Hooking up to the corpus: the *Viennese Lexicographic Editor's* corpus interface

Gerhard Budin[1,2], Karlheinz Mörth[1]
1 Institute for Corpus Linguistics and Text Technology (ICLTT), Austrian Academy of Sciences
Sonnenfelsgasse 19/8, 1010 Vienna, Austria
2 Center for Translation Studies, University of Vienna
Gymnasiumstraße 50, 1190 Vienna
E-mail: karlheinz.moerth@oeaw.ac.at, gerhard.budin@oeaw.ac.at, gerhard.budin@univie.ac.at

Abstract

The paper addresses the issue of interfacing between digital corpora and a new dictionary writing application being developed at the ICLTT (Institute of Corpus Linguistics and Text Technology of the Austrian Academy of Sciences). It deals with issues of dictionary creation, software design, usability and interoperability in relation to the example of this fairly new piece of software, the *Viennese Lexicographic Editor*. In addition, it outlines the ICLTT's projects which are using the new tool and explains the role of these undertakings as part of the ICLTT's involvement in the Austrian CLARIN-AT initiative. The focus of the discussion will be on the implementation of efficient workflows designed to streamline the transfer of corpus examples stemming from distributed online corpora into dictionary entries. An important additional topic is the access mode of digital corpora available on the internet and the issue of service-oriented software design. As a last point, we will also touch on the important issue of standards and de-facto standards used in these projects.

Keywords: dictionary editor; dictionary software; CLARIN; LRT standards; TEI

1. Introduction

The Institute for Corpus Linguistics and Text Technology (ICLLT), which was founded in early 2010, is among the youngest departments of the Austrian Academy of Sciences. It is the successor of the Austrian Academy Corpus (AAC) and has been designed to pursue research in a number of cross-disciplinary projects which cover a wide range of interests. Some of these projects focus on issues of corpustechnology, some on computational lexicography, yet others belong to the sphere of humanities computing, conducting investigations into digital editing and text encoding. Traditionally, research questions have been dealt with in an interdisciplinary manner, as the staff of the department is made up of scholars with a varied background in social sciences and humanities. As the department's name suggests, its primary mission is to conduct empirically based research into human language, in particular written language. Many of the department's projects are driven by lexicographic and terminological interests.

2. Print dictionaries

Over the past decades, the department has been involved in longstanding lexicographic projects which – in the recent past – have transformed into smaller, more diversified projects. Products of the earlier endeavours have been, for the most part, print dictionaries. The two "*Fackel* dictionaries"[1] came into existence as results of a large-scale text lexicographic experiment which was, in part, carried out in cooperation with the Academy's Commission for the Publication of a *Fackel*-Dictionary.

Another dictionary product to which the ICLTT contributed NLP and corpus technology is the hitherto largest German-Russian dictionary, which was published as a cooperative project of the Austrian and the Russian Academies of Sciences.[2]

3. Computational lexicography

The following paragraphs are meant to give a short overview of our lexicographic endeavours and to showcase some of the experiments in eLexicography that are currently being undertaken at the ICLTT.

3.1 Digitised historical dictionaries

Apart from the traditional print dictionary line described above, the department's activities in eLexicography derive from a second source: smaller historical dictionaries, which are part of the Austrian Academy Corpus[3], a digital German language corpus comprising of approximately half a billion tokens. The texts contained in this corpus were collected with a both literary and a lexicographic perspective in mind. The corpus also contains a considerable number of functional and informational texts. Roughly half of the data is made up of periodicals, not large-size daily newspapers but rather medium- and small-size weekly and monthly publications. There are many collective publications such as yearbooks, readers, commemorative volumes, almanacs, and anthologies covering a wide range of writers, topics, types of texts, and genres. While at a first glance the collection might appear heterogeneous, it actually represents a unique collection of historical German texts, many of which cannot be found elsewhere in digital form.

[1] W. Welzig (ed.): *Wörterbuch der Redensarten zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«* (Austrian Academy of Sciences Press, Vienna 1998) and *Schimpfwörterbuch zu der von Karl Kraus 1899 bis 1936 herausgegebenen Zeitschrift »Die Fackel«* (Ibid. 2008).

[2] D. O. Dobrovol'skij (ed.) Neues Deutsch-Russisches Grosswörterbuch. Moskau 2008-2010.

[3] Henceforth, AAC is used in the sense of corpus, not as the institutional name of the predecessor of the ICLTT.

As the corpus was built on a wide concept of literature, which principally included anything reduced to written form, a few monolingual and bilingual dictionaries were also incorporated in the collection. These constituted another starting point for our digital dictionary ambitions.

3.2 Digitally created lexicographic data

Furthermore, the ICLTT also holds some data that came into existence in the digital medium. These were the result of experiments with dictionary creation and dictionary enhancement through automatic and semi-automated procedures. All of the dictionary activities have been closely intertwined with the department's numerous corpus activities.

So far, the largest amount of manually created digital dictionary data produced at the department stem from a project investigating contemporary Arabic varieties. In this project, the *Viennese Corpus of Arabic Varieties*, four small dictionaries have been created so far and a number of others are to follow. These electronic dictionaries are meant to serve a twofold purpose: first, they will furnish the basis of comparatistic dialectological research and will be used to set up a specialised interface (in the department's nomenclature this kind of software component is called a resource viewer) to visualise particular salient linguistic features across a number of linguistic varieties. Second, these dictionaries will also be put to didactic use in language teaching courses at the university. To our knowledge, there are no other machine readable dictionaries of these linguistic varieties so far.

Another project aims at the creation of a machine readable dictionary of Early Modern German (EMG). This is an undertaking that is being carried out on the basis of a small corpus, which has been compiled and annotated at the department. The texts–all of them of Austrian provenance–were automatically tagged with POS and lemmas. In a second step, this data was manually verified. The list of lemmas which were enriched with automatically extracted corpus data will serve as the basis of a small machine-readable dictionary of Austrian EMG. It is planned to complement this dataset with data from other available corpora of the same period in a second step in order to obtain a larger basis for studies of historical variational linguistics.

While the amount of currently available data at the ICLTT is not very large, the number of entries in our dictionary database keeps growing. Yet, it goes without saying that projects such as those described before need a great deal of cooperation and cannot be carried out by individuals. For this reason, we are increasingly focusing on strengthening ties with other interested institutions and intensifying our efforts towards setting up infrastructures that allow collaborative working on dictionaries.

3.3 TEI for dictionaries

In digital text encoding, the guidelines of the Text Encoding Initiative[4] (TEI) have long been considered the de-facto standard and are widely used. While the digital collections of the Austrian Academy Corpus were encoded in a system that might at best be described as TEI inspired, the ICLTT has started to convert all its holdings to TEI P5 as of this year. This also includes the few historical dictionaries contained in this corpus of written German.

When looking for an encoding standard for machine readable dictionaries, LMF (Lexical Markup Framework, ISO 24613:2008) is probably the first thing one might think of. While using the TEI dictionary module to encode digitised print dictionaries has become a fairly uncontested and very common standard procedure, using the very same system for NLP purposes is quite another story. The ICLTT has attempted to make use of TEI's dictionary module to create machine readable dictionaries by imposing a number of constraints on the comparatively flexible structure of the original specifications[5].

The adaptability of digitised dictionary data to TEI P5 is currently being tested in several smaller and larger projects. In one of these experiments, we are converting the above mentioned German-Russian dictionaryinto a machinereadable TEI conformant version.

Another lexicographic experiment is being conducted on the German language version of the collaborative dictionary project Wiktionary. The scarcity of freely available digital multilingual lexical data makes such resources a valuable treasure trove for computational linguists and lexicographers to experiment with. Regrettably, the content of this steadily growing resource, which is not being produced by professional lexicographers but enthusiastic volunteers, is formatted with a lightweight markup system used in different Wiki applications. It is neither standardised nor very structure-oriented. Attempts at preparing Wiktionary for use in NLP applications have been made before, but we have created a freely available tool furnished with a graphical user interface–the first such application to our knowledge–that converts the Wiktionary database dump into a technically reusable XML format, i.e. TEI P5.

4. Digital corpora and dictionary writing

While digital corpora keep growing, they are playing an increasingly important role in modern linguistics. Although representatives of many fields of these

[4]The current version of the guidelines is usually cited with the suffix P5 (i.e. proposal number 5) and can be accessed at http://www.tei-c.org/Guidelines/P5/

[5]We presented a paper entitled *Creating lexical resources in TEI P5. Experiences from building multi-purpose digital dictionaries* at the TEI Members' Meeting 2011 in Würzburg (Germany).

disciplines have remained reluctant to use digital corpora as basis for their investigations, making use of native speaker intuition has become outmoded in contemporary lexicography. Quite to the contrary, lexicographers have long since adopted modern corpus technology; and lexicography has become something like a prototypical field of application for digital corpora.

When creating dictionaries, lexicographers often rely on one particular corpus, very often a collection that has been set up for the particular purpose. The many relevant issues regarding the type of corpus needed for a certain kind of dictionary, the appropriate size of corpus for a particular dictionary, and the corpus features required for a particular lexicographic project will not be touched upon here. It is, however, important to stress the fact that dictionary makers of the 21st century rely heavily on corpus data to build and improve their products. To achieve this end, they need software that allows them to access digital corpora while working on their projects.

When creating dictionaries, lexicographers often rely on one particular corpus, very often a collection that has been set up for the particular purpose. Unfortunately, creating high quality corpora is still costly and time-consuming. It is therefore the most natural thing for those working in the field to look for existing available resources instead. The number of usable corpora has increased considerably over the past years. However, there are still a number of issues that need to be resolved. First of all, freely available does not necessarily imply being ready to be used for lexicographic projects; accessing such corpora often involves some troublesome procedures. Furthermore, federated search in more than one corpus is usually not feasible in optimised dictionary creating workflows.

5. Viennese Lexicographic Editor (VLE)

There are a number of well-established dictionary editing applications. Some of the best-known products include:

- ABBYY Lingvo Content[6],
- DEBII,[7]
- IDM DPS[8],
- Shoebox and the Field Linguist's Toolbox[9],
- iLex[10],
- Lexique Pro[11],
- LEXUS[12],
- TshwaneLex[13]

This list is by no means meant to be exhaustive. Actually, one could make it much longer. Some of these products provide a wide range of functionalities which can be put to use for collecting, refining and enhancing lexicographic data. Some packages are fully integrated systems, others are built in a modular way. Some are being used for particular purposes such as preserving endangered languages, some offer specialised multi-media support. Technically, dictionary writing software is often built around RDBM systems, and often makes use of client-server or multi-tier architecture.

The module presented in this paper is part of a fairly new piece of software that first came into existence as a by-product of an entirely different development activity: the creation of an interactive online learning system for university students. It was first used in a collaborative glossary editing project carried out as part of university language courses at the *University of Vienna*. As the tool proved to be remarkably flexible and adaptable, it was put to work in other projects and is now being used in this research effort designed to fathom out the potential of a more direct integration of corpus data in the dictionary creating process.

At the heart of our dictionary writing system is a dictionary writing client, a standalone application[14], which for the time being has been dubbed – in default of a more adequate name – *Viennese Lexicographic Editor* (VLE). Over the past few months, the client and the associated server scripts have been continually adapted and improved. The whole system relies heavily on XML and cognate technologies such as XSLT and XPath.

5.1 Architecture

The module discussed in this paper is integrated into the above mentioned VLE. The current version only supports web-based editing; the dictionary entries are stored on a web-server. All additional software components (PHP and MySQL) are open source and freely available. On many operating systems, in order to setup the dictionary, simply copying four PHP scripts will suffice to get a working installation of the dictionary server. PHP and MySQL are usually part of the basic installation of such systems.[15] Communication between the dictionary client and the server has been implemented as a RESTful web service.

The distributed architecture has a number of obvious advantages. Being able to work on the data wherever one has access to the internet is unquestionably a useful feature. Lexicographers can then work when they are on vacation without having to carry all data around with

[6] http://www.abbyy.com/lingvo_content/
[7] v. Horak 2006
[8] http://www.idm.fr/products/
[9] http://www.sil.org/computing/shoebox/mdf.html, http://www.sil.org/computing/toolbox
[10] v. Erlandsen 2004
[11] http://www.lexiquepro.com/
[12] v. Ringersma 2007
[13] v. Joffe 2004 and http://tshwanedje.com/tshwanelex/

[14] The software was written in Delphi 2010. Writing software in Pascal dialects is part of a long-standing tradition in Humanities computing. Over the past two decades, many programs have been written at our department making use of this high-level programming language. The large number of reusable libraries allowed us to keep programming overhead to a minimum.
[15] Currently, our main dictionary server is running on *openSuse* 11.3.

themselves. In addition, this system also allows for collaborative working on the dictionary data.

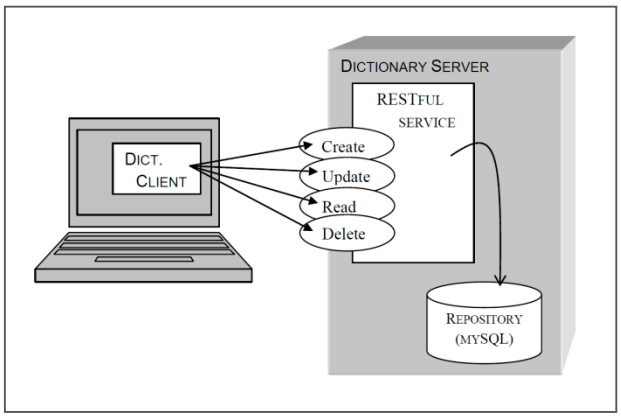

Figure 1: System architecture

5.2 Input validation

One of the main reasons for working with XML is the possibility of ensuring the formal correctness of input. VLE offers the usual two levels of input control: well-formedness, which can be described as the basic compliance of an XML document or data snippet with the syntax of the XML recommendation. When validating a document, the data is checked against a so-called document type definition.

The well-formedness of data in an entry is verified by the VLE tool every time the dictionary entry is saved. Users can also trigger the process manually or make the program check this status with every modification of the entry.

Validation is the process of matching the data on a higher level. When validating the structure of a document, its contents are checked against another document which contains definitions of permissible elements and information as to where these elements may appear in the document. Currently, our tool expects document type definitions in form of an XML Schema which is, like XML, a W3C recommendation. On the to-do-list of the dictionary tool, there is also the implementation of an option to validate against RELAX NG, which is an ISO standard and has found much support in the TEI and OpenDocument communities.

5.3 Editor modes

The user of VLE has two basic options for editing the dictionary data: working in XML mode, which may be considered the expert mode, or working in an editor form with predefined entry controls that function like traditional database input fields. While working on an entry, it is possible to switch between the two modes at random. The second option, i.e. making use of edit controls for particular XML elements, is especially useful when working on the same field across a number of dictionary entries. Navigating in the XML expert mode is more cumbersome than in the edit controls mode, since lexicographers have to position themselves in every entry they work on.

When working on large dictionary entries, keeping track of the entry details or even just the current position within an article can, at times, be a troublesome undertaking. Actually, XML encoded data is of great advantage in this respect, as the structure of the entries can help the software in tackling these problems. In particular, the TEI system with its intuitive and not too verbose element names eases the task. The software allows lexicographers to navigate to particular cognate points within the text.

Figure 2: XML mode

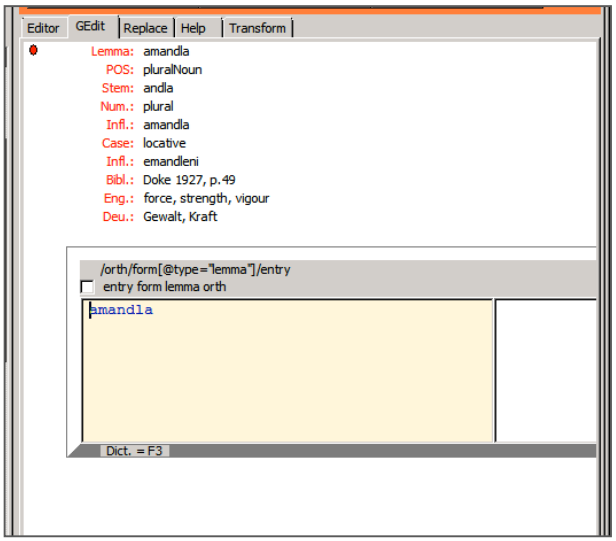

Figure 3: Database-like input mode

5.4 Visualisation of entries

An additional useful feature of the tool is its capability to visualise the data. This task is achieved by means of XSLT stylesheets which are freely configurable. While

this functionality is quite commonplace in many applications today, our tool proves to be particularly versatile. Making use of different styles, allows switching between different views of the same set of data. When working on very large entries, stylesheet transformations tend to be restrictively slow. This is not the case, however, when they are only applied to particular parts of an entry.

Automatically applied links in the output data (HTML) allow navigation from these visualisations back into the editor control, which again makes navigating copious dictionary entries a considerably more agreeable task.

5.5 Data export

VLE stores all data on a server. In addition, it has also been provided with the functionality to store output on the local machine. Making use of the export control, all data can be saved into one document. Usually, all metadata and production related data such as the configuration profile are inserted in these documents.

5.6 Web-publishing

Dictionary entries created by this tool can be published on the internet through a simple PHP script. Adapting an HTML template to create a new dictionary web-site is a matter of minutes. The resulting web-page has a query control and is able to display the results of dictionary queries.

6. From *corpus* to *dictionary entry*

When digital corpora are used to compile new dictionaries or to enhance existing ones, more often than not interfacing between the dictionary writing software and the respective corpus poses considerable problems. When accessing corpora in tandem with producing dictionary entries, the issue at hand is transferring the results of corpus queries in an acceptably comfortable manner. Very often this process involves rather cumbersome steps that require a series of manual manipulations. The focus in this project has been on streamlining this process, on speeding up the import of corpus data into dictionary entries.

When accessing digital corpora, lexicographers might be interested in a broad spectrum of data including, but not limited to:

- lists of collocations,
- multiword units,
- statistical information on lemmas, particular word forms or any of the afore mentioned categories,
- corpus examples.

Although this particular module of the dictionary writing editor can perform many other tasks, this paper focuses on the issue of corpus examples. The principal idea when preparing this module was optimizing access to digital corpora in order to allow lexicographers to glean sample sentences and to integrate them into dictionary entries in a reasonably comfortable manner. The focus of our work was on ease of use and direct access to the data. The corpus interface of the new dictionary writing application presented in this paper was supposed to enable lexicographers to launch corpus queries and to offer functionalities for inserting them into existing dictionary entries without needing to use the clipboard to copy-and-paste, which inevitably results in a lot of inefficient typing or clicking.

The new corpus browser module was designed to be a principally universal web-interface and to allow lexicographers to query not just one particular corpus, but any digital resources accessible via a web-browser. One of the important perspectives of the new corpus browser was its integration with evolving CLARIN[16] infrastructures, in particular *federated content search* facilities, a project which was initiated by a CLARIN working group this year. Researchers of the ICLTT have taken a keen interest in these activities and have actively contributed to this ongoing project.

For the purposes of our research, the VLE's corpus interface needed to enable lexicographers to copy the selected data into the dictionary writing editor using a single click or keystroke. In addition, the scripting of processes had to be possible and the transfer of data needed to be achieved through a transformer component that could automatically perform predefined modifications of the text and translate the HTML text received by the browser into the target formats required by the dictionary system.

The process of enriching dictionary data with data from digital corpora as performed by our new tool can be described as a workflow made up ofsix basic steps:

- Querying a corpus / corpora
- Optional pre-selecting of data in the browser
- Analysing the data
- Selecting data from a list of candidates
- Converting to the target format of the dictionary
- Inserting the data into an entry

6.1 Querying corpora via the Internet

VLE's integrated web browser allows lexicographers to access and search the internet. It works very much like other such tools, but it does not have some of the extra features such as bookmark management, download management, or a search-engine toolbar, which are unnecessary for our purposes. The component used to realise this part of the programme is a common wrapper, which was placed around Windows' native Internet Explorer component.[17]

[16] CLARIN stands for *Common Language Resources and Technology Infrastructure* (http://www.clarin.eu/external/) and was initiated as an ESFRI project.

[17] *TWebBrowser* is a visual component that allows programmers to create simple web-browser applications in just a few minutes.

The current version of the module can work with any corpus that delivers data through the HTTP protocol. It can not only work with text collections structured as corpora but with any data delivered as HTML or raw text. It cannot deal with PDF documents at the moment, however. When the module's browser receives XML data, they are transformed into HTML using XSLT style sheets. An additional interesting feature of the tool, albeit of lesser importance for the particular purpose being discussed here, is the capability to access other online available dictionaries.

Ideally, lexicographers should be able to trigger queries directly from the entry edit control, the part of the tool where the dictionary text is edited by the lexicographer. Unfortunately, this direct approach is not an option with many corpora or text collections available on the internet, as they offer access exclusively through their own web-interfaces. This means, that users have to navigate to the respective corpus entry points first and then input their queries manually.

6.2 Pre-selecting data

When search results have been retrieved from a corpus and appear in the browser, these data have to be dealt with in some way or another before they can be integrated into a dictionary entry. With our tool, users have two options at this stage: they can either accept the received data in their entirety or they can manually select only part of it.

6.3 Analysing the data

Manual intervention is practically unnecessary after the data have been pre-selected, as the program has the ability to perform a first analysis of the imported data. Usually, the results of corpus queries are delivered as concordance lines, typically in form of Key-Word-in-Context (KWIC) lines, which, when sent over the internet, are commonly transformed into HTML tables. These structures can easily be identified by the software of our tool.

6.4 Selecting data from a list of candidates

Having performed this initial analysis of data, the program passes it on to the selector control, which presents the data to the user in a listbox control. Here, the user can make the final selection for the dictionary entry. This is the only point in the process where manual intervention on part of the lexicographer is inevitable.

6.5 Converting data

After the selection of data, data snippets are passed into the entry editor through a template, in which the exact XML structure of the data to be inserted can be defined. In addition, the tool has also the capability of carrying out data conversions through a service based mechanism. This mechanism allows actions to be performed in a distributed manner. This might make it possible in the future to enrich parts of the data being worked on through services offered elsewhere on the internet.

7. Corpus access

The steps described above can each be performed separately. However, the point of systematically defining this workflow was to allow lexicographers to automate as many of the intermediary steps as possible, thus avoiding any redundant key strokes or clicks in order to stave off carpal tunnel syndrome as long as possible and to make work on external sources more efficient.

The most critical step in the importing process is the query. Circumventing the step "pre-selecting of data" as described in 6.2 is only possible through direct access to the corpora one wants to query. The user (with the help of the software) must be able to launch queries directly via the HTTP protocol. Relieving the user of manually initiating the communication with the remote corpus can only be achieved through a service that allows machine-to-machine communication. The establishment of service based access points for corpora is a fundamental prerequisite for the smooth integration of dictionary client and corpus.
VLE is capable of performing the above described process sufficiently well when accessing our own data servers at the Academy. It allows lexicographers to launch queries directly from the editor control, simply by selecting a string and triggering a function. The problem arises on the other side of the communication, as most other corpora do not offer service based interfaces that allow outside software to interact with them directly. Web-interfaces of corpora are usually geared towards the needs of human users. As a result, queries can only be triggered when text is manually entered into edit controls in web forms.

8. LRT standards

Many activities of the ICLTT have been characterised by a strong commitment to standards and de-facto standards. This awareness of the relevance of standards has been largely motivated by the department's involvement in interdisciplinary projects which involved heterogeneous resources and a wide range of methodologies and tools. The need for harmonising divergent environments has heightened our awareness for issues of interoperability, reusability and LRT standards. This also accounts for the extensive use of XML, Unicode and related technologies in all applications. The AAC's first XML encoded digital objects – a digital version of the Austrian historical magazine "Die Fackel" (6 million tokens) – date back as far as 1998, the year in which the World-Wide-Web-Consortium passed its first XML recommendation.[18]

[18] At that time, our work was based on Extensible Markup Language (XML) 1.2 (http://www.w3.org/TR/1998/REC-xml-19980210).

Our current experiments with the dictionary writing software have been conducted using a combination of TEI's dictionary module (P5) and ISOCat[19]. Other standards relied on in these projects include MAF (Morphosyntactic Annotation Framework, ISO/DIS 24611) and ISO 639 (Language codes).

Since the architecture of the dictionary writing system is built on XML, schemata other than TEI can also be implemented with ease. In addition to systems like OLIF (Open XML Language Data Standard), one might consider using formats such as OWL (Web Ontology Language), RDF (Resource Description Framework), TBX (TermBase eXchange) or LMF (Lexcial Markup Framework, ISO 24613)[20] in future projects.

Actually, the first experiments with VLE were undertaken on the basis of LMF encoded data, as our endeavours have been directed towards creating machine readable dictionaries. While we have not discarded the use of LMF for future projects, the less verbose and for human lexicographers more easily readable structure of TEI (P5) has so far tipped the balance in favour of using this encoding system in our projects. However, LMF continues to plays an important role; in contrast to TEI (P5), it is a full-fledged ISO standard. What has remained of the early LMF experiments is a function in the VLE's entry editor control capable of converting the ICLTT's TEI dictionary entries into LMF entries. However, this TEI to LMF converter is not a universally applicable tool, as it only works with the ICLTT's TEI dictionary format. In the future, this part of the editor might be extended as LMF, probably, will gain more importance.

9. Current status and availability

As we have shown above, our dictionary writing system is made up of easily distributable, easy to set up components: All that is needed is a client, a server (in our case Apache) with a mySql database and four PHP scripts to run the RESTful service. The system has been optimised for ease of use and interoperability, everything is based on XML.

The system has been intended for use by individual lexicographers and small groups of researchers. Currently, it is being tested in several small to medium sized projects and numerous amendments are constantly being applied. We have started to work on a small user guide and there are plans to make a first version of the client available for interested researchers in the course of 2012. The package is still in beta stage, and no decision has been made yet as to the license under which the client software will be available, but the four server scripts can be downloaded from the ICLTT Showcase website (http://corpus3.aac.ac.at/showcase).

[19] A web-application, offering access to ISO TC37's Data Category Registry of widely accepted linguistic concepts.

[20] It is worth mentioning here that we also use the dictionary writing system to manage other types of data such as bibliographies, prosopographic databases and feature catalogues for a project involving comparatistic studies in linguistic varieties. All of this data is TEI conformant XML.

10. Conclusions

In view of all the dictionary software that already exists, one could rightfully ask: why produce yet another tool? To answer this question, one has to consider the fact that software lifecycles have shortened considerably in recent years. Reusable components and libraries allow new products to be created with comparatively small overhead. In addition, there are hardly any state-of-the-art applications that are open source, extensible, comfortably manageable by non-technicians and free of charge. With the creation of the *Viennese Lexicographic Editor,* such reusable components have been combined with a state-of-the-art application that can potentially solve some of the problems of streamlining workflows at the interface between corpus and dictionary writing systems.

There are several simple answers to the question above: because it was possible to do it, because it did not cost much and because it might motivate others to muster courage to go ahead with their own lexicographic ambitions. Researchers are often wary of going digital, individual researchers are particularly confronted with problems which could be remedied to a certain degree by more easily attainable and usable software. As the national coordinator of the two projects CLARIN-AT and DARIAH-AT, the ICLTT sees its role also as a facilitator to enable more researchers and scholars in the Humanities and the Arts to take the digital path.

In conducting these experiments, we have been guided by a vision of a densely knit web of dictionaries, where datasets created by human editors are enhanced by automatically created data, where lexical resources created by automatic routines serve as the basis of an ever renewed and growing lexicographic web. In building this new application, we are envisaging more reusable, standards-based and ideally open-source LRTs being developed by ever growing communities of both individual and groups of researchers.

11. References

Atkins, B.T.S., Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Baumann, S., Burnard, L. & Sperberg-McQueen, C.M. (eds.) (2010). *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Oxford - Providence – Charlottesville - Nancy. Accessed at: http://www.tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf.

Breiteneder, E. (2003). *Austrian Academy Corpus.* In Thomas Städtler (ed.) *Wissenschaftliche Lexikographie im deutschsprachigen Raum*.

Heidelberg, Universitätsverlag Winter, pp. 447-448.

Breiteneder, E. (2003). *Wörterbuch der Fackel.* In T. Städtler (ed.) *Wissenschaftliche Lexikographie im deutschsprachigen Raum*. Heidelberg, Universitätsverlag Winter**,** pp. 295-297.

Erjavec, T., Evans, R., Ide, N., & Kilgarriff, A. (2003). From machine readable dictionaries to lexical databases: the Concede Experience. In *Proceedings of the 7th International Conference on Computational Lexicography. COMPLEX'03*. Budapest.

Erjavec, T., Tufiş, D., & Varadi, T. (1999). Developing TEI-conformant lexical databases for CEE languages. In *Proceedings of the 4th International Conference on Computational Lexicography*. *COMPLEX'99.* Pecs, Hungary.

Erlandsen, J. (2004). iLEX – an ergonomic and powerful tool combining, effective and flexible editing with easy and fast search and retrieval. In *EURALEX 2004, Lorient, France*.

Horák, A., Pala, K., Rambousek, A. & Rychlý, P. (2006). New clients for dictionary writing on the DEB platform. In *DWS 2006: Proceedings of the Fourth International Workshop on Dictionary Writings Systems.* Torino, Italy: Lexical Computing Ltd., U.K., pp. 17-23.

Joffe, D., de Schryver, G.M. (2004). TshwaneLex – professional off-the-shelf lexicography software. In *Third International Workshop on Dictionary Writing Systems: Program and List of Accepted Abstracts,* Brno, Czech Republic, Masaryk University, Faculty of Informatics.

Kilgarriff, A., Kovár, V. & Rychlý, P. (2009). Tickbox Lexicography. In S. Granger, M. Paquot (eds*.) eLexicography in the 21st century: New Challenges, New Applications, Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009*. Louvain-la-Neuve: Presses Universitaires de Louvain, pp. 411-418.

O'Keeffe, A., McCarthy, M. (eds.) (2010). *The Routledge Handbook of Corpus Linguistics*. London and New York: Routledge.

Ringersma, J., Kemps-Snijders, M. (2007). Creating multimedia dictionaries of endangered languages using LEXUS. In H. van Hamme, R. van Son (eds.) *Proceedings of Interspeech 2007*. Baixas, France: ISCA-Int. Speech Communication Assoc., pp. 65-68.

Spohr, D. (2008). Requirements for the design of electronic dictionaries and a proposal for their formalisation. In *Proceedings of the EURALEX International Congress 2008*. Barcelona.

Walter, E. (2010). Using corpora to write dictionaries. In A. O'Keeffe, M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics.* London and New York: Routledge, pp. 428-443.