Language Resources: From Storyboard to Sustainability and LR Lifecycle Management

WORKSHOP PROGRAMME

Sunday May 23, 2010

09:00 – 09:15 Welcome and Introduction
Victoria Arranz and Laura van Eerten, ELDA-ELRA, France and TST-Centrale, The Netherlands

09:15 – 10:00 Invited talk: Sustainability for Open-Access Language Resources
Mark Liberman, LDC, USA

10:00 – 10:30 The Flemish-Dutch HLT Agency: a Comprehensive Approach to Language Resources Lifecycle Management & Sustainability for the Dutch Language
Remco van Veenendaal, Laura van Eerten and Catia Cucchiarini, TST-Centrale and Dutch Language Union, The Netherlands

10:30 – 11:00 Coffee break

11:00 – 11:30 Creating and Maintaining Language Resources: the Main Guidelines of the Victoria Project
Lionel Nicolas, Miguel Angel Molinero Alvarez, Benoît Sagot, Nieves Fernández Formoso and Vanesa Vidal Castro, UNSA & CNRS, France, Universidade da Coruña, Spain, Université Paris 7, France and Universidade de Vigo, Spain

11:30 – 12:00 Laundry Symbols and License Management - Practical Considerations for the Distribution of LRs based on Experiences from CLARIN
Ville Oksanen, Krister Lindén and Hanna Westerlund, Aalto University and University of Helsinki, Finland

12:00 – 13:30 Poster session:

Resource Lifecycle Management: Changing Cultures
Peter Wittenburg, Jacquelijn Ringersma, Paul Trilsbeek, Willem Elbers and Daan Broeder, MPI for Psycholinguistics, The Netherlands

The Open-Content Text Corpus project
Piotr Bański and Beata Wójtowicz, University of Warsaw, Poland

Very Large Language Resources? At our Finger!
Dan Cristea, University of Iasi, Romania

Standardization as a Means to Sustainability
Michael Maxwell, University of Maryland, USA

The TEI and the NCP: the Model and its Application
Piotr Bański and Adam Przepiórkowski, University of Warsaw and Polish Academy of Sciences, Poland

13:30 – 14:30 Lunch break

14:30 – 15:00 The German Reference Corpus: New developments Building on Almost 50 Years of Experience
Marc Kupietz, Oliver Schonefeld and Andreas Witt, Institute for the German Language, Germany

15-00 – 15:30 Sustaining a Corpus for Spoken Turkish Discourse: Accessibility and Corpus Management Issues
Şükriye Ruhi, Betil Eröz-Tuğa, Çiler Hatipoğlu, Hale Işık-Güler, M. Güneş Can Acar, Kerem Eryılmaz, Hümeyra Can, Özlem Karakaş, Derya Çokal Karadaş, Middle East Technical University, Ankara University and Hacettepe University, Turkey

15:30 – 16:00 The BAsic Metadata DEScription (BAMDES) and TheHarvestingDay.eu: Towards Sustainability and Visibility of LRT
Carla Parra, Marta Villegas and Núria Bel, Universitat Pompeu Fabra, Spain

16:00 – 16:30 Coffee break

16:30 – 18:00 Panel Discussion

18:00 Closing

Editors

Victoria Arranz	ELDA - Evaluations and Language resources Distribution Agency / ELRA - European Language resources Association, France
Laura van Eerten	Flemish-Dutch HLT Agency (TST-Centrale), Institute for Dutch Lexicology (INL), The Netherlands

Organising Committee

Victoria Arranz	ELDA - Evaluations and Language resources Distribution Agency / ELRA - European Language resources Association, France
Khalid Choukri	ELDA - Evaluations and Language resources Distribution Agency / ELRA - European Language resources Association, France
Christopher Cieri	LDC - Linguistic Data Consortium, USA
Laura van Eerten	Flemish-Dutch HLT Agency (TST-Centrale), Institute for Dutch Lexicology (INL), The Netherlands
Bente Maegaard	CST, University of Copenhagen, Denmark
Stelios Piperidis	ILSP – Institute for Language and Speech Processing / ELRA - European Language resources Association, France
Remco van Veenendaal	Flemish-Dutch HLT Agency (TST-Centrale), Institute for Dutch Lexicology (INL), The Netherlands

Programme Committee

Núria Bel	Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, Spain
Nicoletta Calzolari	Istituto di Linguistica Computazionale del CNR (ILC-CNR), Italy
Jean Carletta	Human Communication Research Centre, School of Informatics, University of Edinburgh, UK
Catia Cucchiarini	Dutch Language Union (NTU), The Netherlands
Christoph Draxler	Bavarian Archive for Speech Signals, Institute of Phonetics and Speech Processing (BAS), Germany
Maria Gavrilidou	Institute for Language and Speech Processing (ILSP), Greece
Nancy Ide	Department of Computer Science, Vassar College, USA
Steven Krauwer	UiL OTS, Utretch University, The Netherlands
Asunción Moreno	Universitat Politècnica de Catalunya (UPC), Spain
Dirk Roorda	Data Archiving and Networked Services (DANS), The Netherlands
Ineke Schuurman	Centre for Computational Linguistics, Catholic University Leuven, Belgium
Claudia Soria	Istituto di Linguistica Computazionale del CNR (ILC-CNR), Italy
Stephanie M. Strassel	Linguistic Data Consortium (LDC), USA
Andreas Witt	IDS Mannheim, Germany
Peter Wittenburg	Max Planck Institute for Psycholinguistics, The Netherlands

Table of Contents

Author Index

PREFACE

The life of a language resource (LR), from its mere conception and drafting to its adult phases of active exploitation by the HLT community, varies considerably. Ensuring that language resources be a part of a sustainable and endurable living process represents a multi-faceted challenge that certainly calls for well-planned anti-neglecting actions to be put into action by the different actors participating in the process. Clearing all IPR issues, exploiting best practices at specification and production time are just a few samples of such actions. Sustainability and lifecycle management issues are thus concepts that should be addressed before endeavouring into any serious LR production.

When thinking of long-term LRs a number of aspects come to our minds which do not always succeed to be taken into account before development. Some of these aspects are *usability*, *accessibility, interoperability* and *scalability*, which inevitably call for a long list of neglected points that would need to be taken into account at a very early stage of development. Looking further into the *portability* and *scalability* of a language resource, a number of dimensions should be taken into account to ensure that a language resource reaches its adult life in an active and productive way.

An aspect that is often neglected is the *accessibility* and thus *secured reusability* of a language resource. Institutions such as ELRA (European Language resources Association) and LDC (Linguistic Data Consortium), at a European and American level, respectively, as well as BAS (Bavarian Archive for Speech Signals) and TST-Centrale (Flemish-Dutch Human Language Technology Agency), at a language-specific level, have worked on these aspects for a large number of years. Through their different activities, they have successfully implemented a sharing policy which allows different users to gain access to already existing resources. Other emerging programmes such as CLARIN (Common Language Resources and Technology Infrastructure) are also looking into these aspects. Nevertheless, many resources still follow development without a long-term accessibility plan into place which makes impossible to gain access once the resource is finished. This accessibility plan should consider issues such as ownership rights, licensing, types of use, aiming for a wide community from the very beginning. This accessibility plan calls for an optimal co-operation between all actors (LR users, financing bodies, owners, developers and organisations) so that issues related to the life of a LR are well established, roles and actors are clearly identified within the cycle and best practices are defined towards the management of the entire LR lifecycle.

We are aware, though, that these above-presented ideas are but a take-off for discussion. It is at this point that we invited the community to participate in this workshop and share with us their views on these and other relevant issues of concern. A fruitful discussion could lead us to finding new mechanisms to support perpetuating language resources, and may lead us towards a sustainability model that guarantees an appropriate and well-defined LR storyboard and lifecycle management plan in the future.

Among the many issues and topics that were suggested for discussion during this workshop, we could mention the following:

- Which fields require LRs and which are their respective needs?
- What needs to be part of a LR storyboard? What points are we missing in its design?
- General specifications vs. detailed specifications and design
- Annotation frameworks and layers: interoperable at all?
- Should creation and provision of LRs be included in higher education curriculae?
- How to plan for scalable resources?
- Language Resource maintenance and improvement: feasible?
- Sharing language resources: how to bear this in mind and implement it? Logistics of the sharing: online vs. Offline
- Centralised vs. decentralised, and national vs. international management and maintenance of LRs
- What happens when users create updated or derived LRs?

- Sharing language resources: legal issues concerned
- Sharing language resources: pricing issues concerned, commercial vs. non-commercial use
- Do LR actors work in a synchronised manner?
- What should be the roles of the different actors?
- What are the business models and arrangements for IPRs?
- Self-supporting vs. subsidised LR organisations
- Other general problems faced by the community

This full-day workshop is addressed to all those involved with language resources at some point of their research/work (LR users, producers, ...) and to all those with an interest in the different aspects involved, whether universities, companies or funding agencies of some nature. It aims to be a meeting and discussion point for the so many bottlenecks surrounding the life of a resource and which remain to be addressed with a sustainability plan.

The workshop features one invited talk, opening the morning session, as well as oral and poster presentations. It concludes with a round table to brainstorm on the issues raised during the presentations and the individual discussions. This workshop is expected to result in a plan of action towards a sustainability and lifecycle management plan to implement.

The Dutch-Flemish HLT Agency: a Comprehensive Approach to Language Resources Lifecycle Management and Sustainability for the Dutch Language

Remco van Veenendaal[1], Laura van Eerten[1], Catia Cucchiarini[2]
1Flemish-Dutch HLT Agency (TST-Centrale), Institute for Dutch Lexicology,
Matthias de Vrieshof 2-3, 2311 BZ Leiden, The Netherlands
2Dutch Language Union, Lange Voorhout 19, 2514 EB Den Haag, The Netherlands
E-mail: remco.vanveenendaal@inl.nl, laura.vaneerten@inl.nl, ccucchiarini@taalunie.org

Abstract

The Dutch-Flemish Human Language Technology (HLT) Agency is a central repository for (government-funded) Dutch digital language resources (LRs) that operates within an extensive Dutch-Flemish HLT Network and takes care of various phases in the LR lifecycle. Two distinguishing features of the HLT Agency are LR maintenance – keeping them up to date – and knowledge management, which are of major importance for LRs to survive. The HLT Agency also handles a clear licensing, pricing and IPR policy, which is necessary to guarantee the sustainability and availability of LRs for research, education and commercial purposes. Now that the HLT Agency has been operational for five years, we can conclude that having one central repository for Dutch LRs is advantageous in many respects.

1. Introduction

Promoting the development of digital language resources (LRs) and language and speech technology for the Dutch language has been a priority in the Dutch language area for at least ten years (Cucchiarini, et al 2001; Binnenpoorte et al. 2002). Governmental support was considered to be mandatory because since Dutch is a so-called mid-sized language (Pogson, 2005 a, b), companies are not always willing to invest in developing such resources for a language with a relatively small market. On the other hand, the development of language and speech technology is considered to be crucial for a language to be able to survive in the information society. Against this background a number of initiatives were set up through the Dutch Language Union, an intergovernmental organisation established by Belgium and the Netherlands in 1980 with the aim of carrying out a common language policy for the Dutch language (see also http://taalunieversum.org/taalunie/).
The most important initiatives are the STEVIN programme (Dutch Acronym for Essential Speech and Language Technology Resources for Dutch), which is aimed at realizing a complete digital language infrastructure for Dutch and at promoting strategic research in language and speech technology (Spyns et al., 2008, see also http://taalunieversum.org/taal/technologie/stevin/english/), and the Human Language Technology (HLT) Agency (in Dutch: TST-Centrale), a central repository for Dutch digital LRs (Beeken, & van der Kamp, 2004, see also http://www.inl.nl/en/tst-centrale). These activities were set up under the auspices of the Dutch Language Union in co-operation with the relevant ministries and organisations in Belgium and the Netherlands.

This paper is focused on how the lifecycle of digital Dutch LRs is managed by the HLT Agency. The paper is organised as follows: in section 2 we present the context in which the HLT Agency operates. In section 3 we describe the various channels through which LRs reach the HLT Agency, and the various phases in the LRs lifecycle of which the HLT Agency takes care. In section 4 we briefly present our target groups and explain which users resort to the HLT Agency. We then go on to discuss the importance of user feedback and the way in which the HLT Agency gathers information to try and improve its services. In section 6 we finish off with some conclusions on the role of the Dutch HLT Agency now and in the future.

2. The Dutch-Flemish HLT Network, the STEVIN programme and the HLT Agency

The approach to stimulating language and speech technology that has been adopted for the Dutch language is comprehensive in many respects. First of all, because it is based on co-operation between government, academia and industry both in Belgium and in the Netherlands (Spyns et al., 2008). Co-operating saves money and effort, boosts the status of the language and means not having to reinvent the wheel over and over again. Second, because it encompasses the whole range from basic resources to applications for language users. Third, because it concerns the whole cycle from resource development to resource distribution and (re)use.

The resources that are developed within the STEVIN programme are subsequently handed over to the HLT Agency which takes care of their future lifecycle. This is a completely different situation from the one existing before the HLT Agency was established. At that time it was not uncommon that official bodies such as ministries and research organisations financed the development of LRs and no longer felt responsible for what should happen to those materials once the projects were completed. However, materials that are not maintained quickly lose value. Moreover, unclear intellectual property right (IPR) arrangements can create difficulties for exploitation.

To prevent that HLT materials developed with public money become obsolete and therefore useless, the HLT Agency was set up and financed by the Dutch Language Union and hosted by the Institute for Dutch Lexicology in Leiden, the Netherlands (with an auxiliary branch in Antwerp, Belgium). The STEVIN programme is not the only source of LRs that are hosted by the HLT Agency, as will be explained in section 3.

Having a central body that takes care of the LRs lifecycle like the HLT Agency turns out to have several advantages:

- efficient use of persons and means is cost reducing;
- combining resources and bringing together different kinds of expertise creates surplus value (e.g. extra applications);
- offering resources through one window (one-stop shop) creates optimal visibility and accessibility;
- in international projects the Dutch language area can act as a strong partner.

3. Managing the lifecycle of an LR

The mission of the HLT Agency is to manage, maintain and distribute Dutch digital LRs for research, education and commercial purposes, to assure that the Dutch language continues to participate in the information society. To this end, as many as possible Dutch LRs are collected in the central repository of the HLT Agency. The lifecycle of each LR is properly managed and various phases are distinguished in the process. These phases are described in detail below.

3.1 Acquisition and intellectual property rights

LRs that are hosted by the HLT Agency come from funding programmes like STEVIN, from research institutes like the Institute for Dutch Lexicology, third parties like (individual) researchers and associations. In the case of STEVIN and the Institute for Dutch Lexicology the transfer of LRs to the HLT Agency is previously arranged. Concerning LRs from third parties, the HLT Agency keeps an eye on the field and actively approaches researchers to examine the possibilities of making their LRs produced for a specific research goal available for a broader audience through the HLT Agency. Over the years the HLT Agency has gained more recognition in the field, and an increasing number of researchers are able to find and contact the HLT Agency themselves, for keeping their data alive.
In order to guarantee that LRs can be optimally managed, maintained and distributed the HLT Agency checks if the LR's IPR is properly taken care of – the HLT Agency has to be able to protect the interests of the supplier of an LR. IPR of (foreground knowledge of) LRs created within the STEVIN programme are transferred to the Dutch Language Union, which subsequently acts as a supplier to the HLT Agency. The rationale behind transferring the IPR to the Dutch Language Union is the need to make LRs maximally and readily accessible to a large number of users.

In other cases, the property rights remain with the developer or supplier and the HLT Agency merely requests distribution rights. If a developer or supplier can no longer maintain or support their LR, the transfer of more rights to the HLT Agency is an option, again guaranteeing optimal management of the LR's lifecycle by the HLT Agency. This policy is also adopted for open source projects. The HLT Agency can of course act as a central repository for active open source LRs, but can also be asked to take over this role from a community if that community is unable to continue to support an open source LR. In some cases, an agreement already exists between open source projects and the HLT Agency that if the community is no longer able to support a certain LR, the HLT Agency can continue to manage the LR's lifecycle.

3.2 Validation

All LRs that are managed by the HLT Agency have to undergo a validation and quality check. Sometimes this check is performed by the HLT Agency, but in most cases the supplier or the funding programme already arranged for this beforehand. For instance, in the STEVIN programme an external evaluation of the results is required.

The validation and quality check that the HLT Agency performs is mainly a technical one The check for data covers the data formats (against e.g. accompanying XML schemas), the quality and completeness of the documentation and, in general, we gauge the measure in which the LR lends itself for distribution as a reusable product.
In the case of software the binaries are tested on the supported platforms, using the accompanying documentation to create test cases. Source code will be compiled and the resulting binary is tested. Similar to the data, the documentation and product-readiness of the software are evaluated.
If a LR fails one of these tests, it is up to the supplier to fix any problems and resubmit the fixed LR to the HLT Agency.

The (linguistic) value and potential use of language resources is far more difficult to measure. This is where the HLT Agency prefers the Long Tail strategy: selling – licensing – a large number of unique items – LRs – in relatively small quantities. Use and peer review of the LRs will subsequently indicate its actual (linguistic) value to the community. We feel that it is not primarily a task for the HLT Agency to try to predict this value.

3.3 Maintenance

When LRs are transferred to the HLT Agency they are stored and backed up in their original form. If necessary the LR is reworked into a form that is (more) suitable for distribution and use. Periodically, the HLT Agency checks if LRs need maintenance, e.g. for the purpose of standardisation (data) or when they risk disuse due to

incompatibility with new operating systems (software). When maintaining LRs the HLT Agency has a two-fold approach: minor maintenance, keeping a resource usable, is done by the HLT Agency and major maintenance, significantly improving a resource, is done in cooperation with experts (and requires additional funding). Other terms for minor and major maintenance are "maintenance" and "improvement or expansion" – cf. maintaining your garden (mowing the grass occasionally) and improving your garden (adding drainage, new trees or flowers).

Minor maintenance is focused on keeping a resource usable and therefore consists of e.g. fixing major or critical bugs, updating manuals and documentation, upgrading formats to newer versions of the standard(s) used, etc. The HLT Agency periodically checks if LRs require (minor) maintenance and incorporates the work required into the work plans after consulting the owner/supplier of the LR. The result of minor maintenance usually is a "patch" or "update" of an LR.

Major maintenance has to result in (significantly) improved or expanded resources and usually requires a larger budget and cooperation with external experts.

To this end, information and advice on which LRs should be improved or expanded can be gathered through the various advisory committees that assist the Dutch Language Union and the HTL Agency. Major maintenance work usually results in a new version of an LR, rather than a patch or update.

To give an impression of how minor and major maintenance work impacts on LR version numbers we present some examples linked to the Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN):

- fixing several bugs in version 6.0 of the Corex software accompanying the CGN resulted in version 6.1.
- adding (annotations for) 13 speech files to CGN version 1.0 in cooperation with members from the original CGN project, upgrading the IMDI metadata to version 3 and fixing several bugs in the annotations resulted in version 2.0 of the CGN.
- at the time of writing, the Corex 6.1 software is being upgraded to version 7.0, resulting in a significantly improved user interface, the use of an XML database and the possibility of exploring both speech and text corpora. For this work, the HLT Agency cooperated (and cooperates) with external experts. The project also required and received additional external funding.

3.4 IP Management and pricing

To ensure the centralised management of LR lifecycles in the future, arrangements are settled between the user of an LR and the HLT Agency in end user agreements. A Pricing Committee helps establish a clear licensing policy, prices and strategies for optimal usability of LRs: free licenses for non-commercial use of an LR by non-commercial organisations and licenses with market-conform prices for other use(r)s.

Both licensing schemes state that insofar alterations, modifications or additions to the LR result in new IPR with respect to the LR itself, these rights are transferred to the original IPR-holder of the LR. This ensures that the IPR situation with respect to the LR remains stable, which greatly simplifies the future management, maintenance and distribution. In the non-commercial licenses there is also a Right of Option: if new products are created using the licensed LR and the licensee wishes to make those LRs available (rather than just using it for their own educational or research purpose), the original IPR-holder must be offered exclusive distribution rights, making it possible that the new LR is also taken care of by the HLT Agency. The original IPR-holder can give up their Right of Option if e.g. the licensee is willing and able to manage the new LR's lifecycle.

3.5 Knowledge Management

The maintenance of LRs is one of the things that make the HLT Agency unique. Another distinguishing feature of the HLT Agency is knowledge management. When LRs are supplied to the HTL Agency, the knowledge available is stored in a knowledge management system. Although we are aware that we will not be able to collect all knowledge about all LRs, we do think that this knowledge management is very important for at least two reasons: the availability of the knowledge is not limited to the availability of the expert(s) and the available knowledge can easily be used, shared and kept up to date. For the HLT Agency there are three primary sources of knowledge:

- knowledge from external experts, preferably made available in the form of documentation or as a result of interviewing the expert;
- knowledge collected by the HLT Agency while working with or maintaining an LR;
- knowledge gained by our skilled service desk (from question answering).

Most LRs come with accompanying user and technical documentation. Usually also a lot of information is also still available from the project that created the LR, e.g. in the form of progress reports or a project wiki. The HLT Agency asks for this information to be made available by the project team in order to function as an additional valuable source of (background) knowledge. When new LRs are supplied to the HLT Agency we ask the supplier to explain to us what the LR is and how it can be used. In some cases we ask the expert to explain in detail how certain parts of the LR came into existence, e.g. how a lexicon was derived from a text corpus. Without this knowledge – and if the expert is not available anymore for some reason – it could e.g. be impossible to continue to maintain the LR.

The HLT Agency creates knowledge about LRs while using and maintaining the LR. Often the user manual of software resources does not provide a detailed description of all possible functions and by using the

software the HLT Agency might discover functionality that could significantly improve the user's experience when documented. Also a lot of knowledge is gained while maintaining LRs: nothing improves the understanding of (the inner workings of) LRs better than taking a look under the bonnet.

The third main source of knowledge for the HLT Agency is the question answering performed by the skilled service desk. Skilled means that it is an intelligent or learning service desk: answers to questions are stored and re-used when similar questions are asked.

The HLT Agency has agreements with several external experts and/or suppliers regarding question answering: when questions require knowledge that the service desk does not (yet) have, the question is forwarded to an expert. The answer provided by this expert is forwarded (with explicit thanks to the expert) and stored in the service desk. Following this procedure, the expert does not have to come up with the same answer to the same question over and over again and the knowledge collection of the HLT Agency continues to grow.

The knowledge management by the HLT Agency ensures that not only the LRs are kept available, but also the knowledge about those LRs.

4. Target groups and users

As mentioned above, the HLT Agency manages, maintains and distributes Dutch digital LRs for research, education and commercial purposes.

Researchers from various disciplines turn to the HLT Agency to access all sorts of LRs; e.g. general, socio-, computational and forensic linguistics, translation studies, social studies, cognition studies, historical and bible studies, communication and information studies, Dutch studies from all over the world etc. Before the HLT Agency existed, researchers often had to collect their own LRs before being able to start their research proper. The advantages of this new approach in which LRs are made publicly available for researchers cannot be overestimated.

Teachers and students can also access LRs for educational purposes. For instance frequency lists are used as a starting point in second language education, or to implement in educational applications for specific groups like dyslectics. Audio can be also used in e.g. educational games or quizzes.

Another important target group for the HLT Agency are small and medium enterprises (SMEs). SMEs are often willing to develop useful HLT applications, but they are not able to support the costs incurred in developing the LRs that are required for such applications. The availability of LRs at affordable prices through the HLT Agency lowers cost barriers and offers a viable solution. For example, a small company that provides speech solutions for a specific user group like people with reading difficulties. The HLT Agency can offer reference lexicons or a part of a lexicon at a reduced price, for improving the company's speech synthesis system.

In addition to such specific target groups there are a lot of private parties that turn to the HLT Agency: lawyers, language amateurs and even artists.

5. Service optimization: the importance of user feedback

Although it is clear that the existence of the HTL Agency has considerable advantages, we are interested to know whether and how the services offered by the HTL Agency could be improved. To this end evaluations are regularly carried out.

In 2007 a three-fold evaluation was carried out consisting of a self evaluation by the HLT Agency, a digital user inquiry by the Dutch Language Union and interviews with a selected group of users, project partners and suppliers held by an external evaluation committee. The main results of the evaluation were incorporated by the HLT Agency in a plan for improvement. The main focus was on increasing the visibility of the Agency in the field and improving collaboration and communication with suppliers and project partners.

In 2009 again a similar evaluation was carried out, and the results will soon become available. These evaluations are also important to keep partners and users involved and to stay informed about the needs of the field.

6. Conclusions

Since its inception in 2004 the HLT Agency has gradually gained recognition in the HLT field in the Netherlands, Flanders and abroad. The idea of a central repository for digital Dutch LRs is widely supported.

In addition to the Dutch Language Union and the Institute for Dutch Lexicology other parties have started to deposit their LRs at the HLT Agency. The sustainability of LRs is supported by having a clear licensing, pricing and IPR policy, maintaining the LRs, actively managing knowledge about the LRs and providing a skilled service desk for question answering.

The HLT Agency has now grown into an important linchpin for the Dutch-Flemish HLT community. Although the policy and procedures adopted may be subject to change over time, it seems that the core aims of obsolescence avoidance and re-usability will have to be pursued in the future too.

7. References

Anderson, C. (2006). *The Long Tail: Why the Future of Business is Selling Less of More.* New York, NY, USA: Hyperion.

Beeken, J. C., Kamp, P. van der (2004). The Centre for Dutch Language and Speech Technology (TST Centre). In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC04)*, Lisbon, Portugal, pp. 555--558.

Binnenpoorte, D., Cucchiarini, C., d'Halleweyn, E., Sturm, J., Vriend, F. de (2002). Towards a roadmap for Human Language Technologies: the Dutch-Flemish experience. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*

(LREC02), Las Palmas de Gran Canaria, Spain.

Boekestein, M., Depoorter, G., Veenendaal, R. van (2006). Functioning of the Centre for Dutch Language and Speech Technology. In *Proceedings of the 5th International Conference of Language Resources (LREC06)*, Genova, Italy, pp. 2303--2306.

Cucchiarini, C., Daelemans, W., Strik, H. (2001). Strengthening the Dutch Language and Speech Technology Infrastructure. *Notes from the Cocosda Workshop 2001*, Aalborg, pp. 110--113.

Pogson, G. (2005a). Language Technology for a Mid-Sized Language, Part I. *Multilingual Computing & Technology*, 16(6), pp. 43--48.

Pogson, G. (2005b). Language Technology for a Mid-Sized Language, Part II. *Multilingual Computing & Technology*, 16(7), pp. 29-34.

Spyns P., d'Halleweyn E., Cucchiarini C., (2008). The Dutch-Flemish comprehensive approach to HLT stimulation and innovation: STEVIN, HLT Agency and beyond. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC08)*, Marrakech, Morocco.

Creating and maintaining language resources: the main guidelines of the *Victoria* project

Lionel Nicolas[1], Miguel A. Molinero[2], Benoît Sagot[3], Nieves Fernández Formoso[4], Vanesa Vidal Castro[4]

1. Équipe RL, Laboratoire I3S, UNSA & CNRS, 2000, route des lucioles, BP 121, 06903 Sophia Antipolis, France
2. Grupo LYS, Departamento de Computación, Universidade da Coruña, Campus de Elviña s/n, 15071 La Coruña, Spain
3. Alpage, INRIA Paris-Rocquencourt & Université Paris 7, 30 rue du Château des Rentiers, 75013 Paris, France
4. Grupo Cole, Universidade de Vigo, Campus de As Lagoas s/n, 32004 Orense, Spain

lnicolas@i3s.unice.fr, mmolinero@udc.es, benoit.sagot@inria.fr, {vvcastro,nievesff}@uvigo.es

Abstract

Many Natural Language Processing (NLP) tools rely on the availability of reliable language resources (LRs). Moreover, even when such LRs are available for a given language, their quality or coverage sometimes prevent them from being used in complex NLP systems. Considering the attention received from both the academic and industrial worlds and the significant efforts achieved during the past decades for LR development, such a lack of high quality and wide-coverage LR shows how difficult their creation and correction can be. In this paper, we describe a set of guidelines applied within the *Victoria* project in order to ease the creation and correction of the LRs required for symbolic parsing. These generic guidelines should be easy to use or adapt for the production of other types of LRs.

1. Introduction

The efficiency and linguistic relevance of most NLP tools depends directly or indirectly on the quality and coverage of the LRs they rely on. Along the past decades, numerous project, such as MULTEXT,[1] MULTEXT-East,[2] DELPHIN,[3] AGFL,[4] etc., have focused on developing LRs while the ongoing CLARIN[5] and FLARENET[6] initiatives aim at managing and bringing under a common framework many existing LRs. Despite such efforts, few LRs may be considered as complete and correct, except maybe for English, the language that has clearly received the most attention over the last decades.

Nevertheless, complex NLP systems such as automatic translation tools, if they make use of LRs, do require high-quality resources. The creation of LRs with a high level of quality in terms of coverage, quality and richness is therefore an important problem in our research field.

The main contribution of this paper is to propose a list of guidelines for the production of LR. This list has been set up while planning and managing the *Victoria* project (Nicolas et al., 2009).

This paper is organized as follows. In Section 2., we briefly introduce the *Victoria* project. We then explain in Section 3. some reasons why creating and maintaining LRs is still so difficult. Next, we detail in section 4. and 5. a set of guidelines for easing this task. Finally, we quickly highlight in section 6. the objectives the *Victoria* project has achieved, before concluding in section 7.

2. The *Victoria* project

The *Victoria* project, started in November 2008, is funded by a grant from the Galician Goverment.[7] It brings together researchers from four different French and Spanish teams: (i) the COLE team[8] from the University of Vigo, (ii) the LyS team[9] from the University of A Coruña, (iii) the Alpage project[10] from the University Paris 7 and INRIA Paris-Rocquencourt and (iv) the RL team,[11] I3S laboratory, University of Nice Sophia Antipolis and CNRS.

The main goal of the project is to develop techniques and tools for producing and improving the high-quality and wide-coverage LRs required for symbolic parsing.[12] So far, the project has been focusing on French, Spanish and Galician languages.

3. Difficulties when creating and maintaining LR

Several reasons explain why the development of an LR has been and is still such an complex task, most of them being consequences of the intrinsic richness and ambiguity of natural languages. Among them, two can be highlighted:

- the difficulty in describing all linguistic description levels (e.g., morphology, syntax, semantics);
- the difficulty in covering all instances of a given linguistic description level for a given language.

A few decades ago, the available computing power made it impossible to imagine or test the complex formalisms that

[1] http://aune.lpl.univ-aix.fr/projects/MULTEXT/
[2] http://nl.ijs.si/ME/
[3] http://www.delph-in.net/
[4] http://www.agfl.cs.ru.nl/
[5] http://www.clarin.eu/
[6] http://www.flarenet.eu/

[7] Project number INCITE08PXIB302179PR.
[8] http://coleweb.dc.fi.udc.es/
[9] www.grupolys.org
[10] http://alpage.inria.fr/
[11] http://deptinfo.unice.fr/~jf/Airelles/
[12] Morphological rules, morpho-syntactic lexicons and lexicalised grammar.

are used nowadays. Even though, we still lack a global consensus for modeling most linguistic description levels. This is particularly true for the semantic level, but the large range of available syntactic formalisms is another illustration of this difficulty. However, as far as lexical information is concerned, morphological and syntactic notions are now reasonably consensual, and are indeed standardized by various ISO norms such as LMF (Lexical Markup Framework) (Francopoulo et al., 2006).

However, despite the fact that there exist now consensus and therefore formalisms for some levels, it is still difficult to find the corresponding high-quality and wide-coverage LRs for many languages. Indeed, even for languages such as Spanish or French, many well known and widely used resources are still in a somehow precarious state of development. For languages with a smaller speech community, such as Galician[13], LRs are almost non-existent.

Currently, one can consider the efforts required to develop LRs as the main limitation. In other words, the difficulty for some linguistic levels does not lie anymore in how to describe them but in actually achieve a description that has the coverage and precision required by complex NLP tasks. As a matter of fact, whoever has developed an LR knows that, in a reasonable amount of time, one can achieve a certain level of coverage and precision. However, as formalized by Zipf's law, increasing the quality of an LR becomes more and more difficult with time. Thus, the corresponding efforts follow a somehow exponential curve, i.e, the efforts are always more demanding when compared with the resulting improvements.

In order to tackle such problems, we propose an approach that relies on two complementary strategies: sharing the efforts among several people interested in obtaining those resources and saving manual efforts by automatizing the processes of creation and correction as much as possible.

4. Enhancing collaborative work

4.1. Problems limiting collaborative work

If a language receives enough attention from the community, the efforts to describe it by means of LRs can clearly be shared among the people interested in building them. Nevertheless, the greater the workforce is, the more difficult it is to manage since it requires to find agreements on several non-trivial aspects.

Formalisms Nowadays, it is not rare to find various LRs describing a same linguistic description level of a given language. This happens mostly for two reasons.

First, the kind of data described in LRs generally depends on the application they have been created for. Therefore, one can find non-related but similar LRs covering the same sub-parts of a given level.

Second, the way a language is described can change when grounding on different linguistic theories. Therefore, there exist similar LRs that are (partially) incompatible.

In both cases, it implies a loss of manual work by formalizing several times a given knowledge and a waste of precious feedback by splitting the users over various LRs.

[13] A co-official language spoken in the north-west of Spain.

License and free availability The distribution and terms of use of LRs are issues both fundamental and problematic/polemic for their life-cycle. Indeed, since LRs are mostly built manually, they have a high cost. This fact often lead LRs to be distributed under restrictive licenses and/or to not be shared with the public. Such an approach presents the drawback to considerably limit collaborations and reduce the valuable feedback brought by a greater number of users.

Confidence Federating as many people as possible around a common LR does not make sense if the overall quality of the LR is reduced by some collaborators. Therefore, one usually needs to first demonstrate his or her competence before being granted the right to edit an LR. The resulting number of candidate collaborators is thus reduced to a small number of persons who have the linguistic and computer skills required for a shared edition of the LR.

Accessibility Obviously, someone willing to help maintaining an LR needs to access it. This basic statement is sometimes restrained by several reasons that can be technical (some restrictive technologies are required), geographical (the LR is not accessible from anywhere) or even security-related (the LR is located on a server restricted by security policies).

4.2. Guidelines to enhance collaborative work

The lexical formalism used for developing LRs should enable as wide a range of applications as possible, in particular by using general frameworks associated with tools (compilers) that are able to convert the general LR into specialized ones. Indeed, such an approach allows experts to develop and maintain specialized modules as independent modules, hence easing the life cycle of LRs and maximizing feedback. For example, one can develop a core lexicon for a language and provide several branches for developing specialized lexicons on zoology, medicine, etc. In addition, the more general the framework is, the more chance it has to be regularly maintained and updated itself.

Concerning licenses, it mostly depends on the main objectives of the developers of the LR. If the main objective is to bring the LR to a greater level of quality, one should try to maximize feedback and federate people with the skills to collaborate, be it academical or industrial. The licenses used should thus be as non-restrictive as possible.

As regards confidence, the main problem is that granting somebody edit rights on the LR generally means to grant such rights on the whole of it. A simple but straightforward approach to bypass this problem is to grant progressively edit rights on sub-part of the LR. Such a scalable approach can be achieved by designing interfaces with restrictions on what is editable or not according to the confidence level assigned to the user. In addition, interfaces can prevent editing/typing errors and allow users to focus on the data itself without worrying about mastering the underlying formalism or technologies. Finally, interfaces can help controlling more easily the evolution of LRs since they can allow to trace their modifications.

Regarding accessibility, web technologies are a convenient way to provide a direct access to LRs. Indeed, they are

among the most standardized online technologies and thus, are free of the technical, distance and security troubles mentioned above. When used to develop interfaces, they generally constitute an appropriate way to access and edit LRs without any particular additional requirement.

5. Saving efforts

We have explained why it can be useful to federate a community around an LR in order to increase the available workforce. However, it is also necessary to try and reduce as much as possible the need for manual efforts. In order to achieve this goal, several tracks may be considered.

5.1. Using existing frameworks

Even if the NLP community did not release stable frameworks for all linguistic levels, most of them have been studied and (partial) solutions have emerged. Since existing frameworks are usually mature and the libraries/codes provided are often free of errors, a reasonable idea is to use them and, if necessary, extend them.

5.2. Using existing resources

Existing resources are generally valuable sources of linguistic knowledge when building new LRs or extending others. Of course, such an approach depends on the kind of knowledge one is trying to adapt and on the formalisms (and its underlying linguistic theory) the LR is based on. Nevertheless, LRs describing a similar level of language description usually share common points. Thus, adapting parts of the available existing resources is often an achievable objective.

Since related languages share significant parts of their linguistic legacy, such an approach should not be limited to the scope of a single language. Indeed, the proximity between linguistically related languages can sometimes allow to "transfer" formalized knowledge. Thus, one can also consider other existing LRs describing related languages. This approach is particularly useful for languages with smaller speech communities and limited digital resources.

5.3. Automatizing correction and extension

LRs are often built with little (or no) computer aid. This causes a common situation where the resources are developed until a (more or less) advanced state of development where it becomes too difficult to find errors/deficiencies manually. Since it can greatly reduce the need for manual work, automatizing the processes of extension and correction should generally be considered in order to enhance the sustainability of LRs.

Of course, such techniques are specific for each type of linguistic knowledge. Some linguistic description levels (e.g., semantics) are more difficult to process with such an approach than others (e.g., morphology). As far as the morphological and syntactic levels are concerned, one can base a generic approach on research results such as those described in (Sagot and Villemonte de La Clergerie, 2006) and (Nicolas et al., 2008), as we now sketch.

Identifying possible shortcomings in an LR can be achieved by studying unexpected/incorrect behaviors of some tools relying on the resource. To do so, it is necessary to first establish what can be considered as an unexpected behavior. For example, for a parser, an unexpected behavior can be defined as a parse failure. Then, if among the elements of a given LR, some are found when unexpected behaviors occur more often than average, such element can be (statistically) suspected to be incorrectly described in the LR.

This "error mining" step, that already provides an interesting data to orientate the correction of the studied LR, can be completed with the following automatic correction suggestion step. Contrarily to formal languages, natural languages are ambiguous and thus, difficult to formalize. Nevertheless, this ambiguity has the advantage of being randomly distributed on the different levels of a language. Consider two different LRs are interacting within an NLP tool (e.g., a syntactic lexicon and a grammar combined in a symbolic parser). This tool is designed to try and find a joint "match" between both resources and the input of the tool (e.g., a parse that is compatible with both the grammar and the lexicon). In other words, one can view each LR as providing a set of possibilities for each element (e.g., lexical unit) in the input. Depending on their state of development, it can be truly rare for two resources A and B to be incorrect on a same given element, i.e., many unexpected behaviors are only induced by only one resource at a time. Therefore, if of one of the LRs, say A, is suspected by the error mining step to provide erroneous and/or incomplete information on a given lexical unit, it is reasonable to try and rely on the information provided by the other LR, B, for proposing corrections to the dubious lexical entry. For example, let us suppose that a verbal entry in a lexicon A is suspected to provide a sub-categorization frame that is incomplete w.r.t. a given sentence. Using a parser that combines A with a grammar B, it is then reasonable to let the grammar decide which syntactic structures are possible for this sentence, by preventing the parser from using the dubious information provided by A about this verb. Then, correction proposals for A can be extracted from the sub-categorization frame built by the parser.

Of course, among the corrections generated thanks to B there might be correct and incorrect ones. Therefore, such approaches should generally be semi-automatic (i.e., with manual validation). Nevertheless, semi-automatic approaches are a good compromise to limit both human and machine errors since most of the updates done on the LRs are automatically created and manually validated.

Finally, another convenient feature of this approach is the following: if resource B cannot provide any longer relevant corrections for resource A, and thus does not offer a solution for some unexpected behaviors, we can consider the remaining ones as mostly representing shortcomings of resource B since it does not cover them. This defines an incremental and sequential way to identify sentences that instantiate shortcomings of resource B. Thus, correcting resource A thanks to resource B generates useful data to correct resource B. Once resource B has been updated, it can be again used to correct resource A and so on.

5.3.1. Using plain text

The approach described in the previous section requires input corpora. They should be as error-free as possible in or-

der to guarantee that most unexpected behaviors are caused by shortcomings of the LRs, and not by errors in the input. If this input data is an annotated one, only manual annotation can guarantee a certain level of quality. But manually annotated data is only available in limited quantities for a small number of languages and producing such data contradicts the objective of saving manual work.

Therefore, the data used should be raw text, daily produced for most languages and freely available in large quantities on the Internet. So as to guarantee the quality of the data, only linguistically correct (error-free) texts, such as law texts or selected journalistic productions, should be used.

6. Results achieved by the *Victoria* project

Even though the *Victoria* project has not yet reach all its goals, the following results have been already obtained using the above-described guidelines as often as possible.

As regards to formalisms, we have chosen the Alexina framework (Sagot et al., 2006; Sagot, 2010) to develop our morphological and syntactic lexical resources. This framework, compatible with the LMF standard, represents morphological and syntactic information in a complete, efficient and readable way. It has already been used to create LRs for various languages (e.g., French, Spanish, Slovak, Polish, Persian, Sorani Kurdish) and has been combined with several taggers and various parsers based on a range of grammatical formalisms (LTAGs, LFG, Interaction Grammars, Pre-Group Grammars...).

Regarding grammatical knowledge, our resources rely on a meta-grammar formalism which represents the syntactic rules of a language by a hierarchy of classes. Even if in practice, we compile our grammars into a hybrid TAG/TIG parser (Villemonte de La Clergerie, 2005), this meta-grammar formalism is theoretically compilable into various grammar formalisms. Such a formalism is convenient in so far that it allows for an easy adaptation of an existing grammar to a linguistically related language.

As regards license issues, the LGPL-LR[14] and CeCILL-C[15] licenses have been chosen to publish our resources, namely our lexicons, grammars and editing interfaces.

Among the three kinds of resources developed, lexicons are clearly those requiring most collaborative work. The efforts concerning interfaces have thus been orientated to develop a web interface for lexicon based on the portlet technology. Its current version allows us to search for entries with complex logical equations covering any kind of data available in the lexicon. It also allows for a guided edition of the entries and traces every change.

Various techniques have been created or improved, in particular for achieving the following tasks: (i) inferring morphological rules from raw text, (ii) inferring morphological rules from a morphological lexicon, (iii) extending a lexicon thanks to a tagger (Molinero et al., 2009), (iv) extending a lexicon thanks to morphological rules (Sagot, 2005), (v) correcting a lexicon thanks to a grammar (Nicolas et al., 2008; Sagot and Villemonte de La Clergerie, 2006). Most of these techniques follows the guidelines described in section 5.3.

These techniques have allowed us to produce several LRs. Among them, two wide coverage lexicons for Spanish and Galician have been produced along with two sets of morphological rules. The Spanish lexicon L*effe*[16] (Molinero et al., 2009) has been obtained by merging several existing Spanish linguistic resources, and contains syntactic information. A Spanish meta-grammar (SPMG) has also been adapted from a French one (FRMG). For both L*effe* and *SPMG*, we took advantage of the similarity between French and Spanish language while building their first versions.

7. Conclusion

We have presented several guidelines to ease and improve the creation and correction of LRs. These guidelines are the cornerstone methodologies of a project dedicated to this task, the *Victoria* project. When considering the manpower involved in this project and the practical results it has achieved so far, we believe that its guidelines might be of interest for anybody involved in a similar task.

8. References

Gil Francopoulo, Monte George, Nicoletta Calzolari, Monica Monachini, Nuria Bel, Mandy Pet, and Claudia Soria. 2006. Lexical Markup Framework (LMF). In *Proceedings of LREC 2006*, Genoa, Italy.

Miguel A. Molinero, Benoît Sagot, and Lionel Nicolas. 2009. A morphological and syntactic wide-coverage lexicon for Spanish: The L*effe*. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 09)*.

Lionel Nicolas, Benoît Sagot, Miguel A. Molinero, Jacques Farré, and Éric Villemonte de La Clergerie. 2008. Computer aided correction and extension of a syntactic wide-coverage lexicon. In *Proceedings of COLING'08*, Manchester, UK.

Lionel Nicolas, Miguel A. Molinero, Benoît Sagot, Elena Sánchez Trigo, Éric de La Clergerie, Miguel Alonso Pardo, Jacques Farré, and Joan Miquel. 2009. Towards efficient production of linguistic resources: the *Victoria* project. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 09)*.

Benoît Sagot and Éric Villemonte de La Clergerie. 2006. Error mining in parsing results. In *Proceedings of ACL/COLING'06*, pages 329–336, Sydney, Australia.

Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. 2006. The L*efff* 2 syntactic lexicon for French: architecture, acquisition, use. In *Proceedings of LREC'06*, Genoa, Italy.

Benoît Sagot. 2005. Automatic acquisition of a Slovak lexicon from a raw corpus. In *Lecture Notes in Artificial Intelligence 3658 (© Springer-Verlag), Proceedings of TSD'05*, pages 156–163, Karlovy Vary, Czech Republic.

Benoît Sagot. 2010. The L*efff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of LREC'2010*, Valetta, Malta.

Éric Villemonte de La Clergerie. 2005. From metagrammars to factorized TAG/TIG parsers. In *Proceedings of IWPT'05*, pages 190–191, Vancouver, Canada.

[14] Lesser General Public License for Linguistic Resources.

[15] LGPL-compatible, `http://www.cecill.info/`.

[16] Léxico de formas flexionadas del español / Lexicon of Spanish inflected forms.

Laundry Symbols and License Management -
Practical Considerations for the Distribution of LRs based on experiences from CLARIN

Ville Oksanen*, Krister Lindén†, Hanna Westerlund†
*Aalto University
P.O. Box 19210, 00760 Aalto, Finland
†University of Helsinki
P.O.Box 24, 00014 University of Helsinki
E-mail: ville.oksanen@tkk.fi, krister.linden@helsinki.fi, hanna.westerlund@helsinki.fi

Abstract

One of the most challenging tasks in building language resources is the copyright license management. There are several reasons for this. First of all, the current European copyright system is designed to a large extent to satisfy the commercial actors, e.g. publishers, record companies etc. This means that the scope and duration of the rights are very extensive and there are even certain forms of protection that do not exist elsewhere in the world, e.g. database right. On the other hand, the exceptions for research and teaching are typically very narrow.

1. Introduction

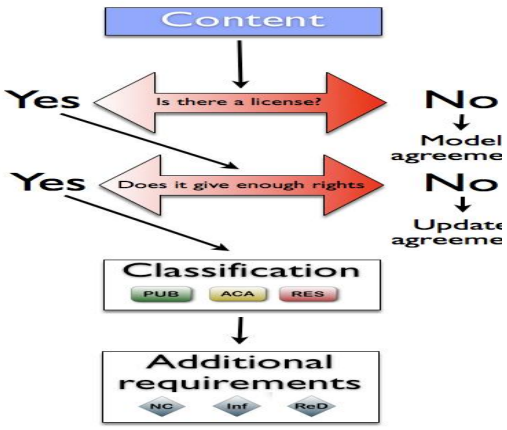
One of the most challenging tasks in building language resources is the copyright license management. For a more general discussion on open access data licensing, see e.g. Klump & al. (2006). There are several reasons for this. First of all, the current European copyright system is designed to a large extent to satisfy the commercial actors, e.g. publishers, record companies etc. This means that the scope and duration of the rights are very extensive and there are even certain forms of protection that do not exist elsewhere in the world, e.g. database right. On the other hand, the exceptions for research and teaching are typically very narrow. To make the situation worse, the possible sanctions for copyright violations are severe, e.g. in Finland the maximum penalty for copyright violation on the Internet is a two-year prison sentence.

This means that there is very little space for errors in the management of copyright licenses - at least in theory. In practice the system mostly "just works" even if the formal agreements are often totally missing or the distribution of material was agreed on in a phone call years ago between persons who no longer work in their respective organizations. The reason for this is that there is typically no commercial interest to start a legal process and high legal fees form an effective preventive factor. However, when building an EU-wide system, one cannot rely on such an informal approach.

In the first part of this article, we describe how we plan to handle the matter in the CLARIN project. In the second part, we describe our early practical experience with the proposed classification. In the last part of the article, we briefly discuss aspects that could be generalized and possible actions for making the use of copyrighted material in research more flexible.

2. CLARIN Resource Distribution Types

In CLARIN the typical flow of the content is the following: A copyright holder, e.g. a newspaper, licenses its content to a CLARIN Content Provider that distributes the content to the End Users through a CLARIN Service Provider. This means that the license chain has to follow a similar structure. Unfortunately even the first step is often difficult because there is a group of resources, for which there are no written license agreements and individuals familiar with the details are no longer available. Another problem from the CLARIN perspective is the variation in the existing license agreements, which makes it hard to offer a centralized service. To tackle these problems, several sets of agreements have to be used. For an outline of the resource classification procedure, see Picture 1.

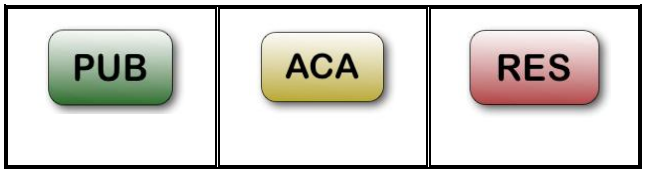


Picture 1. Resource classification task.

Regarding the license variation, we carried out an extensive survey and found that it is possible to categorize the licenses into three different groups:

- Publicly Available Resources
- Resources for Academic Use
- Resources for Restricted Use

We followed the model used by Creative Commons and created simple icons, i.e. *care symbols*, making it easier for the end-user to immediately see under which conditions the resource can be used, see Picture 2. In addition, a deed describes the rights in human readable textual form. Finally, there is also the actual license agreement and the metadata, i.e. the machine readable information. However, Creative Commons is not sufficient as such for CLARIN, because Creative Commons does not allow for distribution restricted to academia or even more limited groups of users, which is essential for many of the older resources to be included in CLARIN.



Picture 2. Symbols for the main distribution classes.

Publicly Available (PUB) is one of the categories endorsed by CLARIN. To belong to this group, the following requirements have to be met:
- the license should allow distribution of the tools and resources from the CLARIN infrastructure,
- there must be no limitations, e.g. based on status or geographical location etc., on who can access and use the tools and resources and
- there must be no limitations on the purpose for which the tools and the resources are used.

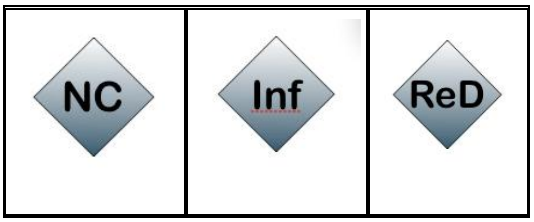
In other words, the license should follow the Protocol for Implementing Open Access Data[1] as closely as possible. For the new tools and resources, the preferable license is either the Creative Commons Zero (CC0)[2] or the Open Database License (ODbL). However, for the previously licensed tools and resources, re-licensing is often not possible, and the submitting party should make a careful assessment of the terms of the existing licensing agreement.

For **Academic Use** (ACA) the license agreement includes an additional requirement that the use is somehow related to an academic institution. Here the problem may arise from the definition of academic use. To qualify under this category, the tools and resources:
- should be available at least for anyone doing research or studying in an academic institution recognized by the Identity Provider Federation and
- should be available for studying, research and teaching purposes.

The last category, **Restricted Use** (RES) includes the resources that do not fulfill the previous requirements but still could be offered to the users if certain additional requirements are met. The most typical reasons for a resource to fall under the scope of RES are:
- a requirement to submit detailed information, e.g. an abstract, on the planned usage or
- specific ethical or data protection-related additional requirements.

In conjunction with the main license categories PUB, ACA and RES, there can also be all or any of three additional requirements:
- A requirement for strictly non-commercial use (NC)
- A requirement to inform the copyright holder regarding the usage of the tools and/or the resources in published articles (INF)
- A requirement to redeposit modified versions of the tools and resources with the Service Provider (ReD)

Picture 3 displays the symbols designed for the additional requirements.



Picture 3. Symbols for additional distribution restrictions.

However, this does not solve all the problems. In some cases there either is no license agreement at all, because such an agreement has never been made. It is also quite common that the existing agreement is somehow problematic, e.g. very low in details, making the categorization impossible. For those situations we created the *CLARIN Update Model Agreements* with the purpose to procure the required rights. The best option is to re-license the content with the CC0-license. See the Berlin Declaration (2003) for best scientific licensing practices. It is well-understood and offers enough rights for all parties in different digital and non-digital environments. It is also compatible with most of the other open content licenses. Unfortunately it is not always possible to use CC0 due to the demands of the copyright holders. Thus Update Model Agreements for Academic and Non-Commercial Use are also available.

These agreements presuppose that there are existing agreements but that the rights are not adequate or too unclear. It should be pointed out that both the terms

[1] http://sciencecommons.org/projects/publishing/open-access-data-protocol/
[2] http://creativecommons.org/choose/zero

non-commercial and academic are relatively ambiguous and it is a relatively demanding task to write generally accepted definitions. See Hietanen & al. (2007) for a discussion on the problems related to the term Non-Commercial in Creative Commons. Especially the scope of accepted commercial use is something that needs first to be solved on a political level and only after that formulated in legal terms.

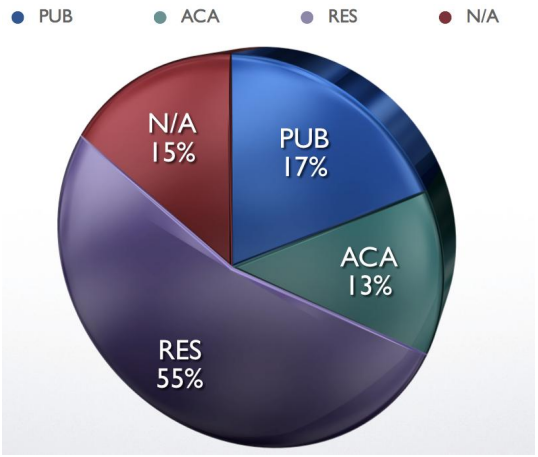
3. Practical Experience

In order to test the usability of the classification system and our classification guidelines, we did an initial classification test. We sent out a request to the custodians of 116 resources found in the CLARIN LRT inventory. The resources and their custodians were located in Finland (91), Denmark (3), Germany (21) and Greece (1). A certain preference was given to our home turf in this initial survey, because we thought that if there were problems in the instructions, it would be easier to correct closer to home. We received an answer for 40 of the resources, i.e. 34.5%. A response rate above 1/3 makes the survey fairly reliable.

Distribution type	Number	Percentage
PUB	7	17.5 %
ACA	5	12.5 %
RES	22	55.0 %
No classification applicable	6	15.0 %
Total	40	100.0 %

Table 1: Distribution of resources according to the CLARIN classification.

In Table 1 and Picture 4, we see that more than half of the resources were classified into the (RES) restricted category. Approximately one third were classified as (PUB) publicly or (ACA) academically available. Finally, one sixth was found to be exceptional.



Picture 4. Main distribution categories.

One of the publicly available resources (PUB) was also classified as non-commercial, whereas all of the academically available resources (ACA) were non-commercial. The restricted resources (RES) were roughly equally divided among no additional restrictions (27.3 %), a non-commercial restriction (31.8 %) and a requirement that the license be personally granted by the content owner (40.9 %).

Only one resource was such that the content provider found no applicable distribution type because there was no formal agreement between the content owner and the content provider. In this case, the content owner had given his consent to the content provider to use the data by email and the data had been further analyzed by a commercial company providing parsing services, but here as well there was no formal agreement regulating the use of the analyzed data. In addition, there were 5 resources for which the research project was still ongoing and the question of distribution would be discussed only after the project had finished. All in all, some kind of classification was received for a total of 40 resources.

A number of feed-back questions concerned the fact that some corpora did not seem to fit a category completely. In this case either an upgrade agreement needs to be concluded with the content owner or the resource will have to be classified into a more restricted category for which it has all the necessary distribution rights. For this reason a number of legacy resources currently fall into the RES category, even if they probably could be brought into the ACA category by procuring some minor additional rights.

An additional question about the classification process was the issue of how to classify commercially available corpora and who should pay for them. Electronic payment is possible and well-regulated within EU so it is more of a political issue than it is a legal issue how the funding should be arranged. Some of the content providers also saw a need for a full blown digital rights management system, and it is technically possible for certain types of resources, so it is also a political decision for a future CLARIN ERIC whether such resources will be included.

4. Future Work

Finally, an important future goal would be to add the necessary research exceptions directly into the national copyright laws as permitted by the EC Infosoc Directive (Directive 2001/29/EC). For this purpose we have created a lobbying message (Oksanen, 2009) aimed at the EU Commission together with DARIAH - Digital Research Infrastructure for the Arts and Humanities and Cessda - Council of European Social Science Data Archives. The purpose of this message is to push the Commisison to make the research exceptions mandatory in the national legislations. Writing such a document is a delicate

balancing act. It should be broad enough to bring some benefits. On other hand, too wide demands just cause strong opposing reactions from publishers and lead nowhere.

The current formulation of the lobbying message includes two main points:

- the legislation should *allow free use of copyrighted works for academic purposes* and
- the legislation should *not unreasonably prejudice the legitimate interest of the rights holder,*

which follow the language of the three step test of the Berne Convention[3]. By using this approach, the benefits are the same as using standardized license agreements – the main actors know at least what the language *most likely* means even if there are some ambiguities (Hugenholtz and Okediji, 2008).

Unfortunately, it is unlikely that any lobbying in this area will bring quick results. Most of the resources of the European Commission are currently dedicated to the directive which aims at extending certain aspects of the copyright duration. That process is currently in a gridlock because many of the member countries oppose the directive and before a solution is found, no new hard law pertaining to copyright will be introduced (Hugenholtz & al, 2008). One option is that there will be some kind of general *open data* regulation that would resolve the situation. The movement for creating open databases is currently very strong in some of the member states, e.g. in UK and Finland, and it is not totally out of the question that EU would step in to harmonize the field further than the PSI directive (Directive 2003/98/EC), which covers only public sector information.

One aspect, which we do not cover in depth in this paper are the questions pertaining to the privacy regulation concerning data enabling recognition of persons. However, in most cases a clear written consent from the research subjects to reuse the data for research solves the problems. Older material containing personal data that have been collected without written consent to reuse can still benefit from the exceptions for scientific, historical or statistical research. It should be pointed out that due to the nature of these exceptions material containing personal data is typically available only for the ACA or RES categories unless it is anonymized in which case it becomes eligible for the PUB category. Anonymizing personal data is often feasible for text data but it may become prohibitively costly for audio and video data.

5. Conclusion

It would be preferable to have most of the resources in the public or at least in the academic domain in order to facilitate sharing, but according to our initial classification test, it seems likely that a sizable portion of the resources for various reasons have restricted access and some will even require a fairly intricate authorization protocol with letters of recommendation and an abstract describing the research purpose. This will hopefully change over time, when researchers realize that they can get citations and fame for making their research material available to others. In addition, some research funding agencies have added the requirement that data collected with their grant funding should be made available to subsequent research projects, which makes sense both from a research financing point of view and from a scientific inter-subjectivity point of view, i.e. the funding agency can avoid paying repeatedly for the same data collection effort and the research results become easier to verify by other research teams.

One obvious problem is that opening resources for research, if they have been created even partially with private funding, should not threaten the business interests of the right holders. There is no easy solution for this and in practice there will always be conflicts of interest when opening databases that have dual usage possibilities, i.e. commercial exploitation and scholarly research, e.g. non-historical news article collections.

6. Acknowledgements

We thank the CLARIN FP7 project for the financing and the numerous content owners that took time to answer our questions regarding the distribution types for their resources.

7. References

Berlin Declaration. (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities*. Berlin. http://www.zim.mpg.de/openaccess-berlin/berlindeclaration.html.

Hietanen, H., Oksanen, V., and Välimäki, M. (2007). *Community created content*. Law, business and policy.

Hugenholtz, P.B., and Okediji, R. (2008). *Conceiving an international instrument on limitations and exceptions to copyright*. http://dare.uva.nl/record/301952

Hugenholtz, P.B., Helberger, Dufft. N., and van Gompel, S.J. (2008). *Never Forever: Why Extending the Term of Protection for Sound Recordings is a Bad Idea*, E.I.P.R., 2008-5, p. 174-181.

Klump, J., Bertelmann, R., Brase, J., Diepenbroek, M., Grobe, H., Hock, H., Lautenschlager, M., Schindler, U., Sens, I., and Wachter, J. (2006). *Data publication in the open access initiative.* Data Science Journal 5: 79–83.

Oksanen, V. (2009). *NEERI Message - Freedom of use of copyrighted works for academic purposes*. http://www.csc.fi/english/pages/neeri09/programme/materials-fri/oksanen2.pdf

[3] The Berne Convention for the Protection of Literary and Artistic Works, usually known as the Berne Convention, is an international agreement governing copyright, which was first accepted in Berne, Switzerland in 1886.

Resource Lifecycle Management: Changing Cultures

P. Wittenburg, J. Ringersma, P. Trilsbeek, W. Elbers, D. Broeder
MPI for Psycholinguistics
E-mail: peter.wittenburg@mpi.nl

Abstract

Proper resource lifecycle management will be essential to not loose our scientific memory in a time where also in the area of language resources the sheer mass of resources and their complex relationships amongst them is exponentially increasing. Proper resource lifecycle management addresses everyone participating: (a) researchers as creators need to deposit their data at trusted repositories and deliver data that adheres to basic criteria with respect to structure, content and metadata; (b) archivists need to take care that quality criteria guide the work of the repositories and that identity and authenticity of the stored objects is ensured; (c) funders need to request for adherence to quality criteria and set funds aside for this and (d) finally developers of software tools need to make use of registered schemas and semantics where possible.

1. Introduction

Although sharing and re-using data are old topics that were already addressed at LREC 1998 in a separate workshop it seems that this issue is seen as being even more important these days. Just recently the report of the Data Management Task Force of e-IRG and ESFRI [1] was published addressing three major issues in its first version: high quality metadata, quality assessment and interoperability. In its next version other topics such as curation, contextual metadata, controlled vocabularies and costs will be addressed as well. Also the communication of the EC on "ICT Infrastructures for e-Science" [2] states that "if (scientific data is) left unmanaged, (this) could undermine the efficiency of the scientific discovery process" and that a solid basis is needed to "develop a coherent strategy to overcome the fragmentation and enable research communities to better manage, use, share and preserve data". All these statements are clear indications for the fact that the state of proper Life Cycle Management (LCM) is not satisfying.

Five groups of persons are influencing LCM: (1) The creators need to be efficient and to meet the egocentric needs of their project in focus, thus often ignoring basic LCM requirements. (2) The repository experts (archivists) need to tackle the issues of long-term preservation and accessibility, of trust and legal and ethical aspects, but often being confronted with dynamic collections and lacking proper LCM procedures to guarantee data authenticity for example. (3) The funders need to establish criteria that ensure re-usability of the data they helped creating. (4) The developers need to develop tools supporting standards to ensure cost-effective curation. (5) The interests of the users cannot be specified in clear terms, since there is a wide variety of users ranging from researchers who want to manipulate the data in their way to those without special computer know how who want to use a readymade and simple search and statistics tool. When talking about long-term accessibility we speak about future users the needs of which we cannot even predict.

From current practice we know that these interests are diverse and that it is therefore impossible to meet all requirements at the same time and that the wishes are partly in conflict with each other. While users are interested in nice and combined presentations for example, archivists need to support neutral and atomic representations. The technologies to bridge this gap are often not trivial and layers of complexity are introduced. But we can describe a few generic IT principles that can serve as agreed meeting points if we are able to change cultures and it is obvious that we need to design a common solution for a data-service e-Infrastructure in Europe to seamlessly help researchers in data management aspects.

2. Nature of Language Resource Collections

First we need to specify where we are talking about when using the term "language resource[1]". It is obvious that we refer to all sorts of (a) primary data dealing about languages such as audio/video/time-series recordings, images of all sorts, texts such as digitized books or newspaper issues or even the content of the web and (b) secondary data such as (structured) annotations, lexica, wordnets, metadata, etc. Collections are aggregations of such resources that have some relations amongst each other. The term "corpus" refers typically to a coherent collection that was created with a certain goal in mind. However we are faced with the fact that creators and users do not share the same goals thus virtual collections are created that combine various resources in unintended ways by new users. Individual resources from collections will be re-purposed increasingly often by viewing differently on them and by setting them into different contexts.

Research collections and partly resources have a dynamic nature, since users are adding, extending and changing resources continuously as part of their research workflows. New resources are added, new versions are created and different sets of relations are overlaid on top requiring on the one hand dynamic management methods and on the other hand a mechanism to uniquely identify a specific version of a single resource, thus a specific object.

[1] We restrict ourselves to digital resources.

It is generally agreed now that such an object is identified with the help of a unique and persistent identifier (PID), proper metadata including provenance information and a checksum information. Yet we lack proper guidelines about the granularity of the objects to be identified. If all kinds of different resources are contained and encapsulated in a relational database management system for example, identification of individual resources needs to be dealt with by application logic which may lead to serious data management problems in the long run.

3. Example of Life Cycle Management

The DOBES program on documenting endangered languages [3] started in 2000 and realized from its beginning that documenting languages that will soon become extinct only makes sense with a clear concept of long-term preservation and accessibility in mind. A number of dimensions were discussed and tackled:

- The researchers have to give a copy of all primary and secondary data to the designated digital archive knowing that only this step will make the state of resources explicit, that only a transfer into a well-organized archive will increase the probability that data will be maintained beyond the time period of project funding and that accessibility for a longer period can be guaranteed.
- For the archive it was agreed that it should only store data in standardized and well-accepted formats and that the teams should describe[2] the semantics of the concepts used, i.e. within the program it was agreed that explicit syntax based on standards and described semantics is the best way to provide interpretability of the data for a longer period of time. The acceptance of standards and best practices also will make it possible to migrate data more easily to new formats and to control whether the transformation is free of data loss. As an example we can refer to media formats where within one decade four standards were introduced (MPEG1, MPEG2, H.264, M-JPEG2000) [4] and where only the last one offers a realistic solution for uncompressed representations that can prevent concatenation effects.
- Metadata creation was taken very serious so that the currently 30.000 records provide rich information according to the IMDI standard [5] allowing to start smart searches and to build attractive community portals, i.e. attractive web-sites where metadata queries are hidden behind menu items.
- Metadata is also the basis for proper organizations of the collections according to specific criteria. Depositors and archive managers need to have a canonical and stable management tree while users need the freedom to create their own virtual collections having separate organizations.
- Still in field research many tools are being used that do not provide data in specified XML formats. Despite high costs for example when converting complex lexica the immediate curation principle is applied where possible, since it is well-known that the costs for late curation grow over time [6]. Curation makes use of open standards where possible such as Lexical MarkUp Framework (LMF) [7] and ISO 639-3 language codes [8] or best practices based on explicit schemas such as EAF [9]. Yet there was no attempt to use a formal framework to register tags which were used to describe linguistic phenomena since ISOcat [10] became available just recently.
- Taking care of long-term accessibility also means to find solutions for preserving the pure bit-streams. Also in this respect agreements were achieved in so far that all data is dynamically replicated 4 times at large computer centers, i.e. in total 6 copies of all resources are maintained at widely distributed locations. In addition the DOBES program managed to establish more than ten so called regional repositories worldwide storing sub-collections [11]. Although versioning is in place and the granularity is appropriate we cannot be satisfied, since the replication protocols widely used do not implement safe authenticity checks and only work at physical level, i.e. do not preserve the resource contexts.

4. Current Discussions

In addition to what we described so far we would like to refer to two recent discussions: (1) The e-SciDR project brought out a report [12] about the perspectives for digital repositories and (2) the US ASIS&T organization just had a meeting [13] which was devoted to issues that have to do with data management. We would like to refer to some results of these two initiatives, since they are closely related with the issue of data lifecyle management.

The e-SciDR report stresses the requirements and tasks of digital repositories that need to take care of lifecycle management, since only handing over data to recognized repositories will make data explicit and allow quality assessments. Data resources that are kept on user notebooks or departmental servers will be lost sooner or later. The report stresses that scientific knowledge extracted from the stored information is part of our cultural heritage and also a strategic and competitive resource. Therefore repositories need to be very serious not only with respect to storing the information, but also by offering adequate deposit, access, searching and visualizing tools. In addition they need to look after a number of critical characteristics such as data integrity, authenticity, usability and discoverability, and also by taking care of interoperability issues.

The current repository landscape is characterized as complex, diverse and by appearing in different forms, sizes and ages. For users the landscape is confusing and obscure due to a complex matrix of technologies, facilities and unfamiliar and fuzzy terminologies. In general it is stated that there is a mismatch between the

[2] The description should be done in prose text lacking a widely accepted formal framework.

availability of funding resources for the creation of data and the development of software compared to the availability of funds for data curation and software maintenance.

It is obvious that this state is not satisfying with respect to the overall quality of the repository landscape and therefore with respect to the state of data lifecycle management. Consequently, the report is suggesting amongst others improved funding schemes stressing the maintenance aspects and more coordination efforts to improve the harmonization between repositories.

The ASIS&T summit brought together librarians, archivists and technologists all being confronted with the problem of maintaining accessibility of research data to not lose our scientific and cultural memory. Librarians are still working on appropriate architectures that can handle the dynamic nature of research data, since they are historically used to static types of data: publications which are the end products of the research process. Traditional approaches are used to make a difference between an access repository and a static archive the resources of which are not being touched. The film industry for example stored copies of all their 35 mm film roles in an old mine with the intention to only access it after perhaps 50 or even more years. In contrast the access repository was established to make money with the copies and to allow re-purposing. In the film industry there is a discussion whether this separation will survive in the area of digital resources. For analogue resources every access and copy action was associated with a degradation of its quality. In the digital area the opposite is true: one has to access and migrate resources at regular times to ensure interpretability. Maintaining a digital master copy costs much more than an old analogue one (factors of 12 were mentioned [14]) and double copy maintenance efforts is adding even more costs. Therefore we can see a trend to overcome this distinction.

Preservation initiatives pushed forward in research driven infrastructure projects are discussing instead the need to make a difference between workspaces and repositories where workspaces are used by researchers to create and manipulate temporary data. Since it is often not evident when a certain temporary resource will get a common value it needs to be very simple for the researchers to upload their data.

The relevance of metadata was stressed by almost all speakers and it is obvious that

- metadata creation is costly if it is not supported by efficient tools and if it is not created from the beginning of a resource's lifecycle;
- DublinCore metadata is in general not sufficient for research purposes and more elaborated schemes are required;
- easy upload facilities similar than YouTube are wished by researchers, but that there is a danger that people stick with the minimal descriptions associated with such procedures;
- metadata needs to be machine readable and structured if advanced concepts such as automatic profile matching need to be supported – something only few people are realizing at this moment;
- incentives are needed to motivate researchers to produce high-quality descriptions.

Quality assessment will increasingly be relevant at various layers such as for resources and for repositories. In the US the TRAC assessment procedure receives much attention while the Data Seal of Approval is hardly known yet.

For the library domain ready-made solutions such as D-Space and Fedora are of high interests, while in the research data world these do not play such an important role. It was briefly discussed whether an object model such as Fedora is useful in a research scenario where individual resources are being re-purposed in various ways by researchers and where the external relations are so heterogeneous. As a consequence object packaging does hardly make sense and simply adds another layer of complexity resulting eventually in performance penalties.

The conference organizers expressed their big concerns about the danger of losing research data, the great efforts of NSF to take actions with two DataNet projects being already granted and concluded with saying: let's be good stewards of our data. Also in the US the goals are set very clearly: (1) Identify good exemplars for proper data management systems; (2) Identify good examples for the federation of collections; (3) Identify solutions for interoperability; (4) Identify possibilities towards a national infrastructure initiative

5. Requirements for Life Cycle Management

Based on the DOBES project and a few others such as the Dutch CLARIN [15] where all participants are requested to fulfill similar requirements we can derive a number of basic guidelines that need to be considered when looking at proper LCM for language resources:

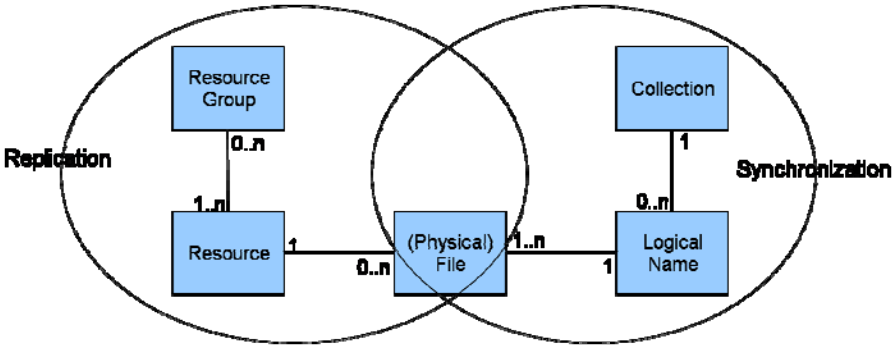
- There needs to be a proper definition of the objects that are the basic units of management and each such object must be associated with a persistent and unique identifier (PID), checksum data and high quality metadata. Changes need to lead to new versions which are new objects that can be referred to and therefore need to be maintained.
- It must be possible to aggregate such objects in arbitrary manners since we cannot anticipate how researchers will combine resources to gain new insights. Thus objects need to be "atomic" and aggregations are identified by more complex metadata descriptions which obviously contain many references to the included objects and which have to be maintained and citable as well.
- Proper bit-stream management with safe replication based on information associated with PIDs is important to guarantee data survival and authenticity.
- Accessibility over long period of times can best be achieved by consequently using open standards and by applying an immediate curation method.
- Explicit syntax and declared semantics all

registered in recognized schema and concept registries are required to guarantee long-term data interpretability.

- Only the transfer of data to a trusted center participating in quality assessment procedures will make the state of a resource explicit.

6. Replication and Synchronization

Since these requirements as far as long-term preservation is concerned are not met by existing systems the REPLIX project [16] was started as collaboration between CLARIN [17], DOBES and DEISA [18] with the intention to build a thin and safe solution for a replication/synchronization layer which is operating at the logical level and independent of the chosen repository system. We speak about replication if there is a master-slave relation in the copying activity, i.e. the bit-stream at the master repository is copied to the slave repository and this repository can only give read access to the collection the resource is belonging to. Synchronization needs to be done when both collections are dynamic, i.e. when also the copied collection can be extended. The copy activity will now happen in two directions. Since modifying an object always will lead to a new version conflicts can be bypassed formally if the versioning system allows unique identification. This is further illustrated in the figure below.



This figure describes the difference between data replication which follows an underlying master-slave relationship and data synchronization which assumes a peer-to-peer relationship.

Copying at the logical level is now getting even more important. It must be defined by rules at collection level which kind of data migration should take place. The node defining a collection in repository A can be associated for example with rules that (a) specify that all objects will be copied to a second repository B, (b) only classified new objects will be copied from B to A since B could be a university where students create many annotations in their classes which will not be archived by A for example.

We need to prevent the propagation of erroneous copies in such a replication network, if we want to take long-term preservation serious. There are different reasons for errors: (a) bit errors can occur when reading from a storage medium; (b) the storage software could create errors of various sorts and (c) the archive manager could per accident manipulate the content of an archived resource. This could happen in both cases, although in the one-directional copying model these errors are less likely to occur as in the bi-directional model. Therefore we need to associate checksum information with the PID that represents a certain object and for each operation carried out at any repository the checksum stored needs to be compared to prove authenticity.

This safe replication layer needs to be lean and independent of concrete implementations of repositories. The reason for this is that we can see that the various institutions setting up repositories for research data are all using different technical solutions which partly grew over many years, i.e. no one can expect that they will change their setup completely at least if they fulfill the basic requirements. The replication layer therefore will have to be based on an abstract API which specifies its basic functionality so that it can be easily adapted to proper repository systems.

7. Changing Cultures

Proper LCM will require a change of culture at various sides as we have indicated. At the ***creator*** side we can state that too little time is devoted for data management problems and that still tools are being used ignoring even the most basic principles such as enforcing the usage of explicit syntax and declared semantics. Creators need to provide high quality metadata and a proper canonical organization of their resources and collections not only to facilitate usage, but also to allow proper resource management.

At the ***archivist*** side proper curation needs to be carried out immediately where possible. This includes proper checks of all resources being uploaded at least with respect to the metadata quality, the presence of a suitable organization scheme and the use of a format which is widely accepted or registered in a schema registry. Yet we cannot establish formal criteria for the quality of metadata, the suitability of an organizational structure and we are at the very beginning to understand how to register concepts. It is obvious that repositories need to understand that regular assessments according to one of the known procedures (TRAC [19], DSA [20]) are a must. To establish trusted digital archives repositories will need to make use of safe replication methods ensuring authenticity of their data.

At the ***funders*** side rules must come into place that request adherence to the mentioned requirements as integral part of each project which also means that funds will be reserved for this.

Developers need to be trained to make use of standards so that users will adhere to standards without noticing. ***Research infrastructures*** such as CLARIN need to take care that open services such as the one offered by EPIC [21] are being established and available for everyone.

8. References

[1] http://www.e-irg.eu/index.php?option=com_content&task=view&id=241&Itemid=22&show=1
[2] ICT Infrastructures for e-Science. Communication from the European Commission COM(2009 108 final. Brussels, 5.3.2009
[3] http://www.mpi.nl/dobes
[4] Trilsbeek, P., Schäfer, R., Schüller, D., Pavuza, F., & Wittenburg, P. (2008). Video encoding and archiving in field linguistics. International Association of Sound and Audiovisual Archives Annual Conference. Sydney, 2008-09-14 - 2008-09-19.
[5] http://www.mpi.nl/IMDI
[6] http://www.alliancepermanentaccess.eu/documenten/beagrieabstract.pdf
[7] http://www.lexicalmarkupframework.org/
[8] http://www.sil.org/ISO639-3/codes.asp
[9] http://www.lat-mpi.eu/tools/elan/manual/ch04s01.html
[10] http://www.isocat.org/
[11] Trilsbeek, P., Broeder, D., Van Valkenhoef, T., & Wittenburg, P. (2008). A grid of regional language archives. In C. Calzolari (Ed.), Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008) (pp. 1474-1477). European Language Resources Association (ELRA).
[12] e-SciDR report: Towards a European e-Infrastructure for e-Science Digital Repositories; http://www.e-scidr.eu; 2008
[13] ASIS&T Summit on Research Data Access and Preservation; Phoenix; April 2010;
[14] The Digital Dilemma; Academy of Motion Picture Arts and Sciences; Hollywood
[15] http://www.clarin.nl/
[16] http://www.mpi.nl/research/research-projects/language-archiving-technology/replix
[17] http://www.clarin.eu
[18] http://www.deisa.eu/
[19] http://www.ais.up.ac.za/digi/docs/giaretta2_present.pdf
[20] http://www.datasealofapproval.org/
[21] http://www.pidconsortium.eu/index.php?page=partner

The Open-Content Text Corpus project

Piotr Bański[1], Beata Wójtowicz[2]

[1] Institute of English Studies, [2] Faculty of Oriental Studies
University of Warsaw
[1] Nowy Świat 4, [2] Krakowskie Przedmieście 26/28
[1] 00-497, [2] 00-927
Warszawa, Poland
{pkbanski,b.wojtowicz}@uw.edu.pl

Abstract

The paper presents the Open-Content Text Corpus, an open-access open-content versatile TEI-XML-encoded resource located at SourceForge and distributed under the GNU Public License. We review the motivation for creating the OCTC and briefly discuss its modular architecture as well as issues relating to its community-building and educational potential. We also touch upon the crucial role that its open-source licensing scheme bears for its development and persistence. Finally, we present some of the OCTC's potential applications for comparable studies of language, linguistic modelling, as well as standards design and testing.

1. Introduction

We present a project that addresses several vital issues arising in the context of Human Language Technology (HLT), and more specifically, in the context of the creation, management and use of Linguistic Resources (LRs). These issues include, among others, long-term accessibility, re-usability and interoperability of resources, as well as matters relating to the role that licensing may play in the persistence and the development of LRs.

The project in question is the Open-Content Text Corpus (OCTC, http://sourceforge.net/projects/octc/), a free open-content resource that combines a collection of monolingual corpora with a parallel corpus component that is created on the basis of subparts of any two or more of the monolingual corpora. The vital parameters of the OCTC that we are going to elaborate on below are the following: it is a freely accessible resource, available from SourceForge.net and protected by the GNU General Public License, its encoding format is TEI XML, and it can only persist and expand as a collective endeavour – and as such, we believe, it has a chance to grow beyond the sum of its parts thanks to the enormous research, collaborative, and educational potential that it has.

Our motivation for starting the project arose partly from frustration with the fact that, while the free software movement has become a thankfully unavoidable part of everyday life, and the open-content movement has already altered the shape of the Net, the two strands are still insufficiently represented in the area of language technology, and especially within the corner of it that deals with the not-too-popular languages. We elaborate on this in section 2 below.

Section 3 looks at the architectural design choices that we have made, planning for the persistence of the OCTC and its interoperability in the area of formats and tools. As already mentioned, the OCTC must be a community effort in order to succeed – this is the issue that we address in section 4. Section 5 motivates our choice of the licensing scheme and highlights a serious issue of two-level licensing that the open-content community should be aware of when creating resources of a similar kind. In section 6, we outline fragments of the research potential of the OCTC, and section 7 wraps up the content of sections 4–6 in a sketch of cost and benefit calculation for an OCTC developer. Section 8 concludes the paper.

2. Not gonna give, not gonna take

The OCTC is our reaction to what we call "data islands" within the linguistic community – separate resources that rarely "talk" to each other and that are either guarded by their creators ("I've sweated over this, so you won't have it the easy way") or by various non-disclosure legacy properties of the data that they include ("there is a noticeable tendency not to transfer linguistic data for fear of breaking the law", as Zimmermann and Lehmberg (2007) note). Some of the resources are placed under a "non-commercial" or "free for research only" restrictions in good faith – because their guardians think that in this manner, their creators' rights will be secured in the best way. Finally, some of them are deemed not worthy of dissemination because they are too small and primitive, being e.g. little lexicons or corpora created as student term projects, or just leftovers from processing larger resources. We thus use the term "data island" as slightly more general than the relatively established term "data silo", which implies a large resource or group thereof. Data silos, even if not easily accessible, can survive for long, while small data islands may easily be lost from sight a few weeks after they have been produced.

While the isolation of data islands has causes in some way intrinsic to them, it may happen that bridges across islands are not built for external reasons due to a version of the "Not-Invented-Here" syndrome, which makes researchers suspicious towards resources non-transparently created by others. Sometimes, this attitude has roots in the expectation that e.g. others' annotations can skew the

interpretation of the primary data, cf. (Thompson, 2010). These are serious issues and show below, in section 3, how the OCTC copes with them.

Rather than build another island – a parallel corpus that we wanted to use for our lexicographic projects – we have decided to plan a versatile resource that would have a chance to become the centre of many research communities and that would have a transparent structure and be easily accessible, for free. In some way, this is a continuation of our quest against data islands initiated in (Bański and Wójtowicz, 2009) within the African Language Technology (AfLaT) community a year ago, with regard to lexical resources.

The project reported here aims at creating a LR centre with a particular attention to minority or "non-commercial" languages (i.e., the vast majority of languages of the world) that are the most exposed to the danger of separating (and eventually sinking) data islands due to the lack of co-ordinated international initiatives, being outside of the sphere of interest of e.g. European or American funding bodies. Access to a common, homogenous platform, and being part of a large project that offers technical backup and peer assistance, might encourage individual researchers working on their own to join, contribute, and, in the long run, to benefit.

3. Architecture and format

The corpus is encoded in TEI P5 XML (TEI Consortium, 2010), one of the prevailing standards for corpus encoding[1], and takes into account practices recommended once by the XML-ised Corpus Encoding Standard (XCES; Ide *et al.*, 2000) and now refined into the emerging ISO TC 37 SC 4 standards (http://www.tc37sc4.org/). In particular, it features stand-off (remote) annotation (Thompson and McKelvie, 1997; Ide and Romary, 2007), with the particular levels of linguistic analysis located in separate annotation layers, usually placed in separate files that may form a hierarchy, whereby one annotation document becomes primary data for another annotation document.

3.1. Stand-off annotation: overview and benefits

The abstract structure of the contents an individual corpus directory is shown below: the headers, containing the metadata, are included by all the other documents: the document containing the primary data (text, lightly marked-up down to the level of paragraph) and each of its annotations, in effect forming what McKelvie *et al.* (1998) call a *hyperdocument*. Each of the documents comprising the hyperdocument is located in a separate file.

The main corpus header and the subcorpus header are located outside of individual corpus subdirectories: at the root of the entire corpus and at the root of the given subcorpus, respectively, but they are included by each corpus file in the same way as the local header is, by means of an XInclude mechanism.

main corpus header (external)
subcorpus header (external)
local header
|------ primary data [single instance]
|------(segmentation)$^{+}$
|------(POS/morphosyntax)$^{+}$
|------(syntax)$^{+}$
|------(discourse)$^{+}$

It can be seen that it is possible for each corpus text to grow layers of annotations both vertically (by adding yet another layer of annotation referring to the existing ones, e.g. discourse annotation referencing the syntactic group annotation layer) or horizontally (by adding additional segmentations, morphosyntactic analyses, POS tagging variants, etc.).[2]

The stand-off annotation approach has numerous advantages described, among others, in (Ide and Romary, 2007; Witt *et al.*, 2009). We have already noted that it allows for incremental building of hyperdocuments, which addresses issues both theoretical and practical. On the theoretical side, stand-off systems allow for numerous and sometimes conflicting analyses of the given text (ranging from segmentation variants through the output of various POS-taggers up to syntactic structures formed according to various theories, etc.). In this way, the motivation for the Non-Invented-Here syndrome cited by Thompson (2010) can be neutralised by allowing researchers to work on the primary data alone and apply their own annotations, without being afraid that the subjective judgements of others will influence their own judgements. On the practical side, the stand-off nature of the OCTC ensures that even a small bit added to the project repository is meaningful: a developer can add a new morphosyntactic layer output by their tagger or just a single, lightly marked-up text, leaving the task of annotating it to others. Researchers decide themselves when and for how long they get involved – in this respect, the project resembles Web 2.0 initiatives, with content created by volunteers.

At the time of submission, the OCTC only has the lowest layer of source text, but we are preparing an example of segmentation, multiple morphosyntactic annotations, and a syntactic chunk annotation in the Swahili part of the corpus, as a demonstration.

3.2. Corpus divisions

The primary division in the OCTC is into the monolingual and the aligned part. Each of them is subdivided into

[1] See http://www.tei-c.org/Activities/Projects/

[2] What the diagram does not show is that each annotation document typically refers to only one other document – either the primary data or, more often, another, lower level, annotation document (and thus, a segmentation document addresses spans of characters in the primary data, but a POS document references segments identified in the segmentation document; a syntactic document typically references groups of elements defined at the POS level, etc.).

subcorpora that are at the same time meant to be subprojects: the monolingual part has one subdirectory for each ISO-639-3-tagged language, and the aligned part has separate directories for pairs (or tuples) of languages.

As shown in the diagram above, each text is located in a separate directory, together with files containing its annotations. Each of these files includes the local header and two other headers: the subcorpus header and the main corpus header. Each header provides a different part of the metadata: the main header provides OCTC-wide characteristics together with the project description and the taxonomy of text features, specific subcorpora, text licenses, etc. The subcorpus header (e.g. the header of OCTC-swh, the monolingual subcorpus of Swahili) provides information specific to the given subcorpus/subproject, including information about the maintainers (we expect that the individual language parts will be maintained by independent teams). Finally, the local header contains the changelog for the individual directory and the metadata of the particular text.

The aligned part contains information on which fragments of which texts from a monolingual subcorpus correspond to similar fragments from (an)other language(s). This is how the OCTC may grow beyond the sum of its parts: each new monolingual text is potentially a part of multilingual alignment information.

At the moment, the corpus contains "seeds" for 55 languages, in the form of the originals and the translations of the Universal Declaration of Human Rights.[3] These texts can at the same time serve as seeds for the aligned part of the OCTC. The present authors concentrate on the Swahili monolingual subcorpus and plan to gradually develop a Polish-Swahili aligned subcorpus.

The project repository (under the control of Subversion) is structured in such a manner that it is easy to check out a single subcorpus together with the trunk of the corpus (header, schemas, XML catalog, tools). This way, each team needs to store locally only a single subcorpus (or only parts thereof).

3.3. Encoding format

The encoding format of the OCTC is TEI P5 XML (TEI Consortium, 2010). The choice of the TEI as the format for linguistically annotated multi-layer corpora against other standards and best practices for text encoding is motivated in (Przepiórkowski and Bański, 2010). The particular customisation of this format elaborates on that of the National Corpus of Polish (cf. Bański and Przepiórkowski, this volume; Przepiórkowski and Bański, forthcoming). We adduce selected XML examples of the OCTC in Bański and Wójtowicz (2010), limiting ourselves here to pointing out that the TEI is an well-known open standard, with 20 years of presence in the HLT, designed for interchange and persistence. Like in typical XML languages, the TEI tags are largely self-descriptive, but this is enhanced by the fact that the individual TEI customisations interweave documentation and schema-building instructions in a literate programming manner, thanks to the mechanism of the so-called ODD configuration files (Burnard and Rahtz, 2004). ODD files (written in the TEI itself) can be turned into a variety of XML schema formats or a uniformly structured and rendered set of guidelines for use of the particular customisation, easy to translate into any language, if such a need arises.

4. Community-related issues

Planning a sustainable free resource of the kind presented here needs to involve planning of the community that will create, maintain and develop it for years to come. On the one hand, the sub-projects should be self-sustaining, on the other – they should be aware of belonging to a single community, grouping people with similar interests and often ready to co-operate beyond the limits of their individual languages.

4.1. Local OCTC community

The OCTC can only grow as a community effort, and a large part of that community should include creators of data islands, now hopefully no longer separated. Thanks to its being located at SourceForge, the OCTC has all the community support that this fantastic platform offers: from an indefinite number of possible mailing lists (e.g., a separate list for each subproject) through a bulletin board that can be extended according to need, to a well-known and tested MediaWiki-based documentation system. A separate, vital issue is the version control system that is hooked to a separate mailing list, in this way facilitating peer review of source code and content. The present authors hope to act primarily as facilitators and coordinators for subprojects that should ideally develop on their own, within the bounds of a common framework. What is important to us is giving the participants a sense of belonging to a project that does some good – for them, for their language, and possibly even for the open-content world in general.

It needs to be stressed that the OCTC is not only a place for creating content – tools are also going to play an important role in it, especially those that can operate on more than a single subproject, i.e. tools that can be taught or configured to handle multiple languages (cf. section 6).

4.2. Reaching out

We recognize the need to include LR-related issues, and especially open-content movement awareness, in academic curricula. The individual subcorpora of the OCTC can become the foci of student projects – we hope that the OCTC may be one of the platforms to be used for that goal – an analogous precedent is the student-project system built into the instance of the Bugzilla system that

[3] http://www.ohchr.org/EN/UDHR/Pages/60UDHRIntroduction.on.aspx

serves mozilla.org projects[4].

It is also important for us to reach out to under-funded research communities targeting languages commonly referred to as "non-commercial" or "under-resourced", "lower-density", "non-central" and by many other names (cf. Forcada, 2006; Streiter *et al.*, 2006), depending on where the focus is placed. Being close to the (Sub-Saharan) African linguistic community, we can observe proposals after proposals bouncing off the European Union's financing institutions because there is too little interest in the creation of systems linking European languages and African languages, even those that are as "commercial" as Swahili, with its estimated up to 80 million speakers. Such research communities have to cope by themselves, and a project such as the OCTC is a way to supply data to them, but also to provide a target for efforts that would otherwise result in the creation of data islands. Planting corpus "seeds", we count on it being much easier for outside researchers to decide to add something to an existing resource rather than build their own from scratch.

In general, the OCTC is intended to play a part in the development of LR-sharing culture. We are certainly not the precursors in this area – other community-based efforts include Kamusi (a collectively-built Swahili dictionary, edited by Martin Benjamin, http://kamusiproject.org/), the well-known Wikimedia projects (Wikipedia, Wiktionary, etc.) or, the closest in spirit, the Crúbadán [5] multi-language Web-as-Corpus project led by Kevin Scannell (Scannell, 2007). An open-source initiative that can serve as an example of how multiple developers can work towards a common goal within a single SourceForge-based project is Apertium (Tyers *et al.*, 2010, http://apertium.org/), producing high-quality Machine Translation systems. A seasoned project that groups researchers producing free annotations for the closed-content American National Corpus is the Open Linguistic Infrastructure (Ide and Suderman, 2006), and one that operates on a much broader scale to store sustainable resources is the OLAC (Open Language Archives Community, (Simons and Bird, 2008)).

It is necessary to make the project members aware that their work must be usable to others, just like the work of others is usable to them. This means, for example, no "magic numbers" in the code, well-commented "kludges" (so that the peers may turn them into well-behaved code), and clear metadata. It is natural to expect some developers to try and avoid "wasting time" on obeying such procedures, and this is where the community has to exert its pressure, by maintaining and requiring quality standards – in the OCTC there is no "their data/tool, which is their problem": the resource in question may become the basis of your or your students' future project, so it is everyone's responsibility to maintain its high standard.

[4] https://wiki.mozilla.org/Education/Projects/MozillaGuidelines
[5] http://borel.slu.edu/crubadan/

5. Licensing issues[6]

The benefit of open licenses for language technology resources has been stressed in numerous publications and its advantages are especially well visible for "non-commercial" languages (cf. Koster and Gradmann, 2004; Pedersen, 2008; Streiter and de Luca, 2003; Streiter *et al.*, 2006). As mentioned in section 2, problems concerning licensing of data very often lead to creating data islands that cannot be distributed even if their guardians would wish otherwise (this is well put by Kilgariff (2001) in the context of the Web-as-Corpus initiative: "the use of the web addresses the hobgoblin of corpus builders: copyright") . Various ways of handling this problem have been implemented in various projects, e.g. providing a corpus of URL links (Sharoff, 2006) or distributing masked corpus data (Rehm *et al.*, 2007). The aim of the OCTC is to handle such problems by accepting only open-content data, using only open-source tools, and by placing all of its resources under a well-tested and popular open-source license: GPLv3.

Our initial plan for the licensing scheme of the OCTC was to make it available under the GNU Lesser General Public License (http://www.gnu.org/licenses/lgpl.html), the so-called "library" license, that allows the resources under it to be made part of closed-content systems. There were two reasons for this: to reach as many end-users as possible, including corporate users, such as publishing houses or Language Technology companies, and not to alienate closed-data corpus-holders by appearing to be a possible threat to them, and to count on their donations of data in return for using the OCTC.

However, the LGPL licensing scheme, while concentrating on the (possibly corporate) end-users, would not have the power to break some of the stereotypes that are responsible for the creation of data islands: those researchers who guard their resources for whatever reason will not get any incentive to add these resources to the OCTC if what they can expect is some company taking their work and making it part of their closed system. The OCTC can only function if it gathers a community, and the community should have a motive for donating their data, tools, time and expertise. Feeding large companies with the fruit of one's work for nothing is not the kind of incentive that would guarantee the growth of the OCTC community. However, a guarantee that whatever its members donate for free will remain free and will possibly come back to them after it has been enlarged and elaborated on, may be a convincing argument. Therefore, we have decided to place the entire corpus under the GNU General Public License (version 3 or any later version, cf. http://www.gnu.org/licenses/gpl.html). This license has the consequence that whatever derived work created on the basis of the OCTC is distributed, it

[6] The decisions reported on in this section arose from personal communication with Kevin Donnelly, whom we thank for taking the time to review the consequences of adopting either LGPL or GPL for the OCTC.

has to be distributed under the terms of the GPL, thereby guaranteeing that what is invested in the OCTC can only grow and evolve, remaining free for all.

It would seem that finding open-content data may be a difficult task, but we are pleasantly surprised that, currently, the amount of data that we are able to get from various projects will keep us busy encoding the Swahili corpus alone for quite a while. Besides, we count on being able to include, with publishers' permission, some closed-content resources, after they are sampled and possibly reshuffled on a paragraph- or even on a sentence-basis. It will be easy to mark such texts as forming a separate category of "jumbled-up" texts that will not be used for testing e.g. anaphora resolution in discourse, but will still provide for frequency and n-gram counts, as well as numerous other NPL/CL methods of research on texts.

Another issue concerns the double layer of licensing that is involved in creating a large resource that includes other resources as its proper subparts: we have to expect the licenses of the data included in the corpus to be varied. A special taxonomy is created in the main corpus header to keep track of the original permissions, and we plan to create release forms that will both state the original licensing of the text or tool added to the OCTC, and the donor's consent that their resource is going to become part of a GPL-ed collection. While there are various ways in which free/open content licenses may conflict with one another,[7] the fact that most texts will be naturally sampled or simplified on entering the OCTC (by stripping them of e.g. graphics and formatting), and then turning them into parts of an XML application, should contribute to the ease of relicensing, although we note this as a problem that needs to be addressed early.

6. Research potential

We have already hinted at the OCTC's research potential. The fact that a single project, with a single uniform encoding format and a centralized tool repository, involves data from numerous languages, creates all sorts of possibilities for embarking on cross-language study, from language-independent algorithm testing and tool creation (cf. e.g. (DePauw and De Schryver, 2009)) to the creation of parallel subcorpora with MT or lexicographic applications – such applications may in fact link the OCTC to the aforementioned Apertium (for MT) or the FreeDict project (http://freedict.org/, for bilingual dictionary creation).[8]

Stand-off annotations can be studied as separate objects. They may be used e.g. for measurements concerning the efficiency of various taggers or tagsets. In order to compare linguistic descriptions, there is a need for an ontology relating the content of annotations, both horizontally within a single hyperdocument (e.g. to express the equivalence of symbols used in various POS-tagging schemes applied to a single text), as well as cross-linguistically, to relate categories used for annotating different languages. This where e.g. the ISO Data Category Registry (Kemps-Snijders *et al.*, 2008) and the GOLD ontology (Farrar and Langendoen, 2003) can be subjected to large-scale tests.

We only sketch the potential for research here, because this sketch is enough to hint at the multitude of possible applications. It is worth stressing that the OCTC makes it possible for a single researcher to create something they would usually have a serious problem creating otherwise: a fragment of a professional-quality corpus, whether monolingual or parallel, with the awareness that after their research is done, their creation will not vanish with their old computer but will remain accessible to others and to them as well, when they decide to re-run some old measurements or just to expand the resource further.

Although the OCTC is by its nature dynamic, it is placed under version control, so that all measurements can be objectivised and made reproducible (anchored to a particular release or even revision). In this way, the OCTC may be used as a common testing ground for tool creators, and for reporting results in professional journals, cf. (Pedersen, 2008).

7. Costs and benefits

The cost of the OCTC will be nothing to some, but time, expertise and/or data to others. Our point is that by putting one's time, skill or data into the OCTC, the researcher in question may in fact make a very beneficial investment: they will get what they could perhaps get on their own computer, except that (i) SourceForge will not fry the way an individual hard drive can, ruining the fruit of several months of sometimes irrecoverable research, (ii) the code or data encoding will be subject to peer review and thus to possible corrections (which in themselves can save enormous quantities of time sometimes), (iii) researchers will be able to advertise their skills through their results, and finally, (iv) the collective nature of the OCTC means that when two researchers each contribute a small amount of their resources, what they get back for their research may be double that amount.

It has to be stressed that the role of SourceForge both as a dissemination platform with multiple mirrors, site-wide backups, robust choice of version control and bug-tracking systems, and as a community centre with all the co-operation-enhancing facilities cannot be underestimated: it is hard to imagine an enterprise such as the OCTC having a chance to start and persevere outside of SourceForge. The costs in terms of infrastructure and maintenance alone would be prohibitive. SourceForge has all the advantages that Streiter *et al.* (2006) and Simons and Bird (2008) point to when talking about sustainability conditions: data is safe due to several mirrors worldwide and constantly accessible with no restrictions, to whoever needs it; it is also disseminated by various downstream

[7] See for example the GNU licenses compatibility page at http://www.gnu.org/licenses/license-list.html.
[8] On interoperability between FreeDict and the OCTC, see also Bański (2010).

projects as e.g. Debian packages.

Finally, there is one more cost-related issue that is most relevant in the kind of environment described here: communicating one's results and finding opportunities for joint incentives: it will be much easier to find a co-author for a conference paper or a partner for a new project among the participants in this kind of collective enterprise, already sharing a common ground: using the same, homogenous encoding format, the same repository, and being able to verify their trust by looking at each other's activity in the project.

8. Conclusion

Data islands sink. This is especially painful in the case of minority languages, where such losses may well turn out to be irreparable. The OCTC is meant to prevent that from happening and offer much more: a way to enhance the individual developer's data and put it to new uses, all for free and with a guarantee of freedom.

The guarantee of freedom and the promise of enhancement is our way of coping with the two "you're not gonna have it" attitude mentioned at the beginning. We cope with the NIH reflex by ensuring fully transparent and version-controlled build process, and by using stand-off annotation, which easily accommodates conflicting analyses.

Free availability need not come at the cost of usability. The OCTC uses an open encoding standard and exploits best practices for the creation of a versatile and at the same time sustainable resource. Sustainability is a general issue, but it is all the more crucial for languages for which electronic data are still scarce. The way the data are exposed by the self-describing XML enriched with metadata from the headers will make the resource usable for many purposes, both mono- and multilingual.

The project assumes scalability along many axes, and thus it does not impose time restraints on the teams or individual researchers working on it. Its usability grows by a small step with each text added to it, with each annotation layer created (whether vertical or horizontal), and with each new alignment document.

Finally, it is worth realising that the present paper does not necessarily concern abstract situations and foreign researchers in faraway countries. It may just as well concern your data and yourself.

9. Acknowledgements

We wish to thank the two anonymous LRSLM2010 reviewers for their remarks.

We are grateful to Kevin Donnelly for his discussion of licensing issues and helpful hints, and to Kevin Scannell for kindly sharing with us information and references to the relevant literature.

We are also grateful to SourceForge, for creating environment conducive to collective research that leads to definitely positive non-zero-sum solutions.

References

Bański, P. (2010). Setting data free – on two open-content, data-sharing, TEI-related projects. In the proceedings of the LREC-2010 CLARIN workshop on "Language Resource and Language Technology Standards – state of the art, emerging needs, and future developments".

Bański, P., Przepiórkowski, A. (this volume). The TEI and the NCP: the model and its application. In the proceedings of the LREC workshop on "Language Resources: From Storyboard to Sustainability and LR Lifecycle Management" (LRSLM2010).

Bański, P., Wójtowicz, B. (2009). 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. (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

De Pauw, G., de Schryver, G.-M. (2009). African Language Technology: The Data-Driven Perspective. In V. Lyding (Ed.) *LULCL II 2008 – Proceedings of the Second Colloquium on Lesser Used Languages and Computer Linguistics*, Bozen-Bolzano, 13th-14th November 2008 (EURAC book 54), Bozen-Bolzano: European Academy, pp. 79--96.

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

Forcada, M., (2006). Open source machine translation: an opportunity for minor languages. In B. Williams (Ed.) *Proceedings of the Workshop "Strategies for developing machine translation for minority languages", LREC'06*. Genoa, Italy, pp. 1--6.

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

Ide, N., Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy.

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

Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P.,

Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In European Language Resources Association (ELRA) (ed), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30, 2008.

Kilgarriff, A., (2001). The web as corpus. In P. Rayson et al. (Eds.), *Proceedings of Corpus Linguistics Conference 2001*. Lancaster. (Available from http://www.itri.bton.ac.uk/techreports/ITRI–01–14.abs.html)

Koster, C.H.A., Gradmann, S. (2004). 'The language belongs to the People!'. In *Proceedings of LREC'04*. Lisbon, Portugal.

McKelvie, D., Brew, Ch., Thompson, H. (1998). Using SGML as a basis for Data-Intensive NLP. *Computers and the Humanities*, 31 (5): pp. 367--388.

Pedersen, T. (2008). Empiricism is not a matter of faith. *Computational Linguistics* 34 (3), pp. 465--470.

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. Goźdź-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.

Rehm, G., Witt, A., Zinsmeister, H., Dellert, J. (2007). Corpus Masking: Legally Bypassing Licensing Restrictions for the Free Distribution of Text Collections. In *Proceedings of Digital Humanities 2007*, June 2–8, University of Illinois, Urbana-Champaign, USA, pp. 6--9.

Scannell, K.P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In Fairon et al. (Eds.) *Building and Exploring Web Corpora*. Louvain-la-Neuve: PUL, pp. 5--15.

Sharoff, S. (2006). Open-source corpora : Using the net to fish for linguistic data. *International journal of corpus linguistics* 11(4), pp. 435--462.

Simons, G.F., Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation: PACLIC 22*. pp. 87--100.

Streiter, O., De Luca, E.W. (2003) Example-based NLP for Minority Languages: Tasks, Resources and Tools. In O. Streiter (ed.): *Proceedings of the Workshop "Traitement automatique des langues minoritaires et des petites langues", 10e conference TALN*. Batz-sur-Mer, France, pp. 233--242.

Streiter, O., Scannell, K. P., Stuflesser, M. (2006). Implementing NLP Projects for Non-Central Languages: Instructions for Funding Bodies, Strategies for Developers. *Machine Translation* 20 (4), pp. 267--289.

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/

Thompson, H.S. (2010). Standards for language resources: what's the right approach?. 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. 153--155.

Thompson, H.S., McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents, Proceedings of SGML Europe. Available from http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Tyers, F.M., Sánchez-Martínez, F., Ortiz-Rojas, S., Forcada, M.L. (2010). Free/open-source resources in the Apertium platform for machine translation research and development. *The Prague Bulletin of Mathematical Linguistics* 93, pp. 67--76.

Witt, A., Heid, U., Sasaki, F., Sérasset, G. (2009). Multilingual language resources and interoperability. In *Language Resources and Evaluation* 43 (1), pp. 1--14. doi:10.1007/s10579-009-9088-x

Zimmermann, F., Lehmberg, T. (2007). Language Corpora – Copyright – Data Protection: The Legal Point of View. In *Proceedings of Digital Humanities 2007*, June 2–8, University of Illinois, Urbana-Champaign, USA, pp. 2--4.

Very large language resources? At our finger!

Dan Cristea
"Alexandru Ioan Cuza" University of Iasi
16, Berthelot St., 700483 – Iasi, Romania
E-mail: dcristea@info.uaic.ro

Abstract

The paper proposes a legislative initiative for acquiring large scale language resources. It militates for raising a large awareness campaign that would allow the storing and preservation for research purpose, in electronic form, of all textual documents which go to print in a country.

1. Introduction

This paper brings into attention a proposal for conserving over long time and using largely, at a national level, for research purposes, the linguistic data which are printed and distributed for public use daily by editorial houses.

It is evident that, without a continuous effort, those languages which are now called "less-resourced" will continue to be viewed like that even when, hypothetically, they will promote to the same amount of resources as the languages that at this very moment are known to be most resourced. Moreover, if the most resourced languages would cease to acquire resources now, on the ground that they have fulfilled their needs, in short time they will lose their leading positions. This is because LRs become obsolete very quickly. Even more, if we look at the annotated resources, the linguistic facts which are subject to automatic annotation could change over time, as the linguistic theories on which the marking conventions are based evolve, and as the automatic annotation processes themselves get improved. So, as the language goes along and evolves and our vision with respect to the language changes, the resources, themselves, get old. There is no end in building LRs.

In many countries a "legal deposit" law is in use. It obliges all providers of printing materials (editing houses, physical or juridical persons which print documents for public, recording houses and studios, the National Bank, the State Mint, the National Post, etc.) – let's call them *resourcers* – to send a number of copies of each printed item intended for distribution to a national library (which could be one physical unit or a consortium of libraries) for long-time preservation. Although the horizon of media production changed dramatically in the last years, to my knowledge, there are only very timid trials for improvement of the juridical aspects.

As resources are needed dramatically and many of them are very expensive, the issue of acquiring them should stop from being accidental or episodic and should become a national policy. Something should be done. A law should defend the linguistic resources of the languages spoken in a country as being of primary interest. This paper discusses one possible solution which, although not simple to implement, could change completely the LRs scene in the near future.

2. Enhancing the legislation on legal deposits

A recent investigation among some of the most important producers of printed information in Romania revealed that many editing houses are keen to donate their resources for research purposes. However, another fraction, which unfortunately makes the majority, is not interested to collaborate. They ignore the importance of the issue, are fearful that donating their data is equivalent to loosing the property control over them, will possibly trigger a loss of profit, or simply do not have time to dedicate to this kind of matters.

In reality, nothing of the kind has to happen. Although we need their linguistic data, we do not want the *resourcers* to be harmed if they give their data to science. The idea is to promote a legislative initiative that imposes the compulsoriness for the *resourcers* to donate their linguistic data for language research. The proper moment has come to try to raise the awareness for a concentrated action in Europe. We need to raise governmental interest towards the promotion of such legislation, simultaneously in many countries.

The following type of resources, produced in series, would be in focus to such a law, irrespective whether the resources are intended for commercial or for free distribution: books, booklets, leaflets, journals, magazines, almanacs, calendars, musical scores, propagandistic materials having a political, administrative, cultural, artistic, scientific, educational, religious, a.s.o. goal, posters, proclamations, any other materials intended for publication on public places, Ph.D. thesis, university courses, documents in electronic format containing linguistic material (CDs, DVDs, etc.), standards and technical norms, publications issued by national and local authorities, collections of norms and laws, any other printed or multiplied material by using graphical or physical-chemical methods.

On the practical level, the initiative presupposes the existence of a national repository, which is an entity (IT center, institute, etc. – let's call it the *Portal*), which, on one hand, has the legal authority to receive and store data contributed by *resourcers*, and, on the other hand, is technically equipped to collect and record, indefinitely

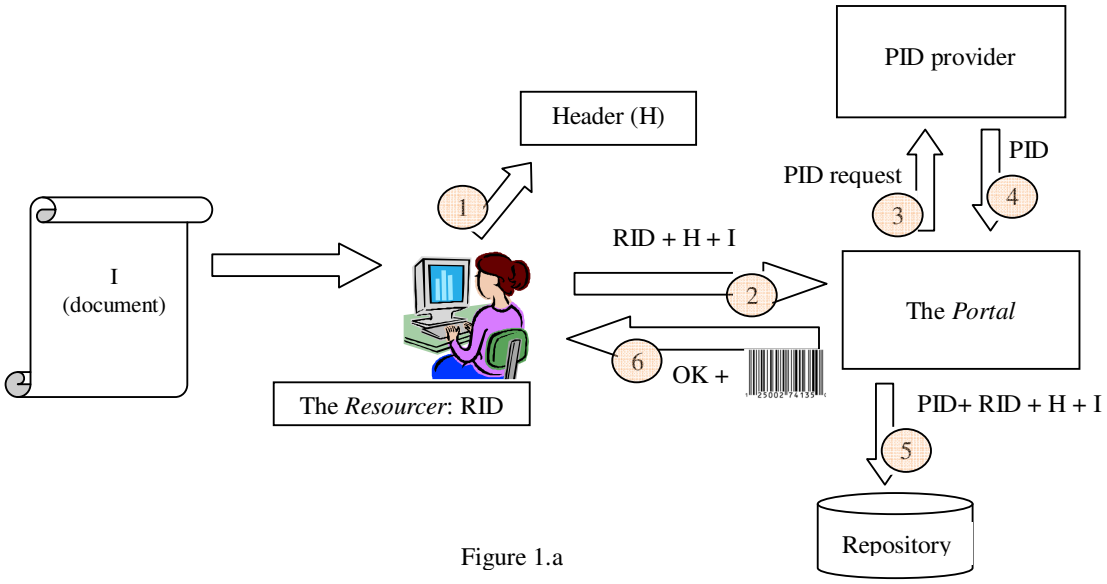


Figure 1.a

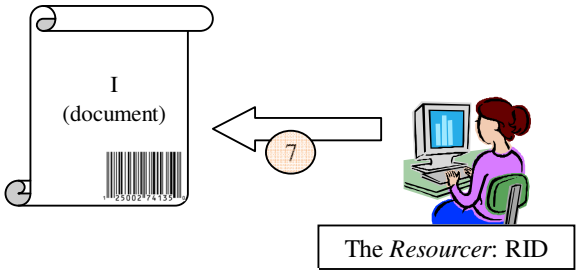


Figure 1.b

long, in electronic format, all data issued for publication, daily, in a country.

The law should state that by sending an electronic copy for long-time preservation to this national repository no authoring rights or commercial benefits are lost by the *Resourcer*. The copy can be used, intermediated by the *Portal*, only for research purposes applied to language and the *Portal* cannot make public the data on internet or on other media, unless it is asked to do so by the owner. It is clear that a fragile IPR chapter will not be acceptable in the text of this law (COM, 2009). A weak statement of IT security measures to protect the authors' rights will also be amendable. All these aspects are very important and should receive full attention in the formulation of this law.

3. The capturing flow

I see the *Portal* as a factory that processes words. The start elements of the data flow should be as follow: before issuing the first publication, or at the moment the law is imposed, the *Resourcer* should have got an identification code (*RID*) from the *Portal*. It will use this code for communication with the *Portal* regarding any publication, during all its juridical lifetime.

Suppose that today the *Resourcer* prepares for publication a new item *I*, which has got the "ready for printing" editorial approval. The Figure 1.a explains the communication initiated by this new item. The *Resourcer* fills in an electronic form (header – *H*), containing identification information of the document, and then interacts with the *Portal*, uploading its *RID*, the header *H* and an editable copy of *I*. The *Portal* receives this data and asks for a persistent identification code (*PID*) to an authority capable of issuing them (Kunze and Rogers, 2003; Schwardmann, 2009). When it gets one, it stores in its repository a bunch of data containing *PID*, *RID*, *H*, and *I*. Then, the *Portal* returns to the *Resourcer* an OK message, containing two parts: a human readable part and a bar code part. The OK box should record a seal of the *Portal*, together with the *PID* and the *RID*. Now the document, which has also this OK box, included on an inner cover or on a sleeve, can be printed (Figure 1.b). This box proves to any authority in charge of controlling the application of the law that the legal deposit was performed by the *Resourcer* on the *Portal*, and all the needed identification information is there.

The above detailed exchange of data between the *Resourcer* and the *Portal*, including also a communication with a third entity responsible for issuing PIDs, seems heavy and time consuming, and if so, totally unacceptable by the editing houses. Indeed, it is a known fact that these entities are most of the time constrained to process data at

great speed, especially, if they print daily newspapers, for example. Nevertheless, the communication which, as is described above, appears to be heavy and cumbersome, can be done as quickly as a blink of an eye by making completely automatic the whole chain, including the fill-in of identification information contained in the header *H*. The content of the header can be extracted by specialized modules from the electronic item *I*. So, practically, the entire chain could be activated by a click on a button on the editing interface. This should end up, almost instantly, with the inclusion of the OK box in a dedicated place of the document going to print.

4. Data processing

Once captured, data on the *Portal* should be processed. In this section I describe a list of processing capabilities that the *Portal* should be able to provide.

First, it is obvious that the Portal should have sufficient storing capacities and that these capacities should be specially designed for preserving data indefinitely long periods of time. Then if should display indexing, search and retrieval capacities, at different levels: header, lexical tokens (words), lexical expressions, as well as contextual information. This means that each document, once placed on the portal, should be submitted to a processing chain that includes, minimally: tokenization, part-of-speech tagging, lemmatization and indexing. It is foreseeable therefore that each document will be recorded as raw text on which the standoff XML annotation will make reference. The XML annotation and the indexing requirements will most probably multiply the size of the initial text documents a couple of times.

Based on these basic functionalities, a different line of processing refers to lexicographic needs. The Portal should be able to perform complex operations such as: detection of foreign words, signaling of new words, recognition of senses of words in context (WSD), detection of new senses, signaling of forgotten (obsolete) words, signaling of senses which are no more used, etc. For instance, signaling of new words and of forgotten (obsolete) words should be triggered by a frequency of occurrence which, over a given interval of time, is above/below certain thresholds, as decided by a linguistic authority. Similarly, signaling of a new sense could be triggered by the fail to align the sense recognized in context to those kept in a repository of senses, like for instance an authoritarian explanatory dictionary, if this happens with a certain frequency recently, and if the pattern of use is sufficiently stable. Forgotten (obsolete) senses are recognized by the occurrence of these senses under a certain threshold.

The process which should be placed at the base of recognizing obsolete words or senses presupposes placing a bag of words under constant surveillance. These are words/senses plausible of becoming under-used because they experience a constantly degrading frequency. Let's note that the criterion of absolute or even relative frequency, over a certain interval, could prove not being relevant, because there are words which are very rarely used, although they could not be in danger of being considered extinguishable (some science neologisms, for instance). The best way to do this is to associate to each word a personal file, recording a set of dynamic features, among which the frequency of occurrence over time (a graphic, from which a gradient of deterioration could be computed), the list of registers that use it (with the associated relative frequencies), etc. So, the problem resides in computing the frequency over a constant interval of time, considered always back from the current day. One could do this by simply searching the spotted word in the repository and counting only the occurrences that fall in the needed interval – a function that would be called only once in a certain long interval – say two to five years (because one cannot expect that the tagging "obsolete" can be updated too frequently, from yesterday to today…).

It is clear that any decision on anyone of these positions should ultimately be taken by a linguistic authority (Academia). Their decisions should investigate the signals transmitted by the Portal, which are rooted on neat statistical evidence.

Different processing flows could implement other functions. A number of resources, which are of increasing importance in keeping a language technologically updated, can be continuously connected onto the Portal. Among these, I see: the main Dictionary of the language, the WordNet (Fellbaum, 1998), the VerbNet (Kipper et al., 2008), the FrameNet (Fillmore, 1976; Atkins et al, 2003) – to name just a few. Supposing all these resources are complete for the language L, at a certain moment, they should be kept updated with the evolution of language. So, any dynamics in language should be mirrored in these resources as well. If, as suggested above, each lexical item of the language has a personal record on the Portal, then if should include references in all these resources. As such, the word *w* is linked to its input in the Dictionary, where the inventory of senses is recorded, and these senses are aligned to those listed in the WordNet for this lexical item, as well to its entry in VerbNet and FrameNet. All these resources are connected among them and kept online with the evolution of language by the Portal.

The Portal can host also a number of services addressed to the *resourcers*, to the language researchers, to the consumers or to the public at large. Public services could be charged to the customers and benefits be returned to the *resourcers*, in amounts proportional to their monthly contribution on the Portal (measured in characters).

Other types of paid services could be imagined, with benefits returned to the *resourcers*, for instance advertising publications and on-line access to parts of their publications, which they are keen to offer on the market. The possibility to develop a set of services from which the *resourcers* could obtain profit is interesting also from the point of view of potentially lowering the

resources' opposition vis-a-vis of a law that would impose the obligation of continuous language preservation, as has been discussed in section 2.

5. Evaluation

It is clear that the type of processing encumbered by such an initiative would bring to the *Portal* a very big amount of linguistic data daily. A rough evaluation of the processing needs and costs encumbered by such a national-wide enterprise should bring into focus parameters such as: the number of editorial houses registered, the average number of publications of a publishing house per year, the average length in pages of a printed item, the average number of characters per page. Leaving aside episodic publications of small size, our enquiry about the average amount of data published in books and journals, in a medium size country of Europe (Romania), at the level of the year 2008, has yielded an amount of textual data which is less than 1Gb daily.

A channel with a bandwidth of 12.5 Mb/sec can lightly face the required transfer described in section 3, avoiding bottlenecks on moments of crowd. Load balancing and mirroring, for safety reasons, should be assured, by storing the data on at least two centers, in different locations. As proved already by data intensive storing houses (Google [1], for instance), software RAID technology, made up of a farm of small computers, is a cheap and appropriate solution for long time preservation and a comfortable processing speed.

6. Conclusions

The advantages of a *Portal* able to process linguistic data at a scale as the one envisioned above are hard to depict now correctly. First of all, it will give a long-time and complete solution to the problem of linguistic data preservation for the language(s) of a nation, as well as an almost complete radiography of its diachronic evolution. Secondly, it will put the basis for an exhaustive research related to language. Thirdly, it could bring into focus a large scale of commercially appealing applications, in the benefit of the authors of the texts or the *resourcers*.

The success of such an initiative at national level depends very much on a large concentrated vision. The new and very fresh breath that is being felt at this moment in Europe with respect to building language processing infrastructures, to establish standards for representation of linguistic data, and to foster large scale initiatives for the acquisition of linguistic resources, as motored by recent consortiums like CLARIN[2], FlareNet[3], T4Me[4], Meta-Net[5], etc. should also move forward a favorable legislation. The proposal advanced in this paper is also in line with other initiatives that try to raise the awareness on the necessity of free access to science[6]. It, however, does not advocate against intellectual property (Stephan, 2001), but is very much in favor of a reconsideration of the IPR legislation, which is too restrictive in many cases of usage of language resources for research. After all, our language, as we use it today, represents a collective contribution and is due to a perpetual reshaping from all its speakers from the beginning of the time… Donating his linguistic creation for language preservation and research, while not harming at all its creator, neither intellectually, nor commercially, represents just the minimum return that an author which uses the language owes to those who have invented it, for the benefit of those which will use it in the future.

[1] http://infolab.stanford.edu/~backrub/google.html
[2] www.clarin.eu
[3] http://www.flarenet.eu/
[4] http://t4me.dfki.de/
[5] http://www.meta-net.eu/

7. References

Atkins, S., Rundell, M. and Sato, H. (2003) The Contribution of Framenet to Practical Lexicography, *International Journal of Lexicography*, Volume 16.3: 333-357.

COM (2009) 532 – Communication from the Commission. Copyright in the Knowledge Economy.

Fillmore, C. J. (1976): Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, Volume 280: 20-32.

Kipper,K., Korhonen, A., Ryant, N., Palmer, M. (2008) A Large-scale Classification of English Verbs, *Language Resources and Evaluation Journal*, 42(1), pp. 21-40, Springer Netherland.

Kunze, J. and R.P.C.Rogers (2003). The ARK Persistent Identifier Scheme. Internet draft at http://www.cdlib.org/inside/diglib/ark/arkspec.pdf

Schwardmann, U. (2009). PID System for eResearch. EPIC – the European Persistant Identifier Consortium, personal communication at NEERI-09, Helsinki.

Stephan, K. "Against Intellectual Property". *Journal of Libertarian Studies* 15.2 (Spring 2001): 1-53.

Fellbaum, C. (1998) WordNet: An Electronic Lexical Database, MIT Press.

[6] See, for instance, the Washington D.C. Principles For Free Access to Science at http://www.dcprinciples.org/statement.pdf, the Open Access initiative http://www.eprints.org/openaccess/, the American Scientist Open Access Forum http://amsci-forum.amsci.org/archives/American-Scientist-Open-Access-Forum.html, The SPARC Open Access Newsletter (see an issue at http://www.earlham.edu/~peters/fos/newsletter/01-02-10.htm), the Budapest Open Access Initiative http://www.soros.org/openaccess,

Standardization as a means to Sustainability

Michael Maxwell
CASL, University of Maryland
7005 52nd Ave., College Park MD 20742 USA
mmaxwell@casl.umd.edu

Abstract

Language resources are expensive. From this simple statement follow two implications: (1) We should try to make resource construction cheaper; and (2) we should try to make resources longer-lasting. This paper deals with the latter implication, in particular for morphological parsers. Our solution to this life cycle problem is to construct grammars, not parsers; if this is done appropriately, a grammar can be straightforwardly converted into a parser. We have developed a linguistic formalism for such morphological descriptions, a formalism which is designed to be adequate for the wide variety of languages known to linguists. We have also developed a program to convert grammars written in this formalism into the programming language of the Stuttgart Finite State Transducer; the program is designed to be interoperable with other morphological parsing engines. Finally, we have tested this system by building grammars and parsers of several natural languages, with good results. We intentionally started with languages that have comparatively simple morphologies (Bengali, Urdu and Pashto), but the method is clearly scalable. The formal grammars are embedded in descriptive grammars, thereby making them maintainable, not to mention useful for such additional goals as language documentation.

1. Introduction

The fundamental problem with most language resources is that they are expensive. This is clearly true of primary (hand-crafted) resources, such as dictionaries. It is less true of secondary resources, because while the secondary resources (such as parsers) are produced from expensive primary resources (such as treebanks), the secondary resources can be re-built at any time from the primary resources at low cost.

Given that resources, especially primary resources, are expensive, once can draw two conclusions:

1. It is desirable to reduce the cost of resource building.
2. It is desirable to increase the usable lifetime of the expensive resources.

This paper discusses the latter point. Our team at the Center for Advanced Study of Language at the University of Maryland has developed a method to increase the longevity of language-specific resources by enabling the automatic derivation of a potentially short-lived secondary resource—morphological parsers—from a long-lived primary resource—morphological grammars. This is a departure from the standard practice, namely building morphological parsers by hand.[1]

Morphological parsers actually require three resources: a parsing engine, a morphological and phonological grammar in the parsing engine's programming language, and a lexicon in the parsing engine's expected format. Until now, because the lexicon and especially the grammar must be written in the format required by the parsing engine, the lifetime of a parser has been driven by the lifetime of the associated engine. Over the past several decades, the lifetime of a morphological parsing engine for practical uses has been no more than a decade or two.[2]

When a parsing engine becomes obsolete, generally its programming language becomes obsolete with it. One of the earlier morphological parsers, AMPLE (Weber, Black, and McConnel 1988) allowed affixes to be defined using an item-and-arrangement descriptive format. Later parsers following the KIMMO model (such as PC-KIMMO, Antworth 1990; see also Sproat 1992) used two-level phonological rules. A more recent parser, xfst (Beesley and Karttunen 2003) added sequentially applied phonological rules, of the kind familiar to most practicing linguists. A grammar written in any of these parsers' programming languages is useless in the other parsers.

Such changes are only natural; the main motivation for implementing a new parsing engine is probably to make it easier to write parsers for natural languages. We therefore expect this short life-cycle to continue. For example, current parsing engines, which are typically finite state transducers (FSTs), make it difficult to work with morphosyntactic features (features like tense, aspect, and number), particularly where those features may have internal structure (such as person and number agreement on verbs; see Copestake 2002 for an implementation of such features in syntax). A logical next generation parsing engine would incorporate an improved mechanism for dealing with such features. Whatever mechanism this is, it seems unlikely that its programming interface will resemble the way such features are treated in current FSTs.

Thus, if the lexicon and grammar used with a parsing engine are treated as primary resources—that is, if they are written specifically for a particular parsing engine—they become obsolete when the engine becomes obsolete. Fortunately, from the standpoint of a morphological parser, lexicons are relatively simple, because lexical complexity tends to be in the semantic information words bear, which is largely irrelevant to parsing.

[1] It is also possible to create morphological parsers as secondary resources derived from annotated text or un-annotated text. At present, annotated morphologically parsed text is very expensive, and we view building practical parsers from unannotated text as more of a research problem than an engineering method.

[2] Longer lifetimes can be achieved in the context of hardware and operating system emulation. But this is unlikely to be the best method of extending lifetime for practical use: emulators are probably more useful for software museums than for mission critical programs.

(However, if inflection classes and irregular forms are not marked in a way conducive to importing this information into the parser, the lexicon format can also be an issue.)

One might hope that morphological grammars are also easy to produce. This turns out not to be true for two reasons. The first reason is that for languages with any significant morphological complexity, converting a grammatical description into computational form is akin to writing a computer program based on a verbal specification, but with the difference that program specifications are (or should be) written so as to be minimally ambiguous (Meyer 1985). Grammars of languages, on the other hand, are generally written so as to be maximally readable, and it appears from our experience that completeness and non-ambiguity are at best secondary aims (David and Maxwell 2008).

The second reason morphological grammars are difficult to write is that many languages are under-documented. This is of course true for many minority languages, but it remains surprisingly true of some "large" languages, as we discovered during the development of grammars and parsers of Bangla (the eighth largest language of the world in terms of number of native speakers), Pashto (between thirty five and forty million speakers), and to a lesser extent Urdu (over sixty million speakers). Indeed, for Pashto we have needed to do field work with native speakers in order to fill in gaps and reconcile differences among grammars. The problem is not that there are no published grammars for these languages; the problem is rather that the published grammars differ, not only in their analyses (e.g. how many declension classes there are in Pashto), but even in the data (such as the existence of certain suffixes in Bangla). Moreover, these grammars often exhibit gaps in grammatical coverage.

Because of these problems with existing grammars of the languages we have worked with, our first step has been to write a descriptive grammar, with the aim of avoiding ambiguity and filling gaps in coverage. This grammar is intended for human readers, and forms the basis for the second step: the creation of a formal grammar which is then automatically converted into the form required by the parser. The mutually supporting relationship between the descriptive and formal grammars is further described in (David and Maxwell 2008).

We have thus developed a way of building parsers which treats the parser as a secondary resource build out of three primary resources. The parsing engine is a language-independent primary resource, while the lexicon and formal grammar are language-specific primary resources. The key to making the lexicon and formal grammar maintainable—that is, the key to giving them a long lifetime—is to make them independent of the short-lived parsing engine. In order for this to work, it must be easy to convert the lexicon and formal grammar into the form required by parsing engines.

2. A Grammar Standard

We therefore now turn to how our formal morphological and phonological grammars are written. These formal grammars are based on a model of morphology and phonology. This model is linguistically based, in that it was designed to handle the range of morphological and phonological constructions known to linguists.

This abstract model has been instantiated as an XML schema defining structures for the description of the morphology and phonology of a language. Grammars written in this form can be readily converted into the form required by morphological parsing engines.[3]

We intend this formalism to be a potential standard for morphological and phonological resources, usable by others outside our group of researchers. A standard for linguistic description must have the following characteristics:

1. The standard must be precise, i.e. not open to varying interpretations.
2. The standard must be stable.
3. It needs to be linguistically sound, i.e. capable of describing relevant linguistic structures.
4. It must be usable by researchers from a variety of theoretical backgrounds.
5. It must be usable for a wide range of language typologies (and preferably for all languages).
6. This standard must be usable together with other standards.

To the above general design criteria, we add the following more specific criteria for standards for morphological and phonological descriptions:

7. Formal morphological descriptions should allow for inter-operation with lexical resources.
8. It should be possible to transform a formal grammatical description plus a lexicon into a parser.
9. The description language defined by the standard should be, insofar as possible, theory-neutral.
10. Where the description language must touch on theoretical issues (e.g. the mechanism for determining the placement of infixes), it should aim for observational adequacy (the ability to describe particular languages), but not necessarily for descriptive adequacy (the ability to model the actual structures in the minds of native speakers, which are in any case not agreed on by different theories), much less explanatory adequacy (the ability to explain how first language learners construct a grammar from the raw data).[4]
11. The description language need not (and probably should not) attempt to rule out the description of linguistic structures which have not (yet) been attested in the world's languages, such as unattested patterns of infixing or reduplication.
12. The standard should allow for linking categories (e.g. parts of speech) to existing definitions of those categories, such as terminology banks or the GOLD linguistic ontology (http://linguistics-ontology.org).
13. Formal grammatical descriptions of a language should come with a structured corpus of test data, so that a parser built from the grammar can be verified against that data.

[3] One might also propose a model of formal grammars for syntax, but our sense is that there is much less agreement among syntacticians about what such a standard might be than there is among morphologists.

[4] On these levels of adequacy, see Chomsky 1965.

Several of the above criteria relate to *interoperability*: criterion (4) is human interoperability; (5) is language interoperability; (6) is interoperability with other standards; (7) is lexical interoperability. Points (6) and (7) together imply interoperability with lexical standards; we are particularly targeting the Lexical Markup Framework (LMF, an ISO standard for electronic lexicons; see http://www.lexicalmarkupframework.org).

Points (9) and (10) have to do with interoperability with linguistic theories. The potentially conflicting requirements that the model be simultaneously theory-neutral and observationally (if not descriptively) adequate are handled in what may appear to be a paradoxical way: the model largely embodies a 1950s-era theory of morphology and phonology, in the sense that it treats phonemes (or graphemes) as phonetic primitives, rather than decomposing these into phonetic features. One motivation for this is the fact that phonological theories differ widely in their treatment of feature structures. Perhaps surprisingly, the absence from the model of phonological features has no effect on observational adequacy (although it might be argued to detract from the descriptive adequacy of grammars written in the model). Natural classes for phonology and morphology can still be defined, but the definitions are extensional (in terms of the phonemes included) rather than intensional (in terms of phonetic features which select certain phonemes). One desirable implication of this is that so-called "unnatural" classes of phonemes can also be defined; the need for this has recently been made clear by Mielke (2008).

At the same time, the linguistic coverage must be scalable, in the sense of being applicable to a wide range of languages, including languages which have not yet been described. Specifically, we wish the standard to cover a wide range of morphological phenomena, not just the phenomena found in typical European languages. Criteria (3) and (5) address this, and motivate aspects of the standard which go beyond a 1950s-era theory. An example of this is the use of a 1980s-era formalism for non-concatenative morphology, including infixation, reduplication, and suprafixation. Even this approach to non-concatenative morphology has since been superseded in the theoretical literature, but largely for reasons of descriptive adequacy, not observational adequacy: alongside "real" rules of reduplication, this older approach allows writing rules which are not attested in natural languages (Marantz 1982). Our treatment of non-concatenative morphology is thus also an example of design criterion (11) above: we aim for observational adequacy, but we do so without ruling out the description of some phenomena which are unattested in natural languages, in the expectation that some of these phenomena will be attested in as yet undescribed languages.

Besides interoperability, we wish resources built according to this standard to be *sustainable*. To use a metaphor, the "adult lifetime" (the lifetime as a usable resource) of a grammar built according to this standard should be long. As discussed above, we have addressed one aspect of this problem by writing our morphological grammars in a parser-independent way, and building a program which automatically converts a grammar into the code needed by a parsing engine. This code, written in Python, operates in two phases: in the first phase, the XML-based grammar is read in and converted to an internal format as objects largely corresponding to the XML elements. In the second phase, these objects are written back out in the format required by the parsing engine. This second phase can be easily re-targeted to a new parsing engine.

However, the Python programming language will some day join other obsolete programming languages like Algol, Simula and PL/I. We therefore consider it necessary to ensure that the formal grammar is understandable by humans, so that a new converter program can be written in the future. One way to try to ensure human understandability is to document the formal grammar by embedding comments into it. But since as described above we are already writing a descriptive grammar of the language, we have chosen instead to embed the formal grammar as fragments into the descriptive grammar, using Literate Programming (Knuth 1992; Maxwell and David 2008). Each piece of the formal grammar is thus supplemented by a description of not only what that fragment means, but also a description of the usage of the construction in the syntax, and by examples. Examples include tables showing the paradigms of example words, and interlinear texts illustrating the usage of the construction.

These examples have turned out to have another use, one which we did not anticipate: they constitute a test set to verify parser correctness. Because the grammar attempts to be exhaustive in its coverage of the morphology, some constructions are quite rare, and might not be found in a typical corpus. The use of test cases automatically extractable from the descriptive grammar ensures that the parser gets tested on a wide range of constructions. (This does not of course obviate the need for testing on a representative corpus.)

3. The Standard in Practice

A standard which has not been implemented is a standard which is likely un-implementable. Accordingly, we have implemented two morphological grammars using this standard already (Bengali and Urdu), and are in the process of implementing a third grammar, for Pashto. These Indo-European languages are far from exhausting the full range of linguistic constructions that our XML schema is intended to cover; for example, we have not yet faced infixes or significant reduplication. This is intentional; we wished to debug the overall methodology beginning with relatively tractable languages.

Nevertheless, the range of morphological and phonological considerations we have had to deal with is not inconsiderable:

- Fusional and agglutinative morphology with both prefixes and suffixes
- Extended exponence (morphosyntactic features realized on more than one affix, implying a sort of agreement between affixes)
- Inflection classes
- Stem and affix allomorphy governed by phonological

rules (with some rules being sensitive to lexical exception features)
- Suppletive stem and affix allomorphy
- Suppletive word forms (requiring blocking of hyper-regular forms)
- Dialectal and spelling variation

All of this has been implemented in the conversion program. Specifically, these constructs in the XML-based formal grammar are automatically mapped into the programming language of the Stuttgart Finite State Transducer (SFST, see http://www.ims.uni-stuttgart.de/projekte/gramotron/SOFTWARE/SFST.html), and are automatically combined with lexical entries extracted from electronic dictionaries to produce a morphological parser.

Non-concatenative morphology is handled in the XML schema, but it has not yet been implemented in the converter program. The first test case will be Pashto, where the oblique case of one noun declension class is realized by a change in a stem vowel, with no overt affix. While this may seem a trivial test case, it will in fact exercise the same machinery required for simple reduplication and infixation.

4. Conclusion

We close with some thoughts on the place of our approach in the field of computational linguistics. Writing the descriptive grammars is very much an exercise in traditional linguistics; writing the formal grammars is much closer to writing a computer program from a specification, that is, to traditional computer programming. The recent history of computational linguistics has seen a turn away from an involvement with linguistics, a change which has been lamented by some observers (Steedman 2008, Wintner 2009). We see our method as a way of bringing the two fields back together, at least in grammar. Moreover, this is potentially a way in which computational linguistics could become relevant to field linguistics, specifically in the documentation of minority and endangered languages (cf. Bird 2009).

Finally, in our opinion it is not absurd to think that grammars built today might have as long a lifetime as grammars of classical Tamil, Sanskrit, Latin, and ancient Greek have enjoyed. Those grammars were of course descriptive grammars, and have their weaknesses; it goes without saying that they were never intended for computational implementation. It is impossible for us to know how linguists decades from now, or even centuries from now, may look at our attempts at grammatical description, or how they might implement morphological parsers. But if a parser can be implemented now based on our formal grammars, then barring the collapse of civilization it seems that future linguists will be able to accomplish this as well. If this is true, then not only we will have overcome the present short life cycle of this kind of language resource, we will have introduced a way to document languages which can in principle be used by field linguists to describe endangered languages, preserving important aspects of languages which will otherwise soon disappear (Bird and Simons 2003).

5. References

Antworth, Evan L. 1990. *PC-KIMMO: A Two-Level Processor for Morphological Analysis*. Occasional Publications in Academic Computing No. 16. Dallas: Summer Institute of Linguistics.

Beesley, Kenneth R., and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. Chicago: University of Chicago Press.

Bird, Steven. 2009. Natural Language Processing and Linguistic Fieldwork. *Computational Linguistics* 35, no. 3: 469-474.

Bird, Steven, and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79, no. 3: 557-582.

Chomsky, Noam. 1965. *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Copestake, Ann. 2002. Implementing Typed Feature Structure Grammars. Stanford: CSLI Publications.

David, Anne, and Michael Maxwell. 2008. Joint Grammar Development by Linguists and Computer Scientists. In *Workshop on NLP for Less Privileged Languages, Third International Joint Conference on Natural Language Processing*, 27-34. Hyderabad, India: Asian Federation of Natural Language Processing. http://hdl.handle.net/1903/7567.

Knuth, Donald E. 1992. *Literate Programming*. CSLI Lecture Notes. Stanford: Center for the Study of Language and Information.

Marantz, Alec. 1982. Re reduplication. *Linguistic Inquiry* 13: 435-482.

Mielke, Jeff. 2008. *The Emergence of Distinctive Features*. Oxford Studies in Typology and Linguistic Theory. Oxford: Oxford University Press.

Maxwell, Michael, and Anne David. 2008. Interoperable Grammars. In *First International Conference on Global Interoperability for Language Resources (ICGL 2008)*, 155-162. Hong Kong. http://hdl.handle.net/1903/7568.

Meyer, Bertrand. 1985. On Formalism in Specifications. *IEEE Software* 3: 6-25.

Sproat, Richard. 1992. Morphology and Computation. Cambridge MA: MIT Press.

Steedman, Mark. 2008. On Becoming a Discipline. *Computational Linguistics* 34: 137-144.

Weber, David J., H. Andrew Black, and Stephen R. McConnel. 1988. *AMPLE: A Tool for Exploring Morphology*. Occasional Publications in Academic Computing Number 12. Dallas: Summer Institute of Linguistics.

Wintner, Shuly. 2009. What Science Underlies Natural Language Engineering? *Computational Linguistics* 35: 641-644.

The TEI and the NCP: the model and its application

Piotr Bański **Adam Przepiórkowski**

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

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

Abstract

We present the National Corpus of Polish (NCP), a TEI-encoded 1-billion-word text corpus with multiple layers of linguistic annotation, the product of co-operation of a consortium of all the major Polish institutions that created their own significant corpora in the past. We review the major properties of the corpus and of its architecture, with an eye to the hot topics of today: interoperability and sustainability. Special attention is paid to the status of the encoding schemes of the corpus vis-à-vis the currently popular annotation standards.

1. Introduction: all the buzzwords

The interrelated issues of sustainability and interoperability are part of the landscape of any system of representation and processing of linguistic knowledge. Helbig (2001), quoted in (Witt *et al.*, 2009:7), lists interoperability, homogeneity, and communicability as three requirements for such a system. Helbig's *homogeneity* requires the same formalism across different levels of linguistic description, and *communicability* refers to the documentation, conditioning successful teamwork and allowing to share resources among different teams.

Our primary focus here is on a particular subset of knowledge inherent in *Language Resources* (LRs)[1], and more specifically, in what Witt *et al.*, (2009) call *static text-based LRs*, i.e., language corpora, although we will also mention *dynamic LRs*, that is tools that manipulate or query corpora. In this context, *interoperability* means, generally, the ability of LRs to "understand" each other and to interact. As Ide (2010) points out, this can take place at several levels, notably at the syntactic level, e.g. via an abstract pivot format that makes it possible to reduce the number of mappings between schemas (cf. also (Ide and Romary, 2007)), and at the semantic level, by reference to a common data model and, crucially, a shared inventory of reference categories.

It is a trivial observation that the practical value of LRs lies in their use, possibly recurrent, and ideally permanent. This takes us towards the concept of *sustainability*, i.e., (very generally and imprecisely) the ability to ensure a prolonged (and ideally permanent) use of LRs. Sustainability, as defined by e.g. (Nathan, 2006) or (Simons and Bird, 2008) requires that a LR be (i) extant (Nathan: permanent), (ii) usable (Nathan: proficiently prepared), and (iii) relevant (Nathan: pertinent). Simons and Bird (2008), whose terminology we adopt here, additionally split usability into sub-conditions: discoverability, availability, interpretability, and portability. Note that the last two properties provide for interoperability.

We will not provide a thorough overview of all these requirements, merely noting them as we proceed in our presentation of the largest linguistically annotated text corpus of Modern Polish, the National Corpus of Polish (NCP, also known under its native abbreviation, NKJP, http://nkjp.pl/).[2] In doing so, we briefly report on the origin and the nearest future of the corpus as well as the design decisions that shaped it.

2. National Corpus of Polish: history and future

The NCP is the deliverable of project number R17 003 03, sponsored by the Polish Ministry of Science and Higher Education. It was launched in December 2007 and will terminate at the end of 2010. The project is carried out by a consortium of four institutions that developed their own significant corpora in the past. These corpora are now joined into a single resource that has been expanded to nearly three times the size of the original corpora put together. The project members are the following:

- the Institute of Computer Science at the Polish Academy of Sciences in Warsaw (ICS PAS),
- the Institute of the Polish Language at the Polish Academy of Sciences in Cracow (IPL PAS),
- the PWN Scientific Publishers in Warsaw,
- the PELCRA group at the University of Łódź.

These four institutions have combined their expertise to merge, uniformly encode, enlarge, and enhance their resources, eventually producing a 1-billion-word corpus (with a carefully balanced 300-million-word subpart), annotated for various levels of linguistic description and designed to last and serve current and future research (for more details, see Przepiórkowski *et al.*, 2010).

After the project has completed at the end of 2010, the

[1] A Language Resource is "any physical or digital item that is a product of language documentation, description, or development, or is a tool that specifically supports the creation and use of such products" (Simons and Bird, 2008). See (Witt *et al.*, 2009) for a suggested taxonomy of LRs.

[2] "NKJP" expands into "Narodowy Korpus Języka Polskiego".

NCP will be used as the empirical basis for the development of a new large dictionary of Modern Polish that is being created at IPL PAS. 1-million-word demo of the corpus may be released under an open license, pending the solving of licensing issues.

The contents of this section, apart from setting the context for the rest of the paper, address some of the requirements mentioned in section 1, for example the requirement of *relevance*: the NCP is the first corpus of its kind in Poland, and the first such corpus of Polish. To our knowledge, it is also the first corpus of this size (10^9 words) with homogeneous encoding of multiple hierarchical layers of linguistic annotation (to be reviewed below), in the world. The entire bulk of the corpus (though, understandably, not the entire set of annotations, some of which are still being created) is already available for searching via two interfaces. The NCP will have a large, carefully balanced subcorpus and it contains nearly 2 million words if informal conversational Polish, which is precious for two reasons: firstly, there has been no corpus of (transcribed) spoken Polish of such a size before, and secondly, most of the digital data is still available for being aligned with the recordings, which opens a further exciting research perspective.

We also address the issue of *permanence*: copies of the corpus are made regularly and its nearest future is secured. Attention is paid to the question of its long-term persistence, which will be reported on in due time.

As for *availability*, the corpus may not be released in the source form due to the numerous legacy restrictions on the use of the data that it contains: many texts have been released to the NKJP Consortium on the condition that they are not distributed further. However, the corpus may already be queried in its entirety, and a 1-million-word part of it (composed of texts carefully selected for their lack of copyright restrictions) will most probably be released publicly – we wish to note that this is much, much more than what many other closed corpora release. *Discoverability* of the corpus is already partially taken care of (it is, naturally, part of the LREC LR Map), and will also be addressed after the project is completed. We look at sustainability and interoperability of the NCP in sections 3 through 5 below.

3. Architecture: stand-off annotation

The NCP is built according to the guidelines for annotating modern LRs and uses the so-called stand-off mechanism of annotation (Thompson and McKelvie, 1997; Ide and Romary, 2007), whereby each annotation document (typically, though not always, containing information pertaining to a single level of grammatical description) is located in a separate file that references another annotation file or the source text by means of various pointing mechanisms. The typical contents of a leaf directory in the corpus are as presented in the list below (see http://nlp.ipipan.waw.pl/TEI4NKJP/ for working versions of these files; the file NKJP_header.xml belongs here only virtually – we look at its role presently):

- text_structure.xml
- ann_segmentation.xml
- ann_morphosyntax.xml
- ann_senses.xml
- ann_words.xml
- ann_named.xml
- ann_groups.xml
- header.xml
- (NKJP_header.xml)

Above, the file text_structure.xml stores the source text – this file contains coarse-grained inline structural annotation, typically down to the paragraph level. The other files contain annotations of other kinds, organized in a hierarchy: the first is ann_segmentation.xml, containing the segmentation layer that identifies the sentence boundaries and the contiguous non-overlapping sequence of individual segments (including segmental ambiguities, cf. (Bański and Przepiórkowski, 2009)), by addressing character spans in the source text.[3] The segmentation layer can be the pivot layer for many other annotation documents, depending on the setup of the particular corpus and the nature of annotations. In the NCP, however, only the morphosyntactic layer (ann_morphosyntax.xml) is built on top of it. This layer contains all the possible morphosyntactic interpretations of each segment together with an optional disambiguation section that points at the most likely interpretation. The morphosyntactic layer serves as the basis for three other layers, namely those (a) identifying syntactic words (ann_words.xml), (b) identifying named entities (ann_named.xml, cf. (Savary *et al.*, 2010)), and (c) disambiguating selected polysemic lexemes (ann_senses.xml, cf. (Młodzki and Przepiórkowski, 2009)). Finally, the level of syntactic chunks (ann_groups.xml, cf. (Głowińska and Przepiórkowski, 2010)) references the syntactic word level. The file header.xml is the local TEI header, included by all the other files in the directory, whether containing the source text or the annotations. The file NKJP_header.xml is the main corpus header, included by all the files in the entire corpus and thus binding them into a single virtual unit.

It has to be pointed out that stand-off architecture is one of the preconditions for *sustainability* and *interoperability*. A stand-off annotated LR preserves the source text in a minimally marked-up form and hence as capable of being easily extracted or processed by future versions of the current tools or by new tools. Such a resource is also easily expandable, which also adds to its attractiveness (Ide and Romary, 2007). Interoperability of stand-off annotated resources can be realised both at the level of the source text and at the level(s) of the annotation layers: it becomes possible to e.g. compare tagsets, conflicting

[3] These spans can be smaller than orthographic words – for the motivation see (Bański and Przepiórkowski, 2009), for their treatment at the level of syntactic words (ann_words.xml), see (Głowińska and Przepiórkowski, 2010).

annotations, or outputs of different tools; it is much easier to map the contents of annotation layers onto different resources. The criterion of heterogeneity becomes important in this regard (Witt et al., 2009), and we shall see in the next section that the NCP fulfils it.

4. Encoding format: Text Encoding Initiative XML

The NCP is encoded in the popular TEI XML encoding standard (TEI Consortium, 2010), a *de facto* standard for resources of many kinds used in the Humanities and for LRs in general.[4] The TEI Guidelines provide a variety of means to encode linguistic information in LRs. When tailoring the TEI model for the NCP, we attempted to follow the existing standards for linguistic annotation. That task was not difficult because of the origin of many of these standards. The current standards that have been or are being established by ISO TC 37 SC 4 committee (http://www.tc37sc4.org/), known together as the LAF (Linguistic Annotation Framework) family of standards, cf. (Ide and Romary, 2007), descend in part from an early application of the TEI, back when the TEI was still an SGML-based standard. That application was the Corpus Encoding Standard (Ide, 1998), later redone in XML and known as XCES (Ide *et al.*, 2000). XCES was a conceptual predecessor of the current ISO LAF pivot format for syntactic interoperability of annotation formats, GrAF (Graph Annotation Framework, (Ide and Suderman, 2007)). GrAF defines an XML serialization of the LAF data model consisting of directed acyclic graphs with annotations (also expressible as graphs), attached to nodes. This basic data model is in fact common to the TEI formats defined for the NCP, the LAF family of standards, and the other standards and best practices such as Tiger-XML (Mengel and Lezius, 2000) – popular for treebank encoding, or PAULA (Dipper, 2005) – a versatile format for multi-modal and multi-layered corpus encoding.[5] The differences pertain to details such as the assumed format of feature structures or the presence or absence of extra mechanisms, such as labelled edges (which can naturally be transduced into nodes when converting formats). We discuss the interrelations between the LAF family of ISO standards, Tiger-XML, PAULA, and the annotation schemas defined for the NCP in (Przepiórkowski, 2009; Przepiórkowski and Bański, 2010).[6] Przepiórkowski and Bański (2010) show that the particular TEI application for the NCP, a result of heavy customisation of the ultra-versatile toolkit that the TEI Guidelines offer, is (a) a concrete ("out-of-the-box") solution subsuming the abstract GrAF, (b) isomorphic with Tiger-XML and PAULA, and often mirroring the devices used there, and (c) equipped with documentation trivially derivable from the literate-encoded ODD files (see below), (d) offering a homogeneous format for a variety of annotation layers and (e) offering well-tested metadata-encoding in the form of TEI headers that not only describe the source text and annotation documents but also (f) virtually link them, by being XIncluded into each of them. All annotation layers from the morphosyntactic layer upwards use the ISO/TEI feature structure representation (FSR) standard (ISO:24610-1). All in all, the NCP application of the TEI is offered for the encoders of complex corpora as a pragmatic solution that allows them to use a *homogeneous* set of well-documented schemas *interoperable* with the currently endorsed standards and best practices.

The above-mentioned ODD ("One Document Does it all") files are the TEI's recipe for what Bauman (2008) calls "literate encoding", by reference to the literate programming paradigm (Knuth, 1984): TEI schemas are defined in TEI documents, with the typical TEI header and a standard text body with the addition of special elements that provide instructions for constructing schemes out of the content models and attribute classes offered by the TEI, cf. (Burnard and Rahtz, 2004). These files are then processed to derive schemas (such as DTD, RelaxNG, Schematron or XML Schema) and/or documentation in various formats. This provides for Helbig's (2001) *communicability*, i.e. sharing uniform documentation across project members and with external entities.

The well-known TEI headers (due to their comprehensiveness and versatility used by many projects that do not use the TEI as such) provide for one of the aspects of sustainability, namely *discoverability*. The NCP headers record the history of the text (in extreme cases, also the entire headers of files that have been converted from the corpora created by the members of the NKJP Consortium) and the history of the annotation documents, classify the text, and provide all the standard information that can be useful in locating or querying the text. A single main corpus header provides information common to all files in the corpus and defines several taxonomies that the local headers use (examples of headers are provided at http://nlp.ipipan.waw.pl/TEI4NKJP/).

5. More on interoperability

In the previous section, we have addressed the issue of interoperability considered in terms of syntactic formats, i.e., from the point of view of what Witt *et al.* (2009) call

[4] See http://www.tei-c.org/Activities/Projects/ for an incomplete list of encoding projects using the TEI.

[5] In the case of Tiger-XML, the genealogy is different: it was created as an independent format and it is now being incorporated into ISO SynAF (ISO:24615). The NCP schema for syntactic annotation is isomorphic to SynAF/Tiger-XML.

[6] The TEI has re-incorporated the (X)CES proposals for corpus encoding (among others, stand-off annotation) and introduced its own schemes for referencing spans of characters and sequences of elements as extensions to the XPointer Framework (http://www.w3.org/TR/xptr-framework/). While the NCP demonstrates that the level of stand-off support in the TEI is sufficient for more technically-oriented users, there are still details that remain to be taken care of in order to ensure a greater level of the TEI's user-friendliness in this regard. Some of them are discussed in Bański (2010).

static text-based LRs. In this section, we look at how the NCP copes with the semantic interoperability and move on to review the dynamic LRs (tools) offered by the project. Recall that semantic interoperability requires sharing a common data model and additionally, a common set of reference categories. In the context of ISO LAF (and LRs in general), one of the places offered to store reference categories is the ISOcat Data Category Registry (http://www.isocat.org/, cf. (Kemps-Snijders *et al.*, 2008)). Recently, as reported in (Patejuk and Przepiórkowski, 2010), the NKJP Tagset has been defined in the ISOcat (totalling in 85 Data Categories) and became the first public ISOcat definition of a complete tagset, available as a public Data Category Selection (keyword: nkjp). This testifies to the *semantic interoperability* potential of the NCP.

The main tools adapted for the NCP, Poliqarp (a search engine and a concordancer, cf. (Janus and Przepiórkowski, 2007))[7] and Anotatornia[8] are offered under the GNU General Public License (GPL). Anotatornia (Przepiórkowski and Murzynowski, forthcoming) is a tool for manual encoding of multi-level corpora (handling word-level and sentence-level segmentation as well as morphosyntax and word-sense disambiguation) that includes inter-annotator conflict-resolution mechanisms. These tools are the NCP's offer in the sphere of dynamic LRs. They are planned to be implemented in projects using TEI XML architecture with the data model based on that of the NCP, namely the Open-Content Text Corpus (OCTC, a multilingual SourceForge resource in the alpha stage of development, cf. (Bański and Wójtowicz, this volume)) and the Foreign Language Examination Corpus (a University of Warsaw Council for the Certification of Language Proficiency project, in the planning phase, cf. (Bański and Gozdawa-Gołębiowski, 2010)).

6. Conclusion

We have presented the National Corpus of Polish – a standards-compliant, scalable, and sustainable language resource with open-source tools designed to be flexible enough to interoperate with other resources of a similar type. The corpus contains a hierarchy of stand-off annotation levels, each of them is encoded in TEI XML, which satisfies the homogeneity requirement of Helbig (2001). Corpus documentation for each annotation layer can be derived from the appropriate ODD configuration files (Burnard and Rahtz, 2004), which fulfils the requirement of communicability.

The NCP demonstrates the usefulness of the TEI XML toolkit configured with an eye towards meeting the challenges that modern Language Resource producers and users face. These design choices have proven to be usable also for other resources of a similar general kind (the OCTC and the FLEC, mentioned above). We believe that this makes both the TEI – on the general plane, and the NCP – on the plane of applications, serious participants in the debate on the current state and future development in the sphere of Language Resources.

[7] http://poliqarp.sourceforge.net/

[8] http://nlp.ipipan.waw.pl/Anotatornia/

7. Acknowledgements

We are grateful to two anonymous LRSLM2010 reviewers for their helpful remarks. Additionally, Piotr Bański wishes to thank Victoria Arranz for her extraordinary patience.

8. References

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

Bański, P. and Gozdawa-Gołębiowski, R. (2010). Foreign Language Examination Corpus for L2-Learning Studies. Submitted for the proceedings of the 3rd Workshop on Building and Using Comparable Corpora (BUCC), "Applications of Parallel and Comparable Corpora in Natural Language Engineering and the Humanities", 22 May 2010, Valletta, Malta.

Bański, P., Wójtowicz, B. (this volume). The Open-Content Text Corpus project. In proceedings of the LREC workshop on "Language Resources: From Storyboard to Sustainability and LR Lifecycle Management" (LRSLM2010), 23 May 2010, Valletta, Malta.

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, pp. 64--67.

Bauman, S. (2008). Freedom to Constrain. In Balisage: The Markup Conference 2008, available at http://www.balisage.net/Proceedings/vol1/html/Bauman01/BalisageVol1-Bauman01.html.

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

Dipper, S. (2005). XML-based stand-off representation and exploitation of multi-level linguistic annotation. In *Proceedings of Berliner XML Tage 2005* (BXML 2005). Berlin, pp. 39--50.

Głowińska, K., Przepiórkowski, A. (2010).The Design of Syntactic Annotation Levels in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta.

Helbig, H. (2001). *Die semantische Struktur natürlicher Sprache. Wissensrepräsentation mit MultiNet*. Berlin: Springer.

Ide, N. (1998). Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora. *Proceedings of the First International Language Resources and Evaluation Conference*, Granada, Spain, pp. 463--470.

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.

Ide, N., Romary, L. (2007). Towards International Standards for Language Resources. In Dybkjaer, L., Hemsen, H., Minker, W. (eds), Evaluation of Text and Speech Systems, Springer, pages 263--284.

Ide, N., Suderman, K. (2007). GrAF: A Graph-based Format for Linguistic Annotations. Proceedings of the Linguistic Annotation Workshop, held in conjunction with ACL 2007, Prague, June 28-29, pp. 1--8.

Janus, D., Przepiórkowski, A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In *Proceedings of the ACL 2007 Demo Session.* Prague. pp. 85--88.

Kemps-Snijders, M., Windhouwer, M.A., Wittenburg, P., Wright, S.E. (2008). ISOcat: Corralling Data Categories in the Wild. In European Language Resources Association (ELRA) (ed), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco, May 28-30, 2008.

Knuth, D.E. (1984). Literate programming. The Computer Journal (British Computer Society), 27(2), pp. 97--111.

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.

Młodzki, R., Przepiórkowski, A. (2009). The WSD Development Environment. In *Proceedings of the 4th Language & Technology Conference*. Poznań, Poland. pp. 185--189.

Nathan, D. (2006). Proficient, permanent, or pertinent: aiming for sustainability. In Barwick, L., Thieberger, N. (Eds.), *Sustainable Data from Digital Fieldwork*. The University of Sydney, pp. 57--68.

Patejuk, A., Przepiórkowski, A. (2010). ISOcat Definition of the National Corpus of Polish Tagset. In proceedings of the LREC 2010 workshop on "LRT Standards". Valletta, Malta. ,

Przepiórkowski, A. (2009). TEI P5 as an XML Standard for Treebank Encoding. In: Passarotti, M., Przepiórkowski, A., Raynaud, S. Van Eynde, F. (eds), *Proceedings of of the Eighth International Workshop on Treebanks and Linguistic Theories* (TLT8), pp. 149--160.

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. Goźdź-Roszkowski (ed.) *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main: Peter Lang.

Przepiórkowski, A., Górski, R.L., Łaziński, M., 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.

Przepiórkowski, A., Murzynowski, G. (forthcoming). Manual annotation of the National Corpus of Polish with Anotatornia. In Goźdź-Roszkowski, S. (ed.) *The proceedings of Practical Applications in Language and Computers (PALC-2009)*. Frankfurt: Peter Lang.

Savary, A., Waszczuk, J., Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*. Valletta, Malta .

Simons, G.F., Bird, S. (2008). Toward a global infrastructure for the sustainability of language resources. *Proceedings of the 22nd Pacific Asia Conference on Language, Information and Computation: PACLIC 22*. pp. 87--100.

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/

Thompson, H. S., McKelvie, D. (1997). Hyperlink semantics for standoff markup of read-only documents, Proceedings of SGML Europe. Available from http://www.ltg.ed.ac.uk/~ht/sgmleu97.html.

Witt, A., Heid, U., Sasaki, F., Sérasset, G. (2009). Multilingual language resources and interoperability. In *Language Resources and Evaluation*, vol. 43 :1, pp. 1--14. doi:10.1007/s10579-009-9088-x

The German Reference Corpus: New developments building on almost 50 years of experience

Marc Kupietz, Oliver Schonefeld, Andreas Witt

Institute for the German Language (IDS)
R5 6–13, 68161 Mannheim, Germany
{kupietz|schonefeld|witt}@ids-mannheim.de

Abstract

This paper describes the efforts in the field of sustainability of the *Institut für Deutsche Sprache* (IDS) in Mannheim with respect to DEREKO (Deutsches Referenzkorpus) the *Archive of General Reference Corpora of Contemporary Written German*. With focus on re-usability and sustainability, we discuss its history and our future plans. We describe legal challenges related to the creation of a large and sustainable resource; sketch out the pipeline used to convert raw texts to the final corpus format and outline migration plans to TEI P5. Due to the fact, that the current version of the corpus management and query system is pushed towards its limits, we discuss the requirements for a new version which will be able to handle current and future DEREKO releases. Furthermore, we outline the institute's plans in the field of digital preservation.

1. Introduction

The Institute for the German Language (IDS) has a long tradition in building corpora. DEREKO (*Deutsches Referenzkorpus*), the *Archive of General Reference Corpora of Contemporary Written German*, has been set off as the Mannheimer Korpus 1 project in 1964. Paul Grebe and Ulrich Engel succeeded in compiling a corpus of about 2.2 million running words of written German by 1967. Since then, further corpus acquisition projects established a ceaseless stream of electronic text documents and let the corpus to grow steadily (Kupietz & Keibel, 2009).

As of 2010 the corpus, which is intended to serve as an empirical basis for Germanic linguistic research, compromises more than 3.9 billion words (IDS, 2010) and has a growth rate of approximately 300 million words per year. In compliance with the statutes of the institute as a public-law foundation that define the documentation of the German language in its current use as one of its main goals, it is declared IDS policy to provide for a long term sustainability of DEREKO. In 2004 a permanent project responsible for its maintenance and further development has been established.

2. Current state

As stated in Kupietz et al. (2010), the key features of DEREKO are the following:

- established and developed in 1964
- contains texts from 1956 onwards
- continually expanded
- contains fictional, scientific and newspaper texts as well as several other types of text
- only complete and unaltered texts (no correction of spelling, etc.)
- only licensed material
- not available for download (due to license contracts and intellectual property rights)
- maximum size, primordial sample design
- allows the composition of specialized samples
- endowed with currently three concurrent annotation layers

Unlike other well-known corpora like, e.g. the BNC (BNC, 2007), DEREKO itself is not intended to be balanced in any way. The underlying rationale is that the term balanced – just as much as the term representative – can only be defined with respect to a specific research question and some statistical population. Thus the composition of a sample should be part of the usage phase and not part of the design phase of a corpus that shall be used as a general basis for empirical linguistic research. As a consequence of this so called *primordial sample* approach, the text acquisition can concentrate on the maximization of size and stratification and as any DEREKO-based samples can be defined an overall boost of versatility and re-usability is achieved. A more detailed view of DEREKO's primordial sample approach and its application scenarios is given in Kupietz et al. (2010).

2.1. Legal aspects of re-usability

To allow for a broad sampling of language data, the IDS has negotiated license contracts with various copyright owners, such as authors, publishing houses and newspapers. The contracts grant non-commercial academic use of the data exclusively and allow access only via software that among other things must prevent the reconstructability of whole texts. Licenses are open-ended, but can be cancelled by the licensor at any time. As a consequence with respect to sustainability, the IDS cannot guarantee the persistency of texts contained in DEREKO as the right holders can in principle withdraw the right of use any of their texts at any time. In the last years, however, this happened only to single newspaper texts. The most frequent reason was that a publisher had undertaken to refrain from the further distribution of an article. As the average frequency of such deletions was less than 50 per year, until now, the replicability of DEREKO-based findings should not have been significantly affected.

At large, the situation concerning usage rights and their sustainability ist not ideal, but like all large-scale corpus

projects, DeReKo, more specifically the IDS as the language resource and service provider, has to walk a tightrope between the interests of its target community and those of the IPR holders. More generally speaking, as the vast majority of digital research resources in linguistics are subject to third parties' rights, the problem boils down to a conflict of basic rights, with freedom of science and research on the one hand and the protection of property and general personal rights on the other. As long as the weighting does not shift dramatically in favor of the freedom of science, there will be no general solutions but only compromises, which are more or less specific to individual resource types and research applications.

The IDS is involved in campaigns for a more research friendly copyright-law, e.g. via the Leibniz Association and in CLARIN. In the context of CLARIN and the German counter-part D-SPIN, the IDS also works on improved licensing models. One approach we follow for example in the context of CLARIN and D-SPIN is to develop upgrade agreement models with a graded transferability of usage rights and to test them with selected licensors of DeReKo-texts in order to improve their re-usability within secure distributed research infrastructures.

3. Annotations

In 1993, the IDS started COSMAS II (IDS, 1991–2009), the Corpus Search, Management and Analysis System, as a first step towards providing an access to linguistic annotations. It was planned in order to specifically be capable of handling multi-layer annotations. In 1995 DeReKo was enriched with annotations from the Logos Tagger and in 1999 the analysis from Gertwol Tagger were added.

The IDS recently has started an extensive corpus annotation venture to provide even more annotations. As described in Belica et al. (to appear in 2010), Machinese Phrase Tagger from Connexor Oy, the TreeTagger from Stuttgart University (Schmid, 1994) and the Xerox FST Linguistic Suite and various custom filters have been applied on DeReKo to produce concurrent stand-off annotations. In a first step only the morphological and the part-of-speech analysis components were considered. This annotation process took about 6 CPU-years and resulted in about 3.5 TB of data. In the meantime, DeReKo was also annotated on the syntactic level with the Xerox Incremental Parser XIP. Currently, however, the IDS has only acquired sufficient licenses to make TreeTagger and Connexor annotations available to the outside world via COSMAS II. Presumably because of the danger of reverse engineering, that would arise when a large annotated corpus was made publicly accessible without restrictions, the problems of for acquiring sufficient licenses for commercial taggers and parsers are comparable to those for copyrighted text.

DeReKo-2009-I (IDS, 2009) was the first release with annotations. These contain part-of-speech and morphological (except TreeTagger) information, provided by the above mentioned tools. A detailed report on the annotation process, an assessment of their reliability, and some thoughts on how to use them methodologically sound in linguistic research can be found in Belica et al. (to appear in 2010).

4. Re-usability and sustainability

4.1. From raw data to corpus representation formalisms

The stream of raw data that constantly feeds DeReKo with currently about one million words per day is supplied by the text donors in many different formats. Mostly, these formats are tailored towards the requirements of the publishing industry. However, for the purpose of analysing the data, it has to be converted to a common format. The IDS has developed a format based on XCES (Ide at al., 2000). The input data is converted through a pipeline of various transformation steps. While due to its funnel-like architecture with many small specialized filters only at the beginning of the processing pipeline, a large part of this transformation system is re-usable also for new data sources, the process is still quite an expensive task because often manual intervention is needed due to the broad variances in the input, even for data coming from a single source. Figure 1 gives an overview of the whole processing pipeline.

Recently, the IDS has started to investigate a migration of DeReKo from the custom XCES variant to TEI. As the TEI P5 guidelines (The TEI Consortium, 2007) provide a sufficient degree of adaptability to encode DeReKo without loss of information, a P5-compliant mapping is scheduled for 2010–2011. Besides the obvious advantages of a most recent version of the standard such a conversion does also have drawbacks: Parts of the processing pipeline as well as a large portion of the quality assurance battery are tailored to the old format and migrating to TEI P5 would not gain an immediate advantage. Since DeReKo is not available for direct download, no one outside the IDS will directly benefit from this conversion. In addition there are currently no tools for processing TEI-P5-compliant data that we know of which could be applied on DeReKo. The vast amount data is also beyond the editing and validation capabilities of any current XML editor. For now, the main immediate advantages concerning interoperability, though also only IDS-internally, will arise from the migration from DTD based to schema based validation, which allows for a finer grained control of data types and better maintainability. In the long run, however, we hope that with a migration to TEI we will contribute to a harmonization and standardization process which after all will also lead to tools that are able to deal with large scale TEI data.

Furthermore, migrating to TEI will save us the reinvention of the wheel for areas that are not yet fully covered by the IDS-XCES formalism. For example, TEI offers the opportunity to exploit the standardized feature structures to describe different annotation layers in a unified representation. Witt et al. (2009) gives a detailed view on how to adopt feature structures to archive this goal and discusses advantages and disadvantages of this approach.

4.2. Persistence and preservation

Unlike more static or monolithic corpora, DeReKo being constantly improved and expanded, also has to deal with challenges in the context of replicability of DeReKo-based research, data persistence, and persistent reference. To ensure that all data states are, in principle, reproducible DeReKo is maintained in a subversion repository since

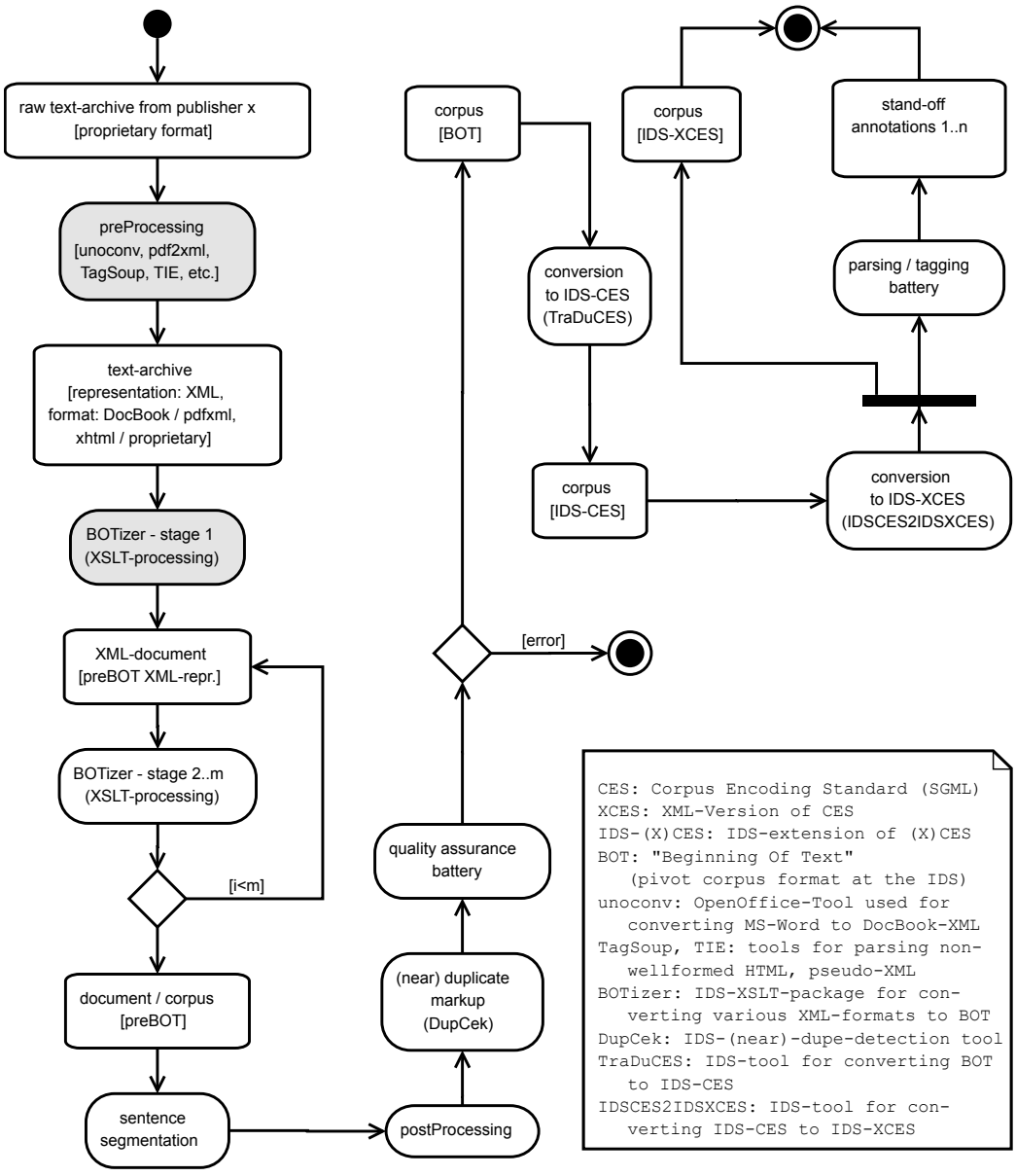


Figure 1: Architecture for processing raw texts. The filter steps highlighted in gray are decreasingly dependent on the input format. Most of the architecture can be reused for new formats. For the migration to TEI P5, first a converter from IDS-XCES will be implemented for testing purposes and evaluation. For a complete migration the following steps will be necessary: (i) insertion of a new conversion routine from preBOT to TEI before the sentence segmentation, (ii) adaption of subsequent steps (quality assurance battery, etc.), (iii) removal of IDS-CES- and IDS-XCES-conversion.

the beginning of 2007. However, with this approach taken alone, the reproduction of old states so that they are actually usable is expensive because a complete version of DeReKo has currently a size of about 5 TB and to make it usable via COSMAS requires at least partial re-indexing. A possible solution to this problem could be to integrate versioning into the core database system. We will consider this in the development of a new corpus search and analysis system (see following section).

To be able to persistently refer to corpora, documents, and texts contained in different DeReKo archive states, internally unique persistent identifiers are used. In the context of the CLARIN initiative, we are currently planning to combine these with globally unique identifiers based on the handle system (Sun et al. 2007), for example to allow for the construction of distributed virtual corpora or resource collections (cf. Kupietz et al. 2010). Together with the ISO standard for the persistent identification of electronic language resources (cf. Broeder et al., 2007; ISO/DIS 24619: ISO/IEC, 2009) this will allow for accurate reference to and citation of DeReKo or parts of it and ensures the traceability of DeReKo-based research.

To further secure the sustainability of DeReKo, the IDS is currently working on a digital preservation strategy. Especially the current the legal arrangements pose a problem for an off-site archiving of the resources, which we regard as a requirement for a proper implementation of such strategy, as most do not allow us to store the data outside of the IDS. We are currently investigating legally in how far storing the data, possibly encrypted, at a co-location would

violate license terms. Eventually, we will have to negotiate license upgrades to explicitly allow storing the data off-site for archival and backup purposes. Moreover the institute is involved in digital preservation activities, e.g., in the context of *nestor*[1] and *WissGrid*[2].

5. Using DeReKo

As DeReKo is not available for download, before even mentioning re-usability and sustainability it is, of course, most important to offer a software to access it that fulfils the needs of the target communities. The current corpus search analysis and management system COSMAS II, with currently about 18,500 registered users, offers a broad range of features. E.g., it allows for the composition of virtual corpora, provides complex search options (including, e.g., lemmatization, proximity operators, search across sentence boundaries, logical operators), can perform complex (non-contiguous) higher order collocation analysis, features various views for search results and different interface clients.

However, COSMAS II was designed in 1993 for a target corpus size of 300 million words and the growth of DeReKo is pushing it towards its limits. Adding more annotation layers to DeReKo will make the situation even worse.

For that reason we currently prepare a new mid-scale project to create a new corpus analysis system. The new system will have to face opportunities and challenges coming from the emerging distributed e-infrastructures as well as, of course, scientific requirements. To mention but a few:

- it must be suitable for performing methodologically sound empirical linguistic research
- observed data and interpretations need to be separable
- more data is better data: it must allow for large amounts of textual data and annotations (target values are 30 billion words with 20 annotation layers)
- the query mechanism shall allow for multi-layer queries
- query, analysis and metadata function should be connectable to e-infrastructures
- virtual corpora should be definable on metadata and text-internal properties
- users should be able to work on previous states of the data
- users should be able to persistently register virtual corpora (/collections)
- users should be able to add cumulative annotations
- users should be able to run own programs on the data
- the system must guarantee that no license terms are violated

In direct comparison to mere information retrieval systems or web search engines, which also have to deal with amounts of data in a petabyte range, a corpus analysis system for scientific linguistic research has to meet some additional requirements, as for example (see also Kilgarriff, 2007):

- results must be exact and reproducible
- function words cannot be ignored
- indexing has to deal with very unfavourable key distributions
- data structures are more complex: multiple layers and relations on and among annotations have to be represented
- query language needs to be more powerful
- the order of the presentation of search hits has to be controllable, in particular random samples of hits are required

With these additional requirements, at least some commonly used technical tricks and shortcuts for handling large-scale text databases will not be applicable.

6. Conclusion

Working on building up corpora since 1964, the IDS has gathered a lot of experience in handling language resources in a sustainable fashion. Despite all difficulties with copyright and licensing, the IDS was and is able to create a large language data resource, which allows for a more empirical approach towards linguistics. The key requirement of sustainability of DeReKo is a continuous maintenance of both the static and the dynamic language resource components and its usefulness for and its usability by its target community, i.e. empirical linguists working on German. To ensure this also for the future, the IDS will start to develop a new corpus management and analysis software. Moreover, the IDS is involved in different infrastructure activities towards sustainability and accessibility of language resources, e.g. in the nestor initiative, WissGrid, TextGrid, and CLARIN.

7. References

Belica, C., Kupietz, M., Lüngen, H., Witt, A. (to appear in 2010). The morphosyntactic annotation of DeReKo: Interpretation, opportunities and pitfalls. In: Konopka, M., Kubczak, J., Mair, C., Šticha, F., Wassner, U. (eds), Selected contributions from the conference Grammar and Corpora 2009, Tübingen. Gunter Narr Verlag.

BNC Consortium (2007). The British National Corpus, version 3 (BNC XML Edition). Distributed by Oxford University Computing. http://www.natcorp.ox.ac.uk

Broeder D., Declerck T., Kemps-Snijders M., Keibel H., Kupietz M., Lemnitzer L., Witt A., Wittenburg P. (2007). Citation of electronic resources: Proposal for a new work item in ISO TC37/SC4.ISO TC37/SC4-Document N366. http://www.tc37sc4.org/new_doc/ISO_TC37_SC4_N366_NP_CitER_Annex.pdf

Ide, N., Bonhomme, P., Romary, L. (2000). XCES: An XML-based encoding standard for linguistic corpora. In: Proceedings of the Second International Language

[1] nestor – German competence network for digital preservation: http://www.langzeitarchivierung.de/eng/

[2] WissGrid – Grid for Science: http://www.wissgrid.de/index_en.html

Resources and Evaluation Conference (LREC'00), Paris. European Language Resources Association (ELRA).

IDS (1991–2008): COSMAS I/II (Corpus search, Managment and Analysis System). Institut für Deutsche Sprache. Mannheim. http.//www.ids-mannheim.de/cosmas2/

IDS (2009). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2009-I (Released 28.02.2009). Institut für Deutsche Sprache. Mannheim. http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html

IDS (2010). Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2010-I (Released 02.03.2010). Institut für Deutsche Sprache. Mannheim. http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html

ISO/IEC (2009). ISO/DIS 24619: Language resource management – persistent identification and access in language technology applications. Technical report, International Organization for Standardization, Geneva, Switzerland, 4. September.

Kilgarriff, A. (2007). Googleology is Bad Science. In: Computational Linguistics 33 (1). S. 147–151.

Kupietz, M., Keibel, H. (2009). The Mannheim Reference Corpus (DEREKO) as a Basis for Empirical Linguistic Research. In: Working papers in corpus-based linguistics and language education. Tokyo University of Foreign Studies (TUFS)

Kupietz, M., Witt, A., Belica, C., Keibel, H. (2010). The German Reference Corpus DEREKO: A Primordial Sample for Linguistic Research. In: LREC 2010 Main Conference Proceedings. Malta

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In International Conference on New Methods in Language Processing, pages 44–49, Manchester, UK.

Sun, S., Lannom, L., Boesch, B. (2003). Handle System Overview. Number 3650 in Request for Comments. IETF, http://www.ietf.org/rfc/rfc3650.txt.

The TEI Consortium, editor. 2007. Guidelines for Electronic Text Encoding and Interchange (TEI P5). The TEI Consortium. http://www.tei-c.org/Guidelines/P5/.

Witt, A., Rehm, G., Hinrichs, E., Lehmberg, T., Stegmann, J. (2009). SusTEInability of linguistic resources through feature structures. In: Literary & linguistic computing : LLC ; journal of the Association for Literary and Linguistic Computing, 24 (2009) 3, pages 363–372

Sustaining a Corpus for Spoken Turkish Discourse: Accessibility and Corpus Management Issues

Şükriye Ruhi[a], Betil Eröz-Tuğa[a], Çiler Hatipoğlu[a], Hale Işık-Güler[a], M. Güneş Can Acar[b], Kerem Eryılmaz[a], Hümeyra Can[c], Özlem Karakaş[a], Derya Çokal Karadaş[a]

[a] Middle East Technical University, [b] Ankara University, [c] Hacettepe University

[a] İnönü Blvd., 06531 Ankara, Turkey; [b] Ankara Üniversitesi İletişim Fakültesi, 06590 Cebeci Ankara, Turkey; [c] Hacettepe University, 06800 Beytepe, Ankara, Turkey

E-mail: sukruh@metu.edu.tr, beroz@metu.edu.tr, ciler.hatipoglu@gmail.com, hisik@metu.edu.tr, acargunes@gmail.com, keryilmaz@gmail.com, hmeyra.can@gmail.com, ozlm.krks@gmail.com, deryacokal@gmail.com

Abstract

This paper addresses the issues of the long-term availability of language resources and the financing of resource maintenance in the context of the web-based corpus management system employed in the Spoken Turkish Corpus (STC), which operates with EXMARaLDA. Section 2 overviews the capacities of the corpus management system with respect to its software infrastructure, online presentation, metadata management, and interoperability. Section 3 describes the plan foreseen in STC for sustaining the resource, and dwells on the ethical issues surrounding the conflicting demands of free resources for non-commercial research and resource maintenance.

1. Introduction

A set of intertwined and pressing issues need to be tackled in the production of corpora that aim to be freely available for non-commercial research (cf. Haugh, 2009) over long periods. Based on our experience emerging from the ongoing construction of the Spoken Turkish Corpus (STC) within the Middle East Technical University Spoken Turkish Corpus Project (METU-STC), we highlight the need to develop corpus management systems that are accessible to (non-expert) corpus production and annotation work teams, the crucial role of open source software with interoperability capacities, and the financing of corpora maintenance and development.

Section 2 very briefly overviews the content of STC. Section 3 describes the web-based corpus management system developed for the corpus. In Section 4, we present the current plans for sustaining STC and suggest ways for reconciling the demands of ensuring that language resources are free for non-commercial research exploitation with those of the financial exigencies of resource maintenance.

2. STC: Design Features

STC stems from the first project in Turkey aiming to produce a relatively large-scale, general corpus of spoken Turkish discourse. In its initial stage, the corpus is designed to consist of one million words of present-day face-to-face and mediated interactions in Turkish in both formal and informal communicative settings. It is a multi-modal resource that presents transcriptions in a time-aligned manner with audio and video files. A more detailed description of its design features are found in Çokal Karadaş and Ruhi (2009).

STC employs EXMARaLDA (Extensible Markup Language for Discourse Analysis), which is a system of data models, formats and tools for the production and analysis of spoken language corpora (see, Schmidt (2004, 2005) and for a detailed description of EXMARaLDA). Informed by the transcription conventions in a previous corpus project, "Interpreting in Hospitals", which includes interactions in Turkish in Germany, the corpus is currently being transcribed and annotated with an adapted form of HIAT for utterances, utterance boundaries, pauses, overlaps, repairs, interruptions, and frequently occurring paralinguistic features such as laughing and certain emotive tones (see Schmidt (2008) for an overview of the basics of the current HIAT system). In its adult stage it will be a corpus annotated for morphology, the socio-pragmatic features of Turkish (e.g. address terms, (im)politeness markers, and a selection of speech act realizations), anaphora, and gestures.

The construction of STC is taking place in a work environment where little standardization in spoken language transcription with computer-assisted tools is available (Hatipoğlu and Karakaş, 2010; Işık-Güler & Eröz-Tuğa, 2010). Furthermore there few to no resources providing quantificational data on the production and reception of spoken domains and genres (Ruhi and Can, 2010; Ruhi, Işık-Güler, Hatipoğlu, Eröz-Tuğa & Çokal Karadaş, 2010), which means that basic research in these areas need to proceed concurrently with the production process. The research, annotation, and recorder teams, on the other hand, involve both expert and non-experts: linguists with a specialization in pragmatics and conversation analysis, IT infrastructure experts and programmers, (under)graduates and professionals in language studies and other areas, and volunteers from the general public interested in supporting the corpus production throughout its various stages. Thus two of the foremost priorities of METU-STC were the development of a workflow and corpus management system that could cater to the needs of this type of environment. More detailed descriptions of STC and its workflow are presented in Acar and Eryılmaz (2010).

3. The Web-based Corpus Management System for STC (STC-CMS)

STC-CMS is a web-based system that was developed and is being improved to make the process of the management and the monitoring of corpus production easy, transparent and consistent for team members who are not specialists in the technology of digital architectures. The system is designed with the goal of maximum automation and validation, as well as a clearly defined, traceable workflow, which enables monitoring of the design parameters of the corpus and the progress of the workflows, and the maintenance of consistency in production (see Fig. 1). The system thus also enables an "agile" (Voorman & Gut, 2008) workflow for controlling representativeness, which as underscored by Leech (2007) and Čermák (2009), remains a central issue in spoken corpora production.

As STC employs EXMARaLDA, a central function of STC-CMS is to achieve integration with its tools. STC-CMS performs this by generating EXMARaLDA compatible transcription and corpus metadata files.

3.1 File creation and interoperability

The system enables smooth control of the media and metadata files through a web interface and a relational (MySQL) database for metadata. Contributors submit recordings and metadata through the web forms, where they are validated and added to the database. At that stage, STC-CMS generates the EXMARaLDA compatible transcription files, which makes it possible to use EXMARaLDA tools and formats in STC. When a transcription file is submitted, it is checked into an SVN system for backup measures.

Being an open source system, EXMARaLDA and its associated tools do not pose the risk of being unavailable in future, and they can be sustained by other programmers. When fully checked for system operation features, STC-CMS will function as an open source project for further enhancement of its capacities and use by other resource producers who may wish to contribute to the STC database or who may wish to develop their own.

Using various file and data formats, STC tries to minimize the risk of digital obsolescence. Amongst its notable features, the system allows any subset of the corpus to be defined and published using EXMARaLDA libraries through a password restricted web site, where anyone with a web browser may access the corpus.

The sustainability of STC is also enhanced by employing EXMARaLDA's various export options (see Fig. 1). Transcriptions can be exported to HTML, PDF, RTF, TEI-compliant and XML-based EXMARaLDA formats, which ensure accessibility, long-term archivability and interoperability (see Schmidt (2005) for a detailed description of the relation between EXMARaLDA and TEI formats). The system also harnesses EXMARaLDA's capabilities for exporting to different transcription systems like Praat, ELAN, and TASX Annotator.

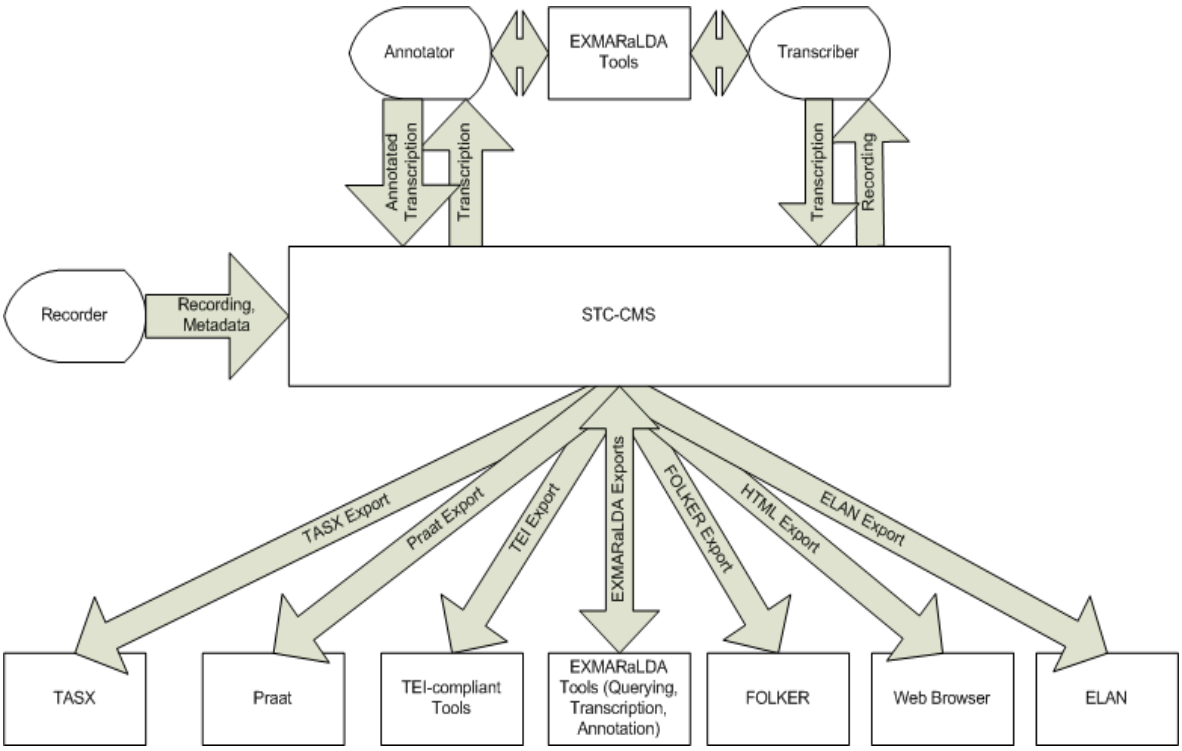


Figure 1: STC workflow and interoperability

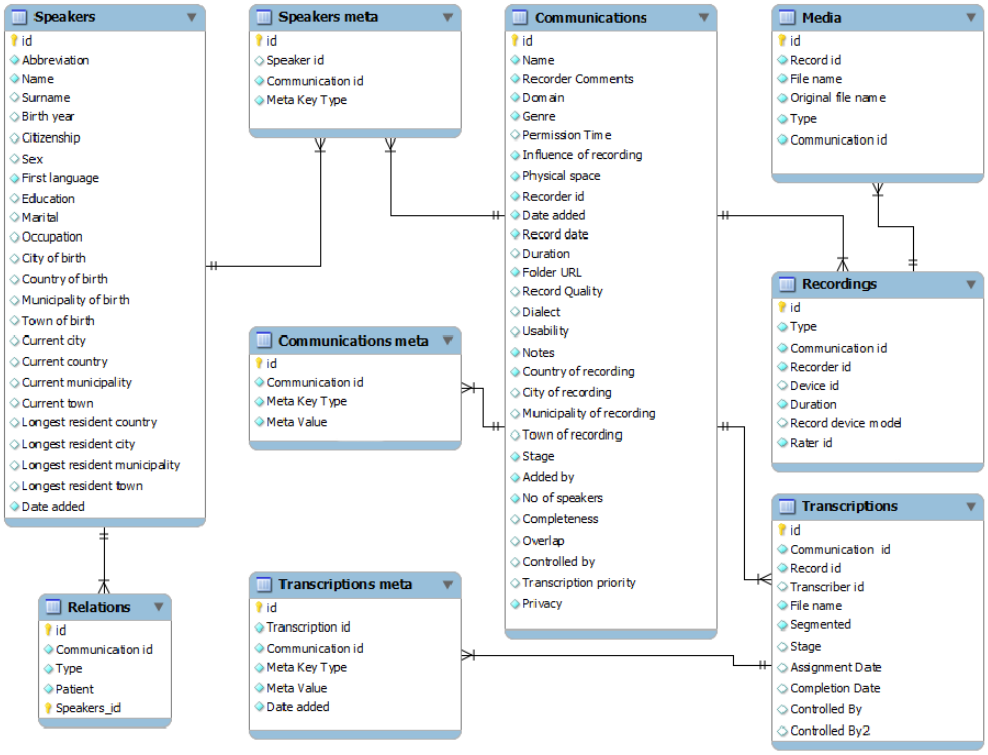


Figure 2: The database structure of STC-CMS

3.2 Metadata in STC

Given the crucial role of standardization in the maintenance of language resources, a few notes on the current state of the metadata in STC are due. The STC metadata fields have been defined through comparisons with the spoken component of BNC, the ISLE Meta Data Initiative (IMDI), sociolinguistic and pragmatics studies in the Turkish context (Ruhi, Işık-Güler, Hatipoğlu, Eröz-Tuğa & Çokal Karadaş, 2010), and the standard fields in COMA – EXMARaLDA's corpus manager tool (see Fig. 2 for an overview of the fields).

In addition to including classificatory and descriptive information on both the recordings and the speakers in the communications, STC follows the practice of providing an overview of the corpus in terms of communication categories, distribution of gender and age (see, e.g., the Spoken Dutch Corpus; Oostdijk, 2000). The METU-STC web site currently presents the overall features and communication types in the DEMO version of STC, along with the projected corpus design, the terms of use and information on copyright holders. Detailed information concerning the corpus design, and the transcription and annotation conventions will be added in its final version.

In regard to the formal properties of the metadata, COMA allows for the addition of Dublin Core (DC) fields to the coma file. Plans are being made to develop a web-based rather than a file-based system for search purposes once a standard metadata format has been decided upon. At present, our experience is that there are considerable differences in guidelines proposed by various spoken language resource initiatives such that an early commitment to any one of these might prove problematic in the long run (see, for example, BNC, IMDI, the Bavarian Archive for Speech Signals, and the parameters discussed in Čermák (2009) for spoken corpora). Given that the purposes for resource production are more varied in the case of spoken language corpora compared to written language corpora, this situation is understandable. In this regard, we find Schmidt's (2010) call for a concerted effort to achieve a "stepwise approximation" between the practices of communities of resource producers, users and language technologists a viable route to be pursued.

4 Accessibility and resource maintenance

STC-CMS and the interoperability capacities of EXMARaLDA are closely linked to the second major issue addressed in the paper: that of ensuring the long-term availability of free resources for non-commercial research. In our context, STC is a product that is being funded over two years by a national institution. It is obvious that this duration is vastly inadequate to achieve the long-term objective of extending the corpus to the size of ten million words. On the positive side, though, the project is hosted by the Dept. of Foreign Language Education at METU, which has a strong incentive to support language research and language resources, and the core research team consists of faculty members at the department. It should thus be possible to maintain the laboratory conditions and the required research activities for the expansion of the

resource. Funding for continued infrastructure maintenance and tool development, for example, will be secured through a variety of long-term projects related to STC.

STC is being constructed with recordings donated by individuals and media institutions. So it is imperative both to protect the royalty rights of contributors and to remain prepared for the possibility of a fluctuating production team. STC is built on the understanding that copyright owners of the various versions of the corpus and its sub-corpora will distribute the corpora freely for non-commercial research purposes. Other uses of the corpus (e.g. materials development in educational settings, NLP commercial applications and products derived therein) will be commercialized for the sole purpose of corpus maintenance and research directly impinging on the development of the corpus. Such commercial uses will be handled through various presentation types at different rates depending on the purpose of commercialization (e.g. internet access and cd/dvd formats; availability of either the whole corpus or sub-corpora; educational vs. non-educational purposes).

To further tackle the challenge of keeping STC a free resource while ensuring its expansion, the present copyright holders are planning to use a combination of CreativeCommons licenses in the forthcoming stages. Amongst the various license options that would allow for expansion of STC it appears that "Attribution Non-Commercial Share Alike" (cc by-nc-sa) provides a practical solution. This option allows for derivative work under the same conditions of the original terms of use. Such multiple availability options may respond to the demands of differing legal strictures and ethical stances both across language resource production communities and across national systems in the sharing and development of resources.

5 Concluding Remarks

Several issues remain to be resolved concerning sustainability, and funding is certainly a pressing issue. However, our experience with STC suggests that standardization in metadata and annotation, and by consequence, the development of tools with interoperability capacities, are by far more crucial in the current state-of-affairs. In this regard, we suggest that collaboration amongst the various stakeholders should involve not only resource producers and experts in digital architectures, but also the user end.

6 Acknowledgements

STC is financed by a research grant from the Turkish Scientific and Technological Research Institution (Türkiye Bilimsel ve Teknolojik Araştırma Kurumu, TÜBİTAK, Grant No. 108K285). We are deeply grateful to Dr. Thomas Schmidt and Dr. Kai Wörner for being a part of the project research team during October 2008–November 2009 and for their ongoing support in the construction and annotation of STC.

7 References

Acar, G.C., Eryılmaz, K. (2010). Sözlü Derlem için Web Tabanlı Yönetim Sistemi. Paper presented at the XXIV. Linguistics Conference, Middle East Technical University.

Bavarian Archive for Speech Signals. http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/

BNC. www.natcorp.ox.ac.uk

Bührig, K., Bernd, M. Interpreting in Hospitals. http://www.exmaralda.org/corpora/en_sfb_k2.html

CreativeCommons. http://creativecommons.org/

Čermák, F. (2009). Spoken Corpora Design: Their Constitutive Parameters. *International Journal of Corpus Linguistics* 14(1), pp. 113--123.

Çokal Karadaş, D. Ruhi, Ş. (2009). Features for an internet accessible corpus of spoken Turkish discourse. *Working Papers in Corpus-based Linguistics and Language Education* 3, pp. 311--320.

EXMARaLDA. http://exmaralda.org/

Hatipoğlu, Ç, Karakaş, Ö. (2010). Sözlü Derlem Çeviriyazısını Standart Dil ve Ağıza Göre Ölçünleştirme. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

Haugh, M. (2009). Designing a Multimodal Component of the Australian National Corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian Corpus*. Somerville, MA: Cascadilla Proceedings Project, pp. 74--86.

Işık-Güler, H., Eröz-Tuğa, B. (2010). Çeviriyazıda Geribildirim, Durak, Sesli Duraklama ve Ünlemlerin Ölçünleştirilmesi. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

ISLE Meta Data Initiative. http://www.mpi.nl/IMDI/

Leech, G. (2007). New resources, or just better old ones? The Holy Grail of representativeness. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, pp. 133--149.

Middle East Technical University Spoken Turkish Corpus Project. http://std.metu.edu.tr/en/

Oostdijk, N. (2000). Meta-Data in the Spoken Dutch Corpus Project. LREC 2000 Workshop, Athens. http://www.mpi.nl/IMDI/documents/2000%20LREC/oostdijk_paper.pdf

Ruhi, Ş. Can, H. (2010). Sözlü Derlem için Veribilgisi Geliştirme: Bağlam ve Tür Kavramlarına Derlem Dilbilimi Açısından Bir Bakış. Paper presented at XXIV. Linguistics Conference, Middle East Technical University.

Ruhi, Ş., Işık-Güler, H., Hatipoğlu, Ç., Eröz-Tuğa, B., Çokal Karadaş, D. (2010). Achieving Representativeness Through the Parameters of Spoken Language and Discursive Features: The Case of the Spoken Turkish Corpus**.** Paper presented at II. International Conference on Corpus Linguistics, Universidade da Coruña.

Schmidt, T. (2004). Transcribing and Annotating Spoken Language with EXMARaLDA. In *Proceedings of the*

LREC-Workshop on XML based richly annotated corpora, Lisbon 2004, Paris: ELRA. http://www1. uni-hamburg.de/exmaralda/Daten/4D-Literatur/ Paper LREC.pdf

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.

Schmidt, T. (2008). Overview of HIAT transcription conventions. http://www1.uni-hamburg.de/exmaralda/files/HIAT_EN.pdf

Schmidt, T. (2010)**.** Linguistic tool development between community practices and technology standards. Paper to be presented at the "Standardising Policies within eHumanities Infrastructures" Workshop at LREC 2010.

Voormann, H., Gut, U. (2008). Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory* 4(2), pp. 235--251.

The BAsic Metadata Description (BAMDES) and TheHarvestingDay.eu: towards sustainability and visibility of LRT

Carla Parra, Marta Villegas, Núria Bel
Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra, Barcelona, Spain
E-mail: carla.parra@upf.edu, marta.villegas@upf.edu, nuria.bel@upf.edu

Abstract

In this paper we address the lack of long-term accessibility plans to ensure the visibility of language resources and tools once they are finished. We believe that a change of strategy is needed: resource and technology providers must be aware of the importance of ensuring the visibility of their resources and tools, as well as the documentation thereof. In order to achieve this aim, we propose the usage of a metadata schema specially designed for describing language resources and tools with a minimum set of metadata, the BAMDES schema (BAsic Metadata DEScription). As the BAMDES *per* se does not solve the problems we intend to address, it is actually exploited by the so-called Harvesting Day initiative, a metadata harvesting routine based on the OAI-PMH protocol for metadata harvesting. This initiative will provide the main resource and tool catalogues and observatories with the harvested BAMDES descriptions and will ensure that these data are up-to-date thanks to a periodic harvesting routine.

1. Introduction

The research reported in this paper is deeply related to the results of a survey carried out within the FLaReNet[1] project as well as to our research activities within the CLARIN Project[2]. As far as the above-mentioned survey is concerned, we gathered information on the different existing linguistic resources and tools, such as the number of words of the different corpora, the tagsets employed, the usage of standards and metadata, etc. Even though this survey was not aimed at assessing the usage of metadata and standards in the description of linguistic resources and tools, we did realize the importance of this fact, as it actually became one of our main obstacles when carrying out our survey[3]. Moreover, these features are actually crucial elements as far as usability, accessibility, interoperability and scalability are concerned.

One of the conclusions we reached is the lack of long-term accessibility plans to ensure the access to the resource once it is finished. Even though usability, accessibility, interoperability and scalability are some of the key issues addressed nowadays in our field, a common strategy and effort shall be made for it to become true. Here we present an initiative, The Harvesting Day, aimed at enabling main resource and tools catalogues and observatories to harvest the key features of the resources and tools available at the same time as we ensure these data are up-to-date thanks to a periodic harvesting routine. This routine will thus ensure the accessibility of resources and tools and will also ease interoperability checks, as all those resources will be described by means of a BAsic Metadata DEScription (BAMDES), common to all of them and providing the main features for each resource and tool.

In what follows, this paper is organized in 3 further sections. In section 2 we explain the BAMDES concept, its main features as well as its relation to other metadata initiatives. In section 3 we present The Harvesting Day initiative and finally in section 4 we discuss the main conclusions that can be reached from the research reported here and why the BAMDES and The Harvesting Day initiative shall be further promoted and supported by researchers.

2. The BAsic Metadata DEScription (BAMDES)

2.1 Rationale and preliminary study

We are conscious of the fact that the problem we intend to solve here is not a novel problem in our field and that it has been addressed several times along time. Gavrilidou and Desypri (2003) already mention the majority of those initiatives in Deliverable 2.2 of the ENABLER project.

In fact, it must be acknowledged that the ENABLER Project did also a major effort as regards to visibility and sustainability. During this project not only were the main

[1] FLaReNet (Fostering Language Resources Network) is a European Project aiming at developing a common vision of the field of Language Resources and Language Technologies and fostering a European strategy for consolidating the sector, thus enhancing competitiveness at EU level and worldwide.

[2] The CLARIN (Common Language Resources and Technology Infrastructure) project is a large-scale pan-European project that aims at fostering the use of technologies in the Humanities and Social Sciences research.

[3] For more information on this survey, please refer to FLaReNet's Deliverable 6.1a: *Survey and assessment of methods for the automatic construction of LRs. Report on automatic acquisition, repurposing and innovative proposals for collaborative building of LRs*.

encoding and metadata initiatives reviewed and a complete metadata schema for describing LRT was proposed, but also a Declaration on Open Access to Language Resources was made. In it, it is stated that "the work of many initiatives and surveys such as the one of the ENABLER project show very clearly that the general information about the existence and the nature of most language resources is very poor. Only a small fraction of them is visible for interested users"[4].

Six years after the completion of ENABLER the situation has improved and initiatives such as ELRA's Universal Catalogue, CLARIN's Virtual Language World, DFKI's Natural Language Software Registry, etc. are involved in gathering information about resources and technologies.

However, and notwithstanding the important efforts that the main resource and tools catalogues and observatories do, the costs of curating and maintaining these catalogues and observatories updated are considerably high, as the data they need to gather is usually hard to be found and resource and tools providers do not usually take the LR lifecycle management into account. As a result of this fact, gathering information and checking whether or not a resource is actually available becomes sometimes a nightmare. This situation is reported in FLaReNet's Deliverable D6.1a:

"The compilation of information for this first survey was harder than expected because of the **lack of documentation** for most of the resources surveyed. Besides, **the availability of the resource itself is problematic**: Sometimes a resource found in one of the catalogues/repositories **is no longer available** or simply **impossible to be found**; sometimes it is **only possible to find a paper reporting on some aspects of it**; and, finally, sometimes **the information is distributed among different websites, documents or papers at conferences**. This made it really difficult to carry out an efficient and consistent study, as the information found is not always coherent (e.g. not every corpus specifies the number of words it has) and sometimes it even differs from the one found in different catalogues/repositories."

The survey mentioned above covered a total of 728 resources coming from two repositories: CLARIN and ELRA , corresponding approximately to 46% of the data in the CLARIN repository and 31% of the ELRA one (as to September 2009).

Only 42 of these 728 resources (5.7%) make use of metadata for their description. Concretely, none of them uses the ENABLER proposal and the majority (4.8 %) uses the IMDI metadata scheme. Furthermore, these resources using IMDI are multimodal and oral corpora, in most cases hosted at the MPI for Psycholinguistics, which explains why the majority used the IMDI schema. Table 1 bellow summarises the usage of metadata according to the results of our survey.

METADATA SCHEME	No. OF RESOURCES
CHAT	1 (0.14 %)
IMDI	35 (4.8 %)
TEI	1 (0.14 %)
CHAT + IMDI	1 (0.14 %)
LDC	1 (0.14 %)
OWN METADATA	3 (0.41 %)

Table 1 - Metadata usage overview

We have acknowledged that even though clear and detailed metadata proposals such as ENABLER exist, resource providers do not use them, and thus there is not a homogeneous and consistent description of the different language resources and every resource provider describes their resources in a different way. Although there are very well documented resources and tools, the user usually has to browse several web pages or documents when trying to figure out whether or not the resource itself fits with his/her needs. Furthermore, as not every resource provider facilitates the same data, some features may be missing or very difficult to find. Thus, it is particularly difficult to find the same information in every resource and carrying out studies as the one we did or locating all the relevant resources for a particular research project becomes particularly tedious as a considerable amount of time has to be invested in order to gather all the information needed and reach useful results. Besides, sometimes this is not always achieved and we cannot guarantee that all relevant resources were consulted, as there does not exist a way to do so.

Another important issue we have considered worth mentioning is the fact that not every language resource produced and referenced to in a repository is easily found or accessible. It was sometimes the case that the website mentioned in a repository was no longer active or that the resource itself was only referenced to at some paper, but not found elsewhere. As we already know, it may be the case that after several years ever since the completion of a project, the resource is lost because no further maintenance is made, or the researcher who used to take care of the resource is no longer at the resource home institution and no one else takes care of its maintenance. Thus, one of the problems encountered during our survey was the persistence of and accessibility to language resources over time. However, as initiatives such as CLARIN aim precisely at tackling this problem, we will not elaborate this point further here.

For all the above mentioned problems, we believe that the way in which metadata like size (i.e. number of words/minutes, etc.), languages covered, annotation, type of annotation, use of standards, etc. is provided should be

[4] ENABLER Declaration (2003): http://www.ilc.cnr.it/enabler-network/documents/ENABLER_Declaration.zip

unified and somehow fixed in order to achieve a homogeneous description of all language resources which facilitates both browsing the different existing catalogues and locating the resources that a researcher needs to carry out his/her research appropriately. Of the metadata schema consulted in our study, we estimated ENABLER to be the most complete and useful one. Furthermore, and as we needed to describe the resources that will be included in the future CLARIN Infrastructure, we addressed our CLARIN-ES and CLARIN-CAT partners explaining them the need for metadata descriptions of their resources and asking them to fill in a form with the ENABLER metadata. Four months have elapsed since then and we only got three answers to our query. Besides, the accompanying comments to the filled-in forms clearly stated that the forms where sometimes too complex and that it was not always possible to find an appropriate value for every metadata attribute.

This situation led us to the conclusion that maybe currently existing metadata forms are too complex and therefore language resource providers do not bother in filling them in, and that the metadata usefulness is still not clear. On the other hand, using a common and detailed metadata schema would simplify both the documentation process of new linguistic resources and tools and their coherent description, so that they can be better compared and assessed. Furthermore, this would also allow for the existence of validation processes to guarantee that all possible resources have been consulted, and would enhance greatly the search through virtual repositories and infrastructures such as CLARIN. This would in turn improve the scope and quality of the research projects in our field, as they could make usage of all the relevant data and thus broaden their scope and obtain better results in terms of scope, quality and relevance. And this is how the idea of setting up at least a minimum metadata description and thus the BAMDES came up.

2.2 The BAMDES - unplugged

The BAMDES is a BAsic Metadata DEScription specially designed for describing language resources and tools. In order to determine which attributes where minimally required to describe any language resource and tool, we took the ENABLER proposal and shortened it up to a minimum list of attributes and values. We then proceeded to create mappings with relevant initiatives such as the Clarin Metadata Infrastructure (CMDI), ISOcat and Dublin Core. Thus, we adopted the attribute names designated by CMDI and ISOcat and created automatic mappings with Dublin Core to be as much standard-compliant as possible.

Similarly to the ENABLER proposal, we divided our metadata into two different types: External (common to every language resource and tool) and Internal (specific to every resource type). In the following figure we show its basic structure. Every resource has the external administrative metadata necessary to identify it, and is assigned a specific resource type, which, in turn, will have its internal metadata.

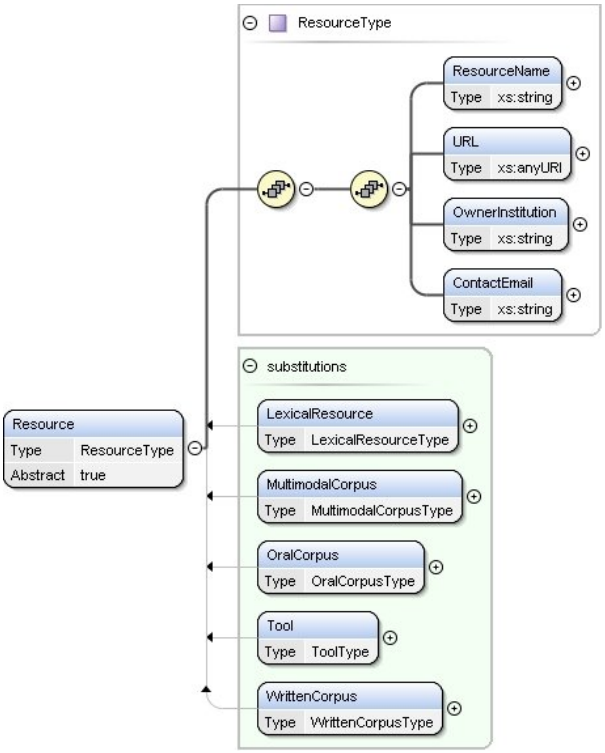


Figure 1 - BAMDES Schema overview

These metadata are the metadata necessary to describe the main features of the resource/tool. Each resource/tool has a different metadata set according to its intrinsic characteristics. The complete schema is available online for consultation/download[5].

3. The Harvesting Day Initiative

From our perspective, the BAMDES *per* se does not solve the problems we intend to address, although it does guarantee that language resources and tools are described in a common and structured manner. Thus, the BAMDES is actually exploited by the so-called Harvesting Day[6] initiative.

The Harvesting Day initiative is based on the OAI-PMH protocol for metadata harvesting. This protocol was developed by the Open Archives Initiative and it is widely used to harvest the metadata descriptions of the records in the archive. OAI-PMH implementations must support the metadata representations in Dublin Core, but may also support additional representations. In our case, The Harvesting Day initiative additionally supports the BAMDES representation. The OAI-PMH protocol is based on a client-server architecture in which harvesters request information on updated records from repositories. We provide some on-line forms that help LRT providers to edit their resource descriptions following the BAMDES schema. The output of these forms consists of

[5] http://gilmere.upf.edu/theharvestingday/schemas/clarin_bamdes.xsd
[6] www.TheHarvestingDay.eu

the corresponding xml descriptions validated against the schema and will constitute the records to be harvested. We also provide a self-executable package to be installed in the provider's server in order to make their metadata harvestable and ensure that every time a new harvesting date is set up, the harvesting robot comes to the provider's server and checks whether new resources have been added or updates to existing ones have been produced.

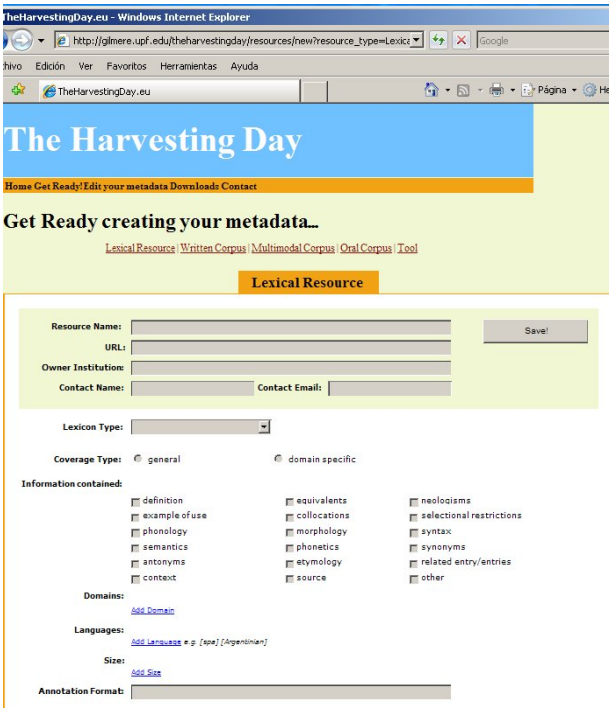


Figure 2 - Sample of online form - Lexical Resources

By means of this initiative, we encourage every resource/tool provider to create their own "metadata farm": an OAI-PMH Server that is ready to be harvested and thus will provide their metadata information when it is prompted to do so.

The OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting) is a standard widely used by digital libraries and other repositories. However, and as we know that not every resource/tool provider may have an OAI-PMH Server available, we also offer a downloadable, package[7] with a step-by-step guide to install the OAI Repository and add the BAMDES records to it.

Once the resource/tool provider has set up his/her own farm, they only have to communicate it to us, so that their farm will be included into the list of farms to be harvested automatically and in a periodic manner.

To sum up, The Harvesting Day initiative cannot be considered a stand-alone and single initiative, but actually a process that will be repeated periodically, thus ensuring not only its continuity, but also the validity and trustability of the data it will be gathering. Resource and tool providers will only need to keep their BAMDES records available at their own farms and the harvesting robot will be in charge of distributing that information. As previously mentioned, all the harvested metadata will be then provided to the main Language Resources and Tools catalogues and repositories thus reducing considerably their creation and curation costs while it is also guaranteed that their data are updated. Resource providers themselves are given a visibility they had not before, as their resources and tools will populate all major repositories and catalogues thanks to the distribution of all harvested BAMDES. In other words, a decentralized and single effort will be made in order to gather relevant data as far as language resources and tools, while at the same time this effort will enhance and guarantee the visibility of the resources and tools available at the different farms around the world.

4. Conclussion

The execution of our survey pointed out the need of at least a minimum set of unified metadata for the description of linguistic resources and tools, as well as the need of fostering the usage of standards in our community. These two factors would increase not only the quality of our resources and their description, but also their usability, accessibility and visibility, as the same features would be used to describe the same kind of resource. The usage of a BAMDES as we propose here to describe the resources and tools and the possibility to harvest those metadata periodically by means of The Harvesting Day initiative will not only ensure that registries and catalogues are provided with the utmost updated data, but also would guarantee their appropriate visibility, thus improving considerably their lifecycle management too.

For all of the above mentioned issues, we believe that a change of strategy is needed: resource and technology providers must be aware of the importance of ensuring the visibility of their resources and tools, as well as the documentation thereof. From our perspective, providers must be responsible for at least a BAsic Metadata DEScription (BAMDES) of their resources and tools. We propose to start a decentralized effort of resource description and to launch an automatic, periodical information gathering routine of those BAMDES. This will eventually become an automatic assessment routine that will enable catalogues and repositories to track every resource and tool that has been harvested once and thus check whether it is still available, new releases have been made, etc.

Finally, we must acknowledge the fact that the success of this initiative greatly depends on the response we obtain from the audience we are addressing. That is why we are making great efforts to achieve that all relevant agents adhere to this initiative and thereby show their interest and support for the BAMDES and The Harvesting Day. It is our responsibility that the language resources and tools

[7] http://gilmere.upf.edu/theharvestingday/farm

of the future are better documented and visible, this initiative is a great step towards this final aim.

5. Acknowledgements

This project has been made possible by the funding of the CLARIN project. The CLARIN project in Spain is co-funded by the 7FP of the EU (FP7-INFRASTRUCTURES-2007-1-212230) and the Spanish *Ministerio de Educación y Ciencia* (CAC-2007-23) and *Ministerio de Ciencia e Innovación* (ICTS-2008-11). Furthermore, the *Departament d'Innovació, Universitats i Empresa* of the *Generalitat de Catalunya* has funded the development of a demonstrator that guarantees the integration of the Catalan language in CLARIN.

The authors wish to thank their colleagues Santiago Bel and Silvia Arano from the Language Resources Technologies Research Group of the Institut Universitari de Lingüística Aplicada at the Universitat Pompeu Fabra their support during the preparation of this paper as well as their active contribution in the development of the initiatives reported here.

6. References

Australian National Data Service (ANDS) : http://ands.org.au/

CLARIN Project : www.clarin.eu

FLaReNet Project : www.flarenet.eu

Gavrilidou, M. and Desypri, E. (2003). *Deliverable D.2.2 - Report for the definition of common metadata description for the various types of national LRs*, ENABLER project. [http://www.ilc.cnr.it/enabler-network/documents/ENABLER_D2.2_FinalVersion.zip]

ISOcat : www.isocat.org

Language Grid : http://langrid.nict.go.jp/en/index.html

Language Resources Metadata Database : http://facet.shachi.org/

McCormick, S. (2005). *The Structure and Content of the Body of an OLIF v.2.0/2.1*. [http://www.olif.net/documentation.htm]

Parra, C., Bel, N., Quochi, V. (2009). D6.1a: Survey and assessment of methods for the automatic construction of LRs. Report on automatic acquisition, repurposing and innovative proposals for collaborative building of LRs. FLaReNet Project. [http://www.flarenet.eu/sites/default/files/D6.1a.pdf]

Project Bamboo : http://projectbamboo.org/

The Harvesting Day Initiative : www.TheHarvestingDay.eu

VV.AA. (2003) *ENABLER Declaration on Open Access to Language Resources*. [http://www.ilc.cnr.it/enabler-network/documents/ENABLER_Declaration.zip]

Wittenburg, P., Broeder, D., and Sloman, B. (2000). *EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources, White Paper*. LREC 2000 Workshop, Athens.