Design Features for the Collection and Distribution of Basic NLP-Resources for the World's Writing Systems

Oliver Streiter 1*, Mathias Stuflesser 2†

*National University of Kaohsiung, Taiwan
ostreiter@nuk.edu.tw
† European Academy of Bozen Bolzano, Italy
mstuflesser@eurac.edu

Abstract

Most of the world's 7000 languages are still lacking freely available language resources. This lack of resources forms a major bottleneck in the processing of those languages and prevents them form being more widely used. To overcome current limitations, researchers might profit from studying the cooperation in free software projects or Wiki-projects. In this paper, we explore such models of 'organic' cooperation for the collection and elaboration of free NLP-resources. We describe the database XNLRDF which has been set up for this purpose. Storing NLP-data for hundreds of languages, we gradually refined and extended the idea of what kind of information has to be included in such a database, how the information is to be stored and how such data might be created in an organic cooperation. A principled distinction we make is that between the data structure used for development and the data structure used for distribution, a relational database and XML-RDF respectively. Taking the advantages of XML for granted, we explain the advantages of a relational database for the development and maintenance of collaboratively developed data. Within the data structure, the notion of 'writing system' functions as pivot. A writing system incorporates a set of metadata such as language, locality, script, orthography, writing standard and assigns them to the NLP resources provided in XNLRDF. An overview over the first data we collected and an outlook on future developments will conclude this paper.

1. Old and New Research Traditions

From the 7000 languages of the world, about 1000 are estimated to have shown up on the Internet[1]. This high number reflects the pride of people in their culture and their willingness to use their language in electronic medias for communication and learning. It also represents the economical and ideological interest in most languages as a means to contact, inform or persuade people. However, many languages are not supported in their digitalized form. Computer users might be able to input the characters of a writing system, sometimes with difficulties (Uchechukwu, 2005), but overall there are no spell-checkers, grammar checkers, information retrieval systems or translation dictionaries.

This deplorable situation is not an exception for one remote language spoken far away, but is reality for more than 99% of the world's languages, a fact not taken into consideration by 99% of the NLP community. This distortion, one might call it even a caricature, is not due, as one might assume, to a lack of money, a lack or scientific interest, a lack of commercial interest or a lack of linguistic knowledge. In fact, many languages have been scientifically described with great care. In addition, each speaker of a language is a potential client for a soft drink, a political movement or a religious community and could thus best be addressed in his or her native tongue. Instead, the misery is rooted in our research tradition.

This research tradition will change however under the influence of free software projects, blogs, Wikis and creative commons licenses (Streiter, 2005; Streiter et al., 2006): Academical hierarchies, the distinction between affiliated scientist, enthusiast and partisans, the attribution of a scientific work to a researcher or a research body, the search for research topics in predefined academic fields and modular models of cooperation in research projects will become less pervasive and thus might pave the way for new models of scientific cooperation.

To explore the potentials of this new modes of research and to bridge the gaps between a) the needs of languages users, b) affiliated research and c) the potential contributions of non-affiliated researchers, we started to create an environment for an organic cooperation through the Internet with the aim of collecting and elaborating NLP-resources for the world's 8000 languages. The created NLP-resources are available in hourly builds under the GNU public license[2] and intellectual insights related to the development of the resources are available under the Creative Commons License[3]. Currently, the discussion of data structures and the collection of the first data is still done by a small circle of volunteers. But we hope that the circle of interested people might gradually enlarge, to open up finally for a free Wikipedia-like cooperation.

2. More is Different

The project XNLRDF (Natural Language Resource Description Framework) thus develops, breadth-first, an NLP-infrastructure for the world's writing systems, and, not tackled for the moment, the world's speech systems. Exploring the world's writing systems in all their differences and particularities, we hope to define a stable framework which can accommodate the most unusual cases without having to redefine the basic model or to compromise the data. While registering 23.000 writing systems, 8200 languages, 150 scripts and textual examples of 700 writing

[1] http://www.guardian.co.uk/GWeekly/Story/0,3939,427939,00.html

[2] http://140.127.211.214/research/nlrdf_download.html

[3] http://xnlrdf.wikispaces.com

systems, we have been forced to rethink and adapt our notions of (i) linguistic metadata, (ii) the nature of basic NLP-data and (iii) the way data are created and managed.

2.1. Different Metadata

In NLP, metadata identify the most suitable resources for the processing of text documents. In document formats like HTML, the language of a document is considered the most important kind of metadata. However, as the same language might be written at different times or in different places, before or after a spelling reform, before or after the adoption of a new alphabet or a new script, NLP-resources and text documents shouldn't use *language* as metadata, but a more specific notion, the ***writing system*** of a language.

To distinguish the estimated 100.000 writing systems of the world and to assign the most suitable resources, metadata have to be much more specific than what is currently used in the HTML header and even more specific than what has been suggested in the framework of OLAC (Simons and Bird (eds.), 2003). In addition, in case the processing of a document needs a resource of a given type, but no such resource has been explicitly assigned to the writing system of the document, inheritance principles are required that allow to assign to the document the most suitable resource from a related writing system. We thus currently define a *writing system* by the n-tuple of the more elementary metadata *language*, *locality*, *script*, *orthography*, the *writing standard*, the *time period* and a *reference* to another writing system. Each of these elementary metadata is identified by an arbitrary ID and ISO-codes, if available. Natural language names for these metadata are provided for convenience or in case standard codes are not defined[4] or ambiguous[5]. The choice of the natural language designators is not crucial as long as they are not pejorative or ambiguous. Designators in XNLRDF are provided in many languages (more precisely writing systems). One of the designators in the writing system *'late modern english_united states of america@latin'* is selected as being the default for generation, e.g. when generating a pick-list of language names.

Supporting evidence for the necessity of these elementary metadata comes from cases like Abkhaz: Abkhaz (*language*) has not only been written with two different Cyrillic alphabets (*script*), but also with two different Latin alphabets (*script*), one between 1926 and 1928 (*time period*) and one between 1928 and 1937. One might want to distinguish these writing systems by their name (the *standard*) or by the *time period*. In such cases, we do not exclude the first solution, although there is frequently no name for the standard. If possible, we prefer the time period, as it offers the possibility to calculate intersections with other time constraints, e.g. on the production date of the document, or the foundation or disintegration of a country or region.

The *writing standard* is best explained with the help of the different, concurring, isochronic writing standards for Norwegian (*language*): Nynorsk, Bokmål, Riksmål and Høgnorsk are different conventions (*writing standards*) to represent basically the same language[6].

[4]e.g. the Ladin variety of Gherdëina Valley

[5]e.g iso-639-1 codes of groups of languages

[6]cf. http://en.wikipedia.org/wiki/Norwegian

The *orthography* is best illustrated by the spelling reform of German with the new orthography coming into force in different *localities* (*'Germany', 'Austria', 'Liechtenstein'* ...) at different times and overlapping with the old spelling for a different number of years. In this case, disposing of the *time period* is a nice feature but it does not allow to dispense with the category of *orthography*. Unfortunately, orthographies, also frequently lack a standard name and are referred to as '*new*' in opposition to '*old*'.

The *reference* is a necessary metadatum to represent transliteration systems, i.e. transliterations in the strong meaning as one-to-one mapping, but also as one-to-many or many-to-many mappings. *'Braille'* is such a transliteration system which changes with the spelling reforms and standards of the referenced writing system. Thus, there exists one Norwegian Braille derived from Nynorsk and a second Norwegian Braille derived from Bokmål. By the same principle, Braille of the new German orthography is different from Braille based on the old German orthography.

Braille changes also when the *locality* of Braille is different from the *locality* of the referenced writing system. For example, Spanish Braille in a Spanish speaking country is different from the Spanish of a Spanish speaking country represented as Braille in the USA. This complexity can be handled when we allow writing systems to refer to each other. Thus Braille, as other transliteration systems, is represented as writing system with its own independent *locality, script* and *standard*, (e.g. '*contracted*' and '*non-contracted*'). The *language* and *time period* of the transliteration and the referred writing system are however the same.

A transliteration is thus marked by a reference to another writing system and mapping tables between the two systems, e.g. between Bokmål and Bokmål Braille. Mappings between writing systems are a natural component in the description of all writing systems, even if they do not represent transliterations of each other, e.g. mappings between *'hanyu pinyin', 'wade-giles'* and *'zhuyin fuhao'*.

Writing systems and, in the future, speech systems are identified by an arbitrary ID. They can be recognized by programmers through the concatenation of default natural language designators of the elementary meta data. Unspecified data are omitted, e.g. *'Uighur@Cyrillic', 'Uighur_Uzbekistan@Latin', 'Norwegian_Norway@Latin#nynorsk'*. NLP-resources are then described with respect to their *function*, their *encoding*, their *copyright*, their *URL*. NLP-resources accumulated within XNLRDF are encoded in UTF-8, distributed under the GNU Public license and associated with one or more writing system. The writing system is thus the pivot notion, which connects metadata and NLP resources.

2.2. Different NLP Data

Given the huge number of writing systems created by mankind, no property of a writing system is universally valid. E.g. the function of a white space, dash or dot varies between scripts, but also between languages of the same script. Writing systems differ also by the characters representing word/syllable boundaries, ciphers and number words, the writing direction (e.g top to bottom left to right for Mongolian) and the sorting of characters in a wordlist.

Unicode which is generally assumed to cover this information fails to provide this NLP relevant information. Unicode refers only to scripts and ignores the notions of language or writing system. Unicode thus assigns properties at a level of the *script*, where these properties can only be understood at best as a default for a writing system.[7] We thus observe a huge gap between what Unicode defines and want NLP resources normally assume to be defined, This gap has to be bridged by XNLRDF.

To test XNLRDF we create, in addition to the data, small applications which are supposed to work for all or most writing systems that have a minimum of data. The set of applications currently includes a language guesser and a spell checker. These and all other applications to follow use only the data available within XNLRDF and thus show whether or not all necessary data types and tokens are included. For instance, while creating the web-interface of the spell checker, we recently discovered that the writing direction has to be provided explicitly by XNRLDF and cannot be left to the discretion of Unicode, the word processor or the Web-browser. Sorting, the function of uppercasing and the relations between characters and writing directions (some Chinese characters change their shape when written vertically or horizontally) are further examples of what kind of writing system-specific NLP information is needed beyond what is traditionally included in NLP.

XNLRDF however will provide more than these very elementary data. We will try to create word lists, dictionaries, corpora, stemmers, morphological analyzers and taggers for each writing systems. The challenge will be to find uniform representations and procedures which can correctly handle the great variety of languages, and, of course, to find or create the necessary data.

2.3. Different Ways of Data Creation

Traditionally, coordinated research is funded by a body which, more often than not, wants its money to be invested in what it perceives to be relevant for the financial resources of that body. Thus, research in France, payed by French tax payers is more likely to create NLP-resources for French than for Khamtanga. This, as natural as it seems, creates however a distortion of the relation between actual requirements and funding. As a consequence of this self-centered perception, those languages, which have the smallest gaps receive most funding.

A second feature of traditional models is that the cooperation between research units is organized in modules. Research units are thus autonomous within their modules and interact with other modules through specific interfaces, standards or protocols. In this way, intellectual properties can be easily assigned to a research unit. In addition, the consistency and coherence of the data within one module seem to be manageable. However, this model cannot take direct advantage of closely overlapping, complementary intellectual competences.

In models of organic cooperation however, volunteers, which may be experts or not, cooperate on the realization of some content, be it software, lingware, translation, images or a new texts, despite the absence of any funding (Bey et al., 2005). The only criterion for setting the research topic is the perceived relevance by the volunteers who, although not free of any self-centered perception, can accommodate more easily to an unbiased view than a funding body can do. Thus, while in the institutional cooperation, no language resources are created for Khamtanga, except in Ethiopia itself, researchers from France and many other countries would contribute to the development of Khamtanga NLP-resource in the model of organic cooperation. As a consequence, the gap between actual needs and research activities becomes smaller.

The cooperation in projects of organic cooperation is not necessarily modular. Different people might work on submodules where the function of the submodule cannot be defined on the basis on its own. This way different knowledge resources can be merged and software can be used to minimize friction and inconsistencies. Especially promising thus seem relational databases as they allow for a maximal fragmentation on the one hand, but guarantee on the other hand consistency and coherence through the usage of uniqueness constraints, references and triggers.

In XNLRDF we attempt to follow a model of organic cooperation, for a number of reasons. Firstly, the sheer amount of data we aim at is far beyond what one even large research team can achieve. Secondly, the many different competences required can only be brought together in an open model of broad cooperation. Third, the cooperation of professional linguists and volunteer experts can help to improve the database infrastructure, keeping it simple at the interface, yet complex and coherent in the data model.

3. A Relational Database as Backbone

While metadata and the organization of linguistic data in XNLRDF is determined bottom-up, we have principled ideas about the overall project design. As backbone for data development serves a relational database, whereas XML is used for the exchange of data in RDF (Powers, Sh., 2003), hence the name XNLRDF. The database can already be downloaded as database dump or as a one-to-one representation of the database in XML. An RDF will be designed which, in order to avoid bulky downloads, will allow for extracts for single languages and writing systems.

The relational database, installed with one command (in `Linux`) and configured with a few clicks in `Webmin` offers a set of features which can hardly be matched by XML. A relational database is integrated in a client-server architecture and designed for collaborative work. A battery of off-the-shelf interfaces is available for different purposes and can be used over the Internet.

A further advantage are the internal checks for data-types, uniqueness, coherence and consistency at a level below the interface so that these checks are effective in all interactions with the database. These checks will be primordial for the quality of the data when a great number of people cooperate blindly on the same database. The checks can can be defined to any level of complexity using *triggers* and *functions*. For example, changing the time period of Middle English in XNLRDF will change the time period for Old English and Early Modern English as well (thus assuring

[7] For a more detailed discussion of the shortcomings of Unicode see (Streiter and Stuflesser, 2005).

the coherence). Any attempt at placing e.g. the writing system of Proto-Norse in the former GDR, however, is most likely to fail due to temporal or local constraints associated with the language and the locality. Organizing data into a network makes singular incorrect data modifications difficult or impossible. *Freezing* an ever growing amount of validated data in this network, will make the space for incorrect modifications smaller and smaller.
Creating ambiguous metadata becomes impossible through *uniqueness constraint*. *References* make it impossible to delete central data, e.g. a language referred to by a writing system. The inclusion of *false positives*, e.g. pejorative language names, marked as deleted, make it impossible to insert or inherit the same value again through the effect of *uniqueness constraints*. Overall, XNLRDF forsees the following hierarchy of collaborators.

All users can enter new data. In this work, users are guided by the XNLRDF-browser[8] which already assisted in the creation of the currently available data in XNLRDF. Step by step the browser will evolve into a Wikipedia-like workbench where linguists can store, elaborate and test linguistic data.

A group of experts in language, linguistic subfields, language groups etc has the power to 'delete' incorrect entries, i.e. to move them into the false positives, or to assign the status of 'unchangeable' to cornerstone data. These experts thus complete and guide the set of control mechanisms provided by the system by controlling the validity of the data.

A third group of language and database experts defines the constraints and inheritance mechanisms to account for the completeness and coherence of the data. All of this is fairly easy to realize within relational databases but probably unreachable for XML.

4. Achievements

We have created a basic architecture for the development of fundamental NLP-resources for the writing systems of the world that might be fit for a model of organic cooperation. The potential of these resources starts to get visible with an automatic writing system (language) recognizer for currently about 700 writing systems and an spelling-checker for more than 700 writing systems. Most of the texts collected in 700 writing systems in XNLRDF are parallel texts, providing a means to create translation dictionaries for thousands of language pairs. Pointers to websites which can be freely downloaded for corpus construction are available for over 150 writing systems. In addition, the database contains a first set of about 2000 number words in 29 writing systems and 900 function words in 25 languages.

5. Further Outlook

While currently the database still requires a password for most modifications, a number of minor modifications have been opened to be changed freely by everyone. A Wikipedia-like cooperation of researchers is thus getting more and more likely. Opening the system step by step we explore techniques for checking new data and the automatic creation of message to the controlling linguists. At the same time we try to estimate the impact of erroneous entries on the quality of the data.
The data collection has focused until now on finding textual examples. We will proceed to a linguistic analysis of these examples to prepare the creation of corpora, stemming, morphological analysis and tagging. This work will be supported by small tools which propose different analysis solutions to be selected by the linguist.
Through the integration of simple applications, which among others test and show the potential of XNLRDF, we want to motivate researchers to enter the required data e.g. to insert open-licensed texts of a language to download shortly later a simple spell-checker, or to enter morphemes to download a better morphological analyzer. Some applications, like the spell checker provide for an inherent feedback function through which more linguistic data can be collected, e.g. the confirmation of unknown words. In addition, as suggested to us by Trond Trosterud, linguists might use integrated parsers or morphological analyzers to test their theories and produce at the same time word lists, classified morphemes and formal linguistic rules.
We hope that the creation and collection of data will speed up and extend to more languages once the system has been opened for organic cooperation. But even now, collaboration, advice and assistance of any kind, related to datastructure, metadata, the creation of applications, designing the final RDF, contributing data etc. are more than welcome.

[8] http://140.127.211.214/cgi-bin/gz-cgi/browse.pl

6. References

Y. Bey, K. Kageurat, and Ch. Boitet. 2005. A framework for data management for the online volunteer translators' aid system QRLex. In *Proceedings of PACLIC 19, the 19th Asia-Pacific Conference on Language, Information and Computation*.

Powers, Sh. 2003. *Practical RDF*. O'Reilly.

G. Simons and St. Bird (eds.). 2003. OLAC metadata set. Technical report.

O. Streiter and M. Stuflesser. 2005. XNLRDF, the open source framework for multilingual computing. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.

O. Streiter, M. Stuflesser, and Q. L. Weng. 2006. Models of cooperation for the development of NLP resources: A comparison of institutional coordinated research and voluntary cooperation. In *Proceedings of the LREC workshop "Strategies for Developing Machine Translation for Minority Languages"*.

O. Streiter. 2005. Implementing NLP-projects for small languages: Instructions for funding bodies, strategies for developers. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.

Ch. Uchechukwu. 2005. The Igbo language and computer linguistics: Problems and prospects. In *Proceedings of the conference "Lesser Used Languages & Computer Linguistics"*.