the conventions naturally also had an impact on the legacy data conversion described above – the converted data now had to be checked for correctness against the new version of HIAT.

Another change of community practice became necessary in the area of *workflows*. As long as corpora were not made available to a larger audience, and no methods existed to automatically query a larger corpus of transcriptions, analyses were usually carried out by a small number of researchers on a small number of transcripts. If errors or inaccuracies in these transcripts were found, they could be corrected immediately without having to take into account how the change would affect the overall corpus or other people analysing the same data. Also, corpora could grow and be completed according to the analysis needs of a single project.

As Bird/Simons (2002) have pointed out, however, the *immutability* of a resource is an important aspect of its usability once it has been made available to a wider audience. Moreover, techniques like standoff-annotation also usually require certain parts of the data to remain unchanged in order for pointers to remain valid. Last but not least, publishing a resource also means agreeing on a certain date at which no further modifications on its current version are allowed. The new technology and new uses for the old data thus required HIAT users to think about issues like version and quality control, and to develop practicable workflows not only for creating, but also for publishing resources.

4.3 Other tools

When the development of EXMARaLDA started, only

Praat and CHAT were available as robust editors for creating transcriptions, and these were, at the time, judged inadequate by the HIAT community for their purposes. This situation has changed fundamentally: tools like ELAN, ANVIL (also Praat in its newer versions) now all run stable and each of them offers interesting features that the others don't. As a further change in community practice, the more innovation friendly members of the HIAT community thus began looking for ways of using different tools side-by-side, exploiting their individual strengths, e.g. doing orthographic transcription in EXMARaLDA, gesture analysis in ANVIL or ELAN and phonetic analysis in Praat. The import and export methods described in 2.2 provide the basis for this. However, given that each of the tools employs a data model that is optimized for its own functionality, data exchange between two of such tools is usually not lossless in both directions. As a further aspect of data creation workflows, processing chains involving different tools and the optimal way of combining them had therefore to be considered.

4.4 Standards

Apart from the fact that they are built on general document standards like XML and Unicode and that they implement specific versions of more general frameworks like AG, neither EXMARaLDA nor the data models and formats of other tools mentioned in the previous sections are "standards" in the strict (ISO) sense of the word. The CLARIN Standardisation Action Plan thus does not list them under the heading of "standards", but under "community practices". It seems to me important to note, however, that they are different from a community practice like HIAT (in its pre-EXMARaLDA version at least) insofar as they have an explicit formal specification and technical realisations that actually exploit this formal basis.⁴

From the frameworks listed under "standards" in this document, at least TEI and LAF are potentially relevant for the users of HIAT and EXMARaLDA. As discussed in 2.3 and 2.4, there seem to be no principal obstacles to converting EXMARaLDA to one of these standards. From the point of view of the HIAT user community, however, these standards currently do not play any important role. Their main reason for this is that they do not yet offer any additional value in terms of data processing or interoperability that would be relevant to the researchers' work. When details of conversion methods have to be worked out for these standards, it might get difficult to motivate the community to further changes of their practices as long as this additional value is not clearly visible to them.

5. Conclusions

This paper has sketched some issues encountered on the way from an informal community practice to more general standards for language resources. It has shown that existing

bodies of legacy data, existing codifications of community practices and existing workflows, as well as parallel development of different tools all co-determine the standardisation process.

The most important lesson learned in the assimilation of EXMARaLDA and HIAT is that, tedious as the method of carefully and iteratively adapting established practices exemplified here may be, it has turned out to be a reasonably successful standardising policy.

If evolving eHumanities infra-structures want to serve a diverse audience, it may be a key requirement that more such community practices with a potential for standardisation are identified. The development of "generic" standards should then ideally be realised as a stepwise approximation between the concrete practices of specific communities and the high-level abstractions underlying current standardisation efforts in language technology.

6. Acknowledgements

Work on EXMARaLDA is financed by a grant from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG).

7. References

- Bird, S. & Liberman, M. (2001). *A formal framework for linguistic annotation*. In: *Speech Communication* 3, 323-60.
- Bird, S. & Simons, G. (2002). *Seven Dimensions of Portability for Language Documentation and Description*. In: *Language* 79, 557-582.
- Laprun, C.; Fiscus, J.; Garofolo, J. & Pajot, S. (2002). *Recent Improvements to the ATLAS Architecture*. Proceedings of HLT 2002, Second International Conference on Human Language Technology, San Francisco, 2002.
- Schmidt, T. (2005). *Time-based data models and the Text Encoding Initiative's guidelines for transcription of speech*. In: *Arbeiten zur Mehrsprachigkeit*, Folge B 62.
- Rehbein, J.; Schmidt, T.; Meyer, B.; Watzke, F. & Herkenrath, A. (2004). *Handbuch für das computergestützte Transkribieren nach HIAT*. In: *Arbeiten zur Mehrsprachigkeit*, Folge B 56.
- Rehm, G.; Schonefeld, O.; Witt, A.; Chiarcos, C. & Lehmsberg, T. (2008). *SPLICR: A Sustainability Platform for Linguistic Corpora and Resources*. In: *Konferenz zur Verarbeitung natürlicher Sprache*, September 30–October 02, Berlin, Germany.
- Schmidt, T. & Wörner, K. (2009). *EXMARaLDA – Creating, analysing and sharing spoken language corpora for pragmatic research*. In: *Pragmatics* 19.
- Schmidt, T.; Duncan, S.; Ehmer, O.; Hoyt, J.; Kipp, M.; Magnusson, M.; Rose, T. & Sloetjes, H. (2009). *An Exchange Format for Multimodal Annotations*. In: Michael Kipp, Jean-Claude Martin, P. P. & Heylen, D. (ed.): *Multimodal Corpora*, Lecture Notes in Computer Science 207-221. Springer.
- Witt, A. (2002). *Multiple Informationsstrukturierung mit Auszeichnungssprachen. XML-basierte Methoden und deren Nutzen für die Sprachtechnologie*. PhD Thesis, Universität Bielefeld.

⁴ The CLARIN document lists CHAT (CHILDES) as a community practice comparable to HIAT insofar as "it is not formally specified as a schema, but a set of widely used tools work on the resources [...]."