CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era

Workshop Programme

09:15-09:30 – Welcome and Introduction

09:30-10:30 – Invited Talk
Steven Moran, *Under-resourced languages data: from collection to application*

10:30-11:00 – Coffee break

11:00-13:00 – Session 1
Chairperson: Joseph Mariani

11:00-11:30 – Oleg Kapanadze, *The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages*

11:30-12:00 – Martin Benjamin, Paula Radetzky, *Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification*

12:00-12:30 – Thierry Declerck, Eveline Wandl-Vogt, Karlheinz Mörth, Claudia Resch, *Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework*

12:30-13:00 – Delphine Bernhard, *Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian*

13:00-15:00 – Lunch break

13:00-15:00 – Poster Session
Chairpersons: Laurette Pretorius and Claudia Soria

Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Varadi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, Sigve Gramstad, *An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"*

István Endrédy, *Hungarian-Somali-English Online Dictionary and Taxonomy*

Chantal Enguehard, Mathieu Mangeot, *Computerization of African Languages-French Dictionaries*

Uwe Quasthoff, Sonja Bosch, Dirk Goldhahn, *Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu*

Edward O. Ombui, Peter W. Wagacha, Wanjiku Ng'ang'a, *InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili*

Ronaldo Martins, *UNLarium: a Crowd-Sourcing Environment for Multilingual Resources*

Anuschka van ´t Hooft, José Luis González Compeán, *Collaborative Language Documentation: the Construction of the Huastec Corpus*

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, Francis M. Tyers, *Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages*

15:00-16:00 – Session 2
Chairperson: Eveline Wandl-Vogt

15:00-15:30 – Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, Joseph Mariani, Claudia Soria, *The LREMap for Under-Resourced Languages*

15:30-16:00 – Dorothee Beermann, Peter Bouda, *Using GrAF for Advanced Convertibility of IGT data*

16:00-16:30 – Coffee break

16:30-17:30 – Session 3
Chairperson: Thierry Declerck

16:30-17:00 – Stefan Daniel Dumitrescu, Tiberiu Boroș, Radu Ion, *Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus*

17:00-17:30 – Katia Kermanidis, Manolis Maragoudakis, Spyros Vosinakis, *Crowdsourcing for the Development of a Hierarchical Ontology in Modern Greek for Creative Advertising*

17:30-18:15 – Discussion

18:15-18:30 – Wrap-up and goodbye

Editors

Laurette Pretorius	University of South Africa, South Africa
Claudia Soria	CNR-ILC, Italy
Paola Baroni	CNR-ILC, Italy

Workshop Organizing Committee

Laurette Pretorius	University of South Africa, South Africa
Claudia Soria	CNR-ILC, Italy
Eveline Wandl-Vogt	Austrian Academy of Sciences, ICLTT, Austria
Thierry Declerck	DFKI GmbH, Language Technology Lab, Germany
Kevin Scannell	St. Louis University, USA
Joseph Mariani	LIMSI-CNRS & IMMI, France

Workshop Programme Committee

Deborah W. Anderson	University of Berkeley, Linguistics, USA
Sabine Bartsch	Technische Universität Darmstadt, Germany
Delphine Bernhard	LILPA, Strasbourg University, France
Bruce Birch	The Minjilang Endangered Languages Publications Project, Australia
Paul Buitelaar	DERI, Galway, Ireland
Peter Bouda	CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal
Steve Cassidy	Macquarie University, Australia
Thierry Declerck	DFKI GmbH, Language Technology Lab, Germany
Vera Ferreira	CIDLeS - Interdisciplinary Centre for Social and Language Documentation, Portugal
Claudia Garad	wikimedia.AT, Austria
Dafydd Gibbon	Bielefeld University, Germany
Oddrun Grønvik	Instituut for lingvistike og nordiske studier, University of Oslo, Norway
Yoshihiko Hayashi	University of Osaka, Japan
Daniel Kaufman	Endangered Language Alliance, USA
Andras Kornai	Hungarian Academy of Sciences, Hungary
Simon Krek	Jožef Stefan Institute, Slovenia
Tobias Kuhn	ETH, Zurich, Switzerland
Joseph Mariani	LIMSI-CNRS & IMMI, France
Steven Moran	Universität Zürich, Switzerland
Kellen Parker	National Tsing Hua University, China
Patrick Paroubek	LIMSI-CNRS, France
Maria Pilar Perea i Sabater	Universitat de Barcelona, Spain
Laurette Pretorius	University of South Africa, South Africa
Leonel Ruiz Miyares	Centro de Linguistica Aplicada (CLA), Cuba
Kevin Scannell	St. Louis University, USA
Ulrich Schäfer	DFKI GmbH, Germany
Claudia Soria	CNR-ILC, Italy
Nick Thieberger	University of Melbourne, Australia
Michael Zock	LIF-CNRS, France

Table of Contents

Index of Authors v

Preface vii

The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages by Oleg Kapanadze 1

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification by Martin Benjamin, Paula Radetzky 9

Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework by Thierry Declerck, Eveline Wandl-Vogt, Karlheinz Mörth, Claudia Resch 17

Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian by Delphine Bernhard 23

An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age" by Georg Rehm, Hans Uszkoreit, Ido Dagan, Vartkes Goetcherian, Mehmet Ugur Dogan, Coskun Mermer, Tamás Varadi, Sabine Kirchmeier-Andersen, Gerhard Stickel, Meirion Prys Jones, Stefan Oeter, Sigve Gramstad 30

Hungarian-Somali-English Online Dictionary and Taxonomy by István Endrédy 38

Computerization of African Languages-French Dictionaries by Chantal Enguehard, Mathieu Mangeot 44

Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu by Uwe Quasthoff, Sonja Bosch, Dirk Goldhahn 52

InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili by Edward O. Ombui, Peter W. Wagacha, Wanjiku Ng'ang'a 56

UNLarium: a Crowd-Sourcing Environment for Multilingual Resources by Ronaldo Martins 60

Collaborative Language Documentation: the Construction of the Huastec Corpus by Anuschka van ´t Hooft, José Luis González Compeán 67

Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages by Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, Francis M. Tyers 71

The LREMap for Under-Resourced Languages by Riccardo Del Gratta, Francesca Frontini, Anas Fahad Khan, Joseph Mariani, Claudia Soria 78

Using GrAF for Advanced Convertibility of IGT data by Dorothee Beermann, Peter Bouda 84

Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus by Stefan Daniel Dumitrescu, Tiberiu Boroș, Radu Ion 90

Crowdsourcing for the Development of a Hierarchical Ontology in Modern Greek for Creative Advertising by Katia Kermanidis, Manolis Maragoudakis, Spyros Vosinakis 95

Index of Authors

Beermann, Dorothee 84
Benjamin, Martin 9
Bernhard, Delphine 23
Boroș, Tiberiu 90
Bosch, Sonja 52
Bouda, Peter 84
Compeán, José Luis González 67
Dagan, Ido 30
Declerck, Thierry 17
Del Gratta, Riccardo 78
Dogan, Mehmet Ugur 30
Dumitrescu, Stefan Daniel 90
Endrédy, István 38
Enguehard, Chantal 44
Frontini, Francesca 78
Goetcherian, Vartkes 30
Goldhahn, Dirk 52
Gramstad, Sigve 30
Ion, Radu 90
Jones, Meirion Prys 30
Kapanadze, Oleg 1
Kermanidis, Katia 95
Khan, Anas Fahad 78
Kirchmeier-Andersen, Sabine 30
Mangeot, Mathieu 44
Maragoudakis, Manolis 95
Mariani, Joseph 78
Martins, Ronaldo 60
Mermer, Coskun 30
Mörth, Karlheinz 17
Moshagen, Sjur 71
Ng'ang'a, Wanjiku 56
Oeter, Stefan 30
Ombui, Edward O. 56

Pirinen, Tommi 71
Quasthoff, Uwe 52
Radetzky, Paula 9
Rehm, Georg 30
Resch, Claudia 17
Rueter, Jack 71
Soria, Claudia 78
Stickel, Gerhard 30
Trosterud, Trond 71
Tyers, Francis M. 71
Uszkoreit, Hans 30
van ´t Hooft, Anuschka 67
Varadi, Tamás 30
Vosinakis, Spyros 95
Wagacha, Peter W. 56
Wandl-Vogt, Eveline 17

Preface

The LREC 2014 Workshop on "Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era" (CCURL 2014) has its origin in the imperative of cultural and language diversity and in the basic right of all communities, all languages and all cultures to be "first class citizens" in an age driven by information, knowledge and understanding. In this spirit, the focus of this first CCURL Workshop is on two strategic approaches by which under-resourced languages can elevate themselves to levels of development that are potentially comparable to well-resourced, technologically advanced languages, viz. using the crowd and collaborative platforms, and using technologies of interoperability with well-developed languages and Linked Data.

Specific questions that the Workshop addresses include the following:

- How can collaborative approaches and technologies be fruitfully applied to the development and sharing of resources for under-resourced languages?
- How can small language resources be re-used efficiently and effectively, reach larger audiences and be integrated into applications?
- How can they be stored, exposed and accessed by end users and applications?
- How can research on such languages benefit from semantic and semantic web technologies, and specifically the Linked Data framework?

All the papers accepted for the Workshop address at least one of these questions, thereby making a noteworthy contribution to the relevant scholarly literature and to the technological development of a wide variety of under-resourced languages.

Each of the sixteen accepted papers was reviewed by at least three members of the Programme Committee, eight of which are presented as oral presentations and eight as posters.

We look forward to collaboratively and computationally building on this new tradition of CCURL in the future for the continued benefit of all the under-resourced languages of the world!

The Workshop Organizers

Laurette Pretorius – University of South Africa, South Africa
Claudia Soria – CNR-ILC, Italy
Eveline Wandl-Vogt – Austrian Academy of Sciences, ICLTT, Austria
Thierry Declerck – DFKI GmbH, Language Technology Lab, Germany
Kevin Scannell – St. Louis University, USA
Joseph Mariani – LIMSI-CNRS & IMMI, France

The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages

Oleg Kapanadze

Tbilisi State University

Chavchavadse av.1,

0162 Tbilisi, Georgia

E-mail: ok@caucasus.net

Abstract

In this paper, we describe outcomes of an undertaking on building Treebanks for underresourced languages Georgian, Russian, Ukrainian, and German - one of the "major" languages in the NLT world (Hence, the treebank's name – **GRUG**). The monolingual parallel sentences in four languages were syntactically annotated manually using *the Synpathy* tool. The tagsets follow an adapted version of the German TIGER guidelines with necessary changes relevant for the Georgian, the Russian and the Ukrainian languages grammar formal description. An output of the monolingual syntactic annotation is in the TIGER-XML format. Alignment of monolingual repository into the bilingual Treebanks was done by *the Stockholm TreeAligner* software. The parallel treebank resources developed in the GRUG project can be viewed at the URL of Saarland and Bergen Universities: http://fedora.clarin-d.uni-saarland.de/grug/ , http://clarino.uib.no/iness.

Keywords: under-resourced languages, annotation, parallel treebanks.

1. Introduction

Naturally-occurring text in many languages are annotated for linguistic information. A Treebank is a text corpus in which each sentence has been annotated with syntactic structure. Building annotated corpora and constructing treebanks lead to improvement of grammars and lexicons (Losnegaard et al., 2012). This is especially relevant for the under-resourced languages.

In this paper we describe an initiative for building German-Georgian, German-Russian, German-Ukrainian and Georgian-Ukrainian syntactically annotated parallel Treebanks.

Parallel corpora are language resources that contain texts and their translations, where the texts, paragraphs, sentences, and words are linked to each other. In the past decades they became useful not only for NLP applications, such as machine translation and multilingual lexicography, but are considered indispensable for empirical language research in contrastive and translation studies.

Treebanks are often created on top of a corpus that has already been annotated with part-of-speech tags. The annotation can vary from constituent to dependency or tecto-grammatical structures. In turn, Treebanks are sometimes enhanced with semantic or other linguistic information and are skeletal parses of sentences showing rough syntactic and semantic information.

2. Multilingual Parallel Corpus Utilized in the Project

The languages (except German) involved in the project are under-resourced languages for which parallel texts are very rare. In this project we used a multilingual parallel corpus appended to a German-Russian-English-Georgian (GREG) valency lexicon for Natural Language Processing (Kapanadze et al., 2002), (Kapanadze, 2010), whose subcorpus has been already utilized partly in the previous project for building the syntactically annotated German-Georgian parallel trees (Kapanadze , 2012).

The GREG lexicon itself contains a manually aligned German, Russian, English and Georgian valency data supplied with syntactic subcategorization frames saturated with semantic role labels. The multilingual verb lexicon is expended with examples of sentences in 4 languages involved in the project. They unfold lexical entries' meaning and are considered as mutual translation equivalents. The size of bilingual sublexicons, depending to a specific language pair, varies between 1200-1300 entries and the number of example sentences appended to the lexicons are different. For example, a German-Georgian subcorpus, used for this study, has a size of roughly 2600 sentence pairs that correspond to possible syntactic subcategorization frames. For the German-Russian language pair there had been extracted more fine grained subcorpus with about 4000 sentences as translation equivalents. A German-Ukrainian subcorpus, specifically created for the GRUG initiative support, is relatively small.

3. Morphological and Syntactic Annotation of a Multilingual GRUG Text

The first two languages addressed in the GRUG project had been a German-Georgian language pair. The later is an agglutinative language using suffixing and prefixing for which text morphological annotation, tagging and lemmatizing procedures were done with a finite-state

morphological transducer that draws on the XEROX FST tools (Kapanadze et al. 2009), (Kapanadze et al., 2010).

We have started syntactic annotation for the Georgian text having an overview of experience in building parallel treebanks for languages with different structures (Megyesi and Dahlqvist, 2003), (Megyesi et al., 2006), (Grimes et al., 2003), (Rios et al., 2009). In this study the most useful information has been collected about Turkish and Quechua languages. For instance, in a Quechua-Spanish parallel treebank, due to strong agglutinative features of the Quechua language, the monolingual Quechua treebank was annotated on morphemes rather than words. Besides, for capturing the Quechua sentences phrase structure peculiarities, a Role and Reference Grammar has been opted that allowed by using nodes, edges and secondary edges to represent the most important aspects of Role and Reference syntax for Quechua (Rios et al., 2009).

Although Georgian is also an agglutinative language, there is no need to annotate the Georgian Treebank on morphemes. The Georgian syntax can be sufficiently well represent by dependency relations. Therefore, the Georgian Treebank was annotated according to an adapted version of the German TIGER guidelines. Nevertheless, the outcomes of an FST morphological analyses, tagging and lemmatizing done by the XEROX Calculus Tool, we had to reformat using a small script written in Python. It converts an output of the Georgian morphological transducer into an input of the *Synpathy* tool developed at Max Plank Institute for Psycholinguistics, Nijmegen (Synphaty: Syntax Editor, 2006).

The morphological features (including POS tags) to the rest two languages, Russian and Ukrainian, were assigned manually in the script encoding process. In the mentioned issue we pursued to the NEGRA-Treebank (STTS) guidelines with the necessary changes relevant to the Russian and Ukrainian formal grammar. The German Treebank annotation follows the TIGER general annotation scheme (Brants et al., 2002), (Smith, 2003).

On the further step the POS-tagged and lemmatized monolingual sentences in 4 languages are fed to the *Synpathy* syntactic annotation engine. It employs a SyntaxViewer developed for TIGER-Research project (Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart) which uses a TIGER-XML script as a source for monolingual syntactic tree visualization (cf. Figure 1).

```
<terminals>
  <t id="s692_1" word="ჯონს" pos="NE" morph="Dat.Sg." />
  <t id="s692_2" word="ლოდინის" pos="NN" morph="Gen.Sg." />
  <t id="s692_3" word="გარდა" pos="ENPOS" morph="--" />
  <t id="s692_4" word="სხვა" pos="IPRN" morph="Nom.Sg." />
  <t id="s692_5" word="არაფერი" pos="NPRN" morph="Nom.Sg."/>
  <t id="s692_6" word="შეეძლო" pos="VMFIN" morph="3.Sg.Past.Ind" />
  <t id="s692_7" word="." pos="$." morph="--" />
</terminals>
<nonterminals>
<nt id="s692_502" cat="S">
    <edge label="SB" idref="s692_1"/>
    <edge label="OO" idref="s692_500"/>
    <edge label="DO" idref="s692_4"/>
    <edge label="HD" idref="s692_501"/>
</nt>
<nt id="s692_500" cat="NP">
    <edge label="CJ" idref="s692_2"/>
    <edge label="HD" idref="s692_3"/>
</nt>
<nt id="s692_501" cat="VP">
    <edge label="MO" idref="s692_5"/>
    <edge label="HD" idref="s692_6"/>
</nt>
<nt id="s692_VROOT" cat="VROOT">
  <edge label="--" idref="s692_502" />
  <edge label="--" idref="s692_7" />
```

Figure 1: An excerpt of the script for the Georgian sentence in TIGER-XML format.

An output of the *Synpathy* tool of a morphologically and syntactically annotated Georgian sentence

ჯონს ლოდინის გარდა სხვა არაფერი შეეძლო.
("John could do nothing, but to wait")

is depicted in Figure 2.

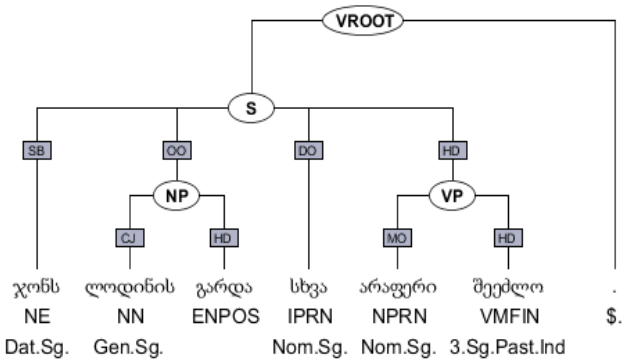

Figure 2: A morphologically and syntactically annotated Georgian sentence in TIGER-XML format.

A syntactically annotated German translation equivalent for the above Georgian sentence is displayed in Figure 3.

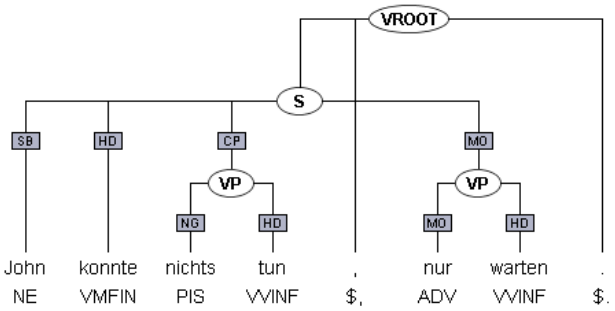

Figure 3: A syntactically annotated German equivalent for the Georgian sentence in TIGER-XML format.

The sentences in Figure 2 and 3 visualize a hybrid approach to the syntactic annotation procedure as tree-like graph structures and integrates annotation according to the constituency and dependency representations. Consequently, in a tree structure the node labels are phrasal categories, whereas the parental and secondary edge labels correspond to syntactic functions.

The same parsed sentences can be also displayed in the INESS treebanking environment as shown in Figure 4 and 5. The INESS-Project is an open system serving a range of research needs, offering an interactive, language independent platform for building, accessing, searching and visualizing treebanks: (http://clarino.uib.no/iness > Treebanks > Selection > Georgian > kat-gego-con)

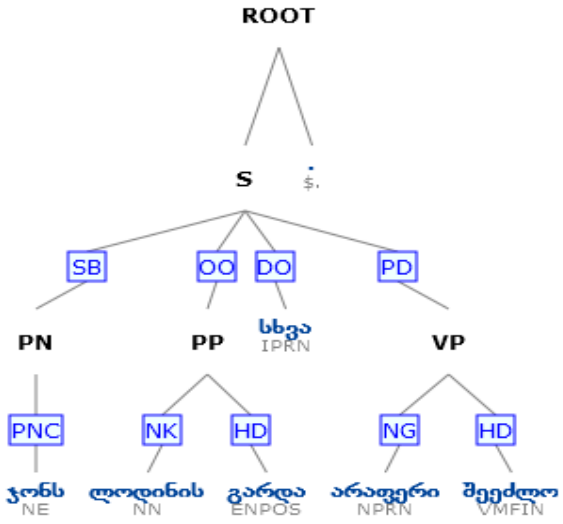

Figure 4: A syntactically annotated Georgian sentence in the INESS format.

The INESS-graph with a slightly different shape, unlike the TIGER –XML trees, is not capable to show a clear liner order of the punctuation marks for the original input sentence, hence visualising it under the ROOT node at the top of the consequent graph.

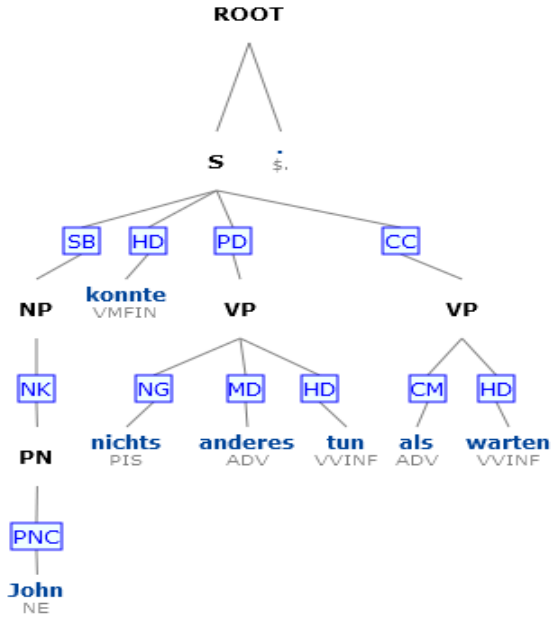

Figure 5: An alternative translation equivalent in German of an annotated Georgian sentence from Figure 1 and 3.

Syntactically annotated graphs for a Russian and an Ukrainian translation equivalents of the source Georgian sentence from Figure 2 are very close to each-other due to structural similarity of those languages:

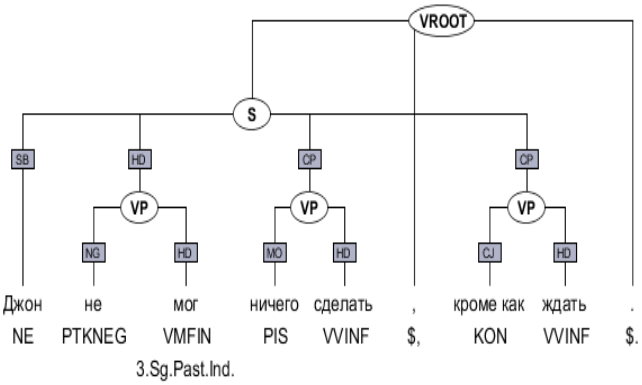

Figure 6: A translation equivalent in the Russian language of an annotated Georgian sentence from Figure 2.

As morphological annotation in those examples we present just grammatical features for the Russian and Ukrainian finite verb forms to fit into the restricted space of current submission format.

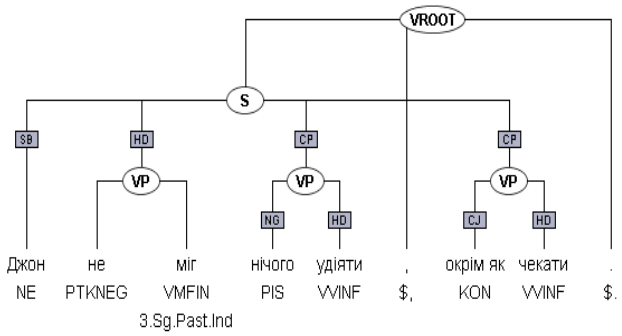

Figure 7: A translation equivalent in the Ukraininan language of an annotated Georgian sentence from Figure 2.

The Russian and Ukrainian languages typologically are more closely related languages also to German than Georgian. Consequently, the TIGER tagsets for these two languages underwent minor changes by incorporated additional POS and CAT features. The changes for the Georgian language tagsets and CAT values are more significant, but, in general, they conform to the TIGER annotation scheme used as a source in compiling the feature sets and their values for three new languages adhering to the mentioned scheme.

4. Alignment of the GRUG Monolingual Resources into Parallel Treebanks

For alignment of the monolingual syntactically annotated trees into parallel Treebank of German-Georgian, German-Russian, German-Ukrainian, Georgian-Russian and Georgian-Ukrainian pairs, we utilized *the Stockholm TreeAligner*, a tool for work with parallel treebanks which inserts alignments between pairs of syntax trees (Samuelsson and Volk, 2005), (Samuelsson and Volk, 2006). *The Stockholm TreeAligner* uses monolingual graph structures in the TIGER-XML format as representations for handling alignment of tree structures. The nodes and words from two languages with the same meaning are aligned as exact translation correspondences using the green colour. If nodes and words from one language represent just approximately the same meaning in the other language, they are aligned as "fuzzy" translation equivalents marked in the red colour.

Phrase alignment, as an additional layer of information on top of the syntax structure, shows which part of a sentence in one language is equivalent to a part of a corresponding sentence in the other language. This is done with help of a graphical user interface of *the Stockholm TreeAligner*. The phrases are aligned only if the tokens, that they span, represent the same meaning and could serve as translation units outside the current sentence context. The grammatical forms of the phrases need not fit in other contexts, but the meaning has to fit. However, syntactic annotation for the Georgian tree structures differ significantly from those of adopted in other languages involved in the GRUG initiative.

The most notable divergence in syntactic description model for the Georgian clause is a phenomenon classified as a mutual government and agreement relations between verb-predicate and noun-actants which number may reach up to three in a single clause. It anticipates control of the noun case forms by verbs, whereas the verbs in their turn, are governed by nouns with respect to a grammatical person. Nevertheless, constituency and dependency relations employed in the TIGER scheme is also powerful for the Georgian syntax description.

A notable structural difference between German and the other three languages involved in GRUG, is absence of articles as grammatical category in those languages. Its general functions in Georgian, Russian and Ukrainian take over as certain lexical items (Pronouns), as well as grammatical means.

From the structural view point, a significant divergence to be discussed, is the word order freedom in GRUG languages. For the German language there is an assumed basic word order, which is postulated to be either SOV in dependent clauses and SVO in main clauses. Quite frequently, within those statements, predictions about the Subject have been replaced by predictions about a general pre-verbal position, yielding XOV/XVO for German.

On a contrary, in Georgian the linguists admit a relative free word order as a result of its rich morphological structure. Nevertheless, a preferred basic word order without a Theme/Rheme bias for Georgian is SOV, which is canonical for the German dependent clauses.

The Russian and the Ukrainian languages are also morphologically rich languages, and, consequently, have a relative free word order, though, to a lesser extent than it is observed in the Georgian language. Despite the mentioned difference, a 1:1 alignment on word, phrase and sentence level can be often viewed in the GRUG parallel trees.

An implication of typological dissimilarity in word order between GRUG languages we can observe in respective syntactic structures. One of the interesting points discussed further concerns prepositional phrases (PP) which in German, Russian and Ukrainian are headed by prepositions standing on the first place in a phrase, whereas in Georgian its translation equivalent is Postpositional Phrase (PSP). In PSP some postpositions, as independent unchangeable words, stand alone and appear after noun. Some others adhere to the noun base form as an enclitic particle. Nevertheless, a German PP in Georgian, Russian and Ukrainian can be also translated by a phrase headed by a noun with a case inflection. This difference is shown on

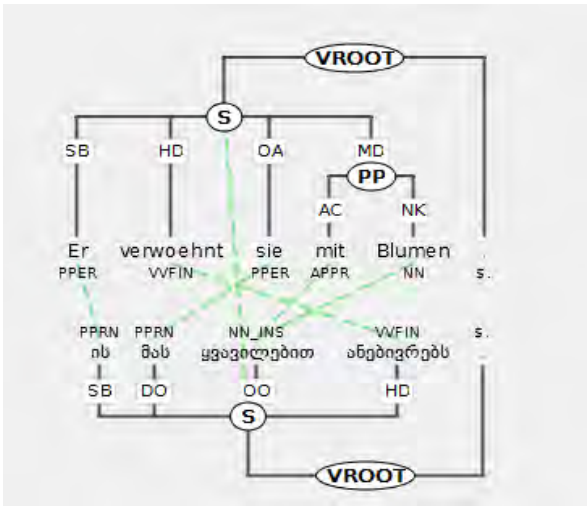

Figure 8: Divergence on a phrasal/constituent structure alignment level for German and Georgian.

an example of a German sentence

"Er verwöhnt sie mit Blumen"

(lit. He cossets her with flowers).

and the alligned Russian counterpart.

Besides the divergence in syntactic category labels, these constituents also differ from functional view point. In the German grammar they are considered as modifiers (MO), whereas in Georgian the PSPs traditionally are qualified as "ordinal objects" (OO). They differ from direct (DO) and indirect objects (IO) also formally, since later two are marked morphologically with specific affixes in verb, which is not the case with OO.

As already has been mentioned, in the Ukrainian and the Russian languages the German PPs in some cases are also translated by means of inflected word forms, but unlike the Georgian syntax trees, they are treated as modifiers (MO).

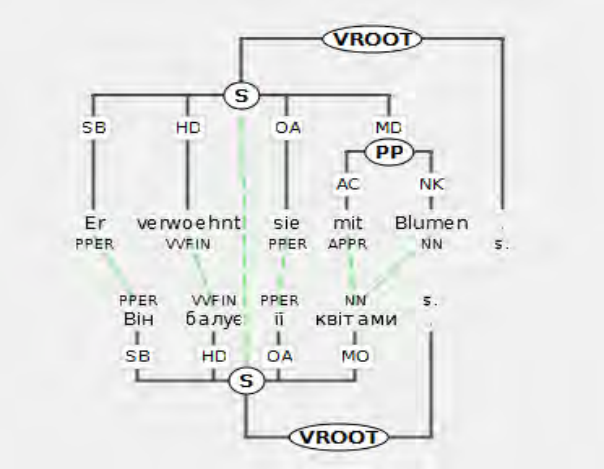

Figure 9: Divergence on a phrasal/constituent structure alignment level for German and Ukrainian.

The dicussed structural difference can be disregarded in the alignment process and the German PP "mit Blumen" (lit. "with/by means of Flower") considered as a "good" translation equivalent, though, a 2:1 alignment on a word level. The suggested solution derives from a prerequisit of " translation equivalence outside the current sentence context". In other words, a German PP

mit +N ("mit Blumen")

is always translated in Georgian as:

N+Instrumental_case („ყვავილებ_ით")

and in Ukrainian and Russian, consequently, as

N+ Instrumental_case ("квіт_ами")

N+ Instrumental_case ("цвет _ами").

Nevertheless, there is also an alternative solution to the alignment approach according to which 2:1 alignment should be regarded as a "fuzzy" equivalence and consequently marked in the red colour in the parallel trees. The second option is presented in Figure 10 with the same where PP is a Modifier (MD).

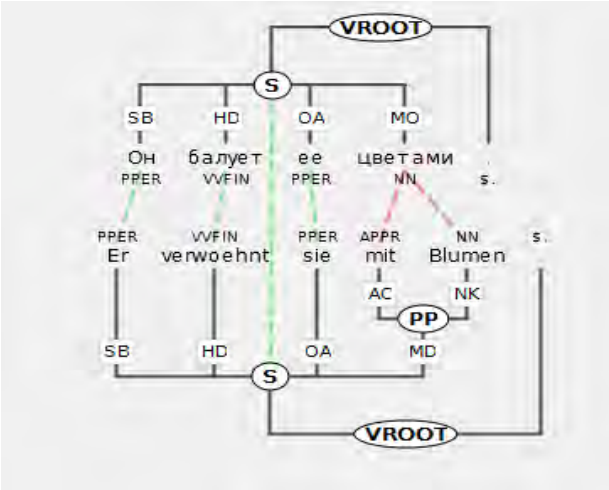

Figure 10: An alternative solution in alignment of a Russian-German parallel tree.

Despite the discussed 2:1 alignment for the PPs, in the Russian and Ukrainian we can also identify a reverse cases, with 1:1 equivalence as it is depicted in Figure 11 for the German Sentence,

"Sie unterhielten sich mit ihm über ihr Problem"

(lit. They discussed with him (about) her problem),

when a PP

mit +N ("mit Blumen")

is aligned to an equivalent one in Ukrainian:

з + N ("з ним").

The same issue can be observed also for the Russian

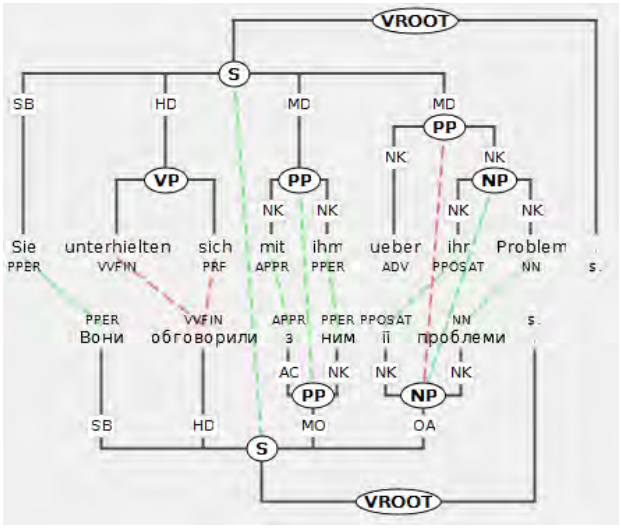

Figure 11: A 1:1 PP alignment on a constituent structure / phrasal level for German and Ukrainian.

parallel tree, whereas Georgian in both cases opts 2:1 alignment on word level, though, with different links on the phrase/constituent layer:

PP[mit ihm] ~ Geo. OO [N+postp "მასთან"]

and

PP[über Problem] ~ Geo.PSP [N+postp "პრობლემაზე"]

Nevertheless, we count them to be the "good" alignment, since they can serve as translation equivalents outside of the context.

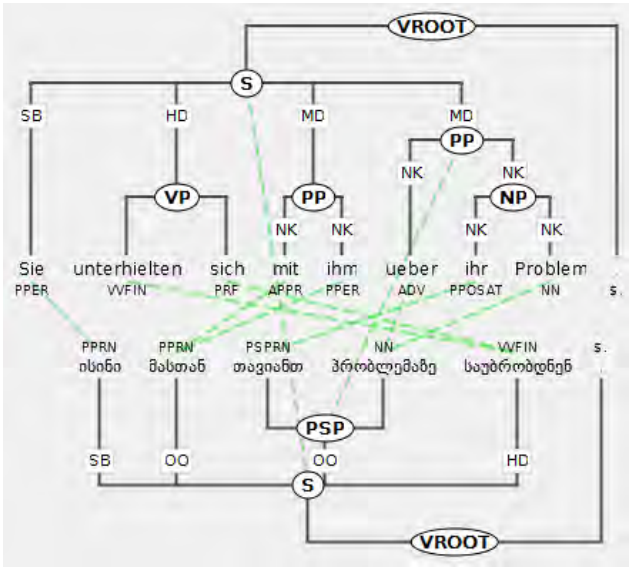

Figure 12: A "good" alignment on a constituent structure / phrasal level for German and Georgian.

In contrary, in Figure 13 is presented an example of a

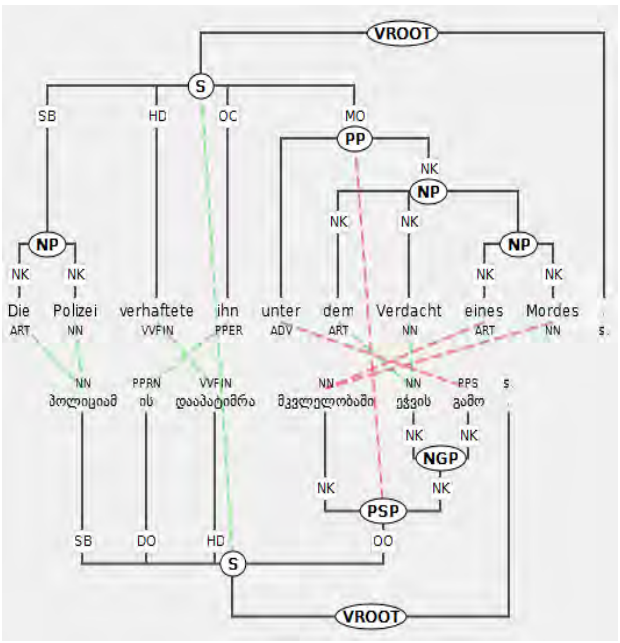

Figure 13: An example of a German PP and a Georgian PSP with "fuzzy" alignment.

"fuzzy" alignment between the German PP and the Georgian PSP counterpart in a sentence

"Die Polizei verhaftete ihn unter dem Verdacht eines Mordes"
(The police arrested him under a suspicion of a murder).

A preposition "unter" and a NP "eines Mordes" from the German PP are aligned to the Georgian counterparts "გამო" and „მკვლელობაში" from a PSP, as "fuzzy" equivalents, since they can not be considered as translation equivalents outside of this sentence context.

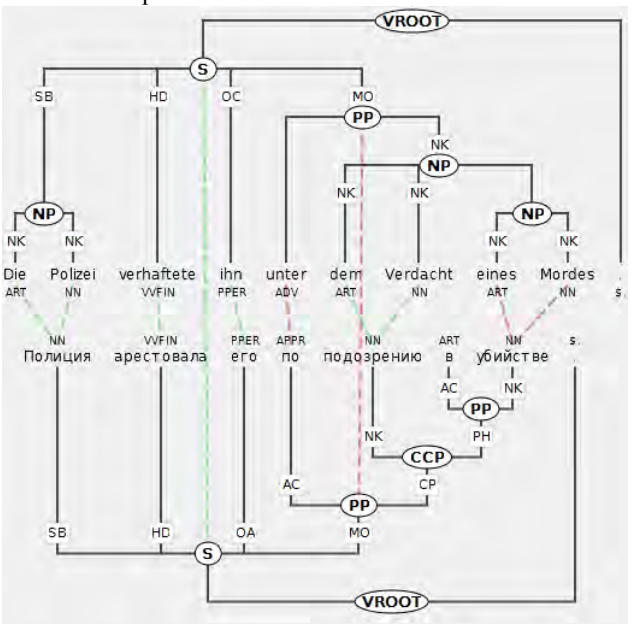

Figure 14: An example of a German and a Russian PP with "fuzzy" alignment.

In the respective German-Russian parallel tree the PPs are aligned also as "fuzzy" equivalents due to the same reason discussed in the above case of the German-Georgian parallel tree. Moreover, in an embedded PP "в убийстве" the preposition "в" remains without a German counterpart in the tree.

As for the same Georgian-Ukrainian and Georgian-Russian parallel aligned trees in Figure 15, we observe an example of a "good" alignment due to a complete match between terminal nodes. The only exception concerns the embedded NGP node of the Georgian tree and the Ukrainian/Russian PP node. NGP denotes a node for the Georgian postpositions that occupy a position immediately after the noun in Genitive case standing separately:

N+*gen*_case "გამო" ("due to", "because of").

Effectively, NGP is a functional equivalent of the Postposition Phrase (PSP) in Georgian. Therefore, in general, NGP could be considered as a tolerant to PSP and accepted as a "good" alignment in this parallel tree.

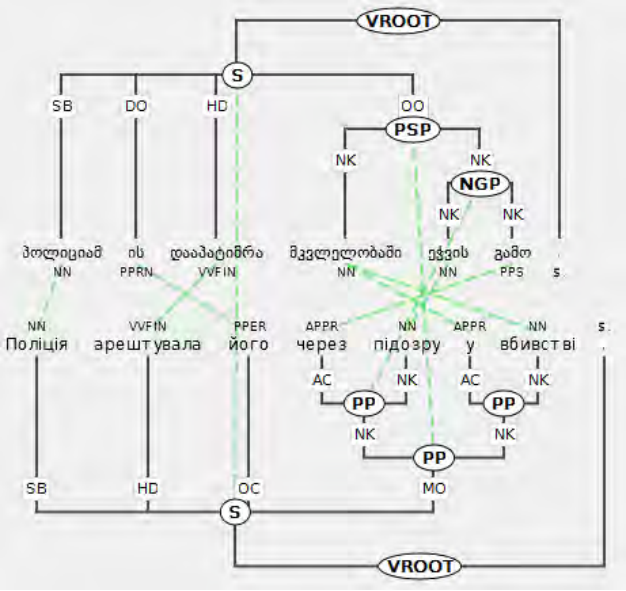

Figure 15: A "good" alignment on all levels for a Georgian-Ukrainian parallel tree.

The examples provided in Figure 14 and 15 may sound as discrepancy to a source statement according to which Russian and Ukrainian are typologically more close related to German than the Georgian language, though, in general the initial assumption remains true.

5. Conclusions and Future Research

In the present paper we gave an outline of an initiative for building multilingual parallel treebanks for three under-resourced languages - Georgian, Russian, Ukrainian and one of the technologically most advanced languages – German. In the second part we have specifically discussed issues of alignment parallel Prepositional/Postpositionsl Phrases in four GRUG languages.

Despite the divergences on syntactic level sketched and depicted in presented examples of PPs, the utilized mono and bilingual alignment tools (*Synpathy*, *the Stockholm TreeAligner*) are capable to cope with divergences in linguistic structures, on the one hand of German and, on the other hand of such typologically different languages as Georgian, Ukrainian and Russian. Therefore, they could be recommended as perfect tools for the other under-resourced languages in compiling parallel linguistically annotated Treebanks.

In the perspective we plan to extend coverage of the GRUG language pairs adding French, which the first experiments has been already done.

We are also considering to enrich linguistic annotation by adding *participant role / semantic case* labels to the constituency and dependency representations used in multilingual tree alignment procedures.

6. Acknowledgments

I would like to thank Prof. Dr. Elke Teich for giving me a chance to conduct this research in the framework of CLARIN-D project at the University of Saarland, Saarbrücken, and Mr. José Manuel Martínez Martínez for assistance in introducing the GRUG repository at the URL: http://fedora.clarin-d.uni-saarland.de/grug/.

References

Brants, S. and Hansen, S. (2002). Developments in the TIGER Annotation Scheme and their Realization in the Corpus. In *Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002),* Las Palmas, pp. 1643–1649,.

Grimes, S, Li, X., Bies, A., Kulick, S., Ma, X. and Strassel, S. (2011). Creating Arabic-English Parallel Word-Aligned Treebank Corpora at LDC. In *Proceedings of the Second Workshop on Annotation and Exploitation of Parallel Corpora*. The 8th International Conference on Recent Advances in Natural Language Processing (RANLP 2011). Hissar, Bulgaria.

Kapanadze, O., Kapanadze, N., Wanner, L. and Klatt, S. (2002). Towards A Semantically Motivated Organization of a Valency Lexicon for Natural Language Processing: A GREG Proposal. In *Proceedings of the EURALEX conference*, Copenhagen.

Kapanadze, O. (2010a). Verbal Valency in Multilingual Lexica. In *Workshop Abstracts of the 7th Language Resources and Evaluation Conference-LREC2010*. La Valletta, Malta.

Kapanadze, O. (2010b). Describing Georgian Morphology with a Finite-State System. In A. Yli-Jura et al. (Eds.), *Finite-State Methods and Natural Language Processing 2009, Lecture Notes in Artificial Intelligence,* Volume 6062, Berlin Heidelberg: Springer-Verlag, pp.114-122.

Kapanadze, O. (2009). Finite State Morphology for the Low-Density Georgian Language. In *FSMNLP 2009 Pre-proceedings of the Eighth International Workshop on Finite-State Methods and Natural Language Processing*. Pretoria, South Africa.

Kapanadze O. (2012). A German-Georgian Parallel

Treebank Project. In *Proceedings of the LREC2012 META-RESAERCH Workshop on Advanced Treebanking*. Istanbul, Turkey.

Kapanadze O. and Mishchenko, A. (2013). *Building Parallel Treebanks for Lesser-Resourced Languages: A Georgian-Ukrainian Treebank Proposal*. In *Cognitive Studies | Études Cognitives*, 13. Warsaw: SOW Publishing House, pp.195–207

Losnegaard, G, L., Gunn, I., Thunes, M., Rosen, V., De Smedt, K., Dyvik, H. and Meurer, P. (2012). What We Have Learned from Sofie: Extending Lexical and Grammatical Coverage in and LFG Parsebank. In *Abstracts* of *LREC-2012 META-RESEARCH Workshop on Advanced Treebanking,* Istanbul, Turkey.

Megyesi, B. and Dahlqvist, B. (2007). A Turkish-Swedish Parallel Corpus and Tools for its Creation. In *Proceedings of Nordiska Datalingvistdagarna (NoDaL- iDa 2007)*.

Megyesi, B., Hein Sa˚gvall, A. and Csato´ Johanson, E. (2006). Building a Swedish-Turkish Parallel Corpus. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*.

Rios, A., Göhring, A. and Volk, M. (2009). Quechua-Spanish Parallel Treebank. In *7th Conference on Treebanks and Linguistic Theories*, Groningen.

Samuelsson, Y. and Volk, M. (2005). Presentation and Representation of Parallel Treebanks. In *Proceedings of the Treebank- Workshop at Nodalida*, Joensuu, Finland.

Samuelsson, Y. and Volk, M. (2006). Phrase Alignment in Parallel Treebanks. In *Proceedings of 5th Workshop on Treebanks and Linguistic Theories*. Prague, Czech Republic.

Smith, G. (2003). *A Brief Introduction to the TIGER Treebank, Version 1*. Potsdam Universität.

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Martin Benjamin†, Paula Radetzky‡
†EPFL — Swiss Federal Institute of Technology, Lausanne, Switzerland
‡Kamusi Project International, Geneva, Switzerland
martin.benjamin@epfl.ch, paula@kamusi.org

Abstract

This paper looks at the challenges that the Kamusi Project faces for acquiring open lexical data for less-resourced languages (LRLs), of a range, depth, and quality that can be useful within Human Language Technology (HLT). These challenges include accessing and reforming existing lexicons into interoperable data, recruiting language specialists and citizen linguists, and obtaining large volumes of quality input from the crowd. We introduce our crowdsourcing model, specifically (1) motivating participation using a "play to pay" system, games, social rewards, and material prizes; (2) steering the crowd to contribute structured and reliable data via targeted questions; and (3) evaluating participants' input through crowd validation and statistical analysis to ensure that only trustworthy material is incorporated into Kamusi's master database. We discuss the mobile application Kamusi has developed for crowd participation that elicits high-quality structured data directly from each language's speakers through narrow questions that can be answered with a minimum of time and effort. Through the integration of existing lexicons, expert input, and innovative methods of acquiring knowledge from the crowd, an accurate and reliable multilingual dictionary with a focus on LRLs will grow and become available as a free public resource.

Keywords: multilingual lexicography, crowdsourcing, gamification

1. Introduction

Humans do a remarkable job of transmitting linguistic data from one generation to the next. Not just parents, but entire communities, transfer innumerable lexical elements, including pronunciation, grammar, syntax, and usage information. We have done a remarkably poor job, however, of downloading such data into forms that can be stored and operated on outside of our Cranial Processing Units (CPUs).[1] This paper looks at the challenges that the Kamusi Project faces in acquiring open lexical data for less-resourced languages (LRLs) of a range, depth, and quality that can be useful within Human Language Technology (HLT).[2] These challenges include accessing and reforming existing data sets into interoperable data, recruiting language specialists to work with new and existing data, locating and working with non-specialist speakers, and funding the requisite work. We lay out the issues facing data collection for LRLs, then look in particular at a crowdsourcing schema, including our mobile phone application, which we have developed to elicit high-quality structured data directly from each language's speakers.

[1] Using your brain, you understood the wordplay with CPU almost immediately. It is unlikely that today's best artificial intelligence could decode the linguistic subtleties embedded in the pun.

[2] The Kamusi Project began as *The Internet Living Swahili Dictionary* at Yale University in 1994. In 2007, the project spun off as an independent non-governmental organization dedicated to the production of language knowledge resources. Kamusi Project USA is registered in Delaware as a 501(c)(3) non-profit corporation, and Kamusi Project International enjoys the equivalent status in Geneva, Switzerland. As of 2013, the informatics aspects of the project are housed at EPFL, the Swiss Federal Institute of Technology in Lausanne.

2. Acquiring Lexical Data for LRLs

Even for well-resourced languages, much recorded lexical data is neither available nor codified in a deeply interoperable form; for example, from the *Oxford English Dictionary* on, no source of English lexical data has been at once open, reliable, well-structured, and richly elaborated. LRLs are even less likely to have comprehensive lexical data. Most LRL dictionaries are small and basic, with few terms and little information beyond a part of speech and a major-language gloss. Exceptions exist in the form of print tomes researched over many decades (e.g., Young & Morgan, 1987; Matisoff, 1988; Hill et al., 1998; Coupez et al., 2005; Cole & Moncho-Warren, 2012), but most such works are not available in machine-usable format, nor are they economically accessible to most LRL speakers. Furthermore, the lexical data published within the past seventy years that has been digitized for LRLs is generally copyrighted, and if the owners can be located, they are often reluctant to share.

In the effort to create a massively multilingual online dictionary, the Kamusi Project has established a system that can accommodate an unlimited amount of lexicographic data within a single, consistent data structure (Benjamin, 2014). The system is designed around the production of monolingual dictionaries for each language, interlinked to other languages at the level of the concept. With each concept in a particular language treated as an individual entity, we are able to elaborate associated data that can be used for natural language processing, machine translation, and other HLTs. Any feature of a particular language, such as the numbers and types of possible morphemes and inflections for each part of speech, alternate scripts, or tone spellings, can be

handled by the project architecture. Over time, each monolingual entry can come to contain a large amount of rich structured data, including intra-language relations, etymologies, examples, and geo-tagged pronunciations and sightings, as well as unstructured information such as usage and cultural notes. Once a monolingual entry has been created, it can be linked to a concept in another language, with a degree of equivalence specified as parallel, similar, or explanatory. Kamusi then shows the train of transitive links from the second language, marking degrees of separation. In this way, each language develops as a full internal monolingual resource that is simultaneously a multilingual communications gateway to every other language in the system. Once an entry is approved into the system, it becomes part of an open-access data set that is available to the public and to machines through a raft of emerging technological tools for online, mobile, and offline use.

Data for Kamusi comes from three types of sources: (1) existing data sets; (2) direct input from language specialists; and (3) controlled input from the crowd. There is substantial interplay among these categories (Nguyen et al., 2013); imported data may be used as part of the process of validating crowd submissions, experts may approve or revise imported or crowd data, and the crowd helps validate imported data and adds details such as pronunciations, examples, and images to entries produced by specialists. A major method for eliciting entries from specialists and the crowd is via reference to a prioritized list of concepts derived from English, using data from both corpus analysis and topical word lists (Benjamin, 2013). Using English as a starting point can be methodologically problematic and is being addressed by ongoing programming, but it is not possible to use corpus approaches to generate wordlists for many LRLs due to a paucity or absence of digitized written material.[3] For languages with a written record, corpus-based lexicon development can occur when a team is in place that can take on the intensive tasks of assembling the records or gaining copyright permissions to an existing corpus; future plans include tools to harvest lexical data from online sources and, when users grant permission, from translation services that interact with Kamusi. In the near term, however, the English-based list gives us a starting point that enables the rapid growth of lexicons that bring together many languages, with the challenges discussed in the following sections.

2.1 Existing Data Sets, Incommensurate Data, and Intellectual Property

Existing data sets offer substantial benefits, but also considerable challenges, to the multilingual dictionary project. The benefits of bootstrapping the project with data that has already been researched and digitized go beyond the obvious savings of time and effort. Much invaluable work currently languishes in isolation, whether in a field researcher's shoebox, a print volume on a library shelf, or even a web page devoted to an individual LRL. The multilingual dictionary provides a central home where all such data can be readily located, and a platform to link the work produced as open data for one language to a great deal more work on the same and other languages (potentially including non-lexical data, such as items in the ELAR and PARADISEC archives),[4] thereby augmenting the utility of previous accomplishments. Lexicography can be the labor of years, often in remote field settings, producing data that cannot be replicated and should not be lost. In many cases, dictionaries from decades past are historical documents that preserve language data prior to contemporary influences such as migration and assimilationist language policies. Preserving data, making it accessible, multiplying the power of what can be done with it, and accelerating the inclusion of LRLs in the multilingual framework are all advantages conferred by mining previous lexicons.

The challenges of existing data, however, are manifold. The Kamusi Project is refining a system for merging existing data sets into our structure—but perhaps "data sets" is a poor description of what is available. Traditionally, the author of a dictionary determines which elements to include, in what format, and in which order for their publication. As Haspelmath (2014) points out, individual dictionaries for LRLs are not readily comparable even for side-by-side perusal. Many entries are composed as undifferentiated text blocks, often without a consistent structure from one line to the next. For example, this is an entry from a Swahili-Mandarin data set that is currently being prepared for incorporation into Kamusi, with evident difference between the type of data that comes after the 1 and the 2: "**-amba I** *kt* 1. 说某人的坏话，议论某人 ***Usikae bure na kuamba watu.*** 你别干坐着说别人坏话。 2. <旧> 说." Determining what the fields are, and converting scanned or text-file dictionary blocks into data that can be categorized in a database, can itself be an enormous undertaking. Furthermore, most dictionaries group polysemous items together under a single headword, while Kamusi's multilingual structure requires each sense to be disaggregated concept by concept, polyseme by polyseme. Prior to merging, many data sets demand a tremendous amount of manipulation, much of which cannot be automated (see Hernández & Stolfo, 1998; Lee et al., 1999; Dong & Naumann, 2009). For instance, in the Swahili-Mandarin case, we have been able to isolate and segment the data (more than 10,000 entries) into individual data points, but not distinguish automatically between glosses, example sentences, and special usage explanations. The lexicographer is left with the task of manually shifting the Mandarin elements to their correct fields within a spreadsheet prior to importing to the online system.

[3] To address these issues, Kamusi is developing a system of "balloons" to levitate concepts that are important in languages related by linguistic family, geography, or cultural features (Benjamin & Radetzky, under review).

[4] http://www.elar-archive.org/index.php and http://www.paradisec.org.au/home.html

Once a data set is ready to be merged, each entry must be reviewed individually. Even in the best cases, when data has been curated using software such as TLex[5] or SIL's Toolbox or FieldWorks[6] and therefore does not need cleansing, it remains impossible to align senses without a human eye. It is not enough to know that a particular term has been defined, for example, by English *light*, which has a great number of homophones. Without disambiguation of the specific sense ('not heavy', 'not dark', 'not serious', etc.), the entry cannot be brought into the multilingual system. The merging engine, still under development, will display the definitions of possible matches to another language of the user's choice, not necessarily English, or offer the option to add a new matching sense in a linking language. This process requires humans who know the language well, whether an expert working for love or money, or a large enough number of crowd members to produce a reliable consensus.

After merging, the data may still be inadequate for the requirements of the multilingual dictionary; in particular, most data sources do not include own-language definitions needed to build the monolingual core for each language. Additionally, most bilingual data sets, which constitute the bulk of existing data for LRLs, include terms that do not yet have translations in Kamusi, so a provisional sense indication in a language already in the system is necessary in order to prevent those terms from hiding as orphans outside of the multilingual framework.

Beyond the technical challenges lie issues of intellectual property. In some cases, ownership of the data cannot be determined. For example, Sacleux (1939) was written by a priest who died in 1943 without heirs. Neither his religious order nor the successor to the museum that published his dictionary wished to prevent use of the data, but neither would take responsibility for authorizing its release. A decade after first attempting to obtain permission, the data is finally in the public domain as of this year (2014). Researching the ownership trail of each LRL data source and then writing letters and awaiting responses, or waiting until seventy years after the death of the author, all to secure permission for works that must then be scanned, cleaned, and converted from text to data, is not a winning strategy for data acquisition.

Even when copyright ownership is clear, acquiring usage rights can be difficult. Publishers do not easily relinquish data that they have obtained under contract, even for an out-of-print work in a small-market language. When publishers are willing some LRL lexicographers (or the organizations they are affiliated with) do not want to share their product. After decades compiling the definitive reference work for a particular LRL, many authors wish to keep rights to hypothetical royalties and retain control over how the data will be presented. Conversations can stretch for months and then break down when the author places an untenable condition on the release of the work, such as the ability by the author to remove data after it has already been merged into the system, or a copyright license different from the Creative Commons Attribution Non-Commercial Share Alike license[7] that has been established for data within the larger Kamusi Project.[8]

It is hoped that authors and organizations will become more interested in sharing their data as the Kamusi multilingual dictionary, maintained by a non-governmental organization with a charter to produce language resources to be shared with the public for free in perpetuity, grows and is able to demonstrate the advantages that joining the project can bring to a language community. For example, work to integrate more than one hundred LRL lexicons developed by the US Peace Corps is expected to begin when the merging engine is complete, after optimizing our mobile app (discussed in §3.2) for low-bandwidth African telecommunications. Again, however, securing the blessing to use existing data only brings it to the point where it must face the technical challenges discussed above.

2.2 Language Specialists

The ideal way to collect lexical data is to have language specialists contribute rich data for every entry, using a comprehensive online edit engine constructed with standard web form elements customized for each language. Such contributions can be considered authoritative (Kleinberg, 1999) and can provide the full range of information needed for the term to be understood by humans and manipulated by HLTs. Specialists can work from the above-mentioned list of concepts derived from English, or they can use another reference language as in the Swahili-Mandarin case above, or bring in terms that are unique to their LRL (Bergenholtz & Nielsen, 2013).[9] The specialists add depth and nuance that cannot come from existing static data and might not be elicited from the crowd. However, working with experts is not without its challenges.

The first problem is identifying people to work on a language. The world's leading authority on a given language may not be the person to bring it into a multilingual dictionary. To begin with, the person may have already published a dictionary that is encumbered by copyright or that they do not wish to share. Additionally, such experts are often academics tied up with other research and teaching. Furthermore, in contrast to books and articles, dictionaries do not weigh highly in tenure and promotion considerations, and participation in a col-

[5] http://tshwanedje.com/tshwanelex/

[6] http://www-01.sil.org/computing/toolbox/ and http://fieldworks.sil.org

[7] http://creativecommons.org/licenses/by-nc-sa/3.0/

[8] Source code is not currently open because we do not have enough staff resources to vet incoming contributions, and it is problematic to release code that would lead to other versions of what must function as a unified project. The code base will be opened when the project has the staff capacity to manage externally-developed programming components.

[9] Kamusi's revised approach to the methodological difficulties of starting with a concept list keyed to English is addressed in footnote 3.

laborative project with indeterminate authorship contributes even less to a CV. Sometimes the leading expert is best equipped to offer guidance and perhaps references to people with the time to do the work.

In addition, knowledge of a language does not necessarily imply the ability to document it within the Kamusi framework. Lexicography is a complicated endeavor to begin with, and Kamusi's multilingual model adds new complexities in the pursuit of creating a detailed matrix of the human linguistic experience. While the current project is built on an editing input system that strives to be clear and user-friendly, aspects remain difficult or non-intuitive. Training is necessary so that contributors, even PhDs with experience in lexicography, can understand the purpose of each field and the formats required for the data to be useable and consistent. It is especially difficult, and particularly important, to teach participants how to write good own-language definitions. Before contributors can be given moderator privileges to confirm data as finalized, they must go through a period of training and observation to determine that they understand the technical and philosophical aspects of producing high-quality data.

It is possible to find volunteer participants who are both interested in, and capable of, rigorous lexicographic work; however, expert contributors are more likely gotten with remuneration. Producing a high-quality entry, including an own-language definition, takes five minutes or more. At that speed, ten thousand entries is a year of labor. Few people have a year or more to donate to their language. Although a volunteer might start out with the best of intentions, financial incentives are a more reliable way of ensuring that the work is accomplished (Bederson & Quinn, 2011). A system is under design to pay experts per lexical term, although, ironically, we have not yet been able to fund the coding through to implementation. Quality control will be a challenge because project management cannot possibly know all the languages in which data is supplied, so this is integrated into the crowdsourcing elements discussed below.

The largest hurdle with language specialists, then, is funding. The costs are not especially high per term, and become infinitesimal when extrapolated to clicks over time, but they are a substantial up-front obstacle when the number of words in a language is multiplied by the number of languages worldwide. Funders have many priorities, among which language resources generally rank low. The Kamusi Project has internal task forces to find funds for particular languages or regions and welcomes all suggestions.

Language specialists are being recruited from a variety of institutions, with more than twenty institutions represented in the multilingual pilot phase completed in February 2013. The invitation is open to academics who are actively working on projects for their languages, or who wish to develop a joint proposal to take advantage of the resources that the Kamusi Project offers. We also solicit citizen linguists, that is, people who are both passionate about their language and have the time and skills to invest. (These citizen linguists using the expansive edit engine are not the same as "the crowd" using the constricted app, discussed in §3.2 below.) One of our models is DEX Online, a monolingual Romanian dictionary, which has built a compelling resource with much volunteer labor from Romanian Scrabble players.[10] Retirees with computer skills and spare time are another community that might be tapped for particular languages, providing a stimulating activity in support of a cultural legacy. In terms of remunerated efforts, the Kamusi Project is currently using NEH grant funds to provide student stipends and training at the University of Ngozi in Burundi, in exchange for data development in the Kirundi language. A related method well-suited for LRLs would be grant support for graduate field researchers. More expensive, but benefiting from contracts and enforceable expectations, is the possibility of working with professional translators. In all cases, the challenge is to match people who can do the work with an appropriate reward for getting it done well.

2.3 Crowdsourced Data Collection

For many languages, reliance on language specialists will be too slow to generate useful data, even if a specialist can be located. Furthermore, specialists do not know and do not have the ability to provide every detail of each word in their language. In fact, certain data elements such as regional pronunciation recordings can only come from a wide assortment of contributors. In order to speed progress and provide greater depth and range, techniques are under development to generate linguistic data from the crowd, as discussed below in §3.2. However, crowd-generated data is notoriously unreliable, so the system is being designed with numerous redundancy and reliability checks. Crowd data must always be subject to rigorous validation procedures, labeled for provenance, and be editable by specialists.

Wiktionary provides a case study in the dangers of crowdsourcing a dictionary. The project is to be commended for seeking a fantastic range of linguistic data. Yet, the open architecture invites mischief and mistakes, and inhibits error-checking. For example, as of this writing, a spam English definition of *spring* as 'erection' has persisted in various forms since being added by an anonymous user in 2006. Definitions are sometimes circular, or one-word synonyms. It is simple to add erroneous translations, which then propagate bad automated data throughout the system. The majority of elements are written in wiki markup language, which is a near-impenetrable barrier to most people's participation. While Wiktionary continues to improve, its laissez-faire approach to crowdsourcing leaves it inconsistent and unreliable as a source for lexical information. As a worst-case example, the Malagasy Wiktionary contains an ever-expanding collection, three million pages and counting, of robot-generated gibberish that the organization has been unable to limit or expurgate (Andrianja-

[10] http://dexonline.ro

nahary, 2013).

Crowdsourcing involves these and several other issues, enumerated here. First, most users prefer to receive information rather than contribute their own knowledge. Second, channeling users to contribute specific types of data requires a well-developed process flow. Third, users can introduce inadvertent errors, such as spelling or formatting problems. Fourth, complex tasks such as writing definitions require training and are not suitable for all users. Fifth, malicious users can intentionally introduce bad data. Sixth, even well-intentioned users can introduce data that turns out to be wrong. Seventh, finding a crowd that is large enough to support the redundancy necessary for validation is difficult for many LRLs, especially those with few speakers or poor communications infrastructure. Eighth, the enthusiasm of individual members of the crowd will be difficult to maintain over the years it takes to collect tens of thousands of terms for a language. With the proper methodology and safety checks in place, however, the crowd can become an important source of data for hundreds of languages. In §3, we present our crowdsourcing model to address these issues.

3. A Preliminary Crowdsourcing Model for LRLs

LRLs face a special challenge: With few existing resources, most LRL Internet users do not expect to encounter their own language, nor do they have a history of participating in its resource development. The crowdsourcing model we are developing is designed to change that by making lexicon development fun, easy, and rewarding. Here, crowdsourcing denotes the completion of specific targeted tasks, as distinct from making use of the in-depth editing system that is anticipated to be mostly for citizen linguists and language specialists.

3.1 Motivating Crowd Member Participation

The first incentive of the system will be to channel users to register for the site. Users will have two options, registering for free access or paying a subscription fee. Free access will come with an asterisk—people can earn usage points by answering questions. This "play to pay" system will give users points for proposing translations, writing definitions, or providing other data such as usage examples culled from online sources. Points can be exchanged for dictionary lookups, and high earners may also win material prizes or rewards that appeal to the psyche, such as congratulatory postings on Facebook. Start-up points will be awarded for registering and providing initial survey data, including language experience.

Points will also be awarded for participating in games (Castellote et al., 2013; Paraschakis, 2013; Hamari, Koivisto, & Sarsa, 2014). One game will be a word race, where an English term and definition will be sent to the players, who will be competing both individually and as part of a same-language team against those working on other languages. When players receive the term and definition, they will send back a translation of that term in their language. When ten answers agree, the person who sent in that answer first will get ten points, the next will get nine, etc. Additionally, all the members of the same-language team will get points based on the order in which their language has completed this task (and slower teams will be given an easy form to recruit more members). Another game will then put the term out for own-language definitions, which will be voted on, with points awarded to the winning author and the people who voted on that definition. Similar gamification will be designed to flesh out other data elements. These games will evolve from the logic of the mobile application discussed below.

Motivation will also be stimulated through social rewards. Users who contribute winning definitions will have their accomplishments posted on their favorite social media (Antin & Shaw, 2012). They will also appear on leader boards on the site, with rankings shown within a language and among languages.

Finally, when we can find sponsors to cover the costs, material prizes such as shirts and clocks will be periodically awarded to the winners of specific limited-time competitions. Competitions for these prizes will often focus on quickly augmenting lexicons for new LRLs as they join the multilingual dictionary. This combination of motivations will be experimented with and and successful approaches expanded, in order to stimulate as much participation as possible.

3.2 Steering the Crowd with Targeted Questions

The researchers at the Kamusi Project have developed a mobile phone application with targeted questions that direct users to provide data in exactly the format required. With the working name "Fidget Widget," the app is envisioned to be used by language aficionados in the small moments when they look to their phones to fill time. The app is in testing as of this writing, with the expectation that it will be demonstrated for LRLs at the May 2014 LREC workshop, "CCURL 2014 – Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era." Increased functionalities will be added to the app over the course of time, including additional types of questions and features for field researchers to collect oral data for talking dictionaries (Benjamin & Radetzky, 2014).

Initial testing of the app will provide data that we can use for determining the thresholds at which we accept crowd data as good enough to either queue for expert moderation (lower threshold) or publish as world-viewable, validated data (higher threshold). While the relative costs of majority decisions versus control groups in crowds have been modeled (Hirth, Hoßfeld, & Tran-Gia, 2013), a numerical standard does not yet exist for statistically determining the point at which different types of crowdsourced data can be considered valid. We expect experiments will show crowd validation can accurately indicate that an item is either very good or very bad, but

that ambiguous evaluation results from the crowd will be useful mostly to indicate entries to be queued for specialist review.

For the initial version of the app, we are interested in two types of information: (1) What is the target language equivalent of a defined source language concept? (2) What is the definition of the term in its own language (i.e., its own-language definition)? We are consciously postponing using crowdsourcing to address lexicographic questions that require subtle understanding of complex ideas, such as the degree of equivalence between the source and target term—even such basic questions as the part of speech of terms proposed by the crowd might be better left to specialist review.[11] However, we are interested in seeing whether this method yields independently-generated own-language definitions of target terms—ones that will allow readers to understand, for example, the subtle differences between *connaissance* and *savoir* in French—or whether crowd definitions tend to be close translations of the source definition, in this case the definition we provide for *knowledge*, which would not be fine-grained enough to distinguish *connaissance* from *savoir*. (See also Haviland (2006) and Svensén (2009) for a discussion of such issues.) The data generated by the app in the current stage is intended to provide a starting point for richer dictionary entries and deeper lexicons that will be expanded later.

To find a target equivalent of source language term, we first ask an open-ended question to several users. The ideal crowd member is a native speaker who is also comfortable in the source language, but people who have acquired the target language later in life cannot be excluded, on the premises that (1) language learners have much to offer based on the concerted efforts they often make to master concepts that native speakers might never notice, and (2) errors will be weeded out by the bulk of the crowd and by contribution analysis. We present the source language term and definition (e.g., *light* 'low in weight') and ask, "What word would you use in [your language]?"[12] If we receive a critical mass of identical answers (we have not yet defined the precise number), then the term will be advanced to the next level of moderation or crowd review. However, if we obtain differing responses to the same question, we next show another set of users the source term (here, *light*) and definition ('low in weight'), and ask, "Is [term] a good translation in [your language]?" For this question, counting thumbs up or thumbs down will allow us to evaluate whether a submission is a near-synonym or a mistake.

After a translation term passes the validation threshold, we seek the target language definition by displaying the original term *(light)*, its definition ('low in weight'), and the target language term, and ask, "Can you give a definition in [your language]?" It is important to show the source original in order to ensure that we do not get a definition for a homophonous term in the target language. (Before writing a definition, each user sees a screen that explains the basic lexicographic requirements, with the choice to opt out of the task.) After receiving the first submission, we display the term and proposed definition to other users and pose the question, "Is this a good definition?" If subsequent users approve, then we advance the definition to the moderator or to validated status. However, if other members of the crowd are dissatisfied, then we solicit the definition anew. When we have two definitions in competition, we show both and ask, "Which definition is better?" Through a formula that will be established when test data is available for evaluation, a winning definition will be advanced to moderation or validated status after reaching a certain level of satisfaction among the crowd.

In the future, many questions for the app and games will be generated by information arriving from existing data sets. For example, if an imported bilingual lexicon indicates that a word in a user's language matches a word that has multiple English senses, the user will be asked to choose the most appropriate sense or suggest a new one. Once enough users have agreed on a sense disambiguation for imported data, the system will steer toward adding definitions, morpheme information, and other elements to fill out the entry. Other questions will seek to group and rank search results that yield multiple entries. On the premise that many crowd members will use the app in short bursts (for example, to answer a single question in order to unlock a device from idle), the questions will be designed to elicit either very short text answers, evaluations of mined data, or judgments about other users' contributions through yes/no or X vs. Y questions. As the system grows, it will be possible to expand questions to demonstrated user interests—for example, asking about other terms in a semantic field that a user has accessed in their current session. Tailoring questions will require some experimentation to discern what strategies are effective (Bergenholtz & Johnsen, 2013).

3.3 Contribution Analysis

Central to the crowdsourcing model will be the analysis of user contributions. It is important to know which users provide consistently good answers versus who comes in wide of the mark. Good answers are those that climb toward a consensus opinion. Bad answers are those that are severely rejected by the crowd. Some answers may be ambiguous—for example, if contributors propose essentially synonymous translations for the same term. (In the model, competing answers that both gain upvotes have equal opportunity to move toward incorporation into Kamusi, with the more popular answer winning primacy in the display hierarchy.) Users who consistently produce good answers will earn trust; trust levels will

[11] While Kamusi has a simple method for matching concepts represented by different parts of speech, such as linking the Swahili verb *–furahi* with the English adjective *happy* via the translation bridge 'be happy', this nuance is not obvious to untrained users. Parts of speech are given provisionally based on the source language, but flagged as questionable until confirmed by a moderator.

[12] In principle, all questions will be localized to the target language.

be displayed on site and optionally on a user's chosen social media. These participants will have their votes on other users' contributions weighted more heavily, and they will have their answers integrated more quickly: their submissions will require fewer votes for validation. On the high end, trusted users will earn the right to moderate contributions that correspond to their demonstrated skill sets, gaining the rank of language specialists with the authority to finalize data as valid for incorporation into the master database.

Conversely, users who consistently score poorly will be diverted to questions that more closely match their skill sets. Easier questions might include evaluation of illustrative photos for appropriateness; voting on whether other users' comments are useful or spam; or recording the pronunciation of a word in their native language. The objective will be to find a level for each user at which they provide useful data and feel comfortable. Having multiple users effectively scoring each other's contributions will result in error checking that builds in good data, weeds out the bad, and creates incentives for users to submit their best possible answers.

Some users are intentionally malicious, and refinements to Kamusi's crowd system are on the drawing board to ferret out these out. Intentional subversion of the system is expected to be less than in previously-studied crowd situations, where paid contributors benefited financially by quickly submitting sloppy work (Kittur, Chi, & Suh, 2008; Suri, Goldstein, & Mason, 2011). However, our ongoing battle against spam registrations and discussion posts shows that some maliciousness is inevitable. In addition to normal channels to flag suspect submissions, including wrong data submitted in good faith, analysis of crowd responses will alert moderators to patterns consistent with abuse. Vandalism might sometimes be difficult to detect because malicious users can mix in valid responses with their spam. They might also attempt to slide their handiwork into obscure concepts in low-volume LRLs, as happens in Wikipedia, or distribute their damage across languages. Algorithms for monitoring ill intent will need to evolve. What is certain is that users who are determined to be vandals will be banished, and all of their submissions will be removed or, if their items have been expanded on subsequently, isolated for further review.

Contribution analysis will require us to keep careful track of the interacting histories of users and entries. This is an informatics challenge rather than a linguistic one, the design of which will be tasked to computer science partners.

4. Conclusions

In order to transfer human linguistic knowledge from people to their machines in a massively multilingual data resource, a number of integrated strategies must be implemented. Existing data sets offer a starting point but require extensive manipulation and human review. Language specialists bring much-needed expertise but can be difficult to locate and engage. Crowd sources have a great diversity of knowledge, but that knowledge is extremely difficult to collect in a systematic and structured fashion. A system to elicit and validate the maximum amount of high-quality linguistic data must therefore combine tools for data import and merging, detailed expert contributions, and regulated crowdsourcing. The Kamusi Project has implemented a web platform and mobile app to address these issues for any language, with refinements constantly in progress. The project is now beginning to use these tools for the collection of reliable data for numerous languages. Through this integrated approach, it will be possible to build in-depth, open lexical data sets and related HLTs for any language, and in particular for currently under-resourced languages where data, specialists, and crowd members can come together in a common resource working toward shared goals.

5. References

Andrianjanahary, R. (2013). My history of the Malagasy Wiktionary. http://terakasorotany.wordpress.com/2013/03/27/my-history-of-the-malagasy-wiktionary.

Antin, J., Shaw, A. (2012). Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India. In *Proceedings of the 2012 ACM Annual Conference on Human Factors in Computing Systems*, pp. 2925--2934.

Bederson, B., Quinn, A. (2011). Web workers unite! Addressing challenges of online laborers. In *Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems*, pp. 97--106.

Benjamin, M. (2013). How we chose a priority list for dictionary entries. http://kamusi.org/priority-list.

Benjamin, M. (2014). Collaboration in the production of a massively multilingual lexicon. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14).*

Benjamin, M., Radetzky, P. (under review). Small languages, big data: Multilingual computational tools and techniques for the lexicography of endangered languages.

Bergenholtz, H., Johnsen, M. (2013). User research in the field of electronic dictionaries: Methods, first results, proposals. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 556--568.

Bergenholtz, H., Nielsen, S. (2013). The treatment of culture-bound items in dictionaries. In R. Gouws, U. Heid, W. Schweickard, & H. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: de Gruyter, pp. 469--481.

Castellote, J.; Huerta, J.; Pescador, J. and Brown, M. (2013). Towns conquer: A gamified application to collect geographical names (vernacular names/ toponyms). In *Proceedings of the 15th AGILE Internation-*

al Conference on Geographic Information Science.

Cole, D., Moncho-Warren, L. (2012). *Macmillan Setswana and English Illustrated Dictionary*. Johannesburg: Macmillan South Africa.

Coupez, A.; Kamanzi, T.; Bizimana, S.; Sematama, G.; Rwabukumba, G.; Ntazinda, C. and collaborators. (2005). *Dictionnaire rwanda-rwanda et rwanda-français / Inkoranya y ikinyarwaanda mu kinyarwaanda nó mu gifaraansá.* Butare, Rwanda: Institut de Recherche Scientifique et Technologique and Tervuren, Belgium: Musée Royal de l'Afrique Centrale.

Dong, X., Naumann, F. (2009). Data fusion: Resolving data conflicts for integration. In *Proceedings of the VLDB Endowment* 2(2), pp. 1654--1655.

Hamari, J.; Koivisto, J. and Sarsa, H. (2014). Does gamification work? – A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*.

Haspelmath, M. (2014). Dictionaries as open-access databases: A vision. http://dlc.hypotheses.org/676.

Haviland, J. (2006). Documenting lexical knowledge. In J. Gippert, N. Himmelmann, & U. Mosel (Eds.), *Essentials of Language Documentation.* Berlin: de Gruyter, pp. 129--162.

Hernández, M., Stolfo, S. (1998). Real-world data is dirty: Data cleansing and the merge/purge problem. *Data Mining and Knowledge Discovery* 2(1), pp. 9--37.

Hill, K.; Sekauquaptewa, E.; Black, M. and Malotki, E. (1998). *Hopi Dictionary/Hopìikwa Lavàytutuveni: A Hopi Dictionary of the Third Mesa Dialect with an English-Hopi Finder List and a Sketch of Hopi Grammar*. Tucson: University of Arizona Press.

Hirth, M.; Hoßfeld, T. and Tran-Gia, P. (2013). Analyzing costs and accuracy of validation mechanisms forcrowdsourcing platforms. *Mathematical and Computer Modelling* 57(11), pp. 2918--2932.

Kittur, A.; Chi, E. and Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *CHI '08 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 453--456.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5), pp. 604--632.

Lee, M.; Lu, H.; Ling, T. and Ko, Y. (1999). Cleansing data for mining and warehousing. In T. Bench-Capon, G. Soda, & A.M. Tjoa (Eds.), *Database and Expert Systems Applications: 10th International Conference, DEXA'99 Florence, Italy, August 30 – September 3, 1999 Proceedings.* Berlin: Springer-Verlag, pp. 751--760.

Matisoff, J. (1988). *The Dictionary of Lahu*. Berkeley: University of California Press.

Nguyen, Q.; Nguyen, T.; Miklós, Z. and Aberer, K. (2013). On leveraging crowdsourcing techniques for schema matching networks. In W. Meng, L. Feng, S. Bressan, W. Winiwarter, & W. Song (Eds.), *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, Wuhan, China, April 22-25, 2013, Proceedings, Part II.* Berlin: Springer-Verlag, pp. 139--154.

Paraschakis, D. (2013). *Crowdsourcing Cultural Heritage Metadata Through Social Media Gaming.* Master's Thesis in Computer Science. Malmo University.

Sacleux, C. (1939). *Dictionnaire swahili-français.* Paris: Institut d'Ethnologie.

Suri, S.; Goldstein, D. and Mason, W. (2011). Honesty in an online labor market. In *Human Computation: Papers from the 2011 AAAI Workshop*, pp. 61--66.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge: Cambridge University Press.

Young, R., Morgan, W. (1987). *The Navajo Language: A Grammar and Colloquial Dictionary*, revised ed. Albuquerque: University of New Mexico Press.

Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework

Thierry Declerck[1], Eveline Wand-Vogt[2], Karlheinz Mörth[2], Claudia Resch[2]
[1]DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg, 3 – D-66123 Saarbrücken
[2]Institute for Corpus Linguistics and Text Technology, Austrian Academy of Sciences
Sonnenfelsgasse, 19 – A-1010 Wien
declerck@dfki.de, {Eveline.Wandl-Vogt| Karlheinz.Moerth|Claudia.Resch}@oeaw.ac.at

Abstract

We describe actual work on porting dialectal dictionaries and historical lexical resources developed at the Austrian Academy of Sciences onto representation languages that are supporting their publication in the Linked (Open) Data framework. We are aiming at a unified representation model that is flexible enough for describing those distinct types of lexical information. The goal is not only to be able to cross-link those resources, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way the dialectal and historical lexical resources to the same "digital dignity" as the mainstream languages have already gained.

Keywords: Dialectal dictionaries, historical corpora and lexicons, Linked Open Data

1. Introduction

We describe actual work based on former experiments made with porting a dialectal dictionary[1] of the Austrian Academy of Sciences [2] onto representation formats supporting their publication in the Linked Open Data (LOD) framework[3] (Wandl-Vogt & Declerck, 2013). The extension of this former work concerns two in TEI[4] encoded dictionaries of Arabic dialects (Mörth et al., 2013) and historical lexical data extracted from a corpus of sacred texts written in Early New High German (Mörth et al., 2012). Dealing with those different types of data calls for a unified approach for their encoding in LOD compliant representation formats.

The ultimate goal of our work is not only to be able to cross-link all the lexical resources described in this paper, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way our dialectal and historical lexical resources to the same "digital dignity" as the mainstream languages have already gained.

We briefly describe in this paper the different types of lexical resources we are dealing with, their commonalities, and how we use those commonalities as the basis for the unified encoding in RDF, SKOS-XL[5] and *lemon*[6]. We show how this encoding allows to enrich our lexical data with additional information, mainly senses, available in the LOD.

2. The different types of lexical Data

In this section we present briefly the three types of lexical resources we are dealing with in our experiments.

2.1 The Austrian Dialects Dictionaries

The starting point for our work was given by two Austrian dialectal dictionaries: The Dictionary of Bavarian dialects of Austria (*Wörterbuch der bairischen Mundarten in Österreich*, WBÖ)[7] and the Dictionary of the Viennese dialect (*Wörterbuch der Wiener Mundart*, WWM)[8]. Both dictionaries have been made available to us in an electronic version. Figure 1 below partially shows an example of an entry in the printed version of WBÖ, while Figure 2 is giving a related example taken from the electronic version of the WWM.

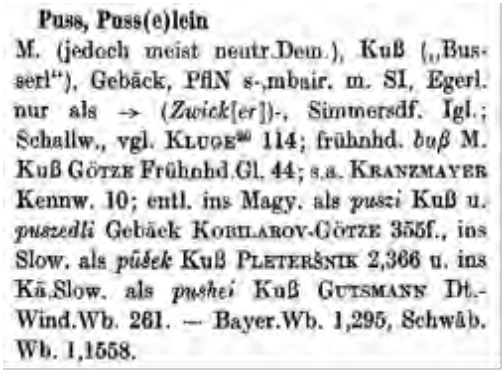
Puss, Puss(e)lein
M. (jedoch meist neutr.Dem.), Kuß („Busserl"), Gebäck, PflN s.-mbair. m. SI, Egerl. nur als → (*Zwick[er]*)-, Simmersdf. Igl.; Schallw., vgl. KLUGE[20] 114; frühnhd. *buß* M. Kuß GÖTZE Frühnhd.Gl. 44; s.a. KRANZMAYER Kennw. 10; entl. ins Magy. als *puszi* Kuß u. *puszedli* Gebäck KOMLAROV-GÖTZE 355f., ins Slow. als *pûšek* Kuß PLETERŠNIK 2,366 u. ins Kä.Slow. als *pushei* Kuß GUTSMANN Dt.-Wind.Wb. 261. — Bayer.Wb. 1,295, Schwäb. Wb. 1,1558.

Figure 1: An example entry of the printed edition of the WBÖ.

[1]We are talking about the "Dictionary of Bavarian dialects of Austria" (http://www.oeaw.ac.at/dinamlex/WBOE.html). We are adapting our work also to external dialectal dictionary resources, like the Dictionary of the Viennese dialect; see (Hornung & Grüner, 2002).

[2]More specifically, the work is carried out at the "Institute for Corpus Linguistics and Text Technology", see http://www.oeaw.ac.at/icltt/

[3] See http://linkeddata.org/

[4] See http://www.tei-c.org/index.xml and (Romary, 2009)

[5] See http://www.w3.org/TR/skos-reference/skos-xl.html

[6] See (McCrae & al., 2012).

[7] http://verlag.oeaw.ac.at/Woerterbuch-der-bairischen-Mundarten-in-Oesterreich-38.-Lieferung-WBOe

[8] See (Hornung & Grüner, 2002).

Bussal, Bussi, Bussl, das, 1) Kuss (Syn.: *Schm$tss*); 2) kleines Süßgebäck; Pl. *Bussaln;* viele Komp. wie *Nussbussal* usw. –

Figure 2: Example entry from the WMM, corresponding to the entry in Figure 1.

In both examples, the reader can observe that the meanings of each entry are given by using words in the standard languages corresponding to the dialects: either High German ("Kuss", *kiss*) or High Austrian ("Busserl"). But the meanings of the entries are not explicitly given by a definition. Linking to the linguistic resources in the LOD is partially motivated by this issue: providing by semi-automatic means to those entries a definition (or more than one definition in case of ambiguities) by pointing to senses encoded in the LOD. As senses in the LOD are many time associated to multilingual entries, we can also take benefit of this and propose a multilingual extension to the words expressing the meaning(s) attached to the entries of the dialect dictionaries we are dealing with.

2.2 TEI encoded dictionaries of Arabic Dialects

Our more recent work on porting under-resourced lexical resources available at ICLTT was applied to two dictionaries of Arabic dialects, encoded in TEI and called "ar-apc-x-damascus" and "ar-arz-x-cairo".

```
<entry xml:id="baab_001">
  <form type="lemma">
    <orth
xml:lang="ar-apc-x-damascus-vicav">bāb</orth>
  </form>
  <gramGrp>
    <gram type="pos">noun</gram>
    <gram                                    type="root"
xml:lang="ar-apc-x-damascus-vicav">bwb</gram>
  </gramGrp>
  <form type="inflected" ana="#n_pl">
    <orth
xml:lang="ar-apc-x-damascus-vicav">bwāb</orth>
  </form>
  <sense>
    <cit type="translation" xml:lang="en">
      <quote>door</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>gate</quote>
    </cit>
    <cit type="translation" xml:lang="en">
      <quote>city gate</quote>
```

Figure 3: An example taken from the TEI encoded "Damascus" lexicon

Figure 3 gives an example of the TEI encoding of the so-called "Damascus" dictionary. The reader can observe that the TEI encoding is explicitly marking up what is implicit in the Austrian dialect dictionaries: the senses of the entries are given using words in other languages.

The building and update of those dictionaries are done in the context of the VICAV project[9], and the approach implemented for gathering relevant lexical data from the Web and correcting/adjusting these data with the help of NLP resources is described in (Mörth et al., 2013).

2.3 The ABaC:us lexicon

The Austrian Baroque Corpus (ABaC:us) at ICLTT is a digital collection of printed German language texts dating from the Baroque era, in particular the years from 1650 to 1750. The ABaC:us collection holds several historical texts specific to religious instruction and works concerning death and dying, including sermons, devotional books and works related to the dance-of-death theme. All Baroque prints that served as the input for the resource have been fully digitized, transcribed, and transformed into an XML format according to the guidelines of the Text Encoding Initiative (version P5).[10] ABaC:us currently contains more than 210.000 running words. The tokens have been mapped automatically to a word class with the tool *TreeTagger*[11], using the *Stuttgart-Tübingen-TagSet* and its guidelines (1999)[12], and have been enriched with a modern High German lemma or canonical form (according to "Duden"[13] or "Deutsches Wörterbuch"[14] as a reference). The results of those processes have been manually corrected and validated.

In order to support our work in the field of Linked Data, we had first to re-organize the structure of the stored lexical data: the result of the work described just above was stored along the lines of the findings in the corpus: one entry in the data base per occurrence of a token in the corpus. While this is essential for keeping track of the context of the word forms[15], we want to reduce the representation of the tokens to their types, as displayed in Figure 4, where the modern High German nominal lemma "Fegefeuer" (*purgatory*), used here as an example, is a unique entry, pointing to the list of form variants that have been detected and marked up in the corpus. So that all variants of the modern High German lemma form "Fegefeuer" (*purgatory*) are associated – and thus identified – with this lemma. Our aim is then to link the correlated unique lemma form to available semantic information sources in the LOD.

In the example in Figure 4 we also include the frequency information for each word forms in the corpus. We observe here a similar property as in the other examples of lexical/dictionary data we used so far: The meaning of an

[9] VICAV stands for "Vienna Corpus of Arabic Varieties". See http://www.oeaw.ac.at/icltt/node/59

[10] See http://www.tei-c.org/Guidelines/P5/ for more details.

[11] See Schmid, 1995.

[12] http://www.ims.uni-stuttgart.de/forschung/ressourcen/lexika/TagSets/stts-1999.pdf

[13] See https://www.duden.de/

[14] See http://dwb.uni-trier.de/de/

[15] Alternatively, but compatible, to this form of storing the data would be a stand-off annotation schema, in which each word form (token) in the corpus is carrying an index.

historical word is given by its corresponding lemma in High German.

```
"Fegefeuer" => {
    "NN" => {
        "Feeg=Feuer"  => "6"
        "Feegfeuer"  => "100"
        "Feegfeuers"  => "4"
        "Fegfeuer"  => "80"
        "Fegfeuers"  => "24"
        "Fegfeur"  => "2"
        "Fegfewer"  => "4"
    }
}
```

Figure 4: The actual form of the High German lemma entry "Fegefeuer" (*purgatory*) with pointers to the historical variants as found in the ABaC:us corpus.

3. A unified Approach to the SKOS-XL encoding of the different lexical and dictionary Data

A motivation for the extension of the work described in (Wandl-Vogt & Declerck, 2013) was to investigate if the SKOS-based model described there can support the (automatized) cross-linking of the Bavarian dialectal dictionary data (ÖBW) with other dialectal dictionaries and historical lexicons. In this particular case, we take advantage of a property of dialectal dictionaries concerning the expression of meanings of entries: Although conceived as monolingual reference works, dialectal dictionaries share with bilingual dictionaries the fact that they express the meanings of their entries in a different language. The meta-language for expressing the meanings of entries in both WBÖ and WWM is standard German, sometimes accompanied by Austrian German, as we already noticed in section 2.1. We observed the same property for the two dialectical lexicon of Arabic and for the historical lexicon: all relate their specific entries to High German words.

3.1 Representation Formalisms used

Based on the Resource Description Framework (RDF)[16], SKOS (Simple Knowledge Organization System) [17] seemed to offer an appropriate modeling language. Our experiment with SKOS is kind of novel, since we apply it to dictionaries, although one can for sure consider dictionaries as being very close to thesauri. In our approach we first encoded elements of entries of the dictionaries as concepts being part of a skos:ConceptScheme. But more recently we decided to encode the strings introducing the entries of a dictionary as being a member of a skos:Collection, while the associated senses are encoded as skos:Concept, being members of a specific skos:Concept:Scheme.

With the use of SKOS (and RDF), we are also in the position to make our dictionary and lexical resources compatible with other language resources available in the LOD cloud. Examples of such resources are the DBpedia instantiation of Wiktionary[18] or the very recent release of BabelNet[19]. Since, contrary to most knowledge objects described in the LOD, we do not consider strings (encoding lemma and word forms as part of a language) as being just literals, but as knowledge objects, we considered the use of SKOS-XL and of the *lemon* model[20] for encoding the lemmas and the associated full forms listed in the lexical resources.

3.2 Porting the lexical Data onto the Linked Open Data Cloud

While the lexical data we are dealing with in all reported resources is mainly about establishing correspondences between dialectal or historical variants of words and their related High German forms, we are also aiming at providing for a semantic description for the entries. For achieving this, we started to semi-automatically link our corpus and lexical data to both domain knowledge (e.g. religion in the case of the ABaC:us data) and to (lexical) senses available in the LOD.

In order to automatically cross-link entries from our dictionaries and lexicons, we wrote first a program for extracting the strings expressing the meanings for each entry and applied an algorithm for comparing the extracted strings. For this latter task, it is necessary to first linguistically analyze the strings expressing the meanings, since pure string matching cannot provide accurate comparisons: lemma reduction and PoS tagging are giving additional indicators for matching strings expressing meanings. To mark the linguistically analyzed string expressing meanings we also use *lemon*.

3.3 RDF/SKOS/*lemon* Representation

We present now in certain details our actual RDF, SKOS-XL and *lemon* based model. Since our more recent extension work deals with the ABaC:us lexical data, we present the LOD compliant model we developed for this lexicon, which is in fact the same for all other considered lexicons. In the case of the lexical data extracted from the ABaC:us collection, we do not deal with a classical dictionary[21] as our source, but rather with a selection of word forms used in a corpus and associated with modern High German lemmas. We introduce for this a special

[16] http://www.w3.org/RDF/

[17] http://www.w3.org/2004/02/skos/

[18] See http://dbpedia.org/Wiktionary. There, *lemon* is also used for the description of certain lexical properties.

[19] http://babelnet.org/

[20] See http://www.monnet-project.eu/lemon

[21] With this, we mean that a dictionary typically lists entries of a specific language and relates those to a definition and meanings (*senses*). But the ABaC:us lexicon is closer in form to a dialectal dictionary which introduce meanings by the use of the words used in the corresponding standard language. An example of a mapping from a dialectal dictionary into SKOS is described in (Wandl-Vogt & Declerck, 2013).

owl:Class[22]:

```
icltt:Corpus_Lexicon
        rdf:type owl:Class ;
        rdfs:comment "Lexicon extracted from a
corpus"@en ;
        rdfs:label "Corpus Lexicon"@en ;
        rdfs:subClassOf owl:Thing .
```

An instance of this owl:Class is the ABaC:us data set, displayed just below:

```
icltt:abacus
        rdf:type
                skos:Collection ,
                icltt:Corpus_Lexicon ;
        rdfs:label "ICLTT lexicon for Baroque
        l          anguage"@en ;
        skos:member icltt:concept_fegefeuer .
```

We consider such data sets as a skos:Collection rather than a skos:ConceptScheme, since we are listing entries and not describing hierarchical or associative relations between those. We use the "skos_member" object property to mark the entries belonging to this collection, as can be seen in the example just above (for reasons of place, we include here only the entry "fegefeuer" as a member of the collection).
Entries are introduced at the schema level by the owl:Class "Entry", which is a subclass of skos:Concept:

```
icltt:Entry
        rdf:type owl:Class ;
        rdfs:label "Entry"^^xsd:string ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:LexicalEntry .
```

In our model, the variants of the High German lemma form are encoded as single lexical forms, bearing just xsd:string information. The owl:Class for this is:

```
icltt:Form
        rdf:type owl:Class ;
        rdfs:label "Form"^^xsd:string ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:Form .
```

And instances of this class look like the example displayed just below, introducing the language tag "fnhd" for "Frühneuhochdeutsch" (*early new High German*):

```
icltt:Feeg_Feuer
        rdf:type lemon:Form , icltt:Form ;
        rdfs:label "Feeg=Feuer"@fnhd ;
        lemon:formVariant icltt:Feegfeuer .
```

The corresponding instance for the High German entry:

```
icltt:concept_fegefeuer
        rdf:type
                lemon:LexicalEntry ,
                icltt:Entry ;
        rdfs:label "Fegefeuer"@de ;
        lemon:lexicalForm
                icltt:Feegfeuer ,
                icltt:Feeg_Feuer ;
        skosxl:prefLabel icltt:entry_Fegefeuer .
```

This instance is pointing, in the last line of the code, to a skos object via the property skosxl:prefLabel. We use this property to link the basic entry (as a string belonging to the corpus-lexicon) to a complex linguistic object, which is displayed just below:

```
icltt:entry_Fegefeuer
        rdf:type icltt:Lemma ;
        rdfs:label "Fegefeuer"@de ;
        icltt:hasPos icltt:noun ;
        lemon:sense icltt:fegefeuer ;
        skosxl:literalForm "Fegefeuer"@de .
```

In this case the corpus_lexicon entry gets associated with PoS information, but more importantly, we add a link to a "sense". As mentioned earlier in this submission, no "meaning" is given to us from the corpus, therefore we are querying for senses in available semantic resources in the Web, more specifically in the Linked Open Data environment.
The strategy is here to send sparql queries to DBpedia and to see how much of our modern High German entries are present in this semantic resource. For this we use the publicly available "virtuoso sparql query editor", in its specialization for the German data sets.[23] Our example in this submission, "Fegefeuer", is indeed included as a concept in DBpedia[24], and from there we get a lot of interesting additional information: So for example all the "redirects of", which in this case are:

- dbpedia-de:Purgatorium,
- dbpedia-de:Fegfeuer
- dbpedia-de:Reinigungsor

3.4 Expressing the Meaning of an Entry by linking to senses in DBpedia

Our aim is to associate senses to the entries of our lexical resources. For this we started to link the entries to senses explicitly encoded in DBpedia. As a preliminary step, we need to introduce in our model for the lexicon an owl:Class "Sense", the instances of which the property "lemon:sense" can point to :

[22] In the other case, we have the owl:Class "Dictioanry". The examples from our ontology model are given in the turtle syntax (see http://www.w3.org/TR/turtle/ for more details)

[23] http://de.dbpedia.org/sparql

[24] See http://de.dbpedia.org/page/Fegefeuer

```
icltt:Sense
        rdf:type owl:Class ;
        rdfs:label "Sense"@en ;
        rdfs:subClassOf skos:Concept ;
        owl:equivalentClass lemon:LexicalSense .
```

Different to the case of entries for the lemmas, we encode the senses as part of a skos:ConceptScheme, since in the case of senses more relations between the items are possible (and wished):

```
icltt:Senses_ICLTT
        rdf:type skos:ConceptScheme ;
        rdfs:comment "Senses that are used in ICLTT
dictionnaries"@en ;
        rdfs:label "Senses"@en .
```

The instance for the sense to be associated with "Fegefeuer";

```
icltt:fegefeuer
        rdf:type lemon:LexicalSense , icltt:Sense ;
        rdfs:label "Purgatory"@en , "Fegefeuer"@de ;
        skos:exactMatch
<http://wiktionary.dbpedia.org/page/Fegefeuer-German-
Noun-1de> ;
        skos:inScheme icltt:Senses_ICLTT .
```

We make use in this case of the skos:exactMatch property to link to a sense of the DBpedia version of Witkionary. One of the advantages of this approach lies in the fact that we can re-use existing semantic resources, without having to invent our own catalogue of senses. Second we get a list of multilingual equivalents, as those are listed in the LOD version of Wiktionary. In the case of "Fegefeuer" we get the equivalents for English, French, Italian, Latin, Swedish, and Catalan exactly for this one sense! And in fact, going to the corresponding page for the English term[25]: we get much more equivalents: ca 50 equivalent terms in ca 40 languages.

This sense-based access to lexical resources available in the LOD is thus supporting the creation of a multilingual net of terms relevant to a domain. In our case, we manage to link old form variants of religious terms (and other relevant terms used in the ABaC:us corpus). For the particular example we have been discussing, we can thus get not only many multilingual equivalents for the word "Fegefeuer" (and its historical German variants), but also for related words that are classified under the DBpedia categories "Eschatology" etc. As mentioned earlier, this approach is valid for all other lexical data we are dealing with: we link those to both an encyclopedic resource and a lexical one in the LOD, so that we can retrieve not only the senses of the entries for our lexical data, but also multilingual equivalents.

[25] http://wiktionary.dbpedia.org/page/purgatory-English-Noun-1en

4. Conclusion

We have described the actual status of our work dealing with the modeling of dialectal and historical lexical data using semantic web standards, and how this supports the linking of entries of the lexicons to lexical senses and multilingual equivalents available in the Linked Open Data framework. We will also publish some of our data in the LOD so that it can be linked to from other resources in the web of data.

Future work will consist in porting automatically all lexical entries to the unified SKOS-XL and *lemon* model, whereas the links to senses in Wiktioanry and DBpedia in a first phase will be established on a manual basis, if there are ambiguities in the LOD resources. We would like to thank our

5. Acknowledgements

We would like to thank our colleagues Eva Wohlfahrter for her assistance in annotating the ABaC:us texts and Barbara Krautgartner for creating the gold standard for our textual analyses in the context of the ABaC:us corpus. The ABaC:us corpus and lexicon have been developed in the context of the project "Text-Technological Methods for the Analysis of Austrian Baroque Literature", supported by funds of the Österreichische Nationalbank, Anniversary Fund, project number 14738.

6. References

Chiarcos, C., Cimiano, P., Declerck, T., McCrae, J.P. (2013). Linguistic Linked Open Data (LLOD) - Introduction and Overview**.** In: Christian Chiarcos, Philipp Cimiano, Thierry Declerck, John P. McCrae (eds.): *2nd Workshop on Linked Data in Linguistics, Pages i-xi,* Pisa, Italy.

Declerck, T., Lendvai, P., Mörth.K. (2013) Collaborative Tools: From Wiktionary to LMF, for Synchronic and Diachronic Language Data. In Francopoulo, G. (ed) *LMF Lexical Markup Framework.* Wiley 2013.

Francopoulo, G. (2013) LMF -- Lexical Markup Framework. Wiley.

Hornung, M., Grüner, S. (2002) Wörterbuch der Wiener Mundart; Neubearbeitung. öbvhpt, Wien

McCrae, J., Aguado-de-Cea, G., Buitelaar P., Cimiano P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2012) Interchanging lexical resources on the Semantic Web. In: *Language Resources and Evaluation.* Vol. 46, Issue 4, Springer:701-719.

Mörth, K., Resch, C., Declerck, D. and Czeitschner, U. (2012). *Linguistic and Semantic Annotation in Religious Memento Mori Literature*. In: Proceedings of the LREC'2012 Workshop: Language Resources and Evaluation for Religious Texts (LRE-Rel-12). ELRA: pp. 49-52.

Mörth, K., Procházka, S., Siam,O., Declerck T. (2013) Spiralling towards perfection: an incremental approach for mutual lexicon-tagger improvement. In: *Proceedings of eLex 2013,* Tallinn, Estonia.

Moulin, C. (2010) Dialect dictionaries - traditional and modern. In: Auer, P., Schmidt, J.E. (2010) (eds) Language and Space. An International Handbook of Linguistic Variation. Volume 1: Theories and Methods. Berlin / New York. pp: 592-612. (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science / Manuels de linguistique et des sciences de communication 30.1).

Miles, A., Matthews, B., Wilson, M. D., Brickley, D. (2005) SKOS Core: Simple Knowledge Organisation for the Web. In *Proc. International Conference on Dublin Core and Metadata Applications*, Madrid, Spain,

Romary, L. (2009). Questions & Answers for TEI Newcomers. *Jahrbuch für Computerphilologie 10*. Mentis Verlag,

Schreibman, S. (2009) The Text Encoding Initiative: An Interchange Format Once Again. *Jahrbuch für Computerphilologie 10*. Mentis Verlag.

Wandl-Vogt, E. (2005) From paper slips to the electronic archive. Cross-linking potential in 90 years of lexicographic work at the Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In: *Complex 2005. Papers in computational lexicography*. Budapest: 243-254.

Wandl-Vogt, E. and Declerck, T. (2013). Mapping a Traditional Dialectal Dictionary with Linked Open Data. In *Proceedings of eLex 2013,* Tallinn, Estonia.

Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian

Delphine Bernhard
LiLPa - Linguistique, Langues, Parole
EA 1339, Université de Strasbourg
dbernhard@unistra.fr

Abstract

This article presents a method to align bilingual lexicons in a resource-poor dialect, namely Alsatian. One issue with Alsatian is that there is no standard and widely-acknowledged spelling convention and a lexeme may therefore have several different written variants. Our proposed method makes use of the double metaphone algorithm adapted to Alsatian in order to bridge the gap between different spellings. Once variant citation forms of the same lexeme have been aligned, they are mapped to BabelNet, a multilingual semantic network (Navigli and Ponzetto, 2012). The mapping relies on the French translations and on cognates for Alsatian words in the English and German languages.

Keywords: lexicon alignment, spelling variants, Alsatian

1. Introduction

Linked Open Data Resources have recently emerged as a new way to represent linguistic knowledge in many languages, by linking resources represented using standard formats. In practice, many of these resources are based either on existing word nets or on collaboratively built encyclopaedias or dictionaries such as Wikipedia or Wiktionary. As a consequence, not all languages are covered and even automatic approaches which acquire knowledge from e.g. Wikipedia or Wiktionary are not always usable because of the lack of information available for under-resourced languages.

In this article, we focus on a dialect, namely Alsatian, and propose to make use of resources which are more easily exploited and readily available, i.e. bilingual lexicons, to provide additional lexicalisations to existing linguistic linked open resources.

The Alsatian dialects are spoken in the Alsace region, located in the North-East of France. They belong to the Franconian and Alemannic language families (Huck et al., 2007). According to a recent study, 43% of the Alsatian population still speak the regional dialect (OLCA / EDinstitut, 2012). However, the proportion of Alsatian speakers is decreasing regularly since the 1960s, to the benefit of the French language. Moreover, the Alsatian dialects are mostly oral and there is no standard written norm.

There have been some initiatives aimed at defining spelling conventions. The ORTHAL system (Zeidler and Crévenat-Werner, 2008) refers to standard German spelling while allowing the transcription of phenomena which are specific to the Alsatian dialects. The GRAPHAL-GERIPA system (Hudlett and Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe, 2003) defines a set of rules to go from sound to grapheme. However, it is difficult to estimate the actual dissemination and use of these systems. Moreover, they accommodate for the various geolinguistic variants encountered in Alsace and thus do not guarantee a unique spelling for the citation form of a given lexeme.[1]

To sum up, Alsatian dialects pose several important challenges for NLP:

- There is no standard and widely acknowledged spelling convention ;
- The Alsatian dialect is actually a continuum of dialects, with geographic lexical and pronunciation variants ;
- There are no large amounts of digital text corpora available.

In this article, we present a first step towards building digital lexical resources for the Alsatian dialects which consists in (i) aligning several bilingual French-Alsatian lexicons and (ii) mapping the Alsatian words to BabelNet, a multilingual semantic network which is connected to the Linguistic Linked Open Data cloud (Navigli and Ponzetto, 2012).

The proposed method relies on the following observations:

- The spelling conventions adopted in the French-Alsatian lexicons are very variable, and thus an Alsatian lexeme may have a different citation form in each lexicon, and even several different citation forms in a given lexicon, to accommodate for geolinguistic variants. Also, many of the Alsatian words are similar to their translation into standard German and even sometimes English.
- Different lexicon authors may choose different translations into French for a given Alsatian lexeme. This complicates the alignment, which cannot only rely on a simple mapping using French lemmas.

[1] We use *lexeme* in the sense given by Bauer (2003): "A lexeme is a dictionary word, an abstract unit of vocabulary. It is realised (...) by word-forms, in such a way that the word-form represents the lexeme and any inflectional endings (...) that are required. (...) The citation form of a lexeme is that word-form belonging to the lexeme which is conventionally chosen to name the lexeme in dictionaries and the like."

We address these issues as follows:

- We propose a variant of the double metaphone algorithm adapted to the Alsatian dialects, in order to identify spelling variants. The algorithm also tackles standard German and English spelling in order to find cognates;
- We use external resources to obtain information about synonyms in the French language and translations into German and English.

The article is organised as follows: in the following section we review previous work on the identification of spelling variants and the alignment of lexical resources. Section 3 details the lexical resources used in our work. We present our alignment and mapping method in Section 4 and the evaluation results in Section 5.

2. State of the Art

2.1. Identification of Spelling Variants

Non-standard writing is an issue when dealing with different kinds of texts, e.g. data from the Web, in particular Web 2.0, historical texts and languages which are mainly oral and thus non-written.

A first family of methods target normalisation, i.e. transforming a minority variant to a given standard. Scherrer (2008) uses orthographic Levenshtein distance and trained stochastic transducers in order to build a bilingual lexicon for a Swiss German dialect and standard German. Hulden et al. (2011) present two methods which automatically learn transformations from a dialectal form to the standard form using a limited parallel corpus for the Basque language and the Lapurdian Basque dialect. The first method relies on an existing tool, `lexdiff` (Almeida et al., 2010), which detects spelling differences. The spelling differences identified are used to obtain replacement rules which are compiled as transducers. The second method is inspired by ILP (Inductive Logic Programming) and tries to select the best set of replacement rules, using both positive and negative examples. Salloum and Habash (2011) describe a rule-based method to generate paraphrases of dialectal Arabic in standard Arabic. The paraphrases are used for Arabic-English statistical machine translation. For historical language variants, Porta et al. (2013) propose a method to map historical word forms to their modern counterparts. The approach is based on a Levenshtein transducer and a linguistic transducer implementing sound change rewrite rules.

In a different vein, Dasigi and Diab (2011) present a clustering algorithm which aims at grouping orthographic dialectal variants. They experiment with several word similarity measures and conclude that string similarity metrics perform better for this task than contextual similarity metrics.

Our work is closest to Dasigi and Diab (2011), in that we cluster dialectal variants and do not resort to normalisation. We preferred this approach as normalisation is not applicable in our case. Fist, there is no consensus on the writing norm for Alsatian dialects and it is thus difficult to decide which form should prevail. Moreover, even though Alsatian is closely related to German, there are a number of lexical and syntactic differences which have to be taken into account. Added to that, considering German as the standard for Alsatian is a very sensitive sociolinguistic issue, which has implications reaching deeper than purely linguistic considerations. Given all these reasons, our proposed method does not attempt to normalise writing variants but preserves their diversity by considering clusters of variants as lexicon entries.

2.2. Alignment of Lexical Resources

The main objective of our work is not only to identify spelling variants of the same Alsatian lexeme, but also to align entries stemming from different bilingual lexicons and map the alignments to a semantic network.

A lot of work has been devoted recently to the alignment of collaborative resources, such as Wikipedia, and classical lexical knowledge bases, such as WordNet.

Niemann and Gurevych (2011) detail a method for aligning senses in WordNet and Wikipedia, which was later employed for creating the UBY lexical-semantic resource (Gurevych et al., 2012). The method relies on a machine learning method which classifies alignments as valid or non-valid. The similarity of aligned sense candidates is computed based on a bag-of-word representation of the senses and then provided to the classifier. For the UBY resource, cross-lingual word sense alignments are induced in the same manner, by first automatically translating the textual representations of the senses.

Navigli and Ponzetto (2012) propose a method to relate Wikipedia pages to WordNet senses used for building the BabelNet resource. The method applies several different strategies sequentially. In particular, it re-uses a technique used for Word Sense Disambiguation which consists in defining a disambiguation context for each Wikipedia page and WordNet sense. The disambiguation context is a set of words obtained from information provided in the resources (e.g. labels, links, redirections and categories in Wikipedia ; synonyms, hypernyms / hyponyms, glosses in WordNet). A similarity score can then be computed based on this context.

When there is no lexical resource in one language, automatic translation of resources in another language is often the best option, in terms of construction costs. In this case, an existing resource is extended with lexicalisations in another language.

The WOLF (Wordnet Libre du Français) has been built by Sagot and Fišer (2008) using the Princeton WordNet and several multilingual resources. The main assumptions underlying their approach are that different senses of an ambiguous word in one language often correspond to different translations in another language and words which are translated by the same word in another language often have similar meanings. They enforce these ideas by collecting a multilingual lexicon with 5 languages from a parallel corpus and by assigning the most likely synset to each lexicon entry, relying on the intersections between the synsets associated to each non-French word in the lexicon in the Princeton WordNet or in wordnets from the BalkaNet project. Hanoka and Sagot (2012) have extended the WOLF resource using a new approach relying on a large

synonymy and translation graph built from Wikipedia and Wiktionary. The graph is queried with literals from synset-aligned multilingual wordnets to get the best translation candidate, based both on translation and back-translation relations.

In our work, we also apply the idea of extending an existing lexical-semantic resource with lexicalisations from another language, namely Alsatian. We use French as a pivot language to obtain a mapping between Alsatian variants and BabelNet. We also exploit the cognacy between Alsatian, German and English in order to enrich the feature vectors.

3. Resources

In this section, we detail the resources used in our work.

3.1. Bilingual French-Alsatian Lexicons

We have retrieved three bilingual French-Alsatian lexicons available on the Web:

- OLCA: the lexicons produced by the OLCA (*Office pour la Langue et la Culture d'Alsace*)[2]. These lexicons are domain-specific (beer, shopping, football, medicine, weather, nature, fishing, pharmacy, vine) and provide variants for the Bas-Rhin (Lower Rhine) and Haut-Rhin (Upper Rhine) Alsatian departments. In the rest of the article, these two variants are identified as OLCA-67 (for Bas-Rhin) and OLCA-68 (for Haut-Rhin);
- WKT: a lexicon retrieved from a Wiktionary user page;[3]
- ACPA: a bilingual lexicon authored by André Nisslé.[4]

These lexicons, though machine-readable, are not available in a standard format. They have been preprocessed with specific parsers to extract French-Alsatian word pairs. When available, information about part-of-speech is kept.[5] Otherwise, we used two heuristics for guessing the part-of-speech : (i) apply the French TreeTagger (Schmid, 1994) to obtain a category for French single words[6] ; (b) for nouns, check the presence of a determiner next to the Alsatian form.

Table 1 lists the number of French entries in the lexicons after preprocessing. The table shows that the coverage of the different parts-of-speech is uneven, and that the lexicons mostly focus on nouns, verbs and adjectives.

The lexicons follow different graphical conventions as exemplified by Table 2, which lists the translations found in the lexicons for several lexemes. Many translations in Table 2 are actually graphical variants of the same Alsatian lexeme (e.g. "Kràb " and "Kràpp"). However, these graphical variants can be very dissimilar if we only consider the characters used.

	OLCA-67	OLCA-68	WKT	ACPA
adjective	194	195	122	1,898
adverb	16	16	49	295
determiner	0	0	20	15
noun	2,628	2,617	1,049	15,770
past participle	45	46	59	476
pronoun	1	1	38	47
verb	276	276	292	3,017
unknown	671	676	393	2,015
TOTAL	3,831	3,827	2,022	23,533

Table 1: Number of French entries in the French-Alsatian lexicons.

French	**corbeau**	**jambe(s)**	**grenier**
English	crow	leg	attic
German	Rabe	Bein	Dachboden
ACPA	Kräje Kràbb	Bai Unterschankel	Behna **Behn** Ästrich Dàchbooda
WKT	Grâb Kràpp **Ràmm**	**Bein** **Baan**	**Behn** Behni Bhena Kàscht Späicher Spicher
OLCA	Kràb **Ràmm**	**Bein** Bei **Baan**	

Table 2: Example translations found in the lexicons. Identical variants found in at least two lexicons are in bold format.

In addition to the bilingual lexicons, we also used two semantic networks: JeuxDeMots and BabelNet.

3.2. JeuxDeMots

JeuxDeMots (Lafourcade, 2007) is a freely available French lexical network built through crowdsourcing games.[7] We used the version dated November 30, 2013,[8] which contains 171,029 occurrences of the synonymy relation (though the network also contains many other types of relations, e.g. association, domain, hypernymy, hyponymy, etc.).

3.3. BabelNet

BabelNet (Navigli and Ponzetto, 2012) is a multilingual semantic network, which integrates knowledge from Word-

[2] http://www.olcalsace.org/

[3] Available from the user page of Laurent Bouvier: http://fr.wiktionary.org/wiki/Utilisateur:Laurent_Bouvier/alsacien-fran%C3%A7ais

[4] http://culture.alsace.pagesperso-orange.fr/dictionnaire_alsacien.htm

[5] We used the following list of POS categories: verb, adjective, adverb, preposition, phrase, conjunction, pronoun, interjection, proper noun, past participle, determiner abbreviation, noun (feminine, masculine, neutral, plural).

[6] We use the TreeTaggerWrapper by Laurent Pointal available at http://perso.limsi.fr/pointal/dev:treetaggerwrapper.

[7] The games can be played on the following website: http://www.jeuxdemots.org

[8] Available from http://www.lirmm.fr/~lafourcade/JDM-LEXICALNET-FR

Net and Wikipedia. BabelNet is composed of Babel synsets, which are concepts with lexicalisations in several languages. The multilingual lexicalisations were obtained either thanks to Wikipedia's inter-language links or to Machine Translation. We used BabelNet version 2.0.[9]

4. Method

In this section, we present our method for aligning the lexicons. It relies on a variant of the double metaphone algorithm, adapted to Alsatian dialects.

4.1. Double Metaphone for Alsatian Dialects

Given the absence of a widely spread writing convention, as well as differences due to geolinguistic variants, it is not possible to align lexicon entries based on their written forms only using classical string similarity measures (consider for instance "Grâb" and "Kràbb" from Table 3). In order to cater for these differences, we have developed a double metaphone algorithm for Alsatian dialects. Double metaphone (Phillips, 2000) was originally proposed for information retrieval, in order to find names spelled differently than the search string, but referring to the same entity. Double metaphone belongs to the class of phonetic encoding algorithms, as it transforms the input string into a key which is identical for words which are pronounced in a similar manner. For instance, for the three given names "Stephan", "Steven" and "Stefan" the resulting key is `STFN`. In order to take ambiguities into account, double metaphone actually returns two keys in some cases. Double metaphone has for instance been used for Web 2.0 text normalisation (Mosquera et al., 2012).
The double metaphone transformations for Alsatian were written based on an analysis of our input lexicons.[10] We also took standard German into account, in order to obtain identical keys for German and Alsatian cognates. Table 3 gives some examples of the double metaphone keys obtained for several Alsatian and German words.

4.2. Lexicon Alignment

Our first objective is to be able to align entries across several bilingual Alsatian-French lexicons. In a first step, all entries in the input lexicons are added to a large graph. The nodes correspond to Alsatian words and their French translations. Alsatian words are connected to their French translations in the lexicons by an edge. Moreover, two Alsatian words are connected by an edge if all of the following conditions are met:

1. they have the same French translation;
2. they share one of their double metaphone keys ;
3. they have the same part-of-speech.[11]

We also use information obtained from the resources detailed in Section 3 in order to relax condition 1.

French Synonyms The JeuxDeMots synonyms list is used to connect two Alsatian words which have synonymous French translations in this resource.

BabelNet French Senses BabelNet French senses are used in the same way as the JeuxDeMots synonyms, to connect Alsatian words which have French translations belonging to the same sense.

4.2.1. Alignment of Alsatian Variants

Alsatian variants corresponding to the same lexeme are retrieved by detecting connected components in the subgraph containing only Alsatian words.
Figure 1 shows a portion of the initial graph. The translations into French, German and English are also shown. In the subgraph formed by the Alsatian words, there are three connected components: (1) ["Winkäller", "Winkeller", "Winkaller"], (2) ["Wikaller"] and (3) ["Kaller"]. The words "Winkäller", "Winkeller" and "Winkaller" are therefore aligned and considered as variants of the same lexeme.

4.3. Mapping to BabelNet Synsets

Our second objective is to map aligned Alsatian variants to BabelNet synsets. For instance, taking the example of Figure 1, the cluster formed by ["Winkäller", "Winkeller", "Winkaller"] should be mapped to the synset with ID `bn:00017041n` (see Figure 2).

Figure 2: Synset `bn:00017041n` in BabelNet's online search interface.

The mapping is achieved by calculating the cosine similarity between binary bag-of-words representations of Babel synsets and aligned Alsatian variants.
In the simplest case, the representation used for Babel synsets consists of their French lexicalisations. Alsatian variants are represented by their French translations: in the example of Figure 1, the cluster formed by ["Winkäller", "Winkeller", "Winkaller"] will be represented by the French words ["chai", "cellier", "cave"].
The bag-of-words representations can be extended by leveraging the translations available in BabelNet. The use of multilingual features has been shown to have a positive effect on the task of word sense disambiguation (Banea and Mihalcea, 2011). However, in looking for translations into English and German for Alsatian lexemes we have to avoid ambiguity. This issue has been addressed in work on the acquisition of bilingual dictionaries for a language pair using a third language as a pivot : in our case, French is the

[9] Available from `http://www.babelnet.org/download.jsp`

[10] Our implementation of Double Metaphone for Alsatian dialects is based on an existing Python module for English `http://www.atomodo.com/code/double-metaphone/metaphone.py/view`.

[11] Adjectives and past participles are considered as the same category.

Word	French translation	English translation	Metaphone key 1	Metaphone key 2
Schloofwàga	wagon-lit	sleeping car	XLFVK	XLFVY
Schlofwaawe			XLFVV	XLFVY
Rüejdàà	jour de repos	rest day	RT	/
Rüaijtààg			RTK	RT
beschtadiga	confirmer	confirm	PXTTK	PXTTY
Uffschtànd	insurrection	insurrection	AFXTNT	/
Iwereinsschtimmung	concordance	agreement	AFRNXTMNK	AVRNXTMNK
bestätigen	confirmer	confirm	PXTTK	/
Aufstand	insurrection	insurrection	AFXTNT	/
Übereinstimmung	concordance	agreement	APRNXTMNK	AVRNXTMNK

Table 3: Example metaphone keys. Alsatian words are in the upper part of the table, while German examples are detailed in the lower part of the table.

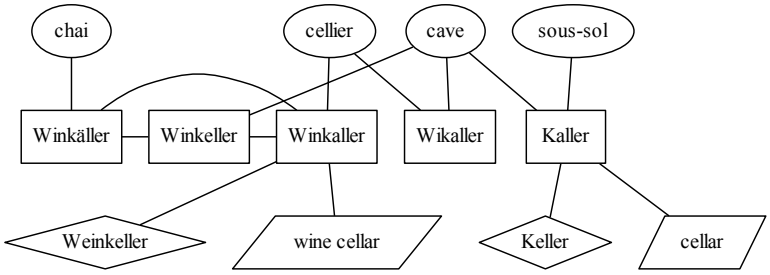

Figure 1: Simplified view of a subgraph. French words are in ellipses, Alsatian words in boxes, German words in diamonds and English words in parallelograms.

pivot language, Alsatian the source language and German and English the target languages. Several methods have been proposed, relying mostly either on the structure of the available bilingual lexicons or on distributional similarity (Tanaka and Umemura, 1994; Saralegi et al., 2011). In our particular case, we exploit the closeness between Alsatian and German, and, to a lesser degree, English. Starting from the French translations, German and/or English translations are added to the bag-of-words representations of Alsatian words if they share one of their double metaphone keys. This constraint performs a sort of disambiguation and ensures that only valid translations are selected. Thus, in the example of Figure 1, the German word "Weinkeller" and English word "wine cellar" will be added to the bag-of-words.

5. Evaluation of the Aligned Lexicon

5.1. Evaluation Methodology

In order to evaluate our method, we manually produced 100 ground-truth alignments between the lexicons and BabelNet. To this aim, we randomly selected entries from a multilingual French-German-Alsatian-English dictionary (Adolf, 2006). This dictionary presents several advantages for the evaluation: several spelling variants are usually proposed for each Alsatian entry, translations into French, German and English are provided, thus facilitating the mapping to BabelNet and, finally, the dictionary focuses on Alsatian lexemes which are very similar to corresponding German and English words.

To produce our evaluation dataset, we excluded BabelNet mappings with no translations into French and chose to limit ourselves to at most two Babel synsets. In case of a tie, the mapping to BabelNet is considered as correct if at least one of the Babel synsets is correct.

The alignment of variants is evaluated in terms of precision, recall and F-measure. For each French word in the evaluation dataset, we count the intersection between its Alsatian variants in the gold standard and in the automatic alignments as true positives (TP). Automatically aligned variants which are not in the gold standard are considered as false positives (FP), while those in the gold standard which are not in the alignments are considered as false negatives (FN). Then, precision (P), recall (R) and F-measure (F) are computed as follows :

$$P = \frac{TP}{TP + FP} \quad ; \quad R = \frac{TP}{TP + FN} \quad ; \quad R = \frac{2 \cdot P \cdot R}{P + R}$$

The mapping to Babelnet is evaluated in terms of the proportion of correct mappings. Since Babel synsets can be ranked according to cosine similarity, we consider the top 1, 2 and 3 mappings and judge the mapping as correct if one relevant Babel synset is found among the top 1, 2 or 3.

	Lexicon alignments			Mapping to BabeNet		
	P	**R**	**F**	**top 1**	**top 2**	**top 3**
baseline	1.00	0.69	0.82	0.52	0.83	0.88
+ BN FR	0.98	0.71	0.83	0.56	0.85	0.89
+ JDM	1.00	0.71	0.83	0.52	0.80	0.86
+ BN FR & DE	0.98	0.71	0.83	0.72	**0.90**	**0.94**
+ BN FR & EN	0.98	0.71	0.83	0.63	0.83	0.91
+ BN FR, DE & EN	0.98	0.71	0.83	**0.76**	0.87	0.93
+ JDM + BN FR & DE	0.98	0.72	0.83	0.71	**0.90**	0.93
+ JDM + BN FR, DE & EN	0.98	0.72	0.83	0.75	0.87	0.92

Table 4: Evaluation results

5.2. Results

The evaluation results for different settings are detailed in Table 4. The baseline corresponds to a setting which does not make use of any external resource. + JDM entails that the JeuxDeMots synonyms have been used. + BN entails that BabelNet has been used, with lexicalisations in French (FR), German (DE) or English (DE).

Overall, the results for the alignment of variants are stable: the use of external resources leads to a slight drop in precision which is compensated by a a slight rise of recall. Also, recall is always lower than precision.

For the mapping to Babel synsets, the use of translations into German and, to a lesser degree, English, lead to clear improvements, in particular for pushing relevant Babel synsets to the first rank. The synonyms provided by JDM actually have a detrimental effect on the performance, most certainly because the synonym sets in this resource are different from those in BabelNet.

5.3. Discussion

The lower recall obtained for the alignment of variants is mainly due to the constraint which demands identical metaphone keys. In some cases, variants have different keys (e.g. "Chilche" - `KLX / XLX` and "Kirche" - `KRX`). This also raises a more fundamental question: can these variants still be considered as alternatives for the same lexeme, or do they form a new lexeme? In our construction of the gold-standard, we grouped variants as found in the multilingual dictionary, even though they might be rather different in some cases. In addition to the metaphone keys, more classical string similarity measures could be used to align variants, as it is done for cognate identification (Inkpen et al., 2005). These measures could help improving recall.

Some errors are also due to problems in retrieving part-of-speech tags for ambiguous dictionary entries. As one of the alignment conditions requires identical parts-of-speech, such entries are not considered as variants.

As shown by the results, adding multilingual features helps improving the mapping to Babel synsets. For the time being, German and English translations are selected based on their metaphone keys, which leads to missing translations for some features vectors. In future work, this could be improved by using additional bilingual lexicons, not necessarily limited to the translations available in BabelNet. Also, the inverse consultation method proposed in the context of pivot based bilingual dictionary construction could be put to use in order to add translations which are not necessarily cognates of the Alsatian variants (Tanaka and Umemura, 1994). However, since there is no monolingual corpus for the Alsatian dialects, methods based on distributional similarity are excluded.

Finally, the method is able to rank Babel synsets, but not to decide which of the synsets are accurate. A threshold for the cosine similarity could be learned, in order to obtain mappings only to relevant synsets.

6. Conclusion and Perspectives

We have presented a method to both align spelling variants of the same Alsatian lexeme found in several lexicons and map the variants to synsets in BabelNet. The alignment of the variants relies on the double metaphone algorithm while the mapping uses multilingual (German and English) features in its best performing setting. The mapping to BabelNet gives access to different kinds of additional information: definitions and glosses, translations into other languages, images, etc. All these could be used to produce language games or didactic resources for Alsatian. Moreover, this method could in principle be applied to many less-resourced languages, as the only needed resource is a bilingual lexicon.

In the future, we plan to provide the aligned lexicon in a standard format, to allow its use as Linked Open Data. SKOS for instance allows for several alternative lexical labels with no preferred label.[12] However, the absence of normalization is an issue for many NLP applications which could use the lexicon, in particular lemmatization. This will require finding solutions for this pervasive problem.

Acknowledgements

This work was supported by a grant from the scientific council of the Université de Strasbourg. We thank the reviewers for their insightful comments.

7. References

Adolf, P. (2006). *Dictionnaire comparatif multilingue: français-allemand-alsacien-anglais*. Midgard, Strasbourg, France.

[12] See `http://www.w3.org/TR/skos-reference/#L1606`

Almeida, J. J., Santos, A., and Simões, A. (2010). Bigorna – A Toolkit for Orthography Migration Challenges. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta.

Banea, C. and Mihalcea, R. (2011). Word sense disambiguation with multilingual features. In *Proceedings of the Ninth International Conference on Computational Semantics*, page 25–34.

Bauer, L. (2003). *Introducing Linguistic Morphology*. Georgetown University Press. 2nd edition.

Dasigi, P. and Diab, M. (2011). CODACT: Towards Identifying Orthographic Variants in Dialectal Arabic. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 318–326, Chiang Mai, Thailand.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY–A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 580–590, Avignon, France.

Hanoka, V. and Sagot, B. (2012). Wordnet creation and extension made simple: A multilingual lexicon-based approach using wiki resources. In *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC 2012)*.

Huck, D., Bothorel-Witz, A., and Geiger-Jaillet, A. (2007). L'Alsace et ses langues. Éléments de description d'une situation sociolinguistique en zone frontalière. *Aspects of Multilingualism in European Border Regions: Insights and Views from Alsace, Eastern Macedonia and Thrace, the Lublin Voivodeship and South Tyrol*, page 13–100.

Hudlett, A. and Groupe d'Etudes et de Recherches Interdisciplinaires sur le Plurilinguisme en Alsace et en Europe. (2003). *Charte de la graphie harmonisée des parlers alsaciens: système graphique GRAPHAL - GERIPA*. Centre de Recherche sur l'Europe littéraire (C.R.E.L.), Mulhouse, France.

Hulden, M., Alegria, I., Etxeberria, I., and Maritxalar, M. (2011). Learning word-level dialectal variation as phonological replacement rules using a limited parallel corpus. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 39–48, Edinburgh, Scotland, July.

Inkpen, D., Frunza, O., and Kondrak, G. (2005). Automatic identification of cognates and false friends in French and English. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, page 251–257.

Lafourcade, M. (2007). Making people play for Lexical Acquisition. In *Proceedings of SNLP 2007, 7th Symposium on Natural Language Processing*, Pattaya, Thailande.

Mosquera, A., Lloret, E., and Moreda, P. (2012). Towards Facilitating the Accessibility of Web 2.0 Texts through Text Normalisation. In *Proceedings of the Worshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250, December.

Niemann, E. and Gurevych, I. (2011). The people's web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS)*, page 205–214.

OLCA / EDinstitut. (2012). Etude sur le dialecte alsacien. Online, visited Feb 11, 2014: `https://www.olcalsace.org/sites/default/files/documents/etude_linguistique_olca_edinstitut.pdf`.

Phillips, L. (2000). The Double Metaphone Search Algorithm. *C/C++ Users Journal*.

Porta, J., Sancho, J.-L., and Gómez, J. (2013). Edit Transducers for Spelling Variation in Old Spanish. In *Proceedings of the Workshop on Computational Historical Linguistics at NoDaLiDa 2013*, volume 87 of *Linköping Electronic Conference Proceedings*, page 70–79.

Sagot, B. and Fišer, D. (2008). Construction d'un wordnet libre du français à partir de ressources multilingues. In *Actes de TALN 2008-Traitement Automatique des Langues Naturelles*.

Salloum, W. and Habash, N. (2011). Dialectal to standard Arabic paraphrasing to improve Arabic-English statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, page 10–21.

Saralegi, X., Manterola, I., and San Vicente, I. (2011). Analyzing methods for improving precision of pivot based bilingual dictionaries. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, page 846–856.

Scherrer, Y. (2008). Transducteurs à fenêtre glissante pour l'induction lexicale. In *Actes de RECITAL 2008*, Avignon.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*.

Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, page 297–303.

Zeidler, E. and Crévenat-Werner, D. (2008). *Orthographe alsacienne: bien écrire l'alsacien de Wissembourg à Ferrette*. J. Do Bentzinger, Colmar, France.

An Update and Extension of the META-NET Study "Europe's Languages in the Digital Age"

Georg Rehm[1], **Hans Uszkoreit**[1], **Ido Dagan**[2], **Vartkes Goetcherian**[3], **Mehmet Ugur Dogan**[4], **Coskun Mermer**[4], **Tamás Varadi**[5], **Sabine Kirchmeier-Andersen**[6], **Gerhard Stickel**[7], **Meirion Prys Jones**[8], **Stefan Oeter**[9], **Sigve Gramstad**[10]

META-NET
DFKI GmbH
Berlin, Germany[1]

META-NET
Bar-Ilan University
Tel Aviv, Israel[2]

META-NET
Arax Ltd.
Luxembourg[3]

META-NET
Tübitak Bilgem
Gebze, Turkey[4]

EFNIL, META-NET
Hungarian Academy of Sciences
Budapest, Hungary[5]

EFNIL, META-NET
Danish Language Council
Copenhagen, Denmark[6]

EFNIL
Institut für Deutsche Sprache
Mannheim, Germany[7]

NPLD
Network to Promote Ling. Diversity
Cardiff, Wales[8]

Council of Europe, Com. of Experts
University of Hamburg
Hamburg, Germany[9]

Council of Europe, Com. of Experts
Bergen, Norway[10]

Abstract

This paper extends and updates the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series. The updated comparison confirms the original results and paints an alarming picture: it demonstrates that there are even more dramatic differences in LT support between the European languages.

Keywords: LR National/International Projects, Infrastructural/Policy Issues, Multilinguality, Machine Translation

1. Introduction and Overview

The multilingual setup of our European society imposes societal challenges on political, economic and social integration and inclusion, especially in the creation of the single digital market and unified information space targeted by the Digital Agenda (EC, 2010). Language technology is the missing piece of the puzzle, it is the key enabler and solution to boosting growth and strengthening Europe's competitiveness.

Recognising Europe's exceptional demand and opportunities, 60 leading research centres in 34 European countries joined forces in META-NET, a Network of Excellence dedicated to the technological foundations of a multilingual European information society. META-NET was partially supported through four projects funded by the EC: T4ME, CESAR, METANET4U and META-NORD. META-NET is forging the Multilingual Europe Technology Alliance (META) with more than 760 organisations and experts representing multiple stakeholders and signed collaboration agreements with more than 40 other projects and initiatives. META-NET's goal is monolingual, crosslingual and multilingual technology support for all European languages (Rehm and Uszkoreit, 2013). We recommend focusing on three priority research themes connected to application scenarios that will provide European R&D with the ability to compete with other markets and achieve benefits for European society and citizens as well as opportunities for our economy and future growth.

This paper extends and updates one important result of the work carried out within the META-VISION pillar of the initiative, the cross-language comparison of LT support for 30 European languages as published in the META-NET Language White Paper Series (Rehm and Uszkoreit, 2012).

2. The Language White Paper Series

Answering the question on the current state of a whole R&D field is difficult and complex. For LT nobody had collected these indicators and provided comparable reports for a substantial number of European languages yet. To arrive at a first comprehensive answer, META-NET prepared the Language White Paper Series "Europe's Languages in the Digital Age" (Rehm and Uszkoreit, 2012) that describes the current state of LT support for 30 European languages (including all 24 official EU languages). This undertaking had been in preparation with more than 200 experts since mid 2010 and was published in the summer of 2012. The study included a comparison of the support all languages receive in four areas: MT, speech, text analytics, language resources. The differences in technology support between the various languages and areas are dramatic and alarming. In the four areas, English is ahead of the other languages but even support for English is far from being perfect. While there are good quality software and resources available for a few larger languages and application areas, others, usually smaller languages, have substantial gaps. Many languages lack basic technologies for text

analytics and essential resources. Others have basic resources but semantic methods are still far away.

The original study was limited to 30 languages (most of them official and several regional languages). These were, in essence, the languages represented by the membership of META-NET at the time of preparing the study. Since then, META-NET has grown and added members in countries such as Israel and Turkey. When we presented pre-prints of the series at LREC 2012 in Istanbul (also elsewhere), volunteers approached us and explained their interest to prepare white papers on additional languages. The first new white paper, reporting on Welsh, has recently been published (Evas, 2014).

The series is available at `http://www.meta-net.eu`. Here, we also present the press release "At least 21 European Languages in Danger of Digital Extinction", circulated on the European Day of Languages 2012 (Sept. 26). It generated more than 600 mentions internationally (newspapers, blogs, radio and television interviews etc.). This shows that Europe is very passionate and concerned about its languages and that it is also very interested in the idea of establishing a solid LT base for overcoming language barriers.

In 2010, META-NET initiated a collaboration with the European Federation of National Institutions for Language (EFNIL) and started presenting its goals at the annual EFNIL conferences. Along the same lines, META-NET approached the Network to Promote Linguistic Diversity (NPLD) and, in 2013, the Council of Europe's Committee of Experts that is responsible for the Charter on Regional and Minority Languages. Representatives of the three organisations were invited to a panel discussion at META-FORUM 2013 (Berlin, Germany, September 19/20) where it was agreed to intensify the collaboration between all organisations.

3. Language Communities

In addition to the update of the cross-language comparison, this paper extends the co-authorship and support of the META-NET study by three organisations representing the language communities.

3.1. EFNIL

Formed in 2003, the European Federation of National Institutions for Language has institutional members from 30 countries whose role includes monitoring the official language(s) of their country, advising on language use or developing language policy. It provides a forum for these institutions to exchange information about their work and to gather and publish information about language use and policy within the EU. EFNIL encourages the study of the official EU languages and a coordinated approach towards mother-tongue and foreign-language learning, as a means of promoting linguistic and cultural diversity within the EU.

There is an increasing awareness among EFNIL members of the relevance and importance of LT on several counts. First, as a vital component and indeed a requirement for the sustainability of their respective national languages in the digital age. Second, as a research and productivity tool that has increasing impact on their daily work. Third, EFNIL members, many representing the central academic institutions for their language, can contribute to the technology support for their language through the invaluable language resources they develop. As a modest homegrown effort, EFNIL is running a pilot project (EFNILEX) aimed at developing LT support for the production of bilingual dictionaries between language pairs which are considered by mainstream publishing houses as commercially unviable.

3.2. NPLD

The Network to Promote Linguistic Diversity is a pan-European network which works with constitutional, regional and smaller state languages. It has 35 members, 10 of these being either member state or regional governments and the others major NGOs who have a role or are interested in language planning and management. NPLD was established in 2007 and has already asserted itself as the main voice of those linguistic communities that are not the official languages of the EU. NPLD's formation is a reflection of the growing interest in lesser used languages in Europe. Many governments from across the continent have established departments charged with the specific task of revitalizing and promoting the use of these languages. Many of these governments are represented within NPLD.

NPLD has two main goals. The first is to take advantage of the growth in knowledge and expertise which is now available in the area of language regeneration by ensuring that it is shared. This is done mainly through meetings and seminars, and is in the process of being further developed through the expansion of a digital library on language planning for its members. The second goal concerns the issue of policy development at a European level. Although much is said by the European Institutions about the importance of linguistic diversity, very few policy initiatives are undertaken and less funding is provided to support European linguistic diversity. We aim to highlight this deficiency and to promote the need for more support for all indigenous languages of Europe to ensure that our rich landscape of languages, many of them highly endangered, survive into the future.

ICT and social media will play a vital role in the future survival of most, if not all of the languages of Europe. Working together on a European stage to develop technical resources in areas such as translation and voice recognition will be vital if we are to avoid the digital extinction of many of our languages.

3.3. Council of Europe Committee of Experts on the Language Charter

The European Charter for Regional or Minority languages is a treaty of the Council of Europe with the purpose to protect and promote the regional and minority languages used in Europe. The two main political goals are the preservation of Europe's cultural heritage and diversity, and the promotion of democracy. The historic cultural and linguistic diversity in Europe is an integral part of European identity, and policies that acknowledge and promote this diversity also facilitate intercultural exchange and the participation in democratic processes. 33 European states have signed the treaty, and 25 states of those have ratified. The Languages Charter is applied to more than 190 regional or minority languages (or language situations), with around 40 million users. Most of these languages are small, less than 50,000 users. Only a handful are spoken by more than a million.

There are three main regional or minority language (RML) situations: 1. A RML in one country is a majority language in another country (as German, Ukrainian and Hungarian); 2. A RML is a minority language in more than one country (as Basque, Romani and Sami); 3. A RML is only found in one country (as Galician, Sorbian and Welsh). The content provisions are found in two parts of the Charter. Part II sets out that the state party shall base its policies, legislation and practise on certain objectives and principles. They cover the acknowledgement of the RML as an integral part of the state's cultural wealth, securing the language area, the use of the RML in public and private life, education, also regarding non-speakers, the elimination of unjustified discrimination, raising awareness and tolerance among the majority population. Part III contains concrete undertakings a state may apply to specific languages in the areas where the languages are in traditional use. Topics covered in Part III are education, judicial authorities, administrative authorities and public services, the media, cultural activities and facilities, and economic and social life. A Committee of Experts (Comex) monitors how the states comply with their obligations under the Charter. The monitoring is primarily based on three-yearly, national reports, visits to the country and information from NGOs.

LT may serve as a vehicle for the protection and promotion also of RML. At present, LT is primarily used in relation to national and large regional languages, partly due to the investment required. However, from the perspective of the Language Charter: To preserve the historical cultural and linguistic diversity of Europe and to facilitate an active participation of all European citizens in our democratic processes, it is also important for the smaller languages in Europe to make use of LT. The challenge to all of us, governments, research, the industry and RML users, is therefore to identify which tools are the most important ones. The development of tools that will serve the needs of these languages, and to make them available in practice, both from an economic and user-friendly perspective, is the task ahead of us.

4. The Set of Languages

The original set covered by the META-NET White Paper Series comprised 30 languages (see table 1). Back then, several of the languages represented by research centres that are members in META-NET could not be addressed because due to a lack of funding for those members (e. g., Hebrew, Luxembourgish). Multiple regional and minority languages could not be taken into account because META-NET's focus were the official EU languages and the official national languages of all partners of the four funded projects.

The extended set of languages addressed in this paper now finally contains *all* official languages represented by META-NET and also by EFNIL. It also contains all regional and minority languages represented by NPLD and many of the languages monitored by Council of Europe's Committe of Experts on Regional and Minority Languages. About 40 of the languages that fall under the mandate of the Committee of Experts were excluded to keep this extension and update of the cross-language comparison manageable. We excluded languages which were not listed in (Ethnologue, 2013), which had less than 100,000 speakers (according to Ethnologue) and also all languages which did not originate in Europe.

5. Cross-Language Comparison

As already reported in the White Paper Series (Rehm and Uszkoreit, 2012), the current state of LT support varies considerably from one language community to another. In the following, we briefly recapitulate how the original cross-language comparison was prepared. In order to compare the situation between languages, we selected two sample application areas (machine translation, speech), one underlying technology (text analytics), and the area of basic language resources. Languages were categorised using a five-point scale: 1. Excellent support; 2. Good support; 3. Moderate support; 4. Fragmentary support; 5. Weak or no support. For the original 30 languages, LT support was measured according to the following criteria:

MT: Quality of existing MT technologies, number of language pairs covered, coverage of linguistic phenomena and domains, quality and size of existing parallel corpora, amount and variety of available applications.

Speech: Quality of existing speech recognition technologies, quality of existing speech synthesis technologies, coverage of domains, number and size of existing speech corpora, amount and variety of available speech-based applications.

	Language	Speakers	White Paper
1.	**Albanian**	7,436,990	
2.	**Asturian**	110,000	
3.	Basque	657,872	(Hernáez et al., 2012)
4.	**Bosnian**	2,216,000	
5.	**Breton**	225,000	
6.	Bulgarian	6,795,150	(Blagoeva et al., 2012)
7.	Catalan	7,220,420	(Moreno et al., 2012)
8.	Croatian	5,533,890	(Tadić et al., 2012)
9.	Czech	9,469,340	(Bojar et al., 2012)
10.	Danish	5,592,490	(Pedersen et al., 2012)
11.	Dutch	22,984,690	(Odijk, 2012)
12.	English	334,800,758	(Ananiadou et al., 2012)
13.	Estonian	1,078,400	(Liin et al., 2012)
14.	Finnish	4,994,490	(Koskenniemi et al., 2012)
15.	French	68,458,600	(Mariani et al., 2012)
16.	**Frisian**	467,000	
17.	**Friulian**	300,000	
18.	Galician	3,185,000	(García-Mateo and Arza, 2012)
19.	German	83,812,810	(Burchardt et al., 2012)
20.	Greek	13,068,650	(Gavrilidou et al., 2012)
21.	**Hebrew**	5,302,770	
22.	Hungarian	12,319,330	(Simon et al., 2012)
23.	Icelandic	243,840	(Rögnvaldsson et al., 2012)
24.	Irish	106,210	(Judge et al., 2012)
25.	Italian	61,068,677	(Calzolari et al., 2012)
26.	Latvian	1,472,650	(Skadiņa et al., 2012)
27.	**Limburgish**	1,300,000	
28.	Lithuanian	3,130,970	(Vaišnien and Zabarskaitė, 2012)
29.	**Luxembourgish**	320,710	
30.	**Macedonian**	1,710,670	
31.	Maltese	429,000	(Rosner and Joachimsen, 2012)
32.	Norwegian	4,741,780	(Smedt et al., 2012a; Smedt et al., 2012b)
33.	**Occitan**	2,048,310	
34.	Polish	39,042,570	(Miłkowski, 2012)
35.	Portuguese	202,468,100	(Branco et al., 2012)
36.	Romanian	23,623,890	(Trandabăţ et al., 2012)
37.	**Romany**	3,017,920	
38.	**Scots**	100,000	
39.	Serbian	9,262,890	(Vitas et al., 2012)
40.	Slovak	5,007,650	(Šimková et al., 2012)
41.	Slovene	1,906,630	(Krek, 2012)
42.	Spanish	405,638,110	(Melero et al., 2012)
43.	Swedish	8,381,829	(Borin et al., 2012)
44.	**Turkish**	50,733,420	
45.	**Vlax Romani**	540,780	
46.	**Welsh**	536,890	(Evas, 2014)
47.	**Yiddish**	1,510,430	

Table 1: Languages included in the updated cross-language comparison (new languages in bold, number of world-wide speakers according to Ethnologue)

Text Analytics: Quality and coverage of existing text analytics technologies (morphology, syntax, semantics), coverage of linguistic phenomena and domains, amount and variety of available applications, quality and size of existing (annotated) text corpora, quality and coverage of existing lexical resources (e. g., WordNet) and grammars.

Resources: Quality and size of existing text corpora, speech corpora and parallel corpora, quality and coverage of existing lexical resources and grammars.

Figures 1, 2, 3 and 4 show that there are massive differences between the 47 languages surveyed. The four updated comparisons can be considered a solid first draft that the authors of this contribution agree upon. The updated tables have been circulated and discussed by the organisations and communities involved in this article in order to arrive at a coherent result that all organisations and language communities are in agreement with.

6. Conclusions

In the original series of white papers, we provided the very first high-level comparison of LT support, taking into account 30 European languages. Even though more fine-grained analyses are needed, the first draft of the extended and updated comparison presented in this paper confirms the original results and paints an alarming picture: in its extended form, the comparison demonstrates that there are even more dramatic differences in LT support between the European languages, i. e., the technological gap keeps widening. While there are good-quality software and resources available for a few languages and application areas only, other (usually smaller) languages have substantial gaps. Many languages lack basic technologies for text analytics and essential resources. Others have a few basic tools and resources, but there is little chance of implementing semantic methods in the near future.

Back in September 2012, the original results were disseminated using a press release with the headline "At least 21 European languages in danger of digital extinction" (Rehm et al., 2014). The updated and extended comparison demonstrates, drastically, that the real number of digitally endangered languages is, in fact, significantly larger; also see (Soria and Mariani, 2013). Overcoming language borders through multilingual language technogies is one of our key goals. The comparison shows that, in our long term plans, we should focus even more on fostering technology development for smaller and/or less-resourced languages and also on language preservation through digital means. Research and technology transfer between the languages along with increased collaboration across languages must receive more attention.

One key problem in this regard is the following: the number of speakers of a certain language seems to correlate with the amount and quality of technologies available for that language. For companies there is simply no sustainable business case which is why they refrain from investing in the development of sophisticated language technologies for a language that is only spoken by a small or very small number of speakers. This is why regional, national and international organisations as well as funding agencies should team up in order to address this issue. META-NET suggests setting up and actively supporting a shared programme to develop at least basic resources and technologies for all European languages (Rehm and Uszkoreit, 2013).

Our results show that such a large-scale effort is needed to reach the ambitious goal of providing support for *all* European languages, for example, through high-quality

machine translation. The long term goal of META-NET is to enable the creation of high-quality LT for all languages. This depends on all stakeholders right across politics, research, business, and society uniting their efforts. The resulting technology will help transform barriers into bridges between Europe's languages and pave the way for political and economic unity through cultural diversity.

Acknowledgments

META-NET was co-funded by FP7 and ICT-PSP of the European Commission through the contracts T4ME (grant agreement no.: 249 119), CESAR (no.: 271 022), METANET4U (no.: 270 893) and META-NORD (no.: 270 899). The work presented in this article would not have been possible without the dedication and commitment of the 60 member organisations of the META-NET network of excellence and the more than 200 authors of and contributors to the META-NET Language White Paper Series.

7. References

Ananiadou, S., McNaught, J., and Thompson, P. (2012). *The English Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Blagoeva, D., Koeva, S., and Murdarov, V. (2012). *Българският език в дигиталната епоха – The Bulgarian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Bojar, O., Cinková, S., Hajič, J., Hladká, B., Kuboň, V., Mírovský, J., Panevová, J., Peterek, N., Spoustová, J., and Žabokrtský, Z. (2012). *Čeština v digitálním věku – The Czech Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Borin, L., Brandt, M.D., Edlund, J., Lindh, J., and Parkvall, M. (2012). *Svenska språket i den digitala tidsåldern – The Swedish Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., Trancoso, I., Quaresma, P., de Lima, V.L. Strube, and Bacelar, F. (2012). *A língua portuguesa na era digital – The Portuguese Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Burchardt, A., Egg, M., Eichler, K., Krenn, B., Kreutel, J., Leßmöllmann, A., Rehm, G., Stede, M., Uszkoreit, H., and Volk, M. (2012). *Die Deutsche Sprache im digitalen Zeitalter – German in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Calzolari, N., Magnini, B., Soria, C., and Speranza, M. (2012). *La Lingua Italiana nell'Era Digitale – The Italian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

EC. (2010). A Digital Agenda for Europe. European Commission. `http://ec.europa.eu/information_society/digital-agenda/publications/`.

Ethnologue. (2013). Ethnologue – Languages of the World. `http://www.ethnologue.com`.

Evas, J. (2014). *Y Gymraeg yn yr Oes Ddigidol – The Welsh Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

García-Mateo, C. and Arza, M. (2012). *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Gavrilidou, M., Koutsombogera, M., Patrikakos, A., and Piperidis, S. (2012). *Η Ελληνικη Γλωσσα στην Ψηφιακη Εποχη – The Greek Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Hernáez, I., Navas, E., Odriozola, I., Sarasola, K., de Ilarraza, A. Diaz, Leturia, I., de Lezana, A. Diaz, Oihartzabal, B., and Salaberria, J. (2012). *Euskara Aro Digitalean – Basque in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Judge, J., Chasaide, A. Ní, Dhubhda, R. Ní, Scannell, K.P., and Dhonnchadha, E. Uí. (2012). *An Ghaeilge sa Ré Dhigiteach – The Irish Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Koskenniemi, K., Lindén, K., Carlson, L., Vainio, M., Arppe, A., Lennes, M., Westerlund, H., Hyvärinen, M., Bartis, I., Nuolijärvi, P., and Piehl, A. (2012). *Suomen kieli digitaalisella aikakaudella – The Finnish Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Krek, S. (2012). *Slovenski jezik v digitalni dobi – The Slovene Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Liin, K., Muischnek, K., Müürisep, K., and Vider, K. (2012). *Eesti keel digiajastul – The Estonian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Mariani, J., Paroubek, P., Francopoulo, G., Max, A., Yvon, F., and Zweigenbaum, P. (2012). *La langue française à l' Ère du numérique – The French Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Melero, M., Badia, T., and Moreno, A. (2012). *La lengua española en la era digital – The Spanish Language in the*

Digital Age. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Miłkowski, M. (2012). *Język polski w erze cyfrowej – The Polish Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Moreno, A., Bel, N., Revilla, E., Garcia, E., and Vallverdú, S. (2012). *La llengua catalana a l'era digital – The Catalan Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Odijk, J. (2012). *Het Nederlands in het Digitale Tijdperk – The Dutch Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Pedersen, B. Sandford, Wedekind, J., Bøhm-Andersen, S., Henrichsen, P. Juel, Hoffensetz-Andresen, S., Kirchmeier-Andersen, S., Kjærum, J.O., Larsen, L. Bie, Maegaard, B., Nimb, S., Rasmussen, J.-E., Revsbech, P., and Thomsen, H. Erdman. (2012). *Det danske sprog i den digitale tidsalder – The Danish Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Rehm, G. and Uszkoreit, H., editors. (2012). *META-NET White Paper Series: Europe's Languages in the Digital Age*. Springer, Heidelberg, New York, Dordrecht, London. 31 volumes on 30 European languages. `http://www.meta-net.eu/whitepapers`.

Rehm, G. and Uszkoreit, H., editors. (2013). *The META-NET Strategic Research Agenda for Multilingual Europe 2020*. Springer, Heidelberg, New York, Dordrecht, London. `http://www.meta-net.eu/sra`.

Rehm, G., Uszkoreit, H., Ananiadou, S., Bel, N., Bielevičienė, A., Borin, L., Branco, A., Budin, G., Calzolari, N., Daelemans, W., Garabík, R., Grobelnik, M., García-Mateo, C., van Genabith, J., Hajič, J., Hernáez, I., Judge, J., Koeva, S., Krek, S., Krstev, C., Lindén, K., Magnini, B., Mariani, J., McNaught, J., Melero, M., Monachini, M., Moreno, A., Odijk, J., Ogrodniczuk, M., Pęzik, P., Piperidis, S., Przepiórkowski, A., Rögnvaldsson, E., Rosner, M., Pedersen, B. Sandford, Skadiņa, I., Smedt, K. De, Tadić, M., Thompson, P., Tufiş, D., Váradi, T., Vasiļjevs, A., Vider, K., and Zabarskaite, J. (2014). The Strategic Impact of META-NET on the Regional, National and International Level. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, May.

Rosner, M. and Joachimsen, J. (2012). *Il-Lingwa Maltija Fl-Era Diġitali – The Maltese Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Rögnvaldsson, E., Jóhannsdóttir, K.M., Helgadóttir, S., and Steingrímsson, S. (2012). *Íslensk tunga á stafrænni öld – The Icelandic Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Simon, E., Lendvai, P., Németh, G., Olaszy, G., and Vicsi, K. (2012). *A magyar nyelv a digitális korban – The Hungarian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Skadiņa, I., Veisbergs, A., Vasiļjevs, A., Gornostaja, T., Keiša, I., and Rudzīte, A. (2012). *Latviešu valoda digitālajā laikmetā – The Latvian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Smedt, K. De, Lyse, G. Inger, Gjesdal, A. Müller, and Losnegaard, G.S. (2012a). *Norsk i den digitale tidsalderen (bokmålsversjon) – The Norwegian Language in the Digital Age (Bokmål Version)*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Smedt, K. De, Lyse, G. Inger, Gjesdal, A. Müller, and Losnegaard, G.S. (2012b). *Norsk i den digitale tidsalderen (nynorskversjon) – The Norwegian Language in the Digital Age (Nynorsk Version)*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Soria, C. and Mariani, J. (2013). Searching LTs for Minority Languages. In *Proceedings of TALN-RECITAL 2013*, pages 235–247.

Tadić, M., Brozović-Rončević, D., and Kapetanović, A. (2012). *Hrvatski Jezik u Digitalnom Dobu – The Croatian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Trandabăț, D., Irimia, E., Mititelu, V. Barbu, Cristea, D., and Tufiş, D. (2012). *Limba română în era digitală – The Romanian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Vaišnien, D. and Zabarskaitė, J. (2012). *Lietuvių kalba skaitmeniniame amžiuje – The Lithuanian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Vitas, D., Popović, L., Krstev, C., Obradović, I., Pavlović-Lažetić, G., and Stanojević, M. (2012). *Српски језик у дигиталном добу – The Serbian Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Šimková, M., Garabík, R., Gajdošová, K., Laclavík, M., Ondrejovič, S., Juhár, J., Genči, J., Furdík, K., Ivoríková, H., and Ivanecký, J. (2012). *Slovenský jazyk v digitálnom veku – The Slovak Language in the Digital Age*. META-NET White Paper Series, Rehm, G. and Uszkoreit, H. (eds.). Springer, Heidelberg, New York, Dordrecht, London.

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	French Spanish	Catalan Dutch German Hungarian Italian Polish Romanian	**Albanian** **Asturian** Basque **Bosnian** **Breton** Bulgarian Croatian Czech Danish Estonian Finnish **Frisian** **Friulian** Galician Greek **Hebrew** Icelandic Irish Latvian **Limburgish** Lithuanian **Luxembourgish** **Macedonian** Maltese Norwegian **Occitan** Portuguese **Romany** **Scots** Serbian Slovak Slovene Swedish **Turkish** **Vlax Romani** **Welsh** **Yiddish**

Figure 1: Machine translation – state of language technology support for 47 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch Finnish French German Italian Portuguese Spanish	Basque Bulgarian Catalan Danish Estonian Galician Greek Hungarian Irish Norwegian Polish Serbian Slovak Slovene Swedish	**Albanian** **Asturian** **Bosnian** **Breton** Croatian **Frisian** **Friulian** **Hebrew** Icelandic Latvian **Limburgish** Lithuanian **Luxembourgish** **Macedonian** Maltese **Occitan** Romanian **Romany** **Scots** **Turkish** **Vlax Romani** **Welsh** **Yiddish**

Figure 2: Speech processing – state of language technology support for 47 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Dutch French German Italian Spanish	Basque Bulgarian Catalan Czech Danish Finnish Galician Greek **Hebrew** Hungarian Norwegian Polish Portuguese Romanian Slovak Slovene Swedish	**Albanian** **Asturian** **Bosnian** **Breton** Croatian Estonian **Frisian** **Friulian** Icelandic Irish Latvian **Limburgish** Lithuanian **Luxembourgish** **Macedonian** Maltese **Occitan** **Romany** **Scots** Serbian **Turkish** **Vlax Romani** **Welsh** **Yiddish**

Figure 3: Text analytics – state of language technology support for 47 European languages

Excellent support	Good support	Moderate support	Fragmentary support	Weak/no support
	English	Czech Dutch French German Hungarian Italian Polish Spanish Swedish	Basque Bulgarian Catalan Croatian Danish Estonian Finnish Galician Greek **Hebrew** Norwegian Portuguese Romanian Serbian Slovak Slovene	**Albanian** **Asturian** **Bosnian** **Breton** **Frisian** **Friulian** Icelandic Irish Latvian **Limburgish** Lithuanian **Luxembourgish** **Macedonian** Maltese **Occitan** **Romany** **Scots** **Turkish** **Vlax Romani** **Welsh** **Yiddish**

Figure 4: Speech and text resources – state of language technology support for 47 European languages

Hungarian-Somali-English Online Dictionary and Taxonomy

István Endrédy
Pázmány Péter Catholic University, Faculty of Information Technology and Bionics
50/a Práter Street, 1083 Budapest, Hungary

MTA-PPKE Hungarian Language Technology Research Group
50/a Práter Street, 1083 Budapest, Hungary
istvan.endredy@gmail.com

Abstract

Background. The number of Somalis coming to Europe has increased substantially in recent years. Most of them do not speak any foreign language, only Somali, but a few of them speak English as well.
Aims. A simple and useful online dictionary would help Somalis in everyday life. It should be online (with easy access from anywhere) and it has to handle billions of word forms, as Hungarian is heavily agglutinative. It should handle typos as the users are not advanced speakers of the foreign languages of the dictionary. It should pronounce words, as these languages have different phonetic sets. It should be fast with good precision because users do not like to wait. And last but not least, it should support an overview of the vocabulary of a given topic.
Method. A vocabulary (2000 entries) and a taxonomy (200 nodes) was created by a team (an editor and a native Somali speaker) in an Excel table. This content was converted into a relational database (*mysql*), and it got an online user interface based on *php* and *jqueryui*. Stemmer and text-to-speech modules were included and implemented as a web service. Typos were handled with query extension.
Results. Although the dictionary lookup process does stemming with a web service and makes a query extension process, it is very fast (100-300ms per query). It can pronounce every Hungarian word and expression owing to the text-to-speech web service.
Conclusion. This dictionary was opened to the public in October, 2013. (*http://qaamuus.rmk.hu/en*) The next step is the creation of a user interface optimised for mobile devices.

Keywords: online dictionary, taxonomy, Somali

1. Introduction

In the past years, the number of immigrant Somalis in Hungary has increased. Most of them do not speak any language except for Somali, but a few of them speak English, too. They can not manage their business without being able to communicate effectively, so they need local help.
Some of the immigrants asked for help at the Reformed Mission Center (*Református Missziói Központ*), where they got the opportunity to learn Hungarian as a foreign language. This helps them a lot in becoming independent.
Somali dictionaries are not easily accessible, especially not in the Hungarian–Somali direction. Therefore an online Somali dictionary was developed that can be used almost from everywhere. This project was started in the framework of the School Integration Programme of the Refugee Mission, funded by the European Refugee Fund (*Menekültmisszió Iskolai Integrációs Programja, Európai Menekültügyi Alap*).
László Joachim (a native Hungarian) created a Hungarian-Somali-English dictionary in the form of an Excel spreadsheet with the help of a native Somali speaker, Tukale Hussein Muhyadin. The dictionary contained a basic vocabulary of about 2000 entries, and a taxonomy that had 200 nodes. This database served as a basis for the online dictionary.

2. Available solutions

There are very few online Somali-English dictionaries[1]. They do not use stemming and they cannot correct spelling errors on input. At the time of development there was only one Somali–Hungarian dictionary[2] on the web, which was a community built word list with 865 translations.
The situation changed on 10th December 2013, when Google introduced[3] Somali on its popular Google Translate service. Although this application can even translate full sentences, erroneous translations are very frequent (Table 1).
Google Translate is based on statistical machine translation. In theory, the more example sentence pairs there are in the training corpus of translation system, the better the translations are. The system may improve in time, but if a language is highly agglutinative, this method can not learn every possible phrase and sentence. Hungarian words have many forms. Nouns might have several thousands word forms, verbs might have thousands of different word forms (cf. Section 4 below). Owing to the practically infinite number of possible word forms and the relatively free word order of Hungarian, the quality of conventional statistical machine translation for Hungarian will never be perfect. For example, Hungarian *elmehettetek* 'you may have left' is unknown for Google Translate, it is an infected word form of *elmegy* (last row of Table 1). Futhermore, Somali is also heavily agglutinative. This causes even more difficulty in translation.
Google usually translates between different languages through English. For example, it first translates from Somali to English, then from English to Hungarian. These

[1] http://www.afmaal.com/dictionary, http://www.freelang.net/online/somali.php?lg=gb

[2] http://en.glosbe.com/hu/so

[3] http://www.webpronews.com/google-translate-hits-80-languages-milestone-adds-9-new-ones-2013-12

input words	Google English	Google Somali	Google Hungarian	Our English	Our Somali	Our Hungarian
big		weyn ✓	nagy ✓		wayn ✓	nagy ✓
nice		- ✗	szép ✓		macaan ✓	finom, kedves ✓
went		u galay ✗	ment ✓		tagid ✓	megy ✓
high		sare ✓	nagy ✗		dheer ✓	magas ✓
degaan	residence ✓		tartózkodás ✓	accommodation ✓		szállás ✓
isku eeg	to see ✗		hogy ✗	similar ✓		hasonló ✓
jön	come ✓	yimaado ✓		come ✓	kaalay ✓	
nagy	high ✗	sare ✗		big, large ✓	weyn ✓	
elmehettetek	- ✗	- ✗		leave ✓	tagid ✓	

Table 1: Google translate test 17/12/2013

steps may include errors, a single error (in any of the steps) may impair the final translation. This type of error can be seen in Table 1, at the input word *nagy*. Google translates this Hungarian word as *high*, but this is not correct: it means *big, large*. This error occurs in the Hungarian–Somali direction as well: *nagy* is translated to Somali *sare* just like English *high*. As we can see Google uses English as intermediate language between Hungarian and Somali. This solution may impair the quality of translation.

To sum it up, Google Translate has errors in Somali-Hungarian translations according to tests. It is due to the fact that both languages are agglutinative, and GT translates with English as an intermediate language. A small mistake at any level may result finally in a poor translation. It should be used carefully in case of these languages. Consequently, a dictionary tool is needed for accurate translation between Somali-Hungarian-English instead of GT.

3. Architecture and modules

This project aimed to create an online dictionary for Somali immigrants. The content of the dictionary was imported into a relational database (*mysql*), with a few tables (details in Figure 3). The web interface was developed in *php* and *jqueryui*. The database design, data migration and the development of the web interface were done in this project. English and Hungarian stemmer and text-to-speech modules were used out of the box as a web service. The stemmer we used in this project is based on the morphological analyzer engine HUMOR ('High speed Unification MORphology') developed at MorphoLogic (Prószéky and Kis, 1999). The stemmer was implemented by the author. The TTS engine we used is Profivox (Olaszy et al., 2000) with Microsoft Speech API. The TTS and the stemmer service is provided by *morphologic.hu*. The basic dataflow of the system is illustrated on Figure 1.

4. Content of the dictionary

Size and structure. There are 2,000 entries, and 200 taxonomy nodes in the dictionary. Each entry has several fields, as shown in Table 2.

The Taxonomy field contains nodes which are related to the entry. For instance, "vegetable" belongs to the "food" and "vegetables" groups. These connections help students to explore or refresh the vocabulary of a given topic, by listing the child nodes of the taxonomy.

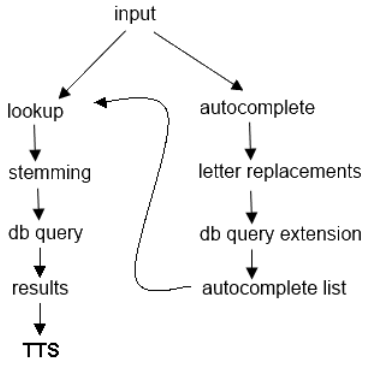

Figure 1: Dataflow of the system

field name	example value
English	would you mind, if...?
Hungarian	*baj, ha...?*
Hungarian keyword	*baj*
Somali	dhib male, hadii
part of speech (keyword)	noun
pronunciation	
other forms, grammatical information	
usage	
examples	Hu: *Nem baj, ha kinyitom az ajtót?* So: Dhib malah, hadaan daqada furo? En: Would you mind if I opened the door?
taxonomy	So: qalab wax sahlaya/karaan, awood, mug/ogolaansho, rukhsad Hu: *lehetőség/ képesség/ engedély* En: opportunity/ ability/ permission

Table 2: Example entry from the Excel table of the dictionary

Importing entries into the database. The editors of the dictionary created entries in an Excel table and a taxonomy hierarchy in a MS Word document. The entries were exported in csv (comma separated value) format. In this form they can be easily imported into a relational database, such as MySQL. The SQL table structure reflects that of

the columns of the Excel table (columns are illustrated in Table 2). The key columns have been indexed as well (Hungarian keyword, Somali, English). These columns became searchable.
The taxonomy did not have such a strict format, therefore it was parsed with a php script, and each taxonomy entity was put into a database table. The connection between words and their connections to the taxonomy were defined with the help of the "taxonomy" column of the Excel table. A word may have several taxonomy connections, it is a one-to-many relation in the database.
Some taxonomy entries were poorly formatted (missing a delimiter between different languages, or other syntax errors). In such cases, errors were corrected one by one or with the help of the editors.

An example fragment from the taxonomy Word document illustrates (Figure 2) that its format was not computer friendly. The content is bilingual without delimiters, therefore structure and content were not easy to parse (English translation is only for illustration)

1. DAD – *AZ EMBER* MAN
macluumaadka shakhsiga *személyes adatok:* personal data
1 macluumaadka shakhsiga guud ahaan *személyes adatok általában* personal data in general
2 magaca-qofka *név* name
3 ciwaan *lakcím,* address
4 da' *életkor* age

Figure 2: Example taxonomy entry

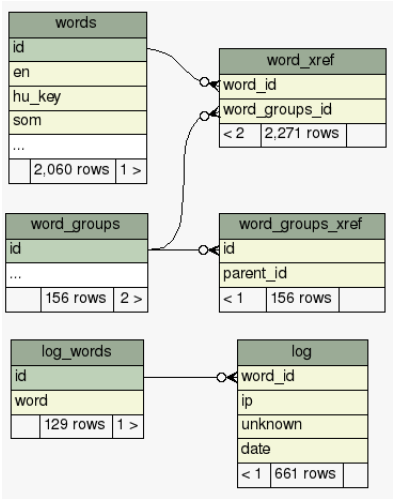

Figure 3: Relations between datatables

5. Features

The development of this online dictionary focused on the features which are important for foreign speakers. The following sections present the main features which are not available in other Somali online dictionaries.
A basic requirement of the application was to be user friendly and fast with good precision, despite the differences between the three languages (different phonetic sets, stemming rules and different types of frequent typing mistakes). The vocabulary should help Somali users to solve the most typical situations of everyday life.
Correction of typical typos. Hungarian has some digraphs and trigraphs which are difficult to write for a foreign speaker. The application replaces the typical typos with the correct word forms; otherwise the users will not find the searched word and will not learn the correct spelling. Typical typos were collected from all the three languages involved in this project, and search terms are completed with suggestions. This operation is triggered when the user types into the search input field. At this point, an autocomplete list is shown, and the user can click on a suggestion. (This process is described in detail in Section 6.)

Figure 4: The autocomplete feature with suggestions

Table 3 contains the typical letter replacements which are used for the creation of the autocomplete suggestion list.

Typical consonant replacements		
Consonants		Vowels
tsz → c	j → ly	i+<vowel > → ij+<vowel> ("fiatal" → "fijatal")
tz → c	nj → ny	e,é → i,í
dj → gy	f → v	a → e
dzs → gy	d → t ("fáratt")	a → o
cs → gy	s → sz	o → u
b → p	z → sz	fel → föl
ts → cs	sz → ssz	with and without accent: aeoiu → áéúöőüűóí
tj → ty	zs → sz	o → öőó
th → t	sz → s	u → üúű
lj → ly	sz → z	
l → ll ("szálás")		

Table 3: Letter replacements at query time

The application is capable of handling Hungarian di- and trigraphs, typical mistypings and phonetic mistakes, so it can find words in a great distance. For example for Hungarian "fijatal" the program will find "fiatal", or for "tsolad" the program will find "család". Of course these strings seem to be similar for a human, but for the computer these strings are very different. It is not trivial to find them based on these inputs.

Our application is capable of finding the correct spelling form even for strings with multiple errors.
Why does not provide the user interface a phonetized input option, a keyboard with phonetic input which may solve the problem of spelling errors and orthographic variations? In some languages, for instance French, phonetic input may help the users when a phoneme may have several letter combinations. Specifically if you do not know the spelling of 'éléphant', you can type with phonetic input 'elefan', and it will find the word correctly. In this case, Hungarian phonemes and letters are unknown for a Somali speaker, in addition Hungarian di- and trigraphs have different pronunciation (*gy, ty, ny, sz, zs, dz, dzs*, etc.). Consequently, a phonetic keyboard could not help, because the user does not know which letter or phoneme is necessary in the given word. We found it to be more comfortable for the user just to type the word, and correction is done on the fly with the autocomplete list. The Somali phonemes and letters are replaced with the possible Hungarian equivalents on each key press and the user may choose the correct form from the list.

Input stemming. Hungarian is an agglutinative language: one word (especially verbs) may have more than one thousand word forms (Oravecz and Dienes, 2002). In addition, Somalis most probably cannot type Hungarian words correctly. That is why the online user interface has to support typos and handle word stems: it has to find the entries by any word form of a given word. At query time the stem of the word is also searched in the dictionary. For example, if the user searches for *vagyok* 'I am', then its stem *van* 'is' will also be looked up. This feature increases the recall of the query results. Stemming is available in English and Hungarian as web services at *morphologic.hu*. The dictionary makes a web service call each time it needs to stem a word in these languages, and stems will be looked up in the dictionary as well. This way the user has the opportunity to copy/paste words in the form they occur in the original context, and the dictionary can find them easily.

Pronunciation: the text-to-speech module. Hungarian and Somali letter-to-sound rules differ considerably. For example, several sounds are marked by a single consonant letter in Somali while by a digraph in Hungarian. (e.g. the sound /s/ is marked by 's' in Somali, while by 'sz' in Hungarian). Due to these differences the application should be able to pronounce words as well. This application makes language learning easier. A text-to-speech module is available for Hungarian as a web service at *morphologic.hu*. If the user clicks on the icon "Listen", a web service call will be made, and the text can be listened to.

Multilanguage options. Visitors can select the language of the online user interface: it can be Somali, English or Hungarian. (The default setting is Somali since it is intended for Somali speakers.) The user can also set the language of the query.
At the beginning, the default source language of the search was Hungarian. But the first experiences showed that visitors type words in all three languages. Therefore the default setting is now to search in all three languages. The dictionary looks up words in each language, thus the 'not found' message became rarer.

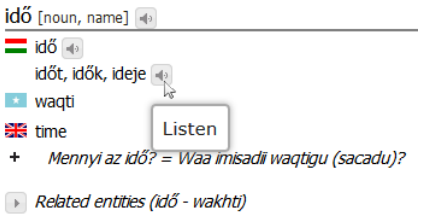

Figure 5: Text-to-speech module on the user interface

Taxomony. Our dictionary had a requirement that it should facilitate the overview of the vocabulary of a given topic. To attain this goal, we used a taxonomy. Although it would have been possible to use an existing semantic resource, we decided to create one of our own, as we found the hierarchy in the existing resources too detailed. The main consideration of the taxonomy nodes was the everyday usability from the apects of Somali immigrants.
DBpedia (Auer et al., 2007) has large scope with many nodes, but our project needs a taxonomy supporting at least two languages of the dictionary. DBpedia has English, but neither Hungarian nor Somali is included among the supported languages.
Although YAGO (Suchanek et al., 2008) has labels in Hungarian besides English, it has an extremely high granularity: several types of relations and levels, just like WordNet. YAGO covers a huge amount of concepts, people, organizations, geographical locations. For our project, only a basic subset of nodes, about 2% of the knowledge in YAGO would be needed.
Lexvo (de Melo and Weikum, 2008) has English and Hungarian translation as well. Although the taxonomy it is based on is almost a detailed as YAGO, we consider it as a potential source of extension for our taxonomy. As a first approach, Lexvo is connected to the dictionary in a light way. Each entry has a *related Lexvo taxonomy* link which may show the related nodes, translations and definitions from Lexvo. The connection is lazy: Lexvo content is downloaded on the fly based on the Hungarian keyword. Therefore an entry may show the related Lexvo nodes on the front end, with its sisters and parents. Further possibilities are discussed in Section 9 below.

Exploring the taxonomy in two ways. During this project, a taxonomy was also built, which represents topic nodes and their semantic connections. For instance, root nodes are *man*, *communication*, or *properties of things*. These nodes have child nodes; moreover each node may have connections to other nodes.
Entries of the dictionary may also have connections to these taxonomy nodes. These connections can be used to show related content. Entries with strong semantic connections can be listed or explored. There are two entry points to viewing the taxonomy: a bottom up and a top down approach.
Moreover, topic nodes can be explored in a hierarchical view, and each topic (or node) may list its children. This

way the vocabulary of a special topic can be listed. Topic-driven exploration helps language learners to look for a word or to revise the vocabulary of a semantic field fast. This function is available in a separate menu. In this case, the root nodes are shown by default (e.g. *man, things, habits*), and each node can be opened to reveal its child nodes.

On the other hand, an entry may show which other entries are connected to the same topic node. For example the *primary school* entry has a connection to the *school types* taxonomy node. Then every connected entry of *school types* taxonomy node will be shown when displaying the *primary school* entry.

Every word entry has a "related entries" link. At this point the user can view its sibling and parent nodes in the taxonomy hierarchy. This is illustrated in Figure 5. There is also a "more related entries" link, which displays the parent node of the entry. This possibility may give a bigger overview of the semantic group of the given entry.

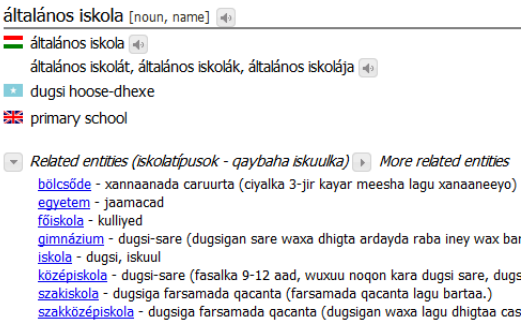

Figure 6: Related (sibling) entries

Feedback. The project had a requirement that the users should have the ability to report if a word is missing or they have any problem with the dictionary. Therefore a feedback user interface was developed, which sends an email to the administrator with the user's message.

6. Online administrative features

The content is constantly enlarged by the editors, therefore an online administration user interface was developed. It is more powerful than importing Excel tables. On the one hand, importing fails if a delimiter is missing or a new column appears: it is not a fault-tolerant process. On the other hand, online editing has the benefit that every modification is immediately ready for use by the public.

Editable entries + taxonomy. Each property of the entry can be edited. Some of them have an autocomplete feature: if the administrator starts to type in the field 'part of speech' or 'taxonomy connections', potential suggestions are displayed. This method fastens the process of editing, and it keeps these fields more consistent (see Figure 7).

There is an option to upload images or videos to an entry.

Editing can be started from the administrator's interface and from the public interface as well. (If one is logged in as an administrator, an edit icon appears next to each entry.)

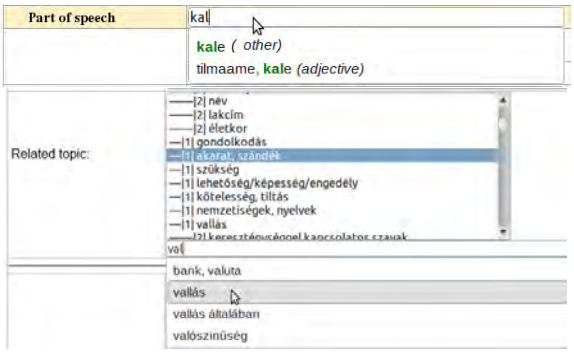

Figure 7: Autocomplete features on the admin interface

Logging queries. It is useful for editors to look at the searched words. It can answer such questions as: what is important for users, what is missing, which topic is the most popular this month, what kind of words were interesting for a user in one session?

Therefore each searched word is logged with the following pieces of information: known or unknown word, timestamp, ip address (just for identifying the user session).

Google analytics is also used independently, to analyse visitor information.

7. Example of usage

A user would like to find the meaning of the word *'young'*. He can not spell this word correctly, and types *'yuung'* into the input field. The autocomplete feature of the dictionary application suggests words for this string, in other words, it corrects the input to the forms which are known to this dictionary. This step is illustrated in Figure 8.

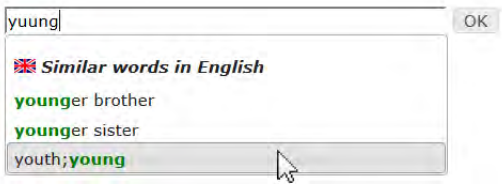

Figure 8: Autocomplete feature

Even if the user types the word correctly, the autocomplete feature makes the typing and inquiry faster. As it saves time, so users usually like it. At this point, the user may choose from the suggestion list. In this case, the intended word is in the last line (Figure 8). The search is started, and the user gets the results, the screenshot presented in Figure 9. It is important to mention that at this point (when the user clicks on an option in the autocomplete list), no automatic correction is done. The suggestion list contains only correct words, consequently it would be unnecessary to make corrections or suggestions on this input as well. If the user chooses a word from the suggestion list, the application takes the user's input as it is and looks it up without any automatic correction. Otherwise, similar entries would be noise in the result.

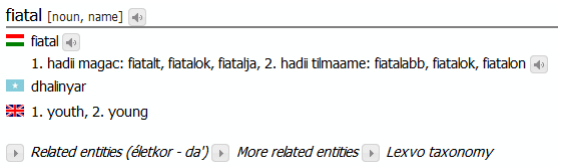

Figure 9: Example entry

8. Discussion

Other online Somali dictionaries lack autocomplete search and they do not handle typos either, and most of them have no text-to-speech option. Our solution goes beyond the earlier attempts. It is a big advantage that one can practice the pronunciation of a given word: it is practical for foreign speakers, especially for language learners. If the user can not remember the exact spelling of a word, and (s)he types a similar word (in other words: (s)he spells it incorrectly), our application will find it despite the errors. When you paste an unknown and inflected word from a text, a dictionary without stemming can not find its entry, especially in Hungarian where words may have very different forms. Our application includes stemming, so inflection is not a problem.

As for direct feedback, the editors of the dictionary are satisfied with the administrative features. Users have just started to use the application, therefore it is early to evaluate the project.

9. Future plans

The next step in the project could be the creation of a mobile application or a mobile-optimized web page. This way the dictionary could be used easily from anywhere. The dictionary service would be accessible in a comfortable way.

However, the exact improvements and changes will be based on feedback from the users, so that the program could satisfy real needs. Therefore the service will follow the requirements.

The present content of the dictionary is tuned to beginners' needs, with a basic vocabulary. The size of the vocabulary may be increased in the future. A wider entry set might serve professional needs as well.

Input stemming is done only in English and Hungarian. A Somali morphology and stemmer would increase the precision of the dictionary.

The taxonomy used in the dictionary is connected and completed with information from the Lexvo system. But this connection is created only on the fly, the related nodes are downloaded from Lexvo.org when user clicks on it. As a further step, Lexvo may be integrated in a deeper way. It can also be used as a source of additional nodes to our taxonomy and the corresponding dictionary nodes by importing English and Hungarian labels from Lexvo (possibly with manual correction in case of mistranslations) and opening up the possibility of supplying a Somali translation to users who have some knowledge of English in addition to Somali.

10. Conclusion

An online Somali-English-Hungarian dictionary was developed in this project for the Somalis who started to live in a foreign language environment (*http://qaamuus.rmk.hu/en*). The main aim was to help them in the most common situations, such as settling an administrative issue in an office, or shopping. It is important for them to be able to manage their business on their own, to live as ordinary citizens.

The features and the structure of the application were designed to serve the typical needs of language learners: assisting them in the process of learning how to write, pronounce and use words correctly. Entries were also selected for beginners, thus the vocabulary is composed of a basic vocabulary of everyday usage.

Administrators of the dictionary can edit the contents online, which is comfortable and the entries are ready for the public immediately after the modification.

Users can send feedback to the editors with a single click. This kind of direct feedback may result in a better and more usable dictionary.

11. Acknowledgements

I would like to express my gratitude to Dr Nóra Wenszky and her husband Attila Novák for their constructive and patient suggestions during the writing of this article. This work was partially supported by TÁMOP – 4.2.1.B – 11/2/KMR-2011-0002 and TÁMOP – 4.2.2/B – 10/1–2010–0014.

12. References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference*, ISWC'07/ASWC'07, pages 722–735, Berlin, Heidelberg. Springer-Verlag.

de Melo, G. and Weikum, G. (2008). Language as a foundation of the Semantic Web. In Bizer, C. and Joshi, A., editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, volume 401 of *CEUR WS*, Karlsruhe, Germany. CEUR.

Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Zainkó, C., and Gordos, G. (2000). Profivox — A Hungarian text-to-speech system for telecommunications applications. *International Journal of Speech Technology*, 3(3-4):201–215.

Oravecz, C. and Dienes, P. (2002). Efficient stochastic part-of-Speech tagging for Hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation, LREC2002*, pages 710–717, Las Palmas.

Prószéky, G. and Kis, B. (1999). A unification-based approach to morpho-syntactic parsing of agglutinative and other (highly) inflectional languages. In Dale, R. and Church, K. W., editors, *ACL*. ACL.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. *Web Semant.*, 6(3):203–217, September.

Computerization of African Languages-French Dictionaries

Chantal Enguehard, Mathieu Mangeot

Laboratoire LINA
2 rue de la Houssinière, BP 92208
F-44322 NANTES CEDEX 03 France

Laboratoire GETALP-LIG
41 rue des mathématiques BP 53
F-38041 GRENOBLE CEDEX 9 France

Email: Chantal.Enguehard@univ-nantes.fr, Mathieu.Mangeot@imag.fr

Abstract

This paper relates work done during the DiLAF project. It consists in converting 5 bilingual African language-French dictionaries originally in Word format into XML following the LMF model. The languages processed are Bambara, Hausa, Kanuri, Tamajaq and Songhai-zarma, still considered as under-resourced languages concerning Natural Language Processing tools. Once converted, the dictionaries are available online on the Jibiki platform for lookup and modification.

The DiLAF project is first presented. A description of each dictionary follows. Then, the conversion methodology from .doc format to XML files is presented. A specific point on the usage of Unicode follows. Then, each step of the conversion into XML and LMF is detailed. The last part presents the Jibiki lexical resources management platform used for the project.

Keywords: DiLAF, dictionary, Jibiki

1. Introduction

The work behind this paper has been done during the DiLAF project to computerize African languages-French dictionaries (Bambara, Hausa, Kanuri, Tamajaq, Zarma) in order to disseminate them widely and extend their coverage. We present a methodology for converting dictionaries from Word .doc format in a structured XML format following the Unicode character encodings and Lexical Markup Framework (LMF) standards. The natural language processing of African languages is in its infancy. It is our duty to help our colleagues from the South in this way. This requires, among other things, the publication of articles, that are a valuable resource for under-resourced languages.

Many studies have been conducted in the past in this area. However, it seemed interesting to redefine a new methodology taking into account recent developments such as the Open Document Format (ODF) or LMF standards. On the other hand, we wanted to develop the simplest possible method based solely on free and open source tools so that it can be reused by many. This method can also be used for other dictionaries and by extension, any text document (language resource at large) to be converted to XML.

2. Presentation of the DiLAF project

If access to computers is considered as the main indicator of the digital divide in Africa, we must recognize that the availability of resources in African languages is a handicap with incalculable consequences for the development of Information Technology and Communication Technologies (ICT). Most languages in francophone West Africa area are under-resourced (π-language) (Berment, 2004): electronic resources are scarce, poorly distributed or absent, making use of these languages difficult when it comes to introducing them into the education system and especially develop their use in writing in the administration and daily life.

Dictionaries are the cornerstone of processing natural language, be it in the mother tongue or in a foreign language. The primary function of communication is conveying meaning, yet meaning is primarily conveyed through vocabulary. As David Wilkins, a british linguist (1972) wrote “so aptly “While without grammar little can be conveyed, without vocabulary nothing can be conveyed".

Thus, to help bridge this gap, we are engaged with colleagues from North and South to improve the equipment of some African languages through, among others, the computerization of printed dictionaries of African languages.

The DiLAF project aims to convert published dictionaries into XML format for their sustainability and sharing (Streiter et al., 2006). This international project brings together partners from Burkina Faso (CNRST), France (LIG & LINA), Mali (National Resource Centre of the Non-Formal Education) and Niger (INDRAP, Department of Education, and University of Niamey).

Based on work already done by lexicographers we formed multidisciplinary teams of linguists, computer scientists and educators. Five dictionaries were converted and integrated into the Jibiki lexical resources management platform (Mangeot, 2001). These dictionaries are therefore available on the Internet[1] under a Creative Commons license:

- Bambara-French dict. Charles Bailleul, 1996 edition;
- Hausa-French dict. for basic cycle, 2008 Soutéba;
- Kanuri-French dict. for basic cycle, 2004 Soutéba;
- Tamajaq-French dict. for basic cycle, 2007 Soutéba;
- Zarma-French dict. for basic cycle, 2007 Soutéba.

1 http://dilaf.org/

The aim of these usage dictionaries is to popularize the written form of the daily use of African languages in the pure lexicographical tradition (Matoré, 1973) (Eluerd, 2000). Departing from interventionist approaches of normative dictionaries (Mortureux, 1997), the present descriptive dictionaries remain open to contributions and their online availability online will, hopefully, develop a sense of pride among users of these languages. Similarly, they will participate in the development of a literate environment conducive to increase the literacy whose low level undermines the achievements of progress in other sectors.

3. Presentation of the dictionaries

Four of the five dictionaries have been produced by the Soutéba project (program to support basic education) with funding from the German cooperation and support of the European Union. These dictionaries for basic education have a simple structure because they were designed for children of primary school class in a bilingual school (education is given there in a national language and in French). Most terms of lexicology, such as lexical labels, parts-of-speech, synonyms, antonyms, genres, dialectal variations, etc. are noted in the language in question in the dictionary, contributing to forge and disseminate a meta-language in the local language, a specialized terminology. The entries are listed in alphabetical order, even for Tamajaq (although it is usual for this language to sort entries based on lexical roots) because the vowels are written explicitly (this mode of classification was preferred because it is well known by children).

3.1. Hausa-French dictionary

The Hausa-French dictionary includes 7,823 entries sorted according to the following lexicographical order: a b ɓ c d ɗ e f fy g gw gy h i j k kw ky ƙ ƙw ƙy l m n o p r s sh t ts u w y y̆ z (République du Niger, 1999a).
They are structured with different patterns according to the part-of-speech. All entries are typographical, followed by the pronunciation (tones are marked with diacritics placed on vowels) and part-of-speech. On the semantic level, there is a definition in Hausa, a usage example (identified by the use of italics), and the equivalent in French. For a noun, the gender, feminine, plurals and sometimes dialectal variants are noted. For verbs, it is sometimes necessary to specify the degree to calculate morphological derivatives. Morpho-phonological variants of feminine and plural adjectives derivations are also written.
Example:

> **jaki** [jàakíi] *s.* **babbar dabbar gida mai kamar doki, wadda ba ta kai tsawon doki ba amma ta fi shi dogayen kunnuwa**. *Ya aza wa jaki kaya za ya tafi kasuwa. Jin.:* n. *Sg.:* **jaka**. *Jam.:* **jakai, jakuna**. *Far.:* âne

3.2. Kanuri-French dictionary

The Kanuri-French dictionary includes 5,994 entries sorted according to the following lexicographical order: a b c d e ə f g h i j k l m n ny o p r ɾ s sh t u w y z (République du Niger, 1999b).
The orthographic form of the entry is followed by an indication of pronunciation targeting rating tones. The part-of-speech is shown in italics, followed by a definition, a usage example, a French translation and meaning in French. Additional information may appear as variants.
Example:

> **abərwa** [àbə̀ɾwà] *cu.* **Kəska təngəri, kalu ngəwua dawulan tada cakkidə.** *Kəryende kannua nangaro, abərwa cakkiwawo.* [Fa.: ananas]

3.3. Soŋay Zarma-French dictionary

The Zarma-French dictionary includes 6916 entries sorted according to the following lexicographical order: a ã b c d e ẽ f g h i ĩ j k l m n ŋ ɲ o õ p r s t u ũ w y z (République du Niger, 1999d).
Each entry has an orthographic form followed by a phonetic transcription in which the tones are rated according to the conventions already set for the Kanuri. The part-of-speech specify explicitly the transitivity or intransitivity of verbs. For some entries, antonyms, synonyms and references are indicated. A gloss in French, a definition and an example end the entry.
Example:

> **ɲagas** [ɲágás] *mteeb.* ● *brusquement (détaler)* ● sanniize no kaŋ ga cabe kaŋ boro na zuray sambu nda gaabi sahã-din ● *Za zankey di hansu-kaaro no i te ɲagas*

3.4. Tamajaq-French dictionary

The Tamajaq-French dictionary includes 5,205 entries sorted according to the following lexicographical order: a â ă ə b c d ḍ e ê f g ğ h i î j ǰ ɣ k l ļ m n ŋ o ô q r s ṣ š t ṭ u û w x y z ẓ (République du Niger, 1999c)
The orthographic form of the entry is followed by the part-of-speech and a gloss in French displayed in italics. For nouns, morphological information about the state of annexation is often included, the plural and gender are also explicitly stated. A definition and an example of usage follow. Other information may appear as variants, synonyms, etc. As Tamajaq is not a tonal language, phonetics does not appear.
Example:

> **əbeɣla** *sn.* mulet ♦ **Ag-anɣer əd tabagawt.** *Ibeɣlan wər tăn-tăha tămalăya.* anammelu.: **făkr-ejăḍ.** təmust.: **yy.** iget.: **ibəɣlan**.

3.5. Bambara-French dictionary

The Bambara-French dictionary of Father Charles Bailleul (1996 edition) includes more than 10,000 entries sorted according to the following lexicographical order: a b c d e ɛ f g h i j k l m n ɲ ŋ o ɔ p r s t u w y z.
This dictionary is primarily intended for French speakers wishing to improve Bambara but it is also a resource for Bambara speakers. In the words of the author himself, the

dictionary "plays the role of a working tool for literacy, education and Bambara culture." To date, it can be considered as the most comprehensive dictionary of the language. It is also used by specialist of other varieties of this language like Dyula (Burkina Faso, Côte d'Ivoire) and Malinké (Guinea, Gambia, Sierra Leone, Liberia, etc.).

4. General conversion methodology and tools

The main objective is to convert dictionaries from a word processor format adapted for human use into XML and explicitly mark all the information so one can use them automatically in natural language processing tasks. The constraints are, on the one hand, working with free, open and multi-platform tools and on the other hand define a simple process that can be understood and then performed independently by linguists having no computer knowledge except regular expressions.

4.1. Conversion methodology

The conversion methodology follows these steps:

1. conversion of problematic characters to Unicode;
2. conversion of OpenOffice format to XML;
3. identification and explicit tagging of each part of information (headword, part-of-speech, etc.);
4. XML validation and manual correction of errors in the data (closed lists of values, references);
5. entries structuring following the LMF standard.

4.2. Tools used

The first tool allows one to edit files in original format and then convert them into XML. For this step OpenOffice (or LibreOffice) is ideal. It is free and open source. Furthermore, besides the ISO standard Open Document Format (odt), it can open many Microsoft formats (rtf, .doc as well as .docx). Finally, the XML produced is simple, especially compared to Office Open XML Microsoft.

OpenOffice has a regular expression engine for search/replace functions. This tool can be used for many steps before converting to XML. However, the search/replace may be problematic because during replacements, the boundaries of text styles can be changed (part of a word is suddenly in italics). Because we rely on styles to convert to XML, we limit the conversion to Unicode characters to keep styles intact.

Then, we need an editor to modify the files. For these operations, the XML editors are not very useful because they do not directly change the plain text with regular expressions and most are not able to edit large files like dictionaries. We recommend using a simple "raw" text editor supporting regular expressions and syntax highlighting.

For XML validation and verification steps, a web browser such as FireFox does it very nicely. It is able to detect and display the XML validation errors and can interpret CSS and XSLT style to enhance the display.

4.3. Incremental backups needed

The methodology intends to make backups at each stage and keep track of all search/replace operations done in order to go back when errors resulting from improper action are identified. Sometimes it happens that an error is noticed long after being made. If an error can not be corrected simply by a new search/replace, it is possible to go back from a previous version.

Despite all precautions, sometimes errors are detected very late and it is very difficult to go back. If the error can not be detected automatically, it will require manual correction. One must keep in mind that nobody is perfect and yet others even better trained had to forget the possibility to automatically correct all the errors in the conversion process.

5. Use of Unicode

5.1. Characters conversion to Unicode

Although the alphabets of languages on which we have worked (Enguehard 2009) are mainly of Latin origin, new characters needed to note specific sounds in some languages with a single character has been adopted by linguists in a series of meetings. Thus, each of the alphabets we previously presented comprises at least one of these special characters: ɓ ɗ ƎɛɣƘɲŋɔƴ. Characters composed of a Latin character and a diacritical mark were also created: âêîôûãẽĩõũḍḷṣṭẓǧǰšr.

Although most of these characters are present for several years in the Unicode standard (based on the work of the ISO 10646 (Haralambous 2004)), dictionaries were written using old hacked fonts. A methodology has been defined to identify and replace the inadequate characters

abirillu [àbìríllù] *m.* • *avril* • annasaara handu taacanta kaŋ go marsu nda me game ra • *Abirillu, 15, 1974 no Sayni Kunce na hino sambu* • *f/b.* abirillo, abirilley

abiyanso [àbìyànsôo] *m.* • *aéroport* • batama kaŋ ra abiyey ga zumbu • *Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri ga zumbu* • *f/b.* abiyansa, abiyansey

abiyo [àbíyò] *m.* • *avion* • naarumay hari no kaŋ ra i ga boro nda jinay daŋ a ma deesi nd'ey • *Jidda no abiyey ga alfujaajey zumandi* • *him.* beene-hi • *f/b.* abiya, abiyey

abunaadam [àbúnăadàm] *m.* • *être humain, personne* • *f/b.* abunaadamo, abunaadamey • *di.* adamayze

Figure 1: Excerpt of the Zarma-French dictionary in original format.

with the ones defined in the Unicode standard. It implies that all identified characters are recorded in a file so one can easily repeat this operation if necessary. Table 1 shows part of the list for Zarma. There is no automatic method that will detect these problematic characters. It is imperative to look at the data.

Origin	*Unicode*
§	ã
é	ẽ
$	ɲ
ù	ŋ
£	Ɲ

Table 1: Partial view of the Unicode correspondence table for Zarma.

5.2. Digraphs lexicographical order

Digraphs can be easily typed using two characters but their use changes the sort order which determines the lexicographic presentation of dictionary entries. Thus, for Hausa and Kanuri, the digraph 'sh' is located after the letter 's'. So, in the Hausa dictionary, the word "sha" (drink) is located after the word "suya" (fried), and, in Kanuri, the word "suwuttu" (undo) precedes the name "shadda" (basin).

These subtle differences can hardly be processed by software and require that digraphs appear as a proper sign in the Unicode repertoire. Some used by other languages are already there, sometimes under their different letter cases: 'DZ' (U+01F1), 'Dz' (U+01F2), 'dz' (U+01F3) are used in Slovak; 'NJ' (U+01CA), 'Nj' (U+01BC), 'nj' (U+01CC) in Croatian and for transcribing the letter " Њ " of the Serbian Cyrillic alphabet, etc.

It would be necessary to complete the Unicode standard with digraphs of Hausa and Kanuri alphabets in their various letter cases.

fy	Fy	FY
gw	Gw	GW
gy	Gy	GY
ky	Ky	KY
kw	Kw	KW
ƙy	Ƙy	ƘY
ƙw	Ƙw	ƘW
sh	Sh	SH
ts	Ts	TS

Table 2: Hausa and Kanuri digraphs missing in Unicode.

5.3. Characters with diacritics

Some characters with diacritics are included in Unicode as a unique sign, others can only be obtained by composition.

Thus, vowels with tilde 'a', 'i', 'o' and 'u' can be found in Unicode in their lowercase and uppercase forms while the 'e' with a tilde is missing and must be composed with the character 'e' or 'E' followed by the tilde accent (U+303), which can cause renderings different from other letters with tilde when viewing or printing (tilde at a different height for example).

Letter ȷ with caron exists in Unicode as a sign ǰ (U+1F0), but its capitalized form Ɉ must be composed with the letter J and caron sign (U+30C).

The characters ẽ, Ẽ et Ɉ should be added to the Unicode standard.

5.4. Letter case change

Word processors usually provide the letter case change function, but do not always realize it the correct way.

Thus, we found during our work that OpenOffice Writer software (3.2.1 version) fails in transforming 'ɾ' to 'Ɽ' from lowercase to uppercase or vice versa (the character remains unchanged) while Notepad++ (5.8.6 version) fails in transforming ǰ in Ɉ.

6. Conversion of the format towards XML

Figure 1 shows an excerpt of the Zarma-French dictionary in the original .odt format. All the following examples are based on this dictionary.

The Open Document Format has the great advantage of being based on XML. Instead of a conversion, we will actually retrieve the contents of the XML document, then transform it to get what we want.

A document in ODF format is actually a zip archive containing multiple files including the text content in XML. This content is stored in the “content.xml” file in the archive. To retrieve this file, some clever manipulations must be followed. On MacOs, one has to create an empty folder and then copy the .odt file inside. Then, with a terminal, the “unzip” command must be launched to unzip the file. On Windows, the .odt file extension must be changed into .zip and then the. zip archive can be opened.

The file “content.xml” can now be extracted from the archive and then renamed and placed in another location. It becomes the base file on which we will continue our work. The next step consists in editing this file with a “raw” text editor.

One may first think that since the source file “content.xml” is already in XML, it may be enough to write an XSLT stylesheet to convert the file into an XML dictionary, but the XML used in the source file is completely different from the XML targeted. Indeed, the source file comes from a word processor. It is designed for styling a document and not for structuring a dictionary entry. Therefore, it is finally easier to convert the XML file “by hands” with regular expressions than to write an XSLT stylesheet for automatically converting the source file.

7. Explicit tagging of the information

```
<text:span text:style-name="Phonetic_20_form">
<text:span text:style-name="T7">[àbìyànsôo]</text:
span></text:span>
```

Figure 2: Part of an entry (prononciation) in XML ODF format

This step consists in tagging explicitly all pieces of information. Each piece of information is usually distinguished from others in the original file with a different style. Figure 2 shows a part of the "abiyanso" entry (airport) in the Zarma-French dictionary. The style used to indicate the pronunciation is "Phonetic_form".

After locating the pieces of information, one must choose a set of tags to mark them.

This raises the question of the choice of the language used for tags. The choice of English as the international language of research may be privileged. But in our case, English is not a language present in our dictionaries and furthermore, it is not mastered by all linguists colleagues working on the project. The use of French solves this problem since all partners master the language. However, in the case of under-resourced languages computerization projects, we believe that it is important to encourage partners to use the words of their language to define the name of the tags. This may possibly give rise to the creation of new terms that did not exist in these languages. From a political perspective, it helps to move away from a post-colonial vision of the social status of African languages and brings new value to these languages.

The set of tag now defined, the next step is to replace the ODF markup by this new “homemade” tagset.

Simply perform search/replace operations for each type of information. For the example, the following regular expression (perl syntax) removes the tag "T7":

```
s/<text:span text:style-name="T7">([^<]+)
<\/text:span>//g
```

The second expression replaces the tag "Phonetic_form" with "ciiyaŋ":

```
s/<text:span text:style-name="Phonetic_20_form">
([^<+)<\/text:span>/<ciiyaŋ>$1<\/ciiyaŋ>/g
```

```
<sanniize>abiyanso</sanniize><ciiyaŋ>[àbìyànsôo]
</ciiyaŋ><kanandi>m.</kanandi><bareyaŋ>aéroport
</bareyaŋ><feeriji>batama kaŋ ra abiyey ga
zumbu</feeriji><silmaŋ>Tilbeeri nda Dooso sinda
abiyanso kaŋ ra abiyo beeri ga zumbu</silmaŋ>
<f>abiyansa</f><b>abiyansey</b>
```

Figure 4: Entry converted with « homemade » tags

Replacing all tags leads to the result in Figure 4.

8. Correction of the data

At this stage, several corrections are performed on the data.

8.1. XML Validation

In order to use XML tools, our file must be well formed. The manipulations of the previous step almost always introduce XML syntax errors. FireFox includes an XML parser and is also able to indicate exactly where the errors are located in the file.

abarba [ábàrbà] *m.* type de banane **banaana dumi no kaŋ i ga haagu ga ŋwa** *Abarba gani ŋwaayaŋ ga hin ga te boro se gunde-kuubi* budde abarbaa abarbey
abirillu [àbìrílìù] *m.* avril **annasaara handu taacanta kaŋ go marsu nda me game ra** *Abirillu, 15, 1974 no Sayni Kunce na hino sambu* abirillo abirilley
abiyanso [àbìyànsôo] *m.* aéroport **batama kaŋ ra abiyey ga zumbu** *Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri ga zumbu* abiyansa abiyansey
abiyo [àbíyò] *m.* avion **naarumay hari no kaŋ ra i ga boro nda jinay daŋ a ma deesi nd'ey** *Jidda no abiyey ga alfujaajey zumandi* beene-hi abiya abiyey
abunaadam [àbúnăadàm] *m.* être humain, personne abunaadamo abunaadamey adamayze

Figure 3: Compact view in a browser

Once the error is located, one has to check if it is not repeated elsewhere in the file. If this is the case, a regular expression must be written to correct the error in a systematic way instead of doing it by hand. In our case, the following regular expression can solve the problem:

```
s/<sanniize([^<+)<\/sanniize>/<sanniize>$1<\/sanniize>/g
```

The XML file is now well formed. It is then possible to manipulate it with XML tools.

8.2. Verification of closed lists of values

The stage of verification of information taking their value in a closed list is important. Some errors come from bad handling in the previous steps, while others were present in the original file before conversion. For example, a dictionary uses parts-of-speech, a termbase uses a list of domains, etc. Make a copy of the file and keep only the values to check is a systematic approach for verification. In the example of Figure 4, the part-of-speech marked by "kanandi" can be extracted with the following expression:

```
s/^*<kanandi>([^<]+)<\/kanandi>*$/$1/
```

The resulting list must then be sorted alphabetically. TextWrangler and Notepad + + plugin with its TextFX have the necessary commands. If the editor does not offer this option, OpenOffice Calc spreadsheet can be used. This approach is then used to quickly detect irregularities. If a value appears only once, it is very likely that this is a mistake. In the dictionary used in the examples, we corrected "alteeb" to "alteeb.", "Dah." to "dab.", "m/tsif." to "m / tsif.", etc.

8.3. Simple corrections

A CSS style sheet can be set to view the data directly in a browser. A compact display with a different style for each type of information helps to detect structuring errors in an entry. In the example of Figure 3, we see immediately that definition (in bold) and example (in italics) are lacking for the entry "abunaadam".

With an XSL stylesheet, one can modify the data before display like adding a unique identifier for each entry, then,

for each reference define a hypertext link to the corresponding entry. When the linguist browses the file, s/he can click on the hyperlinks to verify that the references are also entries of the dictionary like the entry "abunaadam" in Figure 3 with a reference to the entry "adamayse".

It is essential to scrutinize the data to detect some errors, even if they can be fixed automatically thereafter with regular expressions. The data visualization step is also very important from a pedagogical point of view. It allows to show the benefits of XML encoding the data, in particular that several forms (style) can be associated with the same information (data). By learning the basics of CSS, the lexicographers can modify the style sheets themselves.

9. Structure of the entries

The entries can now be restructured. In files coming from word processors, the data structure is usually implied. We will have to add new structural elements to move towards a more standardized structure, allowing subsequent reuse. Concerning standards, LMF (Romary et al., 2004) became an ISO standard in November 2008 (Francopoulo et al., 2009). It suits ideally our goals. As it is a meta-model and not a format, we can apply the principle of the LMF model to our entry structure and keep our tags without using the LMF syntax. The core meta-model LMF is shown in Figure 5. The object "Lexical Entry" contains a "Form" and one or more "Sense" objects.

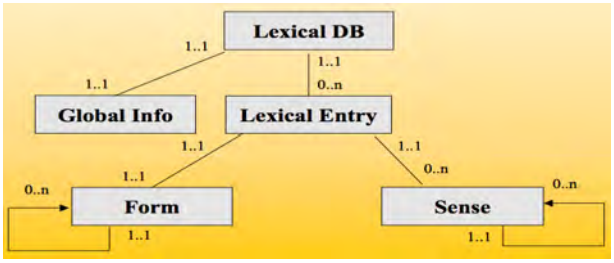

Figure 5: LMF kernel meta-model

Our lexical entries must now follow this meta-model. Figures 6 and 7 show an example of an entry before and after the addition of structuring tags. The "article" tag corresponds to the "Lexical Entry" object; the “bloc-vedette” tag correspond to the "Form" object and the "bloc-semantique" tag is the "Sense" object.

```
<sanniize>abiyanso</sanniize>
<ciiyaŋ>[àbìyànsôo]</ciiyaŋ>
<kanandi>m.</kanandi>
<bareyaŋ>aéroport<bareyaŋ>
<feeriji>batama kaŋ ra abiyey ga zumbu.</feeriji>
<silmaŋ>Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri
ga zumbu</silmaŋ>
<f>abiyansa</f><b>abiyansey</b>
```

Figure 6: Zarma entry before structuring

```
<article>
 <bloc-vedette>
  <sanniize>abiyanso</sanniize>
  <ciiyaŋ>[àbìyànsôo]</ciiyaŋ>
 </bloc-vedette>
 <bloc-grammatical>
  <kanandi>m.</kanandi>
  <f>abiyansa</f><b>abiyansey</b>
  <bloc-sémantique>
   <bareyaŋ>aéroport<bareyaŋ>
   <feeriji>batama kaŋ ra abiyey ga zumbu.</feeriji>
   <silmaŋ>Tilbeeri nda Dooso sinda abiyanso kaŋ ra abiyo beeri
ga zumbu</silmaŋ>
  </bloc-sémantique>
 </bloc-grammatical>
</article>
```

Figure 7: Entry after structuring following LMF model

```
<LexicalEntry id="abiyanso">
 <Lemma>
  <feat att="writtenForm" val="abiyanso"/>
  <feat att="phoneticForm" val="àbìyànsôo"/>
 </Lemma>
 <feat att="partOfSpeech" val="m."/>
 <Sense id="1">
  <Equivalent>
   <feat att="language" val="fra"/>
   <feat att="writtenForm" val="aéroport"/>
  </Equivalent>
  <Definition>
      <feat att="writtenForm" val="batama kaŋ ra abiyey ga
zumbu"/>
  </Definition>
  <Context>
   <TextRepresentation>
    <feat att="language" val="dje"/>
      <feat att="writtenForm" val="Tilbeeri nda Dooso sinda
abiyanso kaŋ ra abiyo beeri ga zumbu."/>
   </TextRepresentation>
  </Context>
 </Sense>
</LexicalEntry>
```

Figure 8: Zarma entry in LMF syntax

A simple XSLT stylesheet is then provided for download with each dictionary.

```
<xsl:template match="article">
   <LexicalEntry id="{sanniize}{sanniize/@lamba}">
     <xsl:apply-templates />
   </LexicalEntry>
</xsl:template>
<xsl:template match="bloc-vedette">
   <Lemma>
     <xsl:apply-templates />
   </Lemma>
</xsl:template>
<xsl:template match="sanniize">
      <feat att="writtenForm" val="{.}"/>
</xsl:template>
<xsl:template match="ciiya">
   <feat att="phoneticForm" val="{.}"/>
</xsl:template>
```

Figure 9: excerpt of the Zarma XSL stylesheet for producing LMF syntax

It converts each dictionary into the LMF syntax (see Figure 8). For more detailed information about this part, refer to (Enguehard & Mangeot, 2013).
The next step planned is to convert the resources into the Lemon format[2] and integrate them into dbnary[3] (Sérasset, 2014), the team database for linked data.

10. Web access via the Jibiki platform

10.1. Presentation of the platform

Jibiki (Mangeot et al., 2003; Mangeot et al., 2006; Mangeot, 2006) is a generic platform for handling online lexical resources with users and groups management. It was originally developed for the Papillon Project. The platform is programmed entirely in Java based on a the "Enhydra" environment. All data is stored in XML format in a Postgres database. This website mainly offers two services: a unified interface for simultaneous access to many heterogeneous resources (monolingual or bilingual dictionaries, multilingual databases, etc.) and a specific editing interface for contributing directly to the dictionaries available on the platform.
Several lexical resources construction projects used or still use this platform successfully. This is the case for the GDEF project (Chalvin et al., 2006) building an Estonian-French bilingual dictionary[4], the LexALP project about multilingual terminology on the Alpine Convention or more recently MotÀMot project on southeast Asias' languages[5]. The source code for this platform is freely available for download from the forge of the LIG laboratory[6].
An instance of the platform has been adapted specifically to DiLAF project[1] because, in addition to dictionaries, specific project information must be accessible to visitors:
- presentation of the project and partners;
- general methodology form converting published dictionaries to LMF format;
- stylesheets for different tools or tasks to be performed: tutorial on regular expressions, methodology of converting a document that uses fonts not conform to the Unicode standard to a document conforming to the Unicode standard, list of software used (exclusively open-source), methodology to monitor the project;
- presentation of each dictionary: original authors, principles that governed the construction of the dictionary, language, alphabet, structure of the lexical entries, etc.
- dictionaries in LMF format.

It is also envisaged to localize the platform for each language of the project.

10.2. Lookup interfaces

Three different interfaces are available to the user:
- the generic lookup allows the user to lookup a word or a prefix of a word in all the dictionaries available on the platform. The language of the word must be specified.
- the volume lookup allows the user to lookup a word or prefix on a specific volume. On the left part of the result window, the volume headwords are displayed, sorted in alphabetical order. An infinite scroll allows the user to browse the entire volume. On the right part of the window, the entries previously selected on the left part are displayed.
- the advanced lookup is available for complex multi-criteria queries. For example, it is possible to lookup an entry with a specific part-of-speech, and created by a specific author. On the left part of the result window, the headwords of the matching entries are displayed, sorted in alphabetical order. An infinite scroll allows the user to browse all the matching entries. On the right part, the entries previously selected on the left part are displayed.

10.3. Editing process

The editor (Mangeot et al., 2004) is based on an HTML interface model instantiated with the lexical entry to be published. The model is generated automatically from an XML schema describing the entry structure. It can then be modified to improve the rendering on the screen. Therefore, it is possible to edit any type of dictionary entry provided that it is encoded in XML.
The editing process can be adapted for specific needs through levels and status. A quality level (eg: from 1 to 5 stars, an entry with 1 star is a draft and one with 5 stars is certified by a linguist) can be assigned to each contribution. Similarly, a competence level can be assigned to each contributor (1 star is a beginner and 5 stars is a certified linguist). Then, when a 3 stars level user edits a 2 stars entry, the entry level raises to 3 stars.
Status can also be assigned to entries and roles to users. For example, in order to produce a high quality dictionary, an entry must follow 3 steps: creation by a registered user, revision by a reviewer and validation by a validator.

2 http://lemon-model.net/
3 http://dbnary.forge.imag.fr/
4 http://estfra.ee
5 http://jibiki.univ-savoie.fr/motamot/
6 http://jibiki.ligforge.imag.fr

10.4. Remote access via a REST API

Once dictionaries are uploaded into the Jibiki server, they can be accessed via a REST API. Lookup commands are available for querying indexed information: headword, pronunciation, part-of-speech, domain, example, idiom, translation, etc. The API can also be used for editing entries. The user must be previously registered in the website.

11. Conclusion

We presented a methodology for dictionaries conversion from word processing files to XML format. The DiLAF project does not stop in so good way. Before distributing dictionaries, there are still manual correction steps and possibly data addition. For example, examples of the Zarma-French dictionary will be translated into French. Once dictionaries are converted, we can then extend their coverage through a system of contribution / editing / validation that can be done online live on the Jibiki platform. The low Internet access in Africa will require us to develop alternative methods. We can then use the data as raw material to increase the computerization of these languages: morphological analysers, spell-checkers, machine translation systems, etc.

12. Acknowledgements

The DiLAF project is funded by the Fonds Francophone des Inforoutes of the International Organization of the Francophonie. We also thank all the linguists of the team without whom this project would not have been possible: Sumana Kane, Issouf Modi, Michel, Radji, Rakia, Mamadou Lamine Sanogo.

13. References

Berment V. (2004). Méthodes pour informatiser des langues et des groupes de langues « peu dotées ». Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France.

Buseman A., Buseman K., Jordan D. & Coward D. (2000). The linguist's shoebox: tutorial and user's guide: integrated data management and analysis for the field linguist, volume viii. Waxhaw, North Carolina: SIL International.

Chalvin, A. & Mangeot, M. (2006) Méthodes et outils pour la lexicographie bilingue en ligne : le cas du Grand Dictionnaire Estonien-Français. Actes d'EURALEX 2006, Turin, Italie, 6-9 septembre 2006, 6 p.

Eluerd R. (2000). La Lexicologie. Paris : PUF, Que sais-je ?

Enguehard C. (2009). Les langues d'Afrique de l'ouest : de l'imprimante au traitement automatique des langues. Sciences et Techniques du Langage, 6, 29–50.

Enguehard C. & Mangeot M. (2013) *LMF for a selection of African Languages*. Chapter 7, book "LMF: Lexical Markup Framework, theory and practice", Ed. Gil Francopoulo, Hermès science, Paris.

Francopoulo G., Bel N., George M., Calzolari N., Monachini M., Pet M. & Soria C. (2009). Multilingual resources for nlp in the lexical markup framework (LMF). Language Resources and Evaluation, 43, 57–70. 10.1007/s10579-008-9077-5.

Haralambous Y. (2004). Fontes et codages. O'Reilly France.

Mangeot M. (2001). Environnements centralisés et distribués pour lexicographes et lexicologues en contexte multilingue. Thèse de nouveau doctorat, spécialité informatique, Université Joseph Fourier Grenoble I, Grenoble, France.

Mangeot, M., Sérasset, G. and Lafourcade, M. (2003) Construction collaborative de données lexicales multilingues, le projet Papillon. (Papillon, a project for collaborative building of multilingual lexical resources). special issue of the journal TAL (Electronic dictionaries: for humans, machines or both?, edited by M. Zock and J. Carroll), Vol. 44:2/2003, pp. 151-176. 2003.

Mangeot, M. et Thevenin, D. (2004). Online Generic Editing of Heterogeneous Dictionary Entries in Papillon Project. Proc. of COLING 2004, ISSCO, Genève, Switzerland, 23-27 August, vol 2/2, pp 1029-1035.

Mangeot M. & Chalvin A. (2006). Dictionary building with the Jibiki platform: the GDEF case. In LREC 2006, p. 1666–1669, Genova, Italy.

Mathieu Mangeot (2006) Dictionary Building with the Jibiki Platform. Software Demonstration, Proc. EURALEX, Torino, Italy, 6-9 September 2006, 5 p.

Matoré G. (1973). La Méthode en lexicologie. Paris, France : Didier.

Mortureux, Marie-F. (1997). La lexicologie entre langue et discours. Paris, SEDES.

République du Niger (1999a). Alphabet haoussa, arrêté 212-99.

République du Niger (1999b). Alphabet kanouri, arrêté 213-99.

République du Niger (1999c). Alphabet tamajaq, arrêté 214-99.

République du Niger (1999d). Alphabet zarma, arrêté 215-99.

Romary L., Salmon-Alt S. & Francopoulo G. (2004). Standards going concrete: from LMF to Morphalou. In Proceedings of the COLING Workshop on Enhancing and Using Electronic Dictionaries, ElectricDict '04, p. 22–28, Stroudsburg, PA, USA: Association for Computational Linguistics.

Sérasset, G. (2014) Dbnary: Wiktionary as a Lemon Based RDF Multilingual Lexical Resource. Semantic Web Journal - Special issue on Multilingual Linked Open Data, 2014. (to appear).

Streiter O., Scannell K. & Stuflesser M. (2006). Implementing NLP projects for non-central languages: Instructions for funding bodies, strategies for developers. In Machine Translation, volume 20.

Wilkins, David A. (1972). Linguistics in Language Teaching. Cambridge, MA: MIT Press.

Morphological Analysis for Less-Resourced Languages: Maximum Affix Overlap Applied to Zulu

Uwe Quasthoff[1], Sonja Bosch[2], Dirk Goldhahn[1]
[1]NLP Group, Dept. Comp. Sci., University Leipzig, Germany
[2]Dept. of African Languages, University of South Africa
E-mail: quasthoff@informatik.uni-leipzig.de, boschse@unisa.ac.za, dgoldhahn@informatik.uni-leipzig.de

Abstract

The paper describes a collaboration approach in progress for morphological analysis of less-resourced languages. The approach is based on firstly, a language-independent machine learning algorithm, Maximum Affix Overlap, that generates candidates for morphological decompositions from an initial set of language-specific training data; and secondly, language-dependent post-processing using language specific patterns. In this paper, the Maximum Affix Overlap algorithm is applied to Zulu, a morphologically complex Bantu language. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. With limited training data and a ranking adapted to the language family, the effort for manual verification can be strongly reduced. The machine generated list is manually verified by humans via a web frontend.

Keywords: Complex morphology, Zulu, machine learning algorithm

1. Introduction

The paper describes work in progress. A two-step process is used to generate high quality morphological data for less-resourced languages, especially in the case of languages with complex morphologies. In these cases one cannot expect to have an automatic high quality analysis without extensive training data. The training data are usually expensive and for many languages, cannot be generated.

The approach introduced here is explained for morphological decomposition, but is applicable for solving other challenges as well (as described in section 5). We proceed as follows:

(1) Startingfrom an initial set of training data (i.e. words with their morphological decomposition), a machine learning algorithm generates candidates for morphological decompositions of 'new' words. This training set may be relatively small and also may contain errors or other inconsistencies. For each word, this might be either one in a ranked list of possible decompositions or just one (i.e. the most probable) decomposition.

(2) The machine generated list is manually verified by humans via a web frontend. Their task is to mark the correct decompositions. Alternatively, a word can be marked as "incorrectly analyzed" and the correct analysis can be inserted. A typical result is one correct decomposition per word. In the case of ambiguities, several decompositions might be correct. A word is treated as verified if at least one decomposition is marked as correct or an additional decomposition has been added. It is treated as not verified (and will be presented to another person for verification later) if nothing is marked.

The quality of both the annotated data and machine generated decompositions can be increased using a more complicated process:

- For higher quality and/or measuring agreement of different annotators, some or all entries can be presented to several persons. Additionally, pattern based algorithms may search for inconsistencies in the annotated data.
- The results of the human verification can be regarded as additional training data, with the result that the quality of the data presented in (2) increases steadily.

It should be noted that the task described in (2) is much simpler than decomposition without any suggestions. Choosing from a set of alternatives is less time consuming and needs less proficiency. For these reasons the task is well suited for a collaboration scenario.

The procedure above is demonstrated on Zulu morphology which is representative of many languages with complex morphology: morphological analysis is a prerequisite for POS tagging due to numerous short affixes and roots of possibly only one character.

2. Complex morphology of Zulu

Zulu [ISO 639-3: zul] belongs to the family of Bantu languages which have a complex morphological structure, based on two principles: a nominal classification system, and a concordial agreement system. According to the nominal classification system, nouns are categorized by prefixal morphemes that have been given class numbers for analysis purposes. These noun class prefixes generate concordial agreement linking the noun to other words in the sentence such as verbs, adjectives, pronouns, possessives etc. (cf. Poulos and Msimang, 1998) as illustrated by the bold printed morphemes in the following sentence:

***Aba**ntu **aba**ningi **ba**nga**yi**chitha **i**mali **ya**bo.*
Aba-ntu **aba-**ningi **ba-**nga-**yi-**chitha **i-**mali **ya-**bo.
[Many people may waste their money.]

In this example, the class 2 noun *abantu* [people] determines the subject agreement morpheme*ba-* in the verb *bangayichitha* [they may waste it], as well as the adjective agreement *aba-* in the qualificative*abaningi* [who are many]. The class 9 noun *imali* [money] determines object agreement *-yi-* in the verb and possessive agreement *-ya-* in *yabo* [of them]. We follow the root-based approach in morphological analysis of

Zulu where the root carries the principal semantic load of the word, e.g. *-ntu* and *-mali* (noun roots) and *-chith-* (verb root) in the sentence above. It should be noted that noun and verb roots belong to an open class which may demonstrate continuous growth.
The conjunctive orthography of the Zulu language causes a certain degree of morphophonological complexity. Most of the phonological adjustments at morpheme boundaries are predictable and rule-based. However, there are some exceptions - these are handled in the training data.

Zulu as a less-resourced language
According to Scannell (2007:1), more than 98% of the world's living languages lack most of the basic resources needed as a base for advanced language technologies, and are referred to as less-resourced or under-resourced languages. Zulu can therefore also be regarded as a less-resourced language, considering the unavailability of e.g. large monolingual and bilingual corpora, machine-readable dictionaries, POS taggers, morphological analysers, parsers, etc. Although some corpora exist (cf. University of Pretoria [1], Language Resource Management Agency [2] and Leipzig Corpora Collection[3]), they are limited in size, are not annotated and often not even accessible. Morphological analysers for Zulu are reported on e.g. a finite-state morphological analyser ZulMorph (Bosch et al. 2008), machine learning Zulu analysers (Spiegler et al. 2008; Shalanova et al. 2009), and a bootstrapping approach (Joubert et al. 2005). However, none of these morphological analysers is freely available.
The following algorithm describes a morphological analyser with a strict separation of the language-independent algorithm and the language specific training data. It can be assumed that the algorithm will work for other Bantu languages and possibly other language families as well. Possible language dependent limitations will be treated in a post processing step.
In the following section the algorithm is described without special reference to Zulu. Only the examples are taken from this language.

3. Morphologic Decomposition: The Maximum Affix Overlap Algorithm

Algorithms for morphological decomposition can use training data (so-called supervised algorithms) or use only word forms without any additional information (unsupervised algorithms like Morfessor (Creutz et al. 2006)). Unsupervised algorithms often have problems with complex morphologies; therefore we chose a supervised algorithm. The repeated succession of some of the morphemes will be used to classify the morphemes using the training data. In contrast to a rule-based morphological analyser such as ZulMorph (Bosch et al. 2008) that depends on a word root lexicon for successful analyses, this approach concentrates on affixes and allows the identification of previously unknown roots. The only additional assumption is that there is exactly one central element in the word, namely the root. In the case of compounds with two or more roots we assume that compound decomposition was applied in advance. We do not assume that the morphological analysis is unique. Instead, both the segmentation and the classification of the segments may be ambiguous.

Step 1: Language independent decomposition and ranking
We start with training data containing decompositions for a certain number of words so that we can assume that all possible combinations of prefixes are contained in the data as well as all combinations of suffixes. We do not assume, however, that all combinations of prefixes and suffixes are contained in the training data. Moreover, we do not assume all roots to be known in advance because one of the aims is to detect unknown or 'new' roots. In fact, checking for known roots will be postponed for step 2 of the algorithm. The algorithm returns a ranking of different decompositions and tags which is appreciated for languages with complex morphology because multiple decompositions are possible.
For a word *w* to be analysed, we perform the following steps:
For all segmentations of the word *w* into three segments *w1*, *w2* and *w3* (where *w1* and *w3* might be of zero length) we do the following:

- For each word *x* in the training set having exactly the prefix sequence *w1* we collect the pair (morphological analysis of *w1*, *w2* with the tag of the root of *x*).
- For each word *x* in the training set having exactly the suffix sequence *w3* we collect the pair (*w2* with the tag of the root of *x*, morphological analysis of *w3*).

From this, we form triples by joining on identical root tags: (morphological analysis of *w1*, *w2* with the tag of the root of *x*, morphological analysis of *w3*). Interesting features are the length of *w2* and the frequency of identical triples above. Because the procedure above allows considering affixes next to the root as part of the root, shorter roots should be preferred. In the case of multiple decompositions with the same root (or different roots of the same length) we rank the decompositions according to the frequency of the corresponding triple (morphological analysis of *w1*, tag of the root of *x*, morphological analysis of *w3*). In general, we set a frequency threshold of 2 for decompositions to be considered.
The example in Table 1 shows the analysis of the word *yocwaningo* [of research]. The correct decomposition has the highest frequency, but a shorter root candidate ranks higher.

[1] http://web.up.ac.za/default.asp?ipkCategoryID=1866&subid=1866&ipklookid=9
[2] http://rma.nwu.ac.za/
[3] http://corpora.informatik.uni-leipzig.de/

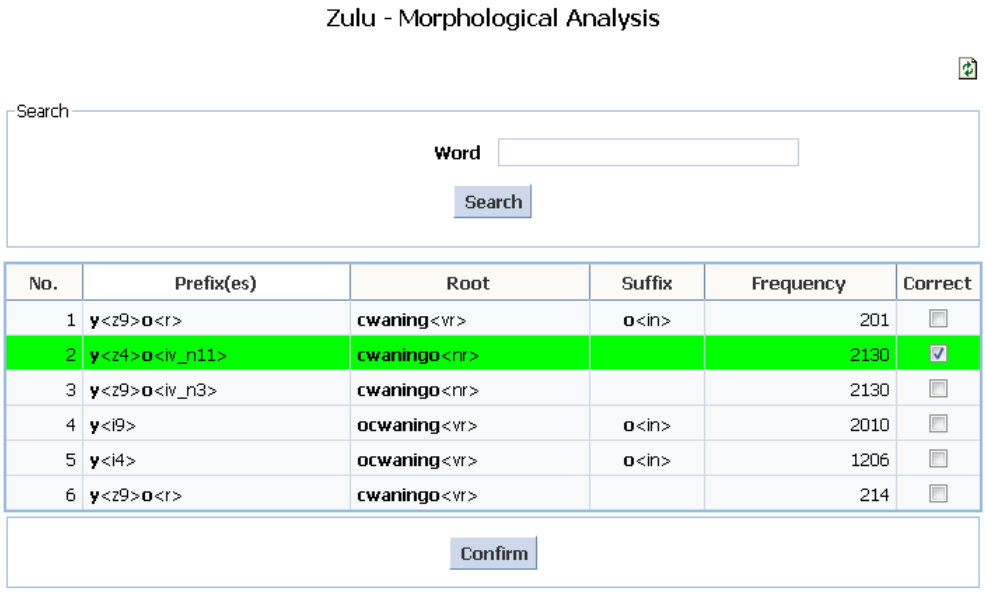

Table 1: Analysis of the word *yocwaningo* in the verification tool: line no. 2 is correct.

Step 2: Language dependent post-processing using special patterns

Step 1 of the algorithm produces too short roots if possible affixes (or parts thereof) are instead part of the root. In this case, it is not the shortest root candidate that will be the correct one. Here some language specific patterns help to exclude root candidates or give them a lower rating:

- Roots might not begin or end with some character or character sequence.
- Some roots can be extended by one character (usually a vowel) which also might be a suffix.
- Blacklisting: Some incorrect very short root candidates will be generated repeatedly. They can be blocked using a blacklist.
- Root transformations: The algorithm fails if the root is not part of the input word. But for Zulu, this happens only in rare cases. The most frequent transformation rule is given here: In a case such as the locative noun *ezandleni* [in the hands] the algorithm incorrectly provides *-andl-* as noun root. The locative prefix *e-* is necessary toensure that *-eni* is indeed a locative suffix and, moreover, that the nounroot is ending in *-a* or *-e*. Hence, this rule generates two possible roots: *-andla* (correct) and *andle* (incorrect).
- Agreement: In some cases, agreement between the prefix tag and the root is required. The noun root in the word *nomndeni* [and the circle of relatives] seems on the surface, to have a locative suffix *-eni*, and therefore the correct noun root-*ndeni* [circle of relatives] (class 3 noun) is not recognised. However, the absence of a locative prefix *e-* is the clue to the fact that there cannot be a locative suffix in this noun. Hence, *-eni*is part of the root.

Using frequency data for re-ranking

The following rules can be used if we have frequency information for roots. Usually, higher frequency should give a higher ranking. Such frequency information is:

- Frequency of a root in the training data (always available)
- Frequency of a root candidate (usually if not in the training data) in analysed corpus data.If, for instance, a correct root might be extended with several different vowels, these extensions will automatically get lower frequencies.

Training data

The training data used is the Ukwabelana (2013) word list consisting of approx. 10,000 words with labelled analyses described in Spiegler et al. (2010).

Evaluation

For 50 words of medium frequency (of frequency class 7, i.e. the most frequent word *ukuthi* [that / so that] is about 2^7 as frequent as the test words), which were randomly selected from a Zulu Newspaper Corpus of the Leipzig Corpora Collection [4], the automatic analyses were manually checked for the first correct analysis. It is counted for how many words the first analysis is correct, a correct analysis is in the top-5 or top-10 analyses provided. Here, both correctness of full analysis and correctness only for the root and its type are distinguished (cf. Table 2).

	Complete analysis		**Only root and its type**	
	absolute	**%**	**absolute**	**%**
total	50	100%	50	100%
correct at pos. 1	26	52%	30	60%
correct within pos. 1-5	32	64%	36	72%
correct within pos. 1-10	41	82%	41	82%
not correct within pos. 1-10	9	18%	9	18%

Table 2: Evaluation of the Maximum Affix Overlap algorithm

[4] http://corpora.informatik.uni-leipzig.de/

4. Sample Application: Identifying new roots

Here we focus in particular on the open classes of morphemes, viz. nounand verb roots. The root guesser using the above methods will facilitate the identification of "new" or adopted noun and verb roots that do not as yet occur in existing dictionaries or lexicons of the language. Such lists of "new" roots can be shared and integrated into existing applications such as ZulMorph and at the same time contribute to language development and orthographic standardisation purposes.

Example: The noun root *cwaningo* [research] (noun class 11) is a derivation from the verb root *-cwaning-* [conduct research] that does not feature in most Zulu dictionaries since it is a relatively "new" coinage. The correct analysis can be found in the Table 1 above.

5. Future work

It is planned to test the Maximum Affix Overlap algorithm for the other official Bantu languages of South Africa. Training data should become available in 2014 from the Language Resource Management Agency[5]. For future work, a more elaborate tag set than that used in (Spiegler et al. 2008, 2010) will be considered since the output of the analysis should be suitable for a POS-Tagger.

Both software and data for additional languages will be made availableunder the creative commons license by-nc.

The procedure described for morphological decomposition can also be applied to other tasks. The common feature is that for each input word the correct output has to be generated using machine learning and manual correction. This scenario applies to several problems such as the following:

- inflection type / baseform reduction, morphological decomposition, compound decomposition
- classification tasks for subject areas or relations (as in WordNet)
- bilingual translation equivalents

The combined data created for several of the above problems can contribute to improve the quality of the machine generated data.

6. References

Bosch, S., Pretorius, L. and Fleisch, A. (2008). Experimental Bootstrapping of Morphological Analysers for Nguni Languages.Nordic Journal of African Studies 17(2):66-88. Available: http://www.njas.helsinki.fi/. Accessed on 20 February 2014.

Creutz, M., Lagus, K. and Virpioja, S. (2006). Unsupervised morphology induction using Morfessor. In A. Yli-Jyrä, L. Karttunen, and J. Karhumäki (Eds.), Finite-State Methods and Natural Language Processing, Finite-State Methods and Natural Language Processing (FSMNLP 2005), Volume 4002 of Lecture Notes in Computer Science, pp. 300–301. Berlin, Heidelberg: Springer-Verlag.

Joubert, L., Zimu, V., Davel, M. and Barnard, E. (2004). A framework for bootstrapping morphological decomposition. Available: http://www.meraka.org.za/pubs/joubertl04morphanalysis.pdf. Accessed on 12 October 2013.

Poulos, G. and Msimang, C.T. (1998). A linguistic analysis of Zulu. Pretoria: Via Afrika.

Scannell, K.P. (2007). The Crúbadán Project: Corpus building for under-resourced languages. In C. Fairon, H. Naets, A. Kilgarriff, and G-M. de Schryver (Eds). Building and Exploring Web Corpora: Proceedings of the 3rd Web as Corpus Workshop. Cahiers du Cental, Louvain-la-Neuve, Belgium, Vol. 4 pp. 5-15.

Shalonova, K., Golenia, B. and Flach, P. (2009). Towards learning morphology for under-resourced languages. IEEE Transactions on Audio, Speech and Language Processing, 17(5):956–965.

Spiegler, S. (2011). Machine Learning for the Analysis of Morphologically Complex Languages. PhD Thesis. Merchant Venturers School of Engineering, University of Bristol.

Spiegler, S., Golenia, B., Shalonova, K., Flach, P. and Tucker, R. (2008). Learning the morphology of Zuluwith different degrees of supervision. IEEE Spoken Language Technology Workshop, pp. 9–12.

Spiegler, S., van der Spuy, A. and Flach, P.A. (2010). Ukwabelana - An open-source morphological Zulu corpus. Proceedings of the 23rd International Conference on Computational Linguistics (COLING), pp. 1020–1028.

Ukwabelana - An open-source morphological Zulucorpus. (2013). Available: http://www.cs.bris.ac.uk/Research/MachineLearning/Morphology/resources.jsp. Accessed on: 11 October 2013.

[5]http://rma.nwu.ac.za/

InterlinguaPlus Machine Translation Approach for Under-Resourced Languages: Ekegusii & Swahili

Edward O. Ombui[1,2], Peter W. Wagacha[2], and Wanjiku Ng'ang'a[2]
[1] Computer Science Dept. Africa Nazarene University (Kenya)
eombui@anu.ac.ke
[2] School of Computing and Informatics, University of Nairobi (Kenya)
waiganjo@uonbi.ac.ke, wanjiku.nganga@uonbi.ac.ke

Abstract

This paper elucidates the InterlinguaPlus design and its application in bi-directional text translations between Ekegusii and Kiswahili languages unlike the traditional translation pairs, one-by-one. Therefore, any of the languages can be the source or target language. The first section is an overview of the project, which is followed by a brief review of Machine Translation. The next section discusses the implementation of the system using Carabao's open machine translation framework and the results obtained. So far, the translation results have been plausible particularly for the resource-scarce local languages and clearly affirm morphological similarities inherent in Bantu languages.

Keywords: Machine Translation, InterlinguaPlus, Ekegusii

1. Introduction

Development of language applications for local languages in Africa requires innovative approaches since many of these languages are resource scarce. By this we mean that electronic language resources such as digital corpora, electronic dictionaries, spell checkers, annotators, and parsers are hardly available. These languages are also predominately spoken rather than written. Moreover, they are generally used in environments where there are other competing languages like English and French which have been well documented over the years with properly defined grammars, unlike the local languages with poorly defined grammars and dictionaries. This has been a major setback in the development of technologies for African languages. The presence of diacritics in most of these languages has also contributed to the complexity involved in the development of language technology applications. (Ombui & Wagacha, 2007). Nevertheless, there is pioneering work with the South African languages, which includes the definition of proper language grammars and development of a national language policy framework to encourage the utilization of the indigenous languages as official languages (NLPF, 2003).

In this paper, we consider two Bantu languages in Kenya namely Ekegusii and Swahili. There are approximately two million Ekegusii language speakers (KNBS, 2009). Swahili is widely spoken in East and Central Africa and is one of the official languages of the African Union with lots of printed resources.

For the work that we are reporting, we have adopted the InterlinguaPlus approach using the Carabao open machine translation framework (Berman, 2012). In this approach, all similar meaning words, synonyms, from each language and across the languages existing in the system are stored under the same category and assigned an identical family number. These words are also tagged with numbered lexical information[1]. For example, *Egetabu* (a book) [1=N;2=SG; 5=No]. Tag1 stands for the part of speech (1-POS), Noun, tag2 for number (2-No.), Singular, and tag5 indicates whether the noun is animate or inanimate etc. An amalgamation of the word's family identification number and tag numbers form a unique ID for the word. In addition, a novel way of only storing the base forms of each word and having a different table containing affixes that inflect the word drastically reduces the lexical database size and development time in general. This approach is implemented through the manual encoding of the sequence rules for the two languages.

Preliminary results are encouraging and clearly reveal similarities in the language structure of Ekegusii and Swahili. The advantage of this approach is that the translation is bidirectional and maintains the semantic approach to translation just as a human translator. In addition, it is suitable for rapid generation of domain specific translations for under-resourced languages.

2. Machine Translation

MT research has had a frustrating past and present in the light of translation quality, speed, and cost (Hutchins, 1996). This is evidenced by the ALPAC report (1966), and the small number of MT research being conducted in universities and software firms across the world. This has even resulted in a traditional view that MT challenges are solely linguistic requiring the translation system to have the intuition and knowledge that only human beings have. Nevertheless, we ought to acknowledge the progress of MT research projects in terms of improved translation speed and higher quality of translation outputs over a wider range of translation domains over the years.

Over the history of MT, several techniques and approaches have surfaced. The major methodologies include: Direct translation and indirect translation (i.e.

[1] Grammatical, Stylistic and Semantic tags

transfer-based and Interlingua-based). With the introduction of Artificial Intelligence technology in MT, more recent approaches have been proposed including Knowledge-based, Corpus-based, Human in loop and Hybrid methods (Pike,2006). Examples of the direct translation systems include the Systran, Logos and Fujitsu (Atlas) systems. Existing transfer-based systems include METAL and Ariane at GETA in Grenoble, The most notable Interlingua projects include the Rosetta project and Eurotra project for the European community languages. One of the great strengths of the InterlinguaPlus approach is that it preserves semantic information of the lexicon. Therefore, translation is primarily based on semantic equivalents between the lexicons of these languages.

As a result, the traditional language pair-based translation is replaced by bidirectional translations between the languages existing in the system. Any language can be the source or a target language. Consequently, the lexical database size is drastically reduced and the task of building multiple dictionaries is concentrated in constructing just one Interlingua lexical database. This kind of approach is evidently advantageous when building machine translation applications for under-resourced African languages because it expedites the process of adding a new language with minimal effort especially when adding languages of similar grammatical makeup, which could reuse some of the existing grammar rules.

3. Implementation

The figure 1 below illustrates the translation process in EMT[2] system. The user inputs a sentence, which is parsed into its constituent tokens. These tokens are then marched and mapped to their equivalent target-language tokens using the Family and mapping Identification numbers respectively. In addition, the sequence[3] e.g. Subject+Verb+Object, is parsed into elements and authenticated against the elements of the analyzed sentence. If it is valid, the elements are mapped according to the sequence and modified by the corresponding sequence in the target language. Some of the features that can be modified include deleting or adding a new element. E.g. He ate a mango.[eng:SVO]. *A+li+kula Embe* . Note that Swahili and generally the local languages do not have determiners. Therefore, when translating from Eng-Swa, the English determiner is dropped. However, it is added if the translation is vice-versa.

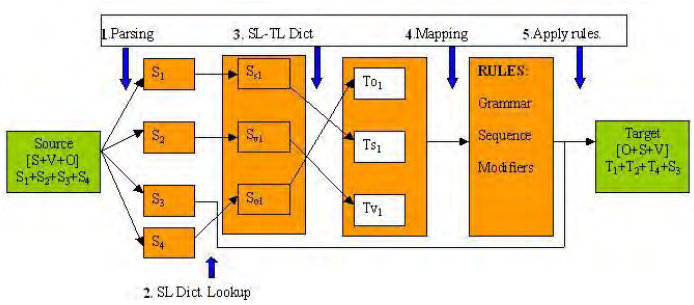

Figure1: EMT's MR-PDF

S1: Subject; v1: Verb;o1: Object; det: determinant; del: delimiter.

The above process, MR-PDF[4], is an acronym for the five translation stages (explained below) with the last two stages shifted at the beginning so as to give it an easy-to-remember name.

We will use example 1, English to Ekegusii SVO phrase to elucidate the process.

Example 1

He ate a mango.

Stage 1:Parsing

The sentence is analyzed syntactically according to its constituent structures i.e. tokens including syntax delimiters like question marks, exclamation marks etc.

He+ ate+ a+ mango.

S_{s1}:[He] S_{v1}:[ate] Det:[a] S_{o1}:[Mango] del:[.]

It is worth noting that at this stage, the parts of speech have not yet been identified.

Stage 2: Source Language Dictionary Lookup

Each token from stage 1 is looked up in the respective source language dictionary to check whether it exists in that language. In case it is not found, the word is left untagged and passed-on as it is to the next stages up to the output.

Stage 3: Family word-match

Every morpheme is examined considering all possible combination of affixes to it and each configuration stored. These are then matched[5] with the corresponding target language dictionary entities.

[He]=[*Ere*]

[ate]=[*ariete*] Past form of eat=*karia*

[a]= [a] yields the same token if equivalent is not found in the target language

[Mango]=[*Riembe*] Singular, noun.

All other delimiters, e.g. question marks (?), comas (,) are presented as they appeared in the source string. From the

[2] Ekegusii Machine Translator, built on Carabao's open MT Framework

[3] Set of elements, which refer to tokens that have specified features.

[4] Mapping, Rules, Parsing, Dictionary look-up, Family word-match

[5] Fuzzy matching is used to find similar meaning words in the target language dictionary

above example, all possible modifiers of the verb " to eat" are generated i.e. eat, ate, eaten, eats, eating, and matched with the corresponding verb in Ekegusii dictionary ie. *Karia, ariete, nkoria,* etc.
The tricky part of it is that one may not always have an equivalent number of modified verbs in the target or source dictionaries. To resolve this ambiguity, the program picks the modified verb with the best match in the target language dictionary i.e. in terms of matching lexical or style information e.g. the type of tense, number, animation, gender etc.
If we refer to the same example above, the following is examined as shown in Table 1and Table 2.

Language	**Morpheme**	**Part Of Speech**	**"Modified Morphemes"**
English	Eat	Verb	Ate; eaten; eating, eats, etc
Ekegusii	Ria	Verb	Karia, ariete, nkoriare, etc

Table 1: *Lexical information*

"Modified Morphemes"	**Tense**	**Number**
Ate	Past	Singular or Plural
Eating	Present continuous	Singular or plural
Mbariete	Past	Plural
Ariete	Past	Singular

Table 2: *Style information*

Language: English
Ate [tense-past; number-any]
It is apparent that both dictionaries are used to provide grammatical information, semantic data and potential equivalents in the target language during this stage.

Stage 4: Mapping

At the mapping stage, the Source text is validated against all existing sequences trees in the language. Only the most complete and detailed tree is picked. From example 1 above, the most appropriate sequence tree will be as follows and illustrated in figure 2.

He ate a mango — *Ri-embe a-rie-te*
[PN] + [V] + [Det] +[N] — [N] + [V]

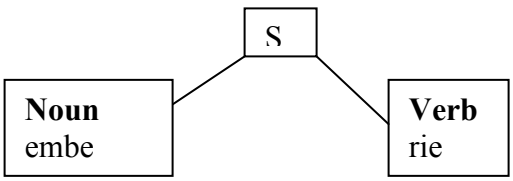

Figure 2: Sequence tree

The elements in the source sequence will map exactly into the [N] + [V] sequence. At this point all the redundant guesses are eliminated and disambiguation occurs. There are more comparisons and checks - like subject and style checks, etc.

Stage 5: Apply Rules.

The elements in the source sequence are modified by the corresponding sequence in the target language. The affixes are attached, or some new elements added or others completely deleted. Each element's unique identity is used to map the source sequence to the equivalent target sequence identities. Remember that Ekegusii does not have determiners and therefore it is dropped.
From the example above, the noun is then modified by adding the singular prefix – *ri*, (noun class 13) while the verb is modified by concatenating the subject- *a* (singular pronoun) to the verb- *rie* and finally adding the suffice-te (Past tense). The final sentence then becomes as shown below
Riembe ariete -> ri-embe a-rie-te
Incase it is converted to plural, the noun prefix will change to- *ama* (noun class 6) and the pronoun to- *ba* while maintaining the past tense suffix- *te*
Amaembe bariete -> Ama-embe ba-rie-te
Finally, the sentence word order is rearranged according to the best fitting sequence tree in the target language sequence table.

4. Results

The results gotten so far are plausible. The word order is correct as per the programmed sequence rules for each language e.g. English: This is a book; Ekegusii. *Eke n'egetabu*; Kiswahili: *Hiki ni kitabu*. In addition, the bidirectional functionality is often more than 50% accurate on the wider domains and about 90% accurate on specific domains, in our case the obituary's domain. Besides, once a text has been translated, it can also be used as the source text and the translator will yield exact translation as the initial source text. This therefore makes a strong case for the high intelligibility of the system.
The idea of storing only the word base forms and having a separate table for the affixes has drastically reduced the lexical database size as well as the building time. It was also noted that there is need for careful configuration of the rule units[6] for the affixes and lexicon otherwise the translation will be inaccurate. If we are to use the example above, the canonic[7] form will be as follows: English : FID-144 Book [POS:N; Number:SG;Animation: No]. However, for Ekegusii, there is need for additional rules

[6] A tag bearing any piece of grammatical data: part of speech, number contrast, gender, conjugation pattern, etc
[7] Base form of the word before any inflection

units to indicate the noun class[8] because the nouns inflection is dependent on the noun class, otherwise the machine translator might concatenate the wrong prefix. Therefore, the English example above will be matched as follows. Ekegusii: FID-144 ***tabu*** [POS: N; Animation: No, EkeNC[9]:8/9]. Consequently, the translator compares the rule units of the word with the rule units of the modifiers[10] in the affixes table and picks the most matching affix, in this case the prefix "*ege"* [POS: N; Number: SG; Animation: No, EkeNC[11]:8/9], ensuing in accurate translated word "ege***tabu***". On the contrary, if the Ekegusii rule units were not added or wrongly configured, the translation will be bizarre e.g. "Omo***tabu***" which is an invalid Ekegusii name. In fact, the prefix "omo" [EkeNC: 1] is often reserved for singular human[12] nouns.

The results obtained also expound the diversity of Ekegusii language linguistic rules[13] as compared to English. Most Indo-European languages, specifically English, espouse the SVO[14] sentence structure rule. However, in Ekegusii both SVO and VOS rules are valid sentence structure rules. For example, English: Mum ate mangoes [SVO]. Ekegusii: 1.*Omog'ina nariete amaembe [SVO]. 2. Nariete amaembe Omong'ina [VOS]*. Interestingly, the Ekegusii sequence and grammar rules that were copied and pasted to Swahili with minimal alteration resulted in almost precise translations between the two languages. This inevitably affirms the similarity in the language structure of the two languages and the ease in defining, constructing and translating between local languages as compared to/or from English.

The project demonstrations made so far to peers and some students have generated a lot of enthusiasm in African languages research and given a good indication of the reception of technology in a familiar language platform.

5. Conclusion

The InterlinguaPlus approach is good particularly for under-resourced languages in terms of generating rapid translations that give a good gist of the meaning in the second language. Although it takes some time to write the grammar rules for a new language at the beginning, it however takes a relatively shorter time when adding languages of similar grammatical makeup. Therefore, the approach is very feasible especially when considering under-resourced languages which may not be afforded the appropriate finances and sufficient political will to have technological resources built for them.

The lexical database building methodology, whereby words and their grammatical data are stored in respective families and assigned a unique identification, provides an excellent way of reducing the chances of ambiguity that may exist in the phonetic disparities inherent in these local languages.

The InterlinguaPlus approach employed in the Carabao Open MT framework forms a good foundation to scale existing language resources to many other under-resourced languages using minimal effort.

[8] There are about 17 Ekegusii noun classes
[9] Ekegusii Noun Class
[10] In this case, Prefixes
[11] Ekegusii Noun Class
[12] Professions, etc.
[13] Sequence and grammar rules
[14] Subject, Verb, Object

6. References

ALPAC (1966). Languages and Machines: Computers in Translation and *Linguistics*. National Academy of Sciences, National Research Council, 1966. (Publication 1416.)

Berman, V. (2012). Inside Carabao: Language Translation Software for XXI Century. *LinguaSys*. http://www.linguasys.com/web_production/PDFs/InsideCarabaoWhitePaper.pdf.

De Pauw, G., Wagacha, P, Atieno, D. (2006). *Unsupervised Induction of Dholuo Word Classes Using Maximum Entropy Learning*. University of Nairobi, Nairobi, Kenya.

Hutchins, W.J. (1986). *Machine translation: past, present, future*. Chichester: Ellis Horwood.

Hutchins, W.J. (1993). Latest Developments In Machine Translation Technology: Beginning A New Era In MT Research. *MT Summit* (1993), pp. 11-34.

Hutchins, W.J. (1994). *Research Methods And System Designs In Machine Translation: A Ten-Year Review, 1984-1994*.

Hutchins, W.J. (1996). ALPAC the (In)famous Report. MT News International, no. 14, June 1996, pp. 9-12.

Pike, J. (2006). *Machine Translation Techniques*. http://www.globalsecurity.org/intell/systems/mt-techniques.htm

Kenya National Bureau of Statistics (KNBS,2009). Ethnic Affiliation http://www.knbs.or.ke/censusethnic.php.

National Language Policy Framework(2003). Department of Arts and Culture. http://www.dac.gov.za/policies/LPD_Language%20Policy%20Framework_English%20_2_.pdf

Ombui, E, Wagacha, P. (2007). Machine Translation for Local Languages. In *Proceedings of COSCIT conference*. Nairobi, Kenya.

Wanjiku, N. (2006). Multilingual Content Development For ELearning In Africa. In *Proceedings of the conference "eLearning Africa: 1st Pan-African Conference on ICT for Development, Education and Training*. Addis Ababa, Ethiopia.

Zhejiang, M.L, Sheen, X,Liu. *Chomsky and Knowledge of Language*. Syracuse University. http://www.bu.edu/wcp/Papers/Lang/LangLiu2.htm.

UNL[arium]: a Crowd-Sourcing Environment for Multilingual Resources

Ronaldo Martins
UNDL Foundation
48, route de Chancy, Geneva, Switzerland
E-mail: r.martins@undlfoundation.org

Abstract

We present the UNL[arium], a web-based integrated development environment for creating, editing, validating, storing, normalising and exchanging language resources for multilingual natural language processing. Conceived for the UNL Lexical Framework, the UNL[arium] provides semantic accessibility to language constrained data, as it interconnects lexical units from several different languages, through taxonomic and non-taxonomic relations, representing not only necessary but also typical associations, obtained from machine learning and human input, in order to create an incremental and dynamic map of the human knowledge.

Keywords: semantic accessibility, crowd-sourcing, UNL

1. Introduction

The Universal Networking Language (UNL) is a mark-up language for organizing, in a machine-tractable and language-independent format, the information conveyed by natural language documents (Martins, 2013; Cardeñosa, Gelbukh, Tovar, 2005; Uchida, Zhu, Della Senta, 1999). Originally proposed in 1996 by the Institute of Advanced Studies of the United Nations University, in Tokyo, Japan, it has been promoted, since 2001, by the UNDL Foundation, in Geneva, Switzerland, under a mandate of the United Nations.

The main assumption behind the UNL is that the information conveyed in natural language documents can be better processed if converted into a semantic hyper-graph. In this sense, the UNL is fore and foremost a semantic network, and can be compared to several other semantic network approaches, such as wordnets (Miller et al, 1990), conceptual graphs (Sowa, 1984) and multinets (Helbig, 2006).

However, as an initiative of the United Nations, the UNL puts emphasis on its commitment with multilingualism: it cannot be bounded to any existing natural language in particular, under the risk of being rejected by the state members of the General Assembly. Accordingly, the UNL of a document in English cannot represent only the semantics or the syntax of English, but must organize and saturate the information so that any other language can easily process it. Furthermore, the UNL must not be circumscribed to major languages, and has a clear commitment with language diversity and under-resourced languages.

In this paper, we address the UNL[arium], the crowd-sourcing environment created by the UNDL Foundation in 2009 to foster the development of lexical resources within the UNL Program. We start by presenting the theoretical background of the system: the concept of "semantic accessibility" (Section 2), the UNL Lexical Framework (Section 3) and the FoR-UNL (Section 4). The UNL[arium] is presented in detail in Section 5. At last, our current challenges are addressed in Section 6.

2. Semantic Accessibility

One of the most outstanding problems in multilingual processing is the lack of isomorphism between vocabularies of different languages, even within the same language family. These lexical divergences are mainly of four different types[1]:

a) Categorial divergence, when the source and the target language represent the same information through different parts of speech (e.g., adjectives being translated as nouns, such as "hungry" from English to Spanish);
b) Conflational divergence, when the source and the target language represent the same information through different lexical configurations (e.g., overt realization of internal arguments, such as "to stab" from English to the Spanish "dar puñaladas", which involves a light verb + a non-verb element);
c) Semantic divergence, when the target language has mutually exclusive candidates for the same source language item (e.g., English has three different possible candidates for the Spanish verb "esperar": "to hope", "to expect" or "to wait"); and
d) Cultural divergence, when the source and the target language organize the world according to different values and categories (e.g., the word "ilunga", from Tshiluba, meaning "person who is ready to forgive any transgression a first time and then to tolerate it for a second time, but never for a third time", does not have any lexical counterpart in English).

[1] Except for the last one, which is normally referred to as a "translation mismatch", these divergences follow the "translation challenges" described by Dorr et al (1999). The authors mentioned some other challenges, such as thematic divergence and structural divergence, but they are rather syntactic and have not been representing actual challenges within the UNL program.

These divergences pose severe problems to multilingual processing, as they may require drastic structural changes (a and b); fine-tuned word sense disambiguation (c); and effective multilingual understanding (d). Our main goal is exactly to explore alternatives to solve these problems, especially the last one, which has been a puzzling obstacle in machine-aided multicultural communication.

In order to address these issues, we have been investigating the idea of "semantic accessibility", i.e., the degree to which a given lexical resource can be extended, at the semantic level, to as many languages as possible. Although related, accessibility is not to be confused with usability, which is rather the extent to which a resource can be (re)used to achieve specified goals with effectiveness and efficiency, and which seems to be the main goal of other language resource management initiatives, such as the Lexical Markup Framework (Francopoulo, 2013).

In the UNL Program, the main concern is the normalisation of content (i.e., meaning) rather than the standardisation of the format of lexical resources, although the latter is also part of the agenda. But our most urgent task is to interconnect lexical units from as many languages as possible, through taxonomic and non-taxonomic relations, representing not only necessary but also typical associations, obtained from machine learning and human input, in order to create an incremental and dynamic map of the human knowledge, which is actually the final goal of the UNL program, and which would grant semantic accessibility to otherwise language-dependent strings of characters.

This semantic accessibility is implemented, in UNL, by the use of a common background, the UNL Lexical Framework, which includes the UNL Dictionary, the UNL Knowledge Base and the UNL Memory, described in the next section.

3. UNL Lexical Framework

The main goal of the UNL is to be a digital link between human languages. In this sense, the UNL is actually a technology for connecting languages. At the lexical level, this cross-language indexation is carried out by the UNL Dictionary, which is planned to be a repository of all concepts that can be instantiated by any natural language. In its latest releases, this common vocabulary is understood, not as a set of semantic primitives supposedly shared by all languages, but as the list of any lexicalized concepts, regardless of their universality[2].

In the UNL Dictionary, concepts are represented by a specific type of uniform resource identifier, the Uniform Concept Identifier (UCI), which consists of two different parts: a Uniform Concept Locator (UCL), the 9-digit address of the corresponding node in the UNL Knowledge Base; and several possible Uniform Concept Names (UCN), the human-readable version of the UCL. For instance, the entry corresponding to the concept "a piece of furniture having a smooth flat top that is usually supported by one or more vertical legs" is represented at the UCL

http://unlkb.unlweb.net/104379964

and may be referred, in addition to the UCL itself, by several different UCN's, depending on the namespaces:

ucn:eng:table(icl>furniture)
ucn:fra:table(icl>mobilier)
ucn:deu:Tisch(icl>Möbel)
ucn:rus:стол(icl>мебель)
etc.

In the UNL Lexical Framework, the UCI plays the role of a pivot-symbol or index used to connect lexical items from different languages. The UCI table(icl>furniture), for instance, is the common link between "tafel" (Afrikaans), "طاولة" (Arabic), "սեղան" (Armenian), "masa" (Azerbaijani), "tabùru" (Baatonum), "টেবিল" (Bengali), "маса" (Bulgarian), "taula" (Catalan), "几" (Chinese) and "stol" (Croatian), only to mention some of the languages for which it has been already mapped.

As the UCI identifies a concept rather than a word, it is also used to connect synonyms from the same language. The same UCI table(icl>furniture), for instance, connects "टेबल" to "मेज़" in Hindi; "ಟೇಬಲ್" to "ಮೇಜು", in Kannada; and "ਟੇਬਲ" to "ਮੇਜ਼ ", in Panjabi.

In that sense, the UCI can be understood as a sort of ILI (interlingual index), as conceived by the EuroWordNet, i.e., as a "universal index of meaning" (Vossen, Peters and Gonzalo, 1998). Indeed, the UCI is also meant to be "the superset of all the concepts occurring in the different wordnets so that we can establish relations between minimal pairs of synsets". The main difference, however, is that the ILI is rather (or still) an "unstructured fund of concepts", whereas UCI's are, by definition, nodes in the UNL Knowledge Base (UNLKB). This means that UCI's are connected, not only to lexical items from natural languages, but to other UCI's, which are used to make them semantically accessible. The UCI table(icl>furniture), for instance, is currently represented, in the UNLKB, in Simplified UNL[3], as:

[2] The concept of "universality", in UNL, must be understood in the sense of "capable of being used and understood by all" (as in "universal adapter" or "universal remote control"). In that sense, the UNL Dictionary brings concepts that may range from absolutely global to absolutely local, provided that they are lexicalized, i.e., acknowledged as a "lexical unit", in at least one language. The only exception are proper names, which have been included in the UNL Dictionary only when accredited by local publishing authorities, such as encyclopedias, almanacs and books of facts.

[3] Sample of a UNLKB entry in Simplified UNL, in the format:
<relation>(<source>,<target>)=<degree of certainty>;
Where:
<relation> is one of the relations of the UNL (icl = is-a-kind-of, pof = is-a-part-of, aoj = is-an-attribute-of, pur = is-used-for, etc.);
<source> and <target> are UCI's (represented only by the root of the corresponding UCN's in Simplified UNL); and

icl(furniture, table)=255;
pof(table, top)=255;
aoj(top, rectangular)= 120;
aoj(top, round)=100;
aoj(top, semi-circular)=30;
pof(table, leg)=255;
aoj(table, rigid)=255;
pur(table, support)=255;[4]

Given that the UCL is actually the address of the entry in the UNLKB, it is not possible to create a UCI without linking it to other existing UCI's (i.e., without providing its definition in the UNL format). This process has been done manually, on demand, whenever a user notices that a given concept, necessary to map a natural language entry to UNL, has not been included yet in the UNL Dictionary. In order to define the exact location of the UCI, users usually UNLize the definition of the concept as presented by ordinary monolingual dictionaries.
In addition to the UNLKB, UCI's are further defined in the UNL Memory, which brings customary relations between UCI's, automatically extracted from annotated corpora. The same UCI table(icl>furniture), for instance, is currently associated, in the UNL Memory, to more than 1,500 other UCI's through place relations, among which we can find the following[5]:

plc(table, room)=63;
plc(table, dining room)=48;
plc(table, office)=37;
plc(table, meeting room)=27;
plc(table, kitchen)=12;
plc(table, living room)=9;

Differently from the UNLKB, which is common to all languages and brings stable relations between UCI's, the UNL Memory is rather corpus- and language-dependent, and therefore much more dynamic and fluctuating. For the time being, it brings mostly relations extracted from English data, which is one of the languages to have achieved the C2 level in the system, as informed below.

4. FoR-UNL

The FoR-UNL (Framework of Reference for UNL) is a guideline used to describe achievements of natural languages in relation to UNL. It was inspired by the Common European Framework of Reference for Languages (CEFR)[6], and its main goal is to provide a method for assessing the availability and quality of natural language resources inside the UNL System.

The FoR-UNL classifies languages in three broad divisions (A, B and C), which can be divided into six levels, according to the recall and precision of the corresponding resources:

*A - Basic Level
**A1 - Breakthrough or beginner
**A2 - Waystage or elementary
*B - Intermediate Level
**B1 - Threshold or intermediate
**B2 - Vantage or upper intermediate
*C - Advanced Level
**C1 - Effective Operational
**C2 – Mastery

In order to classify a language in one of the levels above, we use the following descriptors (always in relation to the UNL):

Level	**Dictionary[7] (base forms)**	**Grammar[8]**
A1	5,000	Morphology: NP
A2	10,000	Morphology: other POS
B1	20,000	Syntax: NP
B2	40,000	Syntax: VP
C1	70,000	Syntax: IP
C2	100,000	Syntax: CP

Table 1: FoR-UNL

As of February 2014, we have 34 languages with more than 5,000 base forms in the Dictionary, as indicated below:

Language	FoR-UNL	
	Dictionary	Grammar
Arabic	C2	B1
English	C2	B1
German	B2	A2
Spanish	B2	A2
Armenian	B1	A2
French	B1	B1
Latin	B1	A2
Portuguese	B1	A2
Russian	B1	A2
Chinese	A2	A2

<degree of certainty> may range from 0 (impossible) to 1-254 (typical) to 255 (necessary).
The Standard UNL represents the same information in XML format, with UCL instead of simplified UCN's.

[4] In the above, it is informed that table is a kind of furniture, that it is rigid, that is used for support, that it has a top and legs, and that its top can be rectangular, round or semi-circular.

[5] We present here, in Simplified UNL, the six most frequent place relations between "table(icl>furniture)" and other UCI's according to the processing of a segment of BNC. In the UNL Memory, the degree of certainty of the relations is normalised (between 1 and 254) by reference to the frequency of occurrence of the relations in the corpus.

[6] http://www.coe.int/t/dg4/linguistic/Cadre1_en.asp.

[7] Number of base forms included in the dictionary and linked to UCI's.

[8] Scope of the grammars in relation to the UNL process. For instance, in order to be classified at the A1 level, languages must have the rules to generate all the possible inflections out of the base forms for nouns, in case of inflectional languages.

Greek (Modern)	A2	A0
Japanese	A2	A0
Slovenian	A2	A2
Telugu	A2	A1
Ukrainian	A2	A2
Afrikaans	A1	A1
Baatonum	A1	A1
Bulgarian	A1	A2
Croatian	A1	A2
Estonian	A1	B1
Hindi	A1	A0
Hungarian	A1	A2
Indonesian	A1	A0
Italian	A1	A2
Kannada	A1	A0
Khmer	A1	A0
Malay	A1	A2
Nepali	A1	A0
Panjabi	A1	A1
Persian	A1	A2
Serbian	A1	A2
Tamil	A1	A0
Thai	A1	A1
Vietnamese	A1	A0

Table 2: Language status as of Feb 2014.

5. UNL[arium]

All the work within the UNL Lexical Framework is carried out in the UNL[arium], available at www.unlweb.net/unlarium. The UNL[arium] is a web-based database management system that allows registered users to create, to edit, to browse, to search, to import and to export dictionary, knowledge base, memory, corpora and grammar entries. Although originally conceived inside the UNL framework, the UNL[arium] does not require any deep knowledge on UNL, and its data may be used in several NLP systems, in addition to UNL-based applications. Furthermore, the system is supposed to be used as a research workplace for exchanging information and testing several linguistic constants that have been proposed for describing and predicting natural language phenomena. One of its main goals is to figure out and validate a language-independent metalanguage for language description that would be as comprehensive, as harmonized and as confluent as required by multilingual processing.

5.1 Projects

The work within the UNL[arium] is organized in many different projects leading to the development of the language resources required by the UNL System. The projects can be open or closed, and funded or non-funded, depending on the language and on the scope. Most projects involving the development of language resources follow the flow defined by the FoR-UNL, and range from A1 (most basic level) to C2 (most advanced level). The projects are grouped in different main types: dictionary, corpus or memory.

Dictionary projects aim at providing entries to UNL dictionaries. There are four subtypes of dictionary projects:

- UNL->NL (Generation) Dictionary projects aim at mapping UCI's into natural language lexical items;
- NL->UNL (Analysis) Dictionary projects aim at mapping natural language lexical items into UCI's;
- NL Dictionary projects aims at treating entries resulting from generation dictionary projects; and
- UNL Dictionary projects aim at analysing, defining and exemplifying UCI's.

Corpus projects aim at annotating corpora for machine learning, for assessing UNL-driven grammars and for extracting UNL Memory entries. There are two subtypes of corpus projects:

- UNL->NL (Generation) Corpus projects aim at converting UNL documents into a natural language; and
- NL->UNL (Analysis) Corpus projects aim at converting natural language documents into UNL.

At last, memory projects aim at providing further lexical resources for UNL-based systems. There are five types of memory projects:

- Knowledge Base projects aim at providing entries for the UNL Knowledge Base;
- UNL Memory projects aim at providing entries for the UNL Memory;
- NL Memory projects aim at providing entries for the NL Memory;
- NL->UNL (Analysis) Memory projects aim at mapping translation units into UNL; and
- UNL->NL (Generation) Memory projects aim at UNL segments into natural language expressions.

5.2 Users

As for February 2014, the UNL[arium] has around 1,100 registered users, working with 50 different languages. The environment is open and free to any participant, and targets language specialists rather than computer experts. The system does not require intensive knowledge of UNL or of Computational Linguistics. Nevertheless, it requires some acquaintance with linguistic terminology, with semantic and syntactic formalisms, and very good knowledge of the working language. For the time being, it also requires knowledge of English, which is the language of the interface and of all the documentation.

In order to join a project and start working within the environment, users have to be approved by VALERIE, the

Virtual Learning Environment for UNL, available at www.unlweb.net/valerie. VALERIE comprises several different certificates. Each certificate is divided into several different levels, which must be overtaken by candidates in order to be approved. Each level consists of a brief theoretical explanation and some exercises, which are evaluated automatically. Successful candidates are granted a permission to work within the UNL^{arium}. At first, they can only work with their native languages. Non-accredited users have access to several facilities of the system, but are not allowed to add entries or rules.
For the time being, there have been four different types of contributors in the environment:

- Volunteers, i.e., those who participate in the project voluntarily;
- Freelancers, i.e., accredited professionals who are paid for their work;
- Partners, i.e., members of affiliate institutions; and
- Employees of the UNDL Foundation.

Freelancers are remunerated according to their level of expertise and to the amount of entries accumulated in a given period of time. The level of expertise (from A0 to C2) is measured in terms of UNL^{dots}, a unit of time and complexity for calculating the effort spent in performing UNL-related tasks. Users have also different permission levels (observers, authors, editors, revisers, managers), which are defined according to several factors, including profile, expertise, institutional status and academic records.

5.3 Workflow

After joining a project, users may create assignments. The assignment is actually a reservation of entries to be treated, which is valid for 30 days. After the deadline, non-treated entries return to the database and become available to other users for reservation. Trainees (i.e., users at level A0) can only create assignments with up to 50 entries; after being promoted to the author level (>5,000 UNL^{dots}), this limit is extended to 250 entries. Assignments from trainees and authors are reviewed before being approved. In case reviewers detect any problem with the assignments, the account is blocked for new reservations until the problems are fixed.
The reservation process and the treatment of entries are done online through the UNL^{arium} interface. For each type of project, there is a special form to be filled in. In generation dictionaries, for instance, users are presented UCI's that they have to map onto their native language. They propose a lemma and some basic features (such as gender and number, for nouns in gender- and number-inflective languages). If the UCI cannot be represented by a lexical item in their native language, they flag the corresponding entry, and reduce its degree of generality or prevalence. The reverse process happens in analysis dictionaries, where users are supposed to map lexical items from their native language into UNL. If the concept is not registered yet, they create the corresponding UCI and define it in the UNLKB. In corpus projects, users are expected to map the whole documents from natural language into UNL or vice-versa, depending on the task.
Each entry created inside the UNL^{arium} is double-checked. This means that there are three different types of actions: to create entries, to verify entries and to review entries. Verification projects are open only to editors; revision projects are open only to reviewers. In this verification/revision process, users are evaluated and may be blocked or promoted depending on their performance.
In Figures 1 and 2 below, we present the screenshots for dictionaries and corpora projects (in the analysis direction).

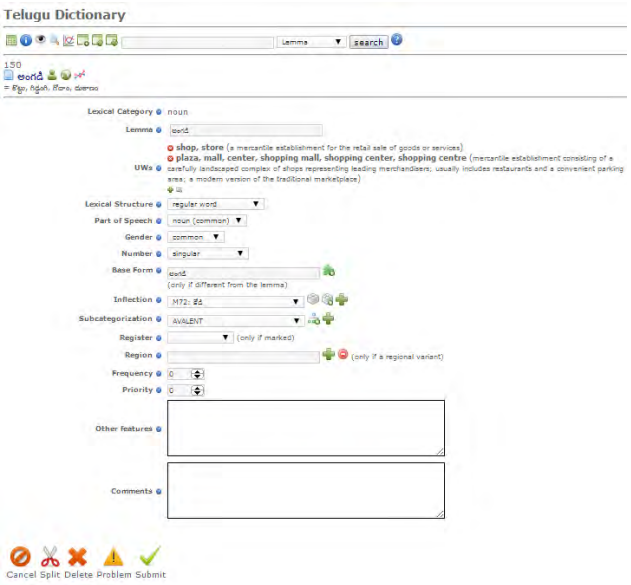

Figure 1: Dictionary Form

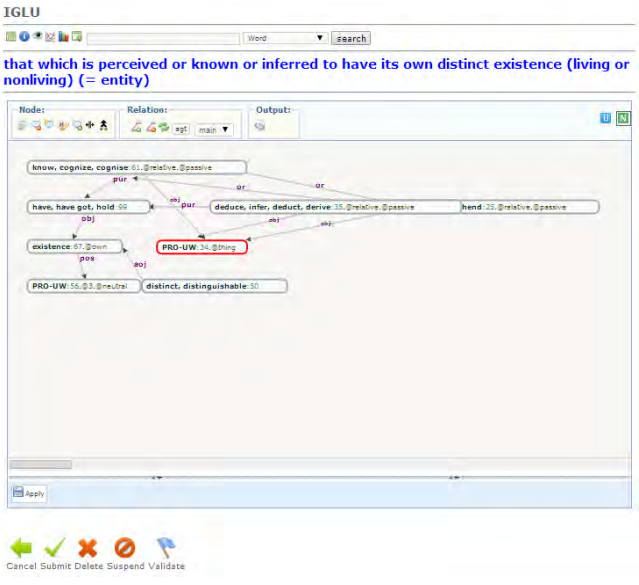

Figure 2: Corpus Form

In natural language dictionary projects, users are expected not only to link natural language entries to UCI's (or vice-versa), but also to analyse and describe the resulting entries. This includes defining the morphological and syntactic behaviour of natural language items, by assigning the proper categories (part of speech, gender, number, transitivity, etc.) and informing the corresponding inflectional paradigms and subcategorization frames, whenever necessary.
For instance, the Croatian grammar contains the following nominal paradigm[9]:

SNG&NOM:=0>"";
SNG&GNT:=1>"e";
SNG&DAT:=1>"i";
SNG&ACC:=1>"u";
SNG&VOC:=1>"o";
SNG&LOC:=1>"i";
SNG&INS:=1>"om";
PLR&NOM:=1>"e";
PLR&GNT:="a"<[-2];
PLR&DAT:=1>"ama";
PLR&ACC:=1>"e";
PLR&VOC:=1>"e";
PLR&LOC:=1>"ama";
PLR&INS:=1>"ama";

Applied to a base form such as "ženidba", this paradigm generates 14 different word forms:

BF=ženidba
SNG&NOM=ženidba
SNG&GNT=ženidbe
SNG&DAT=ženidbi
SNG&ACC=ženidbu
SNG&VOC=ženidbo
SNG&LOC=ženidbi
SNG&INS=ženidbom
PLR&NOM=ženidbe
PLR&GNT=ženidaba
PLR&DAT=ženidbama
PLR&ACC=ženidbe
PLR&VOC=ženidbe
PLR&LOC=ženidbama
PLR&INS=ženidbama

The UNL[arium] is therefore a matrix out of which we may export two different types of dictionaries: a generative dictionary, only with base forms and the corresponding features and inflectional rules, normally used in natural language generation; and an enumerative dictionary, with base forms and word forms, generated automatically by the application of morphological rules, such as the ones indicated above. A sample of each format of the English dictionary is presented below[10]:

Generative dictionary

[foot]{125873}"102153445"(LEX=N,POS=NOU,LST=WRD, NUM=SNG,PAR=M1,FRA=Y0,**FLX(PLR:="feet";)**,ABN=CCT,ALY=NALI,ANI=ANM,CAR=CTB,SEM=ANL,SFR=K0) <eng,5,0>;

Enumerative dictionary

[**foot**]{125873}"102153445"(LEX=N,POS=NOU,LST=WRD, **NUM=SNG**,PAR=M1,FRA=Y0,ABN=CCT,ALY=NALI,ANI=ANM,CAR=CTB,SEM=ANL,SFR=K0)<eng,5,0>;
[**feet**]{125873}"102153445"(LEX=N,POS=NOU,LST=WRD, **NUM=PLR**,PAR=M1,FRA=Y0,ABN=CCT,ALY=NALI,ANI=ANM,CAR=CTB,SEM=ANL,SFR=K0)<eng,5,0>;

In the examples above, all the features (SNG, NOM, GNT, DAT, LEX, POS, LST, etc.) are strongly standardised and harmonised to all languages, i.e., users are not allowed to use any linguistic constant that has not been defined and formalised in the UNDL Tagset, as described in the next section.

5.4 UNDL Tagset

The set of features in the UNL dictionaries depends on the structure of each natural language and may vary a lot. However, in order to better standardize lexical resources inside the UNL framework, we use a harmonised set of linguistic constants, the UNDL Tagset, in order to make the resources as easily understandable and exchangeable as possible. Several of those linguistic constants have been already proposed to the Data Category Registry (ISO 12620), and represent widely accepted linguistic categories. In general, we have tried to stick to the standard abbreviations proposed by the Leipzig Glossing Rules[11], by Cristal (2008) and by SIL International[12].
In most cases, the use of tags is rather unnoticeable and effortless, since users are supposed to make higher-level choices ("adjective", for instance) which will be internally represented through the corresponding authorized labels ("ADJ"). However, in several circumstances, as when creating inflectional paradigms or subcategorization frames, users are expected to address more fine-grained linguistic phenomena that may require a specialized metalanguage. In these cases, they are not authorized to

[9] Inflectional rules follow the general standard <CONDITION>:=<ACTION>;
where <CONDITION> is a set of features (such as SNG&NOM, i.e., singular and nominative); and <ACTION> describes the changes to be performed over the base form (prefixation, infixation, suffixation or circumfixation).
For instance, the rule SNG&GNT:=1>"e"; means that, in order to form the genitive singular, we have to delete the last character and add "e" to the right of it.
The grammar formalism adopted within the UNL[arium] is described at www.unlweb.net/wiki/grammar.

[10] The dictionaries are presented in the raw text format:
[entry]{ID}"UCL"(ATTRIBUTE=VALUE, …)<language,frequency,priority>;
[11] http://www.eva.mpg.de/lingua/resources/glossing-rules.php
[12] http://www-01.sil.org/linguistics/GlossaryOfLinguisticTerms/

create new tags, and must conform to the existing tagset. Nevertheless, the UNL[arium] includes the Tagset module, where users may propose new linguistic constants, which become available only after being approved by a technical committee.

5.5 License

As a result of a collaborative project, all the data stored in the UNL[arium] are available under an Attribution Share Alike (CC-BY-SA) Creative Commons license, which means that anyone may use the resources for any purpose, provided that authors are cited and any derivative work is released under the same or a similar license.

6. Further Work

For the time being, most of the entries introduced in the UNL[arium] have been provided manually, at a very high cost. In order to circumvent the inevitable financial issues associated to the incremental number of users and languages joining the environment, we have been trying to incorporate resources from other open datasets, especially the DBpedia[13]. Additionally, we have been working with the CADMOS[14] consortium in order to extract and to align multiword expressions from several different languages, without any prior lexicon, using comparable (non-parallel) corpora. These initiatives, to be completed in the short term, would release part of the manual work for verification and revision tasks, and would allow us to extend considerably the number of languages and entries already covered in the system.

However, officially open in February 2010, the UNL[arium] involves today more than 1,100 users, working with 50 languages, who have provided around 140,000 UCI's, 1,000,000 KB entries, 2,000,000 base forms and 12,000,000 word forms. As most of the data may be exported in several different formats, from XML to plain text files, it constitutes already an important and free asset for natural language processing.

7. Acknowledgements

The work with the UNL[arium] has been supported by the Arab Fund for Economic and Social Development and by Fondation Hans Wilsdorf.

8. References

Cardeñosa, J., Gelbukh, A. and Tovar, E. (Eds.) (2005). *Universal Networking Language*: Advances in Theory and Applications. Available at http://www.cicling.org/2005/UNL-book.

Crsytal, D. (2008). *A dictionary of Linguistics and Phonetics*. Malden, USA: Blackwell Publishing.

Dorr, B. J., Jordan P. W and Benoit, J. W. (1999). A Survey of Current Research in Machine Translation, in *Advances in Computers*, Vol 49, M. Zelkowitz (Ed),l London: Academic Press, pp. 1—68.

Francopoulo , G. (ed.) (2013) LMF Lexical Markup Framework. ISTE / Wiley.

Helbig, H. (2006). *Knowledge Representation and the Semantics of Natural Language*, Berlin: Springer.

Martins, R. (ed). (2013). *Lexical issues of UNL*. Cambridge Scholar Publishing, Newcastle upon Tyne.

Miller, G. A.; Beckwith, R.; Fellbaum C. D; Gross, D.; Miller, K. (1990). WordNet: An online lexical database. Int. J. *Lexicograp*h. 3, 4, pp. 235–244.

Sowa, J. F. (1984). *Conceptual Structures: Information Processing in Mind and Machine*. Reading, MA: Addison-Wesley.

Uchida, H.; Zhu, M. and Della Senta, T. (1999). *A gift for a millenium*. Tokyo: IAS/UNU.

Vossen, P., Peters, W., and Gonzalo, J. (1999). Towards a Universal Index of Meaning. In *Proceedings of ACL99 Workshop Siglex'99 - Standardizing Lexical Resources*. Maryland, USA: University of Maryland, pp. 81-90.

[13] http://dbpedia.org.
[14] http://www.cadmos.org

Collaborative Language Documentation: the Construction of the Huastec Corpus

Anuschka van ´t Hooft, José Luis González Compeán
Autonomous University of San Luis Potosí
Av. Industrias 101-A, Col. Talleres, 78399 San Luis Potosí, SLP, México
Technological Institute Cd. Valles
Av. Carr. al Ingenio Plan de Ayala km.2, Col. Vista Hermosa, 79010 Cd. Valles, SLP, México
E-mail: avanthooft@uaslp.mx, joseluig@yahoo.com

Abstract

In this paper, we describe the design and functioning of a web-based platform called Nenek, which aims to be an on-going language documentation project for the Huastec language. In Nenek, speakers, linguistic associations, government instances and researchers work together to construct a centralized repository of materials about the Huastec language. Nenek not only organizes different types of contents in repositories, it also uses this information to create online tools such as a searchable database with documents on Huastec language and culture, E-dictionaries and spell checkers. Nenek is also a monolingual social network in which users discuss contents on the platform. Until now, the speakers have created a monolingual E-dictionary and we have initiated an on-going process of the construction of a repository of written texts in the Huastec language. In this context, we have been able to localize and digitally archive documents in other formats (audios, videos, images), yet the retrieval, creation, storage, and documentation of this type of materials is still in a preliminary phase. In this presentation, we want to present the general methodology of the project.

Keywords: collaborative research, language documentation, online repositories.

1. Introduction

The Huastec language is a Mayan language spoken in the Mexican Gulf Coast region, in an area known as The Huasteca. This language has at least 215.500 speakers (INEGI 2010) and is the only Mayan language isolated geographically from the others, which are spoken in the southeastern part of Mexico, in Belize and Guatemala. Huastec language can be roughly divided into a western, eastern and southeastern variant that are present in the states of San Luis Potosí and Veracruz, respectively.

Until now, the generation of Huastec dictionaries and other written materials has resulted in a somewhat slow, disjointed, and static maintenance process for the Huastec language. Also, these sources are completely dispersed and contain local publications with a very limited distribution as well as a few written texts that are available on the internet. Access to both sources is rather difficult. In the Huastec case, the collecting of materials and the construction of dictionaries has almost exclusively depended on individual researchers, who usually perform this task through long-term fieldwork periods. This kind of methodology produces repositories that are mainly based on transcriptions, which commonly have static formats such as compact disks, tapes, books and articles. These repositories are rarely used by the speakers because they are either private or shared with other researchers only.[1]

We developed a collaborative strategy to construct the Huastec corpus through the use of the web, which may store an unlimited source of linguistic data including massive amounts of complete electronic texts that are usually in the public domain (Sinclair 2002). The key factors in the success of this strategy are the constant generation of contents (especially written texts and posts) and the availability of those contents online. However, at this point, the construction of the Huastec corpus through the retrieval of sources on the internet alone cannot be successful: there are still not enough online Huastec materials available, and the variability of their formats and contents do not favor the building of a solid linguistic repository.

This is why we constructed a web-based platform with which to develop a collaborative language documentation project and create, archive and analyze "a comprehensive record of the linguistic practices characteristic of a given speech community" (Himmelmann 1998:166). The platform is called Nenek (www.nenek.mx), which is a colloquial form of greeting in Huastec. Nenek combines digital archiving with language description tasks carried out by native speakers, linguistic associations, government instances and researchers. The way in which we promote the project is through an online monolingual social network in which speakers exchange ideas about their language and culture. At present, more than 1,800 Huastec speakers are actively involved in the project. Their internet activity generates materials in the Huastec language and enables us to retrieve and document different types of sources. At the same time, Nenek aspires to improve the weak situation and position of this language, and aims to strengthen its maintenance and revitalization process. We hope it will also be helpful for students and researchers who want to study themes related to the fields of linguistics or linguistic anthropology on Huastec, in particular to specialists in the natural language processing (NLP) of this lesser-resourced Amerindian language.

[1] We could discuss additional problems that arise when compiling the available Huastec sources, all of which are in tune with the ones described by Bird and Simons (2003) concerning language documentation and description projects.

In this paper we want to describe the design and functioning of the Nenek platform. In particular, we present the collaborative methodology of the project through which speakers, together with the Nenek staff, build different types of repositories.

2. The Nenek Platform

The Nenek platform creates virtual communities of indigenous languages that provide the speakers with a monolingual social network online. The social network includes functionalities such as profiles, work groups and contents management, as well as tools that allow the speakers to create web pages and blogs in which they can contribute to the repository building by sharing and discussing texts, audios, images and videos.

Each registered participant in Nenek has a personal account, with both a private and a public window to access the virtual community. In this private window, the speakers can store materials such as written texts, images, audios and videos. The private window includes both the monolingual social network and a set of linguistic collaborative applications called *Nenek-joined*. We created these specialized computer tools in order to encourage the virtual community to use the language to a major extent, starting with a lexicography tool for the making of E-dictionaries[2] and then constructed a spell checker[3] for the Huastec language. The *Nenek-joined* tools are also used by work groups that are in charge of specific tasks in the language documentation project such as the construction of E-dictionaries, spell checker validation and content evaluation.

Nenek´s public window is for all speakers and those interested in Huastec language and culture. Here one can find published repositories, a dynamic monolingual searchable dictionary, a spell checker and some other materials about Huastec language and culture. This means that the results of the tasks developed by both the work groups and Nenek staff are published here and are freely available. The public window of the virtual community is also interconnected with traditional social networks, such as Facebook, YouTube and Twitter for collecting sentences or small written texts from the speakers (NenekFacebook, 2014).

3. The Collaborative Methodology of Nenek

The language documentation activities are developed in the private window of each user. In this private window, the speakers can store materials such as written texts, images, audios and videos. Here, the user decides whether his or her materials can be consulted publicly, are private or may go into the repository.

Speakers who are interested in participating in Nenek, may choose between two different roles[4]:

- Nenek-User, who is a registered user who has access to the monolingual social network with his or her private account. These persons are mostly students, young workers or teachers who live in the Huasteca region, but a significant segment of participants are migrants who live in different cities in Mexico or the USA. Most of them are between 15 and 40 years old (78% of this age group still speaks the language). They are receptive to the written expression of their language and have internet (HD, 2013);
- Collaborator-User, who is a registered users who participates in a specific language documentation task and has access to both the social network, the private account and the linguistic tools. These users are commonly local linguists, academics and researchers. Like Nenek-users, these participants are registered in the monolingual social network, yet they also have access to Nenek-joined (that is, to the linguistic tools) in order to validate the materials deposited by Nenek-users and other Collaborator-users.

When generating materials (written texts, audio recordings, videos, photos, vocabulary entries), both Nenek-Users and Collaborator-Users decide among three options where to store these items:

- *Japidh:* This Huastec adjective (which means “open” or “disclosed”) represents the public content category. When a speaker introduces a content in the virtual community by choosing this category, the platform stores this content in the Nenek-social repository and it sends an e-mail alert to all participants who have accepted to receive it. This content is now open for viewing, but it is not included in the heritage repository.
- *Mapudh:* This Huastec adjective (which means “closed” or “enclosed”) defines the private content category. When a speaker introduces contents in the virtual community by choosing this category, the platform does not send alerts to the community.

[2] The E-Lexicography tool builds dictionaries attending the demands expressed in the literature about Internet dictionaries (Almind, 2005), since our pilot dictionary is easy to find, the search field is the center of attention, and it gives instant and simple results, which is limited to nine entries per page. Also, it has an autocomplete search function that predicts a word or phrase when the user is trying to type in and it gives alternatives and displays results. Nenek allows several workgroups to develop different dictionaries at the same time.

[3] Huastec speakers are not necessarily familiar with writing in their language. Moreover, there is no standardized alphabet or standardized spelling available for Huastec. In order to provide the speakers with a reference framework to write texts in Huastec, we developed CoTenek. This checker detects new spelling forms and gives multiple writing options for each term, so a speaker can choose whether he or she wants to use one of the options given or not. CoTenek is available for some of the most popular text editors, such as Microsoft Word, OpenOffice, LibreOffice (CoTenek 2014), and a Firefox version has been developed by Kevin Scannell from Saint Louis University by using CoTenek lexicon (CoTenekFirefox 2014).

[4] People who are only interested in consulting Nenek´s public window are called Public-Users. They are not registered and do not participate in any of Nenek´s activities. They can only consult and retrieve the information that is publicly available on the platform.

• *Wejladh K'anilab:* This category (which refers to something "chosen and stored") indicates that a speaker who is participating in a specific task donates content to the community heritage. The speaker offers his or her material to the repository, where it is stored. This time, the platform sends an e-mail notification to all the Collaborator-Users and automatically starts a consensus polling procedure to decide whether that content is valid for repository or not. The results of this evaluation process are reported to Nenek-Users who receive alerts, and thus start a second consensus polling procedure among the Nenek-Users. The basic idea is to emulate the meetings and the members' participation in the decision-making process of real communities. Only when accepted by the community the contents go into the heritage repository, where it is publicly available for all the speakers.

It should be said that while storing the materials, the user has to provide the metadata of each item, so that NLP researchers could make use of it.

Thus, in Nenek, the documentation activities are carried out collaboratively and in a cyclic process that starts when the speakers propose a task for a work group and store their materials in the *wejladh k´anilab* category, the heritage repository. Then, either the speakers' communication or input of materials returns to the virtual community after a categorization and consensus polling procedure (validation process) carried out by speakers, linguists or native linguistic associations who are Collaborator-Users.

Until now, the workgroups have been working on the construction of an E-dictionary of the Huastec language, which includes almost 2,000 entries and is the first dynamic Huastec dictionary online that is constructed in a collaborative manner by the speakers. During the working process, Nenek staff often leads the discussion and poses questions to the virtual Huastec community, for example by sending an image and asking about the forms to describe the item on the image. It then collects as much as ten different proposals, all of which are debated among the members of the work group. Moreover, the community also describes the specific region in which each of the expressions is used. All participants deliver their opinions to our web page by reacting to our question in a public manner, which represents a situation similar to when a researcher obtains terms during fieldwork.

Also, we have initiated the work on the construction of repositories of written texts in the Huastec language. The retrieved materials were obtained from three different sources: on the internet (based on crawler software designed for collecting Huastec texts), through donations of published and unpublished materials by their authors, and through the retrieval of written texts from the Nenek social network. Here too, work groups are created that stimulate speakers to hand in specific types of materials, such as essays, tales, anecdotes, or other written reports.

Nenek handles the contents as a digital library that includes dictionaries and repositories that are automatically categorized according to the type of the role used by the speaker who donates contents. Thus, the materials are stored into three different repositories:

• Nenek-social: This repository includes written texts donated by speakers (Nenek-users) through public blogs of the monolingual social network (NenekBlogs 2014).

• Nenek-academic: This repository includes documents and written texts donated by speakers who are Collaborator-Users. This repository also includes texts written in Huastec that were collected from a special edition of an academic journal coordinated for this project (JournalTenek 2013).

• Nenek-published: Nenek-published is the heritage repository of the virtual community. This repository includes published materials from two different sources. The first source includes books that are automatically recovered from public sites on the internet by using crawler software. The second source includes books that were donated to the project by both government instances and indigenous associations. The last source includes a set of publications that were digitized by the Huastec speakers of the Nenek staff.

As the written texts of the Nenek-published repository have passed through a full reviewing and editorial process, Nenek automatically uses them as valid for the corpus building. The materials from the other two repositories (Nenek-social and Nenek-academic), however, require verification and consensus from the virtual community before considering them for the corpus. It is important to note that all the repositories handled by Nenek are stored in a fault-tolerant cloud including sites in Spain and Mexico to guarantee the contents availability in failure scenarios (González et al. 2012; González et al. 2013).

As a result, and even though there are online sources that offer materials in Huastec (AILLA 2014; OLAC 2014; SOAS 2014; CAILLA 2014), Nenek-published is currently the largest repository of written texts in this language on the internet. It contains materials that belong to the fields of law, education and local oral traditions. Thus, our collaborative approach allowed us to cover a wide social profile for language usage.

It should also be mentioned that these other repositories are based on a depositor scheme in which the volume of contents depends on the activity of few researchers (sources). In addition, the most of the sources of these repositories have defined to deny the access to the contents. Contrastingly, the number of different sources in Nenek is significantly higher because the speakers, associations and government instances are collaborating in the content collection process. Besides, all of its contents is freely accessible online.

4. Conclusions

Before Nenek, Huastec speakers preferred Spanish as their language of communication on the internet because there were no cybernetic spaces where to use their

mother tongue. Now, we have young people joining the project and using the platform to search for friends and discuss various issues in their mother tongue. These speakers are participating in linguistic tasks or debates about Huastec sentences and are gradually creating their own initiatives and debates about their language. This means that they are writing their language –some of them for the first time- and that they do so in a new media, the internet. Consequently, Nenek has been able to expand the use of the language to spheres in which this language was not present before and has contributed to some extent to its revitalization.

The first stage of the language documentation process conducted through collaborative research allowed us to create a searchable pool of information that reflects part of the living language. Nenek concentrated this heritage in a centralized site that can reconstruct contents in failure scenarios. Since this is a collaborative platform, we did not only achieve to store the greatest quantity of materials, but also the most varied ones (at least in regard to written texts in Huastec). Nenek has proved to be an important tool in the language documentation process of the Huastec language.

In the following stage of the project we want to focus on the documentation of other materials, such as audio or video recordings, and images. These materials will give a better view on the living language in its social and cultural context (Gippert, Himmelmann & Mosel, 2006). We are constantly making improvements on the platform (for example, creating mobile applications) in order to make the collaborative work more profitable. Thus, we think, Nenek fosters the empowerment of native peoples in taking care of their linguistic and cultural heritage, making it a project for and by native speakers.

Nenek´s more inclusive process of repository building concentrates efforts and improves the collection results. The collaboration process of a growing social collective as well as the use of the crawler collector software appear to be more time effective than the deposits scheme used by traditional repositories. We believe that researchers who are interested in generating materials for other languages through collaborative approaches may take advantage of the described strategy.

5. Acknowledgements

The Nenek project is sponsored through a grant from the Mexican Secretary of Public Education and the National Council of Science and Technology (SEP-Conacyt research grant CB-2012-180863).

6. References

AILLA (2014). AILLA: The archive of the indigenous languages of Latin America, http://www.ailla.utexas.org/site/welcome.html

Almind, R. (2005). Designing Internet Dictionaries In I. Barz, H. Bergenholtz & J. Korhonen (eds). *Schreiben, Verstehen, Übersetzen, Lernen. Zu ein- und zweisprachigen Wörterbüchern mit Deutsch*. Frankfurt am Main: Peter Lang, pp.103-119.

Bird, S. & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language* 79 (3), pp. 557-582.

CAILLA (2014). Chicago Archive of Indigenous Literatures of Latin America. http://cailla.uchicago.edu/?page=browse_by_language;family=4;language=48

CoTenek (2014). Co-Tenek, Huastec Spell checker, http://www.nenek.mx/huasteco.dic

CoTenekFirefox (2014). https://addons.mozilla.org/addon/huastec-spell-checker/

Gippert, J., Himmelmann, N.P. & Mosel, U. (2006) *Essentials of Language Documentation*. Berlin, New York: Mouton de Gruyter.

González, J.L. et al. (2013). González Compeán, J.L., Carretero Pérez, J. Sosa-Sosa, V. Rodríguez Cardoso, J.F. & Marcelín-Jiménez, R. An approach for constructing private storage services as a unified fault-tolerant system. *Journal of Systems and Software* 86 (7), pp. 1907–1922.

González et al. (2012). González Compeán, J.L., Sosa-Sosa, V., Bergua Guerra, B, Sánchez, L.M. & Carretero Pérez, J. Fault-Tolerant Middleware Based on Multistream Pipeline for Private Storage Services. In *Proceedings of the International conference for internet technology and secured transactions*. London, 10-12 Dec. 2012, pp. 548 – 555.

HD (2013). *Habilidades Digitales para todos*. http://www.hdt.gob.mx/hdt/

Himmelmann, N.P. (1998). Documentary and descriptive linguistics, *Linguistics*, 36, pp. 161-195.

INEGI (2010). *Censo de Población y Vivienda 2010*. Instituto Nacional de Estadística, Geografía e Informática. Aguascalientes (Mexico): INEGI. http://www.inegi.org.mx/est/contenidos/proyectos/ccpv/cpv2010/

JournalTenek (2013). Special Edition for Huastec Speakers, Teczapic ITV Journal. http://www.nenek.mx/esTeczapicITV/index

NenekBlogs (2014). Nenek Speakers In http://www.nenek.mx/prof.php.

NenekFacebook (2014). Nenek in Facebook. https://www.facebook.com/NenekMexico .

OLAC (2014). OLAC resources in and about the Huastec language. http://www.language-archives.org/language/hus

Sinclair, J. (2002). Intuition and annotation - the discussion continues. In K. Aijmer & B. Altenberg (eds.) *Advances in corpus linguistics*. Amsterdam: Rodopi, pp. 39-59.

SOAS (2014). The Endangered Languages Archive at SOAS, London. http://elar.soas.ac.uk/

Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages

Sjur Moshagen⋆, Jack Rueter†, Tommi Pirinen‡, Trond Trosterud⋆, Francis M. Tyers⋆
⋆UiT Norgga árktalaš universitehta, †Helsinki university, ‡Dublin City University
⋆N-9037 Tromsø, †FIN-00014 Helsingin yliopisto, ‡IE-Dublin 9
sjur.n.moshagen@uit.no, jack.rueter@helsinki.fi,
tommi.pirinen@computing.dcu.ie, trond.trosterud@uit.no, francis.tyers@uit.no

Abstract

In order to support crowd sourcing for a language, certain social and technical prerequisites must be met. Both the size of the community and the level of technical support available are important factors. Many language communities are too small to be able to support a crowd-sourcing approach to building language-technology resources, while others have a large enough community but require a platform that relieves the need to develop all the technical and computational-linguistic know how needed to actually run a project successfully. This article covers the languages being worked on in the Giellatekno/Divvun and Apertium infrastructures. Giellatekno is a language-technology research group, Divvun is a product development group and both work primarily on the Sámi languages. Apertium is a free/open-source project primarily working on machine translation. We use Wikipedia as an indicator to divide the set of languages that we work on into two groups: those that can support traditional crowdsourcing, and those that do not. We find that the languages being worked on in the Giellatekno/Divvun infrastructure largely fall into the latter group, while the languages in the Apertium infrastructure fall mostly into the former group. Regardless of the ability of a language community to support traditional crowdsourcing, there is in all cases the necessity to provide a technical infrastructure to back up any linguistic work. We present two infrastructures, the Giellatekno/Divvun infrastructure and the Apertium infrastructure and show that while both groups of language communities would not be able to develop language technology on their own, using the infrastructures that we present they have been quite successful.

Keywords: crowdsourcing, infrastructure, minority languages

1. Introduction

Crowdsourcing (Howe, 2008; Surowiecki, 2005) is often thought of as being the leveraging of a group (or crowd) of non-experts to perform tasks previously only done by experts. This is exemplified by the Amazon *Mechanical Turk* platform.[1] Researchers assign tasks and pay small amounts for each task completed. When working with small language communities (often in the hundreds of people), there is not a sufficient mass of native speakers to be able to harness the power of the crowd in this way.

In this article we describe another approach to crowdsourcing. By our definition, a crowd is a group of people who are united by an interest in the development of language technology for a variety of ends.

This collaborative work is made possible by well defined and technically supported infrastructures. An infrastructure consists of the following components: a pre-established way of laying out linguistic data in files and directories, conventions for encoding the data, pre-defined tools for working with the data and building products, and documentation for working with the tools. It should also facilitate testing of both data and tools.

1.1. Language community size and morphological complexity

Language technology's equivalent of the elephant in the room is *the word*. Many language technology applications reduces this concept to a list, possibly a list of pairs (*walk, walk:walks, mouse, mouse:mice, ...*). For morphology-rich languages, like for example the circumpolar ones, this approach is a showstopper. In these languages, the word forms are, for practical and partly even theoretical purposes, not listable.

This is even more true considering the language community sizes of the languages described in the article. Whereas it is fully imaginable to get a small fraction of the English speaking world to list all word forms of the English language via a Mechanical Turk type of project, convincing 500 speakers of a morphologically-complex language to do the same for a theoretically and practically much larger list of word forms is impossible. That is, any approach targeting these languages must thus provide an analysis of the words.

1.2. Outline of the article

The remainder of this article is laid out in six sections: The first section discusses the limitations of crowdsourcing especially with respect to community size. The following section looks at the viability of crowdsourcing for a set of languages. The next section describes the two infrastructures, and this is followed by a section describing the crowds who are using these infrastructures. We then describe the end-user tools that are produced within our infrastructures. Finally, we draw some conclusions.

2. Language community size and crowdsourcing

Most of the world's minority languages, and in postcolonial societies even many of the majority ones, receive little or no official support. The exceptions to this generalisation are

[1] https://www.mturk.com/mturk/

typically minorities in Western societies. One example of a minority language for which the majority society practices a positive language policy, is North Sámi. North Sámi has a written tradition dating 250 years back, with the present standard in use only since 1979. Sámi language society consists of approximately 22,000 speakers, it is technologically advanced, literate, well off, online, and eager to see their language in use. Pupils in the core Sámi areas have their whole primary and secondary education with Sámi as the language of instruction, pupils outside these areas typically have Sámi lessons in Sámi, but a large part or even the rest of their education in the majority language. Except for Facebook localisation and an early localisation of the Linux KDE environment, there has so far not been any crowdsourcing projects related to language.

North Sámi is hardly a typical representative of a language of its size. Drawing instead a random equally-sized language from Ethnologue may e.g. give us *Dabarre*, a Cushitic language related to, but not mutually intelligible with Somali. Dabarre is a language without a literary language, and with no online resources. Its speakers are probably not connected to the internet. Dabarre is classified by Ethnologue as VIGOROUS.

3. Investigating crowd-sourcing viability

This section presents the Giellatekno/Divvun and Apertium languages, and compares them with respect to what might be called their crowd-sourcing viability. As a yardstick for such a viability, we use the size of the Wikipedia version for each and every language, and their status according to (Kloss, 1967) concept of *Ausbau* and *Abstand* languages (the former sharing a (recent) origin with the majority language, the latter not).

Wikipedia is the archetypal crowd-sourcing project. Using only open-source software and a web browser, more than 30 million articles have been written in close to 300 languages[2] — all of it by volunteers. The size of a Wikipedia for a given language should thus be a good indicator for whether the language community has the resources and interest to support projects through crowd-sourcing. It is also reasonable to assume that all other projects will have lesser visibility and be lesser known, and thus have a harder time than Wikipedia creating a crowd for their projects. It seems reasonable to assume that if there is no Wikipedia for a language, then it will be very hard to build a crowd for creating important natural-language processing tools.

3.1. Giellatekno/Divvun

The languages being actively developed within the Giellatekno-Divvun (*GTD*) infrastructure are listed in Table 1, together with the Kloss classification (b = Abstand, u = Ausbau, m = Majority), the number of Wikipedia articles, speakers ((Lewis et al., 2013), for the two Mari languages, Moksha and Erzya: (Moseley, 2010)) and articles per speaker for each of them.

Only four languages with a population below 50,000 have any Wikipedia at all. For all four it is true that most of the content has been written by non-native speakers. For the Giellatekno/Divvun languages with a bigger population, none of the Wikipedias has more than 10,000 articles[3]. Looking at the three largest Wikipedias in Table 1, we find the following: Faroese is an Ausbau language with a long literary tradition, an autonomous position and a majority position in its own area. The overwhelming majority of the North Sámi Wikipedia is written by non-native speakers[4]. For Hill Mari, the dominating article genre is articles

Language	Cl.	No. of speakers	WP articles	articles / speaker
Cornish	b	-	2 634	-
Liv	b	15	0	0.00
Pite Sámi	b	20	0	0.00
Northern Haida	b	45	0	0.00
Ingrian	b	120	0	0.00
Nganasan	b	130	0	0.00
Plains Cree	b	160	194	1.21
Inari Sámi	b	300	0	0.00
Skolt Sámi	b	300	0	0.00
Kildin Sámi	b	350	0	0.00
South Sámi	b	600	0	0.00
Lule Sámi	b	2 000	0	0.00
Upper Necaxa Totonac	b	3 400	0	0.00
Veps	b	3 610	0	0.00
Chippewa	b	5 000	0	0.00
Kven Finnish	b	5 000	0	0.00
Inupiaq	b	5 580	168	0.03
Khanty	b	9 580	0	0.00
Chipewyan	b	11 900	0	0.00
North Sámi	b	20 700	7 650	0.37
Nenets	b	21 900	0	0.00
Livvi	b	25 600	0	0.00
Hill Mari	b	36 822	5 119	0.01
Greenlandic	m	50 000	1 602	0.03
Võro	u	60 000	5 141	0.09
Faroese	m	66 000	7 951	0.12
Komi-Zyrian	b	156 000	4 141	0.03
Moksha	b	200 000	1 180	0.00
Buriat (Russia)	b	219 000	907	0.00
Udmurt	b	324 000	3 387	0.01
Erzya	b	400 000	1 636	0.00
Meadow Mari	b	414 211	3 932	0.01

Table 1: Table of the languages under active development supported by the Giellatekno-Divvun infrastructure, and the number of Wikipedia articles and speakers for each of them.

[2] http://meta.wikimedia.org/wiki/List_of_Wikipedias

[3] This is the Wikimedia threshold for getting into the page of number of speakers per article, cf. http://meta.wikimedia.org/wiki/List_of_Wikipedias_by_speakers_per_article

[4] None of the 18 most active writers have North Sámi as their mother tongue, cf. http://stats.wikimedia.org/NN/TablesWikipediaSE.htm

on geographical administrative units[5]. Except for Faroese, the most viable of the Wikipedias in Table 1 thus seem to be Võru, Komi-Zyrian, Meadow Mari and Udmurt, these are also language communities with active language movements. But also these language communities have not been able to make a working-size Wikipedia (cf. footnote 3).
That is, for the core languages of our work, and using Wikipedia as an indicator, it seems to be hard to find a crowd to give substantial input for constructing language-technology resources.

3.2. Apertium

Apertium (Forcada et al., 2011) is a free/open-source machine translation project. Its origin on the Iberian Peninsula is clearly reflected in the language coverage, but apart from that, Apertium is community-driven, and the choice of languages is dependent upon whether there are people willing to put in an effort in order to get them off the ground. It currently has 38 released language pairs, and many more in progress.
In the past, Apertium language pairs have been fully funded — by either governments or companies; partially funded — that is some work done with funding and the remainder voluntary; or totally voluntary.
An example of the latter would be the Spanish–Aragonese language pair. Work on the pair was started by Apertium-developer Jim O'Regan, at the request of Aragonese-speaker Juan Pablo Martínez. After three weeks of initial effort, spread over the course of a year, a final week of concentrated effort lead to the release of the first prototype version, translating from Aragonese to Spanish only. The first bidirectional version was completed after another 6 weeks of work by Juan Pablo, spread over the course of another year. The only available resource at the beginning of this work for Aragonese was the Aragonese edition of Wikipedia and a handful of verb templates on the English edition of Wiktionary. The Aragonese–Spanish dictionary was created by hand, but the Spanish morphological analyser/generator and part-of-speech tagger were taken from the Spanish–Catalan pair. No funding was received from any source towards the creation of the system. However, the main developer did receive a substantial amount of assistance from the Apertium "crowd", and was able to, thanks to the free/open-source nature of Apertium, reuse a non-insignificant amount of previous work on the Spanish side.
Language pairs are often started by an interested speaker of an under-resourced language (such as the case of Aragonese), or by an interested linguist with help from native speakers (as the case of Breton).
It is often the case that crowds overlap. For example, the developers of the resources for Aragonese and Breton are also active in Wikipedia. Given the size of the Wikipedias, it should in principle be possible to find people to work as a crowd on language technology. The Apertium languages can be found in Table 2.

[5]Tests using the "random article" function gave 70% for this type of articles. The article on Marmara Ereğlisi, a town in the Tekirdağ Province in the Marmara region of European Turkey, may serve as a representative example.

Language	Cl.	No. of speakers	WP articles	articles / speaker
Manx	b	-	4 700	-
Aragonese	u	10 000	29 707	2.97
Corsican	u	31 000	6 665	0.22
Scots Gaelic	b	63 130	11 940	0.19
Faroese	m	66 150	7 992	0.12
Nogai	b	87 410	0	0.00
Irish	b	106 210	29 095	0.27
Asturian	u	110 000	19 462	0.18
Breton	b	225 000	47 759	0.21
Icelandic	m	243 840	37 020	0.15
Karakalpak	b	424 000	632	0.00
Kumyk	b	426 550	0	0.00
Maltese	m	429 000	3 045	0.01
Tetum	b	463 500	800	0.00
Welsh	b	536 890	53 627	0.10
Basque	b	657 872	165 988	0.25
Avar	b	761 960	1 124	0.00
Chuvash	b	1 077 420	23 441	0.02
Sardinian	u	1 200 000	3 250	0.00
Bashkir	b	1 221 340	31 714	0.03
Latvian	m	1 272 650	52 746	0.04
Macedonian	m	1 710 670	75 690	0.04
Slovenian	m	1 906 630	139 630	0.07
Occitan	u	2 048 310	86 470	0.04
Mongolian	m	2 373 260	12 001	0.01
Kyrgyz	m	2 941 930	27 093	0.01
Lithuanian	m	3 130 970	163 336	0.05
Galician	u	3 185 000	110 443	0.03
Gilaki	b	3 270 000	6 008	0.00
Afrikaans	u	4 949 410	30 423	0.01
Tatar	b	5 407 550	56 856	0.01
Armenian	m	5 924 320	109 758	0.02
Albanian	m	7 436 990	50 674	0.01
Turkmen	m	7 560 560	4 975	0.00
Belarusian	m	7 818 960	69 359	0.01
Kazakh	m	8 077 770	205 153	0.03
Uzbek	m	21 930 230	127 385	0.01
Indonesian	m	23 200 480	333 536	0.01
Azerbaijani	m	24 237 550	98 359	0.00
Ukrainian	m	36 028 490	485 563	0.01
Bengali	m	193 263 700	28 256	0.00
Arabic	m	223 010 130	260 602	0.00

Table 2: Table of languages under active development supported by the Apertium infrastructure, and the number of Wikipedia articles, speakers and articles per speaker.

3.3. Summing up the crowdsourcing potential of the different languages

As can be seen in Table 1 and Table 2, languages with small or non-existing Wikipedias are either small, or they are Abstand languages. The only instances of Abstand languages among the active Wikipedias in our material are Basque, Tatar, Welsh, Breton and Chuvash, these are all quite large languages. For language communities smaller than hundred thousand speakers, especially for Abstand languages, the normal crowdsourcing effect is unlikely to work.
Whereas Giellatekno-Divvun only has a handful of languages with more than 100k speakers, Apertium has only

a handful of languages with *less than* 100k speakers, and a majority of the Giellatekno-Divvun languages have less than 10k speakers.

4. Infrastructure descriptions

Apertium and Giellatekno-Divvun share a couple of core values: both infrastructures assume a grammar-based approach to language technology to be the primary approach, both rely heavily on the principles of free/open source code, and both focus on non-central languages in the sense of (Streiter et al., 2006). This same paper gives an excellent overview of how to set up a working infrastructure for such languages, and the infrastructures described in this article fit quite nicely with their definition of a «language pool».

In the current Giellatekno/Divvun infrastructure there are about 50 languages. For all of them we can automatically produce the same set of tools, ready to be deployed. The quality of these tools will of course vary with the degree of linguistic development, but from a technical point of view, all languages are equally well supported. In the Apertium infrastructure, the situation is slightly more complicated. Many languages are supported only as part of machine translation pairs. Taking into account these pairs, there are approximately 76 languages supported to some degree. Of these 76 languages, 44 are available as monolingual packages which provide at minimum a morphological analyser for the language, and in the most developed case, also provide a constraint grammar or statistical part-of-speech tagger and an installable spell checker.

The implementation of both the Giellatekno/Divvun and the Apertium infrastructure is quite simple, using a centralised version control system (Subversion[6]) to track changes and handle cooperation and interaction on the file level. To configure and create build files for each language, GNU Autotools[7] are used.

Both offer ready-made templates to linguists and developers of language technology tools, where all the hard technical details are taken care off. They get a boiler-plate template for linguistic resources, and can start off directly working on the grammatical and linguistic issues. They can skip the demanding and time-consuming first stretch of the well-known S curve ((Huchzermeier and Loch, 2001) and (Barraza et al., 2004)), meaning they will immediately see real progress as they work. It also means that there is no need for every language to invent the same wheel over and over again, saving both money, time and frustration.

The shared infrastructure also means that shortcomings within it revealed by the needs of one language, will automatically benefit all languages.

The infrastructures facilitate cooperation across languages as everything is organised the same way. This also encourages cross-lingual cooperation and crowd-sourcing. Several of the projects using these infrastructures cover many languages in parallel.

Being a language pool in the sense of (Streiter et al., 2006) also means that continuity is secured even for languages with too few resources to ensure continuity on their own. A common organisation of files and documentation also means that linguists working on different languages can easily help newcomers getting started on a new language.

4.1. Choice of language technology

Given that the languages described here are morphologically complex, any successful attempt at analysing them must be able to analyse and generate the word forms. In order to do that, we use finite-state transducers. For languages with complex morphophonological processes, we combine the concatenative transducers with morphophonological transducers, thereby making it possible to deal with non-linear phenomena like vowel harmony, consonant gradation (Koskenniemi, 1983).

For syntactic analysis we use *Constraint Grammar* (Karlsson, 1990), a robust bottom-up parser framework that makes it possible to do dependency parsing with precision above 95 % for syntactic function, and above 99 % for part of speech.

4.2. Differences between the infrastructures

From a technical point of view, the Giellatekno/Divvun infrastructure is technology agnostic. For historical and other reasons, it has been built to support the Xerox FST tools (Beesley and Karttunen, 2003), but with parallel support for the free/open-source Helsinki Finite State tools (Lindén et al., 2013) (which are source-code compatible with the Xerox tools). Adding support for a third or fourth type of technology for morphological analysis should be no problem whatsoever, and the same goes for other parts of the language tool set as well as for the end user tools. The differs from the Apertium infrastructure, where only free/open-source tools are supported and relied upon. The agnosticity in the Apertium infrastructure comes from also supporting some statistically-based modules, such as for part-of-speech tagging.

The major difference between the two infrastructure is the number of end user tools supported by them. Whereas Apertium was designed to support one — machine translation — and has been extended to support FST-based spellcheckers, the Giellatekno/Divvun infrastructure has always supported a large number of end user tools. For the Sámi languages, and other languages supported in the Giellatekno/Divvun infrastructure, there is no competition. There is no competition because the language communities are too small for there to be a commercially viable market for any language-technology products. Thus, in order to fully serve the language community, the infrastructure must be able to support all of the tools needed by the community.

To add new features and tools to the languages in the Giellatekno/Divvun infrastructure, it is enough to develop the new feature for one language. When the new feature is ready, it is copied over to a build template, and from there distributed to all languages in one operation. With this system, support for new technologies and new features can easily be added to all languages. This is a variant of what (Streiter et al., 2006) describes as leveraging the pool to get upgrades «for free» even in cases where it would not be motivated for a specific language in itself. This differs

[6] `http://subversion.apache.org`

[7] `http://en.wikipedia.org/wiki/GNU_build_system`

from the Apertium method, where each language is developed based on a template, but once the template is copied, changes are only shared by manual copying and merging.

5. Crowds and infrastructure

In this section we try to characterise the groups of people — or the crowds — using the presented infrastructures. The relevant characteristics in this discussion are: paid/unpaid, size (persons/language), and level and type of expertise.

For larger language communities, the crowds consist of a mixture of programmers and language enthusiasts. For all of the languages, and especially for the ones with small language communities, linguists make up an important part of the crowd. One reason for this is that linguists are interested in grammatical analysis of the languages in question, and the linguistic approach makes the projects worthwhile for them.

5.1. The Giellatekno/Divvun crowd

5.1.1. Tromsø

At UiT Norgga árktalaš universitehta the infrastructure and its precursors have been in use from the very beginning of the work on Sámi language technology. It is indeed true that the infrastructure was first developed for the three major Sámi languages in Norway: North, Lule and South Sámi. These language communities vary in size from about 600 to 22,000 native speakers, and none of them have a functioning crowd working on Wikipedia articles — not today, and even much less so when the projects started.

Since the start in the first half of the previous decade, the resources have been developed by native speakers with linguistic education. These have been employed on projects financed through various public funds and institutions and they constitute the first «crowd of experts» using the precursor to the present infrastructure.

Would it have been possible to build a crowd of interested native speakers to help develop these resources? The Giellatekno/Divvun group actually tried a couple of times, and there was genuine interest in both language technology and in our work. But a number of factors caused these attempts to not succeed. One was inexperience, another our lack of understanding of crowd-sourcing and how to make it work in practice. Native speakers often were too occupied with other language-related activities. For several of the candidates the learning curve was too steep, and combined with little to no follow-up afterwards this meant that attendees forgot even the most basic steps in the procedure taught. Often there is also little to no direct feedback (e.g. in the form of seeing your own word available online after the edit). Learning how to master a version control system for submitting changes and edits turned out to be too complex for several of the candidates given the short timeframe.

Nevertheless, a few eager individuals have started to work on other Sámi languages, so that we today cover all the Sámi languages. These individuals are working outside our core group, some at other academic institutions, and some completely on their spare time.

In summary, most of the people working on the Sámi languages are paid, full time workers, native speakers, and except for North Sámi, usually only one person is working actively on any single language.

5.1.2. Nuuk

After a quite expensive — and failed — attempt at making a list-based spellchecker back in 2003, *Oqaasileriffik* (the Greenlandic language secretariat) has since 2005 used the Giellatekno/Divvun infrastructure described here[8]. In 2011 they moved over to the new iteration of the infrastructure, the one presented in this paper. The work has since 2005 involved 7 (mainly 4) people from the Greenlandic language secretariat and 2 people from UiT. Greenlandic was the first language for which we were able to build a spellchecker, *Kuukkinaat* was released in 2006, with the packaging and MS Office integration done by a private company in Finland.

The Greenlandic project has continued using the common infrastructure for the grammatical analysers ever since, but it has chosen other solutions for their practical programs, be it spellchecking, pedagogical programs[9] or online services[10].

This is a perfectly viable way of utilising this infrastructure. The good thing with this solution is that it gives the Greenlandic language secretariat the full control of design and priorities for the end user solution (as for the web services), and that it makes it possible to choose solutions that differ from the other languages when needed (as for the pedagogical programs). Using the common infrastructure for the basic analyser also gives access to the ready-made solutions for them.

The drawback with this solution is that it implies more work for the programmers linked to the Greenlandic project, and that the project is cut off from the synergy effects and possible free rides of the common project.

5.1.3. Pyssyjoki

Kvensk institutt (KI) in Pyssyjoki has since 2012 run a project on Kven language technology, involving 3 employees at KI, two part-time workers at UiT, and one worker at Halti kvenkultursenter.

Kven language technology started out with a 4,000 lemma bidirectional Kven-Norwegian electronic dictionary, written by Terje Aronsen. The dictionary was integrated in the present infrastructure, and paired with a Kven morphological analyser. Still in an initial stage (with a coverage of 71.2 %, measured on a small corpus of 410 words), it is good enough to make the dictionary a reception dictionary[11], allowing the user to click on words in running text cf. also (Haavisto et al., 2013),

The Kven morphological analyser is also the basis for work on interactive pedagogical programs within the *Oahpa* framework (Antonsen et al., 2009). Although not good enough to function as the basis for a spellchecker, the analyser still covers the basic morphological paradigms, and thus make Kven pedagogical programs possible.

[8] http://oqaaserpassualeriffik.org/a-bit-of-history/

[9] http://learngreenlandic.com

[10] http://oqaaserpassualeriffik.org/tools/

[11] http://sanat.oahpa.no/

5.1.4. Helsinki

A Language research funding programme introduced for the years 2012–2016 by the Helsinki-based Kone Foundation is concerned with the retention of a multilingual world. The group at the University of Helsinki has received funding to work on a project of language documentation. The project was initiated to encourage interaction between speakers/users of lesser documented languages and researchers. It involves the construction of morphological parsers for five Uralic languages. The set of languages selected for this project includes Liv (Livonian), Livvi (Olonets-Karelian), Hill Mari, Tundra Nenets, and Moksha Mordvin. The goal was to develop state-of-the-art parsers able to handle extensive inflectional challenges for at least 20,000 lemmas in each of the selected language over a two-year period. At the same time each of the 20,000 lemmas was to be translated into Finnish. With words and ongoing development of both inflection and translation, this small project has been able to utilise several facets of and contribute to the Giellatekno/Divvun infrastructure.

At present (early 2014) the finite-state transducer projects have progressed to the half-way point. Automatically generated reverse-direction dictionaries have also been set up for some of the transducer projects; yet another way to provide access to lesser documented languages.

In Helsinki transducer development coincides with digitisation of 1920–1930 minority Uralic literature at the National Library of Finland[12], and the development of an open-source editor for proof-reading of open-source OCR-ed literature [13]. Transducer descriptions have been used here to enhance text recognition.

5.1.5. Alberta

The cooperation with the University of Alberta in Edmonton is relatively recent, and is thus in a nascent stage. A group of four linguists have started work on Plains Cree, Northern Haida and Dene Suline[14].

For the two first languages, existing dictionaries are being added to the infrastructure, and the grammar is being rewritten in machine-readable form, as a finite-state transducer.

Adding the analysers to the Giellatekno / Divvun infrastructure offers a means of making morphologically-enriched dictionaries[15].

5.2. The Apertium crowd

Members of the Apertium project come from a range of different backgrounds: University researchers in computer science and linguistics, language activists, free-software and language enthusiasts, and students. There is a governing structure in the form of the project management committee,[16] where large decisions are taken democratically, but otherwise this committee takes a largely *laissez faire* approach leaving individual developers to make their own decisions.

The original crowd is based in Alacant in the Valencian Country in Spain. However, the crowd has become increasingly international. Interaction is through fairly low-tech but high productivity tools such as IRC, mailing lists and a Wiki[17]. The project has been working generally with under-resourced languages and communities, rather than endangered-language communities.

As a project, Apertium has participated in the Google Summer of Code and the Google Code-in. The former programme gives students three-month stipends to work on free software during the northern-hemisphere summer. The latter programme offers prizes to school pupils for completing tasks related to the project.[18] These tasks may be programming tasks: implement an algorithm; or linguistic tasks: e.g. lemmatise a wordlist or part-of-speech tag a short text.

5.3. Summary of the crowds and the infrastructure

What we learned form the first attempt at making a Greenlandic spellchecker was that getting a list of word forms from a crowd of language speakers of a morphologically complex language is not going to result in any useful tool. For the languages treated here, the word form is simply not the relevant unit of analysis. What is needed is a system of combining stems, inflectional and derivational affixes, and the set of morphophonological rules to unite them, in short, a grammatical analyser.

Presenting the setup for a morphological analyser to a group of language activists is in itself also not going to result in an analyser. Making grammatical analysers may be achieved by decentralised cooperation, not of language speakers alone, but of different types of experts fulfilling different roles (one of them being the native speakers) in teams working towards a common goal.

6. End-user tools

The Giellatekno/Divvun group have from the beginning focused on proofing tools and language learning. While the type of services and products has been considerably widened, these two are still at the core of the user-oriented activity. For Apertium, the focus has been upon machine translation.

The infrastructures described in this article combine morphological and syntactic parsers with a wide number of end user tools.

6.1. For linguists and researchers

For linguists, the most important tool is the grammatical analysers. Combined with an advanced corpus search interface[19] it is possible to do empirical research, such as distributional studies of syntactic and morphological phenomena.

[12] http://uralica.kansalliskirjasto.fi/

[13] http://ocrui-kk.lib.helsinki.fi/

[14] http://altlab.artsrn.ualberta.ca

[15] http://pikiskwewina.oahpa.no, http://guusaaw.oahpa.no

[16] See for example http://wiki.apertium.org/wiki/Bylaws

[17] http://wiki.apertium.org

[18] Example tasks: http://wiki.apertium.org/wiki/Task_ideas_for_Google_Code-in

[19] http://gtweb.uit.no/korp/

6.2. For language communities

With morphological transducers in place and readily-available bilingual resources, there is a pipeline for creating a wide range of tools: inflecting bilingual dictionaries [20], spellcheckers and morphologically-aware hyphenators[21]. Enriched with syntactic analysis we also are able to make grammar checkers and, with a bilingual dictionary, also machine-translation systems[22].

6.3. For language learners

Most languages dealt with here are inflecting languages. A central part of study is thus mastering the morphological structure of the language. The Oahpa infrastructure (Antonsen et al., 2009) was originally developed for North Sámi, and includes a series of learning programs, including lexical learning, generation of morphological tasks, and open-input dialogue tasks. Oahpa is integrated with the Giellatekno/Divvun infrastructure, so that Oahpa versions for 4 languages are now in use by language learners, and versions for about a dozen additional languages are in the pipeline.

7. Conclusion

We have tried to show that Wikipedia can be a useful indicator of whether it is possible to build a community of crowdsourcing volunteers. We see that for the Apertium languages crowd-sourcing is actually working, whereas it has not been possible for the Giellatekno/Divvun languages. This corresponds quite neatly with the Wikipedia status of those same languages: none of the Giellatekno/Divvun languages have a viable Wikipedia community, whereas most of the Apertium languages do have.

Language technology for morphology-rich languages with few speakers may be done by crowdsourcing of a different kind, by including people fulfilling different roles in a team. With the goal of combining linguistic analysis and functional end-user programs, we have found that finite-state transducers and constraint grammars are effective tools.

For linguists, the possibility of having others write the infrastructure, and themselves concentrate upon linguistic work, while at the same being able to present software to the user community, is clearly an attractive offer. The popularity of the Giellatekno/Divvun infrastructure shows that the possibility of generating a wide range of products while at the same spend the time on working with concrete linguistic problems is attractive enough to really attract linguists to participate in the crowd.

8. Acknowledgements

We would like to thank our colleagues at Giellatekno/Divvun and Apertium, the Norwegian Ministry of Local Government and Modernisation, and all our colleagues from the different cooperating groups.

9. References

Antonsen, L., Huhumarniemi, S., and Trosterud, T. (2009). Interactive pedagogical programs based on constraint grammar. In *Proceedings of the 17th Nordic Conference of Computational Linguistics*, number 4 in Nealt Proceedings Series.

Barraza, G., Back, W., and Mata, F. (2004). Probabilistic forecasting of project performance using stochastic s curves. *Journal of Construction Engineering and Management*, 130(1):25–32.

Beesley, K. R. and Karttunen, L. (2003). *Finite State Morphology*. CSLI publications in Computational Linguistics, USA.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Sánchez-Martínez, F., Ramírez-Sánchez, G., and Tyers, F. M. (2011). Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Haavisto, M., Maliniemi, K., Niiranen, L., Paavalniemi, P., Reibo, T., and Trosterud, T. (2013). Kvensk ordbok på nett – hvem har nytte av den? In *Den tolvte konferansen om leksikografi i Norden*. Nordisk konferanse i leksikografi, Oslo.

Howe, J. (2008). *Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business*. Crown Publishing Group, New York, NY, USA, 1 edition.

Huchzermeier, A. and Loch, C. H. (2001). Project management under risk: Using the real options approach to evaluate flexibility in r...d. *Management Science*, 47(1):85–101.

Karlsson, F. (1990). Constraint grammar as a framework for parsing running text. In *Proceedings of the 13th Conference on Computational Linguistics - Volume 3*, COLING '90, pages 168–173, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kloss, H. (1967). Abstand languages and ausbau languages. *Anthropological Linguistics*, 7(9):29—41.

Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model for Word-Form Recognition and Production*. Ph.D. thesis, University of Helsinki.

Lewis, M. P., Simons, G. F., and Fennig, C. D. (2013). *Ethnologue: Languages of the World*. SIL International, Dallas, TX, USA, seventeenth edition.

Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Silfverberg, M., and Pirinen, T. A. (2013). Using hfst for creating computational linguistic applications. In *Computational Linguistics*, pages 3–25. Springer Berlin Heidelberg.

Moseley, C., editor. (2010). *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris, third edition.

Streiter, O., Scannell, K. P., and Stuflesser, M. (2006). Implementing nlp projects for noncentral languages: Instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289, December.

Surowiecki, J. (2005). *The Wisdom of Crowds*. Anchor.

[20] http://dicts.uit.no

[21] http://divvun.no

[22] http://wiki.apertium.org

The LREMap for Under-Resourced Languages

Riccardo Del Gratta[◇], Francesca Frontini[◇], Anas Fahad Khan[◇], Joseph Mariani[*], Claudia Soria[◇]

[*]LIMSI-CNRS & IMMI, Paris, France, [◇]CNR-ILC
Pisa, Italy
[*]Joseph.Mariani@limsi.fr, [◇]{name.surname}@ilc.cnr.it

Abstract

A complete picture of currently available language resources and technologies for the under-resourced languages of Europe is still lacking. Yet this would help policy makers, researchers and developers enormously in planning a roadmap for providing all languages with the necessary instruments to act as fully equipped languages in the digital era. In this paper we introduce the LRE Map and show its utility for documenting available language resources and technologies for under-resourced languages. The importance of the serialization of the LREMap into (L)LOD along with the possibility of its connection to a wider world is also introduced.

Keywords: LREMap, under-resourced languages, (L)LOD

1. Introduction

1.1. The LRE Map

The LREMap was an initiative devised to address both the problem of a lack of knowledge about existing resources and the need to provide a simple and easy way to encourage the documentation of these resources. It was jointly launched by FLaReNet (the Fostering Language Resources Network)[1] and ELRA (European Language Resources Association) in May 2010 with the purpose of developing an entirely new instrument for capturing community knowledge about language resources, as well as for collecting descriptions both for tools and existing or new resources as applied to NLP research.

The first Map was initially created in conjunction with the LREC 2010 Conference (Calzolari et al., 2010), as the result of a campaign to gather information about the language resources and technologies underlying the scientific works presented during the conference. Authors who submitted a paper were requested to provide information about the language resources and tools either developed or used; the initiative was successful, with close to 2000 resource descriptions collected. The required information was relatively simple and related to basic metadata.

The rationale behind the creation of the LREMap was the indisputable need for accurate and reliable documentation of language resources: the more that these resources are documented the more they "exist" and the more easily that they can be retrieved.

This initiative was so strongly welcomed that it continued with LREC 2012 and LREC 2014 as well as with many other conferences of various kinds and with different audiences such as Association for Computational Linguistics (ACL), Recent Advances in Natural Language Processing (RANLP), International Committee for Co-ordination and Standardisation of Speech Databases (COCOSDA) and so on. To date, the LREMap is a community-based, collaboratively built resource that, so far, contains $\sim$ 7000 records which, on the one hand, reflect the judgments of authors with respect to the language resources they have used or created, and on the other consist of a manually developed/checked normalization of all the data contained in a database. There may not be a lot of data so far, but what there is has a significant specific weight.

The main goal of the LREMap remains to gather information in a bottom up manner and to exploit community knowledge to assist in the discovery and documentation of resources, essentially through a web interface that enables searching using multiple criteria.

In this paper we show how the LREMap can be used to derive (and spread) knowledge about language resources and technologies for less-resourced languages. We also show how the serialization of the LREMap in RDF/XML can help under-resourced languages to become a part of the growing Linguistic Linked Data movement (Chiarcos et al., 2011; Chiarcos, 2012; Lezcano et al., 2013).

1.2. The LREMap Metadata

In the LREMap, each resource is described according to twelve main metadata fields, which provide a minimal amount of relevant and useful information about new and existing resources and their uses. A set of nine initial metadata fields was revised after the abstract submission phase in order to slightly increase the descriptive parameters requested upon submission of the final papers, while an additional set of three was added in the final submission phase.

The first set of metadata represents quite general information available in most language resource catalogs and surveys (e.g. ENABLER). Each of these basic fields has a list of suggested values, which has been deliberately kept short by using only the most frequent and common values. However, the possibility has been left open for the user to select the "Other" field and to specify a more appropriate term in case he/she does not feel that any of the suggested values satisfy his or her requirements.

The set of metadata items contains: a) Resource Type b) Resource Name c) Resource Production Status d) Use of the Resource e) Language(s) f) Modality g) Resource Availability h) Resource URL (if available) and i) Re-

[1] http://www.flarenet.eu

source Description. Three descriptors ("Resource Size", "Resource License" and "Resource Documentation") were added in the final submission phase to allow for extraction of additional information.

2. Knowing about Language Resources for Under-resourced Languages

Not only are language resources essential, but knowledge about those which are missing and required is equally crucial. Indeed, knowledge about existing language resources and technologies is crucial for the overall advancement of research in the field of NLP: it is important to be able to locate and retrieve the right resources for the right applications, and to exploit existing ones before building new ones from scratch. This is particularly true for under-resourced languages, given the limited funding usually available and the often fragmented framework in which development takes place.

Having a clear picture of which resources are available for which languages and for which uses is important in order to identify existing gaps for a given language at a given time and to estimate the amount of investment needed to fill these gaps. Knowledge about the current use of resources is equally important: being able to determine which resources are most commonly used for given applications will help developers and planners to better understand the reasons for their success (e.g., intrinsic quality, wide availability, licensing model, etc.).

Unfortunately, clear and easy-to-access information of this kind about resources and related technologies is still lacking. Several worldwide institutions maintain catalogs of language resources (ELRA, LDC, National Institute of Information and Communications Technology (NICT) Universal Catalog, etc.). However, it has been estimated that only a tiny fraction of existing resources are known, either through distribution catalogs or via direct publicity by providers (web sites and the like).

The majority of language resources are still poorly documented or not documented at all, and use of metadata elements to describe and document resources is still uncommon and often inconsistent. This represents a serious problem that has been repeatedly cited in e.g. the FLaReNet recommendations (Soria et al., 2012). Individual authors can find it difficult to document their own resources, simply because they have a hard time deciding on the relevant set of metadata elements to be used. Moreover, there is insufficient awareness about the importance of documentation which is often regarded as a useless burden.

3. Uses and applications of the LREMap for under-resourced languages

The LREMap represents an observation point from which the landscape of language resources and technologies can be easily observed and analyzed. Although it does not boast of providing absolutely exhaustive coverage, it can nevertheless claim, by its size, to be considered a comprehensive and accurate representation of the current situation.

The Map enables the derivation of information about the available resources in various different European languages; this data can then be interpreted from the viewpoint of anyone interested in designing a *roadmap* for the further development of missing resources, as well as for planning language-sensitive strategies in order to more equally (or more evenly) equip languages with the fundamental tools to be able to function in the digital era.

For instance, the Map contains information for a total of 146 resources for European regional and minority languages, here presented in Table 1 in terms of number of resources for each language listed in the LREMap.

Language	**# of resources**
Basque	45
Catalan	43
Galician	12
Faroese	5
Welsh	4
Aragonese Asturian Breton Frisian Romansh Swiss German Venetan	3
Luxembourgish Corsican Drents Limburgan Lombard Low German Lule, North and South Saami Occitan Scottish Gaelic Scots Sicilian	1

Table 1: Number of resources for under-resourced languages.

Detailed analyses can be performed by combining different parameters. For instance, it is possible to list the particular types of resources available for each language, or the number of newly developed resources vs. already existing ones, or else the number of resources that are freely available in comparison with those available for a fee or not available at all. The interested reader is referred to (Calzolari et al., 2010) for an illustration of the various uses that the Map lends itself to. Here we are particularly interested in showing how the data gathered with the LREMap can be used to derive a picture of the particular types of resources available and missing for the various languages, a concept referred to as a "*Language Matrix*".

3.1. The Language Matrices

The objective of Language Matrices is to provide a clear picture about what exists in terms of language resources, for various languages, and to emphasize which languages are missing which resources. The goal is then to ensure the production of the corresponding resources to fill the gaps for those languages.

A first set of Language Matrices was produced in January

2011. These matrices were constructed from the LREMap that was produced from the information provided by the authors of the papers submitted at the LREC 2010 conference, this comprised close to 2000 entries.

A software application was developed in order to determine the number of resources that exist for each language and give direct access to all the corresponding information (Mariani and Francopoulo, 2012). In this first analysis, we considered the 23 official languages of the EU, together with a category for "Regional European languages" and one for "Non-EU European languages", as well as "Multilingual", "Language Independent" and "Not Applicable" categories.

We produced 8 Language Matrices under the different headings of Multimodal/Multimedia Data and Tools, Written Language Data and Tools, Spoken Language Data and Tools, Evaluation and Meta-resources. Several types of resources were listed for each matrix, either corresponding to the 24 Types that were suggested in the questionnaire or to the author's own wording. This resulted in a total of 160 Language Resource Types, with a variable number for each matrix (from 5 Types for Evaluation to 78 Types for Written Language Tools).

These matrices showed, unsurprisingly, that English was by far the most resourced language, followed by French and German, Spanish, Italian and Dutch. Some languages were clearly under-resourced, such as Irish Gaelic, Slovak or Maltese. Given the large number of types specified by the authors, some of these types existed for only one language, and the matrices therefore showed a large number of zeroes for all other languages. We preferred to keep this information as such rather than to merge these separate types into an "Other Type" category, as these singletons may be weak signals announcing the emergence of new research trends.

We produced the same 8 Language Matrices on the basis of the 2012 LREMap which includes 216 languages. A Language Matrix has also been developed for Sign Languages, covering 21 Sign Languages.

Figures 1, 2 and 3 present simplified Language Matrices for European languages (respectively EU official, non-EU official and regional or minority languages).

		European Union																								
Resource Modality	Resource Category	Bulgarian	Czech	Danish	Dutch	English	Estonian	Finnish	French	German	Greek	Hungarian	Irish	Italian	Latvian	Lithuanian	Maltese	Polish	Portuguese	Romanian	Slovak	Slovene	Spanish	Swedish
Written	Data	40	35	27	53	738	14	16	144	133	29	28	4	92	15	12	4	29	53	36	9	20	116	42
	Evaluation	2	5	3	6	117		1	10	14	1	1	1	5	3	3			5	3			14	3
	Guidelines	2	4	1	6	32	1		6	6	1	2	2	4				3	1	1		1	7	3
	Tools	3	6	2	12	169	4	2	44	37	12	4		23			1	6	12	10	1	5	29	6
Speech	Data		3	4	6	57		1	27	19	2	1		3					2	4	1		1	14	3
	Evaluation					5			7	1												1			
	Tools		1	1	1	7			1	4		1								1	1		1	2
Multimodal	Data		1	3	6	56	3	3	9	17	4	2		4				3	2				11	4
	Evaluation					3																	1	
	Guidelines					1																		
	Tools		1		1	10		1	3	4	1	1		2		1		1	2	1			3	
Sign language	Data		1		1	4		1	3	1				1									1	1
	Guidelines					1																		
	Tools					1			1															
Written/Speech	Data		2	1	3	19	1		5	4				3				4	1				3	2
	Tools				1	5			1	1													1	
Not applicable	Data				3	9			1	1				2					1				2	
	Guidelines								1	1				1										
	Tools					17			2	3	1			2				2					3	
Total		47	60	42	99	1251	23	25	265	246	51	40	7	142	18	16	5	50	82	53	10	28	206	66

Figure 1: Simplified Language Matrix for the 23 EU official languages

		Europe non EU										
Resource Modality	Resource Category	Albanian	Belarusan	Bosnian	Croatian	Icelandic	Macedonian	Norwegian	Russian	Serbian	Serbo-Croatian	Turkish	Ukrainian
Written	Data	8	3	2	21	7	5	9	36	15		13	2
	Evaluation	1			2		1	1	1	2		2	
	Guidelines				1		1		2			1	
	Tools				2	6		2	5			2	
Speech	Data				1	1		4	3			3	1
	Evaluation								1				
	Tools					1			2				
Multimodal	Data							1	3			1	
	Tools											1	
Sign language									1				
Written/Speech	Data			1		1			1			1	
	Tools								1				
Not applicable	Data								1				
Total		9	3	3	27	16	7	17	57	17	1	24	3

Figure 2: Simplified Language Matrix for non-EU languages

		Regional and minority																																
Resource Modality	Resource Category	Aragonese	Asturian	Basque	Breton	Catalan	Cornish	Dhivehi	Faroese	Frisian	Galician	Limburgish	Lombard	Low German	Lule Saami	Luxembourgish	North Saami	Occitan	Romansh	Scottish Gaelic	Scots	Sicilian	South Saami	Swiss German	Venetian	Welsh								
Written	Data	2	1	18	3	23	1	1	2	2	7	1	1	1		1		1	3	1	1	1		1	2	2	76							
	Evaluation			1																							1							
	Guidelines			2		1																					3							
	Tools	1	2	12		11			1		2				1		1						1		1	2	35							
Speech	Data			6		5			1		3					1								2			18							
	Evaluation					1																					1							
	Tools			3					1																		4							
Multimodal	Data					1																					1							
	Tools			1																							1							
Sign language	Data					1																					1							
Written/Speech	Data			1						1																	2							
Not applicable	Data			1																							1							
Total		3	3	45	3	43	1	1	5	3	12	1	1	1	1	2	1	1	3	1	1	1	1	3	3	4	144							

Figure 3: Simplified Language Matrix for European regional and minority languages.

The analysis of Language Matrices confirms the supremacy of (American) English in all categories of Language Resources.

Of the 5218 entries listed in the LREMap, 1364 resources relate to English (close to 26%, compared to 29% in 2011), followed by French (286), German (265), Spanish (222), Italian (158) and Dutch (110), while there are still very few for Slovak (9), Irish Gaelic (5) or Maltese (4). However, it appears that the effort is increasing for other major languages (since most of them doubled their number of resources), while several previously considered "minor" or regional languages have benefited from strong support to recover and some of them (such as Bulgarian, Estonian, Polish or Slovene) more than tripled the number of resources. We note a big increase for the Estonian language (from 7 to 23), for regional languages (67 to 103) and for non-EU European languages (63 to 293), cf. table 2.

The Language Matrices derived from the LREMap also represent an important instrument for assessing the particular types of resources that are missing for any given language, thus usefully complementing, in a data-driven approach, the well known BLaRK (Basic Language Resource Kit) notion (Krauwer, 2003). For instance, figure 4 illustrates the gaps, in terms of missing language resources, for European regional and minority languages on the basis of the available information.

Language	Years	
	2011	2012
English	559	1364
French	143	286
German	132	265
Spanish	111	222
Italian	90	158
Dutch	54	110
Estonian	7	23
Irish	3	5
Slovak	3	9
Maltese	2	4
Eu Regional	67	103
Other Europe	63	293
Total	1889	5218

Table 2: Some examples of evolution of the language coverage of language resources from 2011 to 2012.

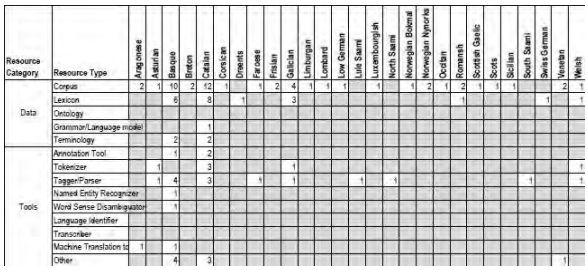

Figure 4: Language Matrix for European regional and minority languages (written resources).

In a similar vein, for instance, the Language Matrices have already begun to be used for identifying the gaps and establishing the Language Tables and Language Status Estimates in the META-NET Language White Papers (Rehm and Uszkoreit, 2012). In addition to a language and resource coverage analysis, we also introduced a measure relating to the popularity of Language Resources, named the Language Resource Impact Factor (LRIF), obtained by counting the number of times a resource is cited. Four LRIF Matrices have been produced based on the papers accepted at the previously mentioned conferences, which provide a glimpse of the most popular resources in each category (Data, Tools, Evaluation and Meta-Resources).

4. The LREMap in the (Linguistic) Linked Open Data Paradigm

In this section we briefly describe the serialization of the LREMap according to the RDF model with an emphasis on the importance of having such a serialization for under-resourced languages. More specific details on how the ontologies have been created and populated can be found in (Del Gratta et al., 2014).

The LREMap is a database consisting of at least three interconnected databases: *Authors*, *Papers* and *Resources* which are linked through an internal unique identifier, the submission code. The record structure is exemplified as follows, cf. Figure 5.

Submission:	S1
Conference:	C1
Year:	Y1
Resource:	R1
Paper:	P1
Author(s):	A1,A2...
....:	....
Submission:	S2
Conference:	C1
Year:	Y1
Resource:	R2
Paper:	P2
Author(s):	A3,A4...
....:	....

Figure 5: Sample records from the LREMap database.

The same example can be differently interpreted as in Figure 6

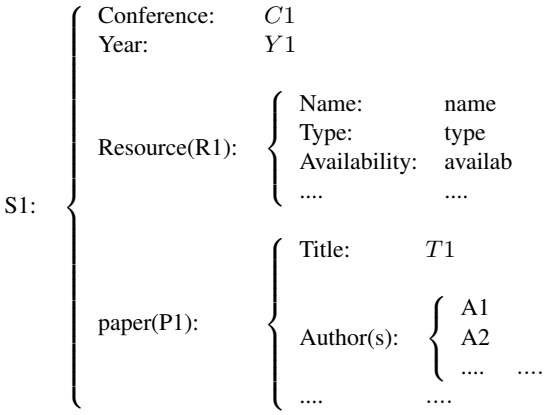

Figure 6: The same example[2]of Figure 5 with objects logically grouped.

We can look at Figure 6 from a different point of view. The fact that the values in the record(s) are interconnected allows us to assign *semantics* to their logical grouping. For example, the connection between the submission *S1* and the paper *P1* can be transformed in a triple:

$$S1 \text{ hasDocument } P1$$

and the connection between *P1* and authors *A1, A2* in:

$$P1 \text{ hasAuthor } A1,\ A2$$

Some attributes of the records have been grouped to define *objects*, such as $A1$, $A2$ which are used to model the *Author* object, and *Paper*, $P1$, cf. Figure 7.

[2]We have added some metadata to the Resource object.

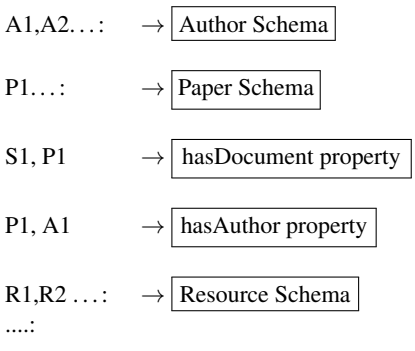

Figure 7: Records, *Objects* and *Relations*.

The complete set of the ontological objects that have been modeled starting from the database structure is reported in Figure 8.

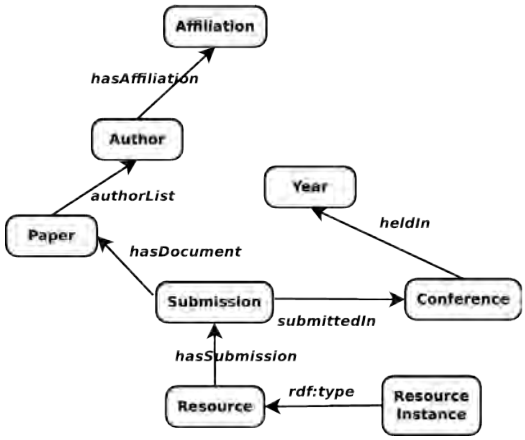

Figure 8: The network scheme for LREMap.

From the data structure we may highlight the following objects:

- *Submission*: representing a submission to a given conference, labeled using a (unique) identifier;
- *Conference*: the name, place and year of a conference, such as LREC 2012, La Valletta, Malta;
- *Paper*: the article connected to the submission;
- *Author*: the list of authors of the paper;
- *Resource*: the list of language resources that the authors of the paper decided to describe. This is the core of the data. Each description represent a single resource instance which describes the authors' ideas and thoughts on a specific resource - the uses, the modalities, languages and other information that are needed to properly describe a language resource.

From Figure 8 we may also notice that Papers, Conferences and Resources are connected to Submissions, while Authors and Affiliations are connected to Papers. It was necessary to distinguish between submission and paper since some conferences only provide data containing anonymous submissions, together with the resources they are linked to, while in other cases (notably LREC) a full description of the papers is also available. Hence the necessity of a simple submission object, which is only identified by a code and by the reference to its related conference, and which may or may not be enriched by further information on the actual paper. In absence of information on the paper, the count of submissions can still provide us with useful information on how many times a certain resource has been used.
The complete formal description of the LREMap schema goes beyond the scope of the present paper; nevertheless Figure 8 provides a general idea on how the modelling of the LREMap using RDF has been carried out using available ontologies, when possible:

- *Resources* are described with an ad hoc ontology that formalizes all aforementioned metadata; furthermore resources are linked to *Submissions*, in turn defined by their belonging to a *Conference*
- Submissions may be linked to an actual *Paper*, which is modeled using the *bibo*[3] ontology.
- Papers in turn have **authors**, that are modeled using *FOAF*[4] for person-related information, such as first and last name, email, ...and *GeoNames*[5] for geographical features related to their affiliation[6];

In this schema the most important link is therefore the *hasSubmission* relation, connecting the resource to one submission. By counting the number of such links, it is possible to gain immediate insight on how many times a certain resource is used. By adding the information on the paper, or the conference that are linked to each submission information may be added on who, when, and how used a resource. The modeling of the LREMap using RDF is particularly interesting for under-resourced languages, for the following reasons:

Connections The LREMap contains many valid pointers to well linked resources already available in RDF, such as Wikipedia, dbpedia and so on. This aspect ensures that under-resourced language resources are directly connected to the cloud with the possibility to link and to be linked to by bigger resources;

Use of important and widely accepted ontologies Authors and papers of the under-resourced language resources will be encoded using *FOAF* and *bibo* ontologies, thus being immediately retrievable by any SPARQL end point which uses these ontologies;

Data are normalized Data gathered by the LREMap are quite noisy, because of the extremely free process of filling the values of the metadata. Before data are distributed they go through a complex process of normalization which involves all metadata of the map. Normalization is fundamental for the language(s) which

[3]http://purl.org/ontology/bibo/
[4]http://xmlns.com/foaf/0.1/
[5]http://www.geonames.org/ontology/ontology_v3.1.rdf
[6]In this respect, the LREMap differs from the Saffron tool (http://saffron.deri.ie), which clusters authors and papers in terms of topics, keywords and shared interests, but it does not provide information about the language resources linked to papers.

are inserted as free text.

To uniform the languages we decided to connect the normalized value of the language[7] to `lexvo`, e.g. the Scottish Gaelic points to http://www.lexvo.org/page/iso639-3/gla;

Data are light The RDF serialization of the LREMap is very light; our way of representing the file is based on the rule "one language resource one RDF file". In this way, even places that aren't well covered by the Internet or that have a poor quality of connection can access and download the language resources they need quite easily. This aspect will help such speakers to be aware of the quantity/quality of language resources are available in the language resource community.

5. Conclusions

The LREMap holds out significant potential for possible applications and uses. It is an instrument for enhancing the availability of information about resources, either new or already existing ones. It is a measuring tool for monitoring various dimensions of resources across place and time, thus helping to highlight evolutionary trends in language resource use and related language technology development. It is able to do this by cataloging not only language resources in a narrow sense (i.e. language data), but also tools, standards, and annotation guidelines.

The potential of the LREMap for becoming a powerful aggregator of information related to language resources was clear from the outset, as was the possibility of deriving and discovering novel combinations of information in entirely new ways. The database underlying the LREMap can yield interesting matrices for the language resources available for different languages, modalities, and applications. Such matrices have been already used, for example, in META-NET to provide a picture of the situation of resources availability for the various European languages.

In the near future the LREMap will continue collecting input about resources in a bottom up manner from authors of papers at other relevant conferences. Providing information about resources could become part of the standard submission process. This will help to broaden the notion of "language resources" and thus attract neighboring disciplines to the field: disciplines that so far have been only marginally involved by the standard notion of language resources. Moreover, it will be extended to authors submitting papers to the "Language Resources and Evaluation" journal.

We believe that the LREMap - and its accompanying Language Matrices - will have an impact as an instrument for documenting, searching and sharing knowledge of available language resources and technologies for less-resourced languages; for highlighting the best tools and resources; and for monitoring the usability of existing resources over a range of different tools and application domains. We therefore strongly advocate the use of the LREMap by organizations, individual researchers, industry, etc. to document resources for regional and minority languages in order to derive an accurate picture of available technologies and an analysis of the current needs and gaps to be addressed.

[7]The normalized value contains the three-letter ISO codes - extracted from http://www.iso.org/iso/language_codes taking into account the tables 639-2 and 639-3- and the complete string of the language: "English" → "English (eng)".

6. References

Calzolari, N., Soria, C., Gratta, R. D., Goggi, S., Quochi, V., Russo, I., Choukri, K., Mariani, J., and Piperidis, S. (2010). The lrec map of language resources and technologies. In Chair), N. C. C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Chiarcos, C., Hellmann, S., and Nordhoff, S. (2011). Towards a linguistic linked open data cloud: The open linguistics working group. *TAL*, 52(3):245–275.

Chiarcos, C. (2012). *Linked Data in Linguistics*. Springer.

Del Gratta, R., Khan, F., Goggi, S., and Pardelli, G. (2014). Lremap disclosed. In *Proceedings of the ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavk, Iceland, May. European Language Resources Association (ELRA).

Krauwer, S. (2003). The Basic Language Resource Kit (BLARK) as the first milestone for the language resources roadmap. In *Proceedings of the 2003 International Workshop Speech and Computer (SPECOM 2003)*, pages 8–15. Moscow State Linguistic University.

Lezcano, L., Sanchez, S., and Roa-Valverde, A. J. (2013). A survey on the exchange of linguistic resources: Publishing linguistic linked open data on the web. *Program: electronic library and information systems*, 47(3):3–3.

Mariani, J. and Francopoulo, G. (2012). The language matrices and the language resource impact factor. In *Proceedings of the Parole Workshop)*, Lisbon, Portugal, October.

Rehm, G. and Uszkoreit, H., editors. (2012). *META-NET White Paper Series*. Springer, Heidelberg, New York, Dordrecht, London.

Soria, C., Bel, N., Choukri, K., Mariani, J., Monachini, M., Odijk, J., Piperidis, S., Quochi, V., and Calzolari, N. (2012). The flarenet strategic language resource agenda. In Chair), N. C. C., Choukri, K., Declerck, T., Doan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Using GrAF for Advanced Convertibility of IGT data

Dorothee Beermann, Peter Bouda

Norwegian University of Science and Technology, Centro Interdisciplinar de Documentao Lingustica e Social
dorothee.beermann@ntnu.no, pbouda@cidles.eu

Abstract

With the growing availability of multi-lingual, multi-media and multi-layered corpora also for lesser-documented languages, and with a growing number of tools that allow their exploitation, working with corpora has attracted the interest of researchers from the theoretical as well as the applied fields. But always when information from different sources is combined, the pertaining lack of interoperability represents a problem. This is in particular a challenge for corpora from endangered and lesser described languages since they often originate from work by individual researchers and small projects, using different methodologies and different tools. Before this material can become a true resource, well-known differences in the physical as well as the conceptual data structure must be leveraged against ways of future data use and exploitation. Working with Interlinear Glossed Text (IGT), which is a common annotation format for linguistic data from lesser described languages, we will use GrAF to achieve Advanced Convertibility. Our goal is to build data bridges between a number of linguistic tools. As a result data will become mobile across applications.

Keywords: Language Documentation, Interlinear Glossed Text, Natural Language Processing

1. Introduction

Convertibility is dependent on the data's physical and conceptual structure. In this paper we would like to focus on the latter. Using the term Advanced Glossing, (Drude, 2002) suggests as a de dicto standard a fixed set of annotation tiers across several annotation tables to allow a conceptually cleaner albeit comprehensive linguistic annotation tailored to the needs of documenting linguists. Using the "Graph Annotation Framework" (GrAF) (as described in (Ide and Suderman, 2007)) we would like to promote the flexible integration of de facto standards instead. The idea is to present a relatively simple model for the presentation of analytic layers. Yet, our approach is confronted with multiple challenges which arise from the particular nature of Interlinear Glossed Texts (IGTs). They do not only differ in terms of the concepts they encode, but also in the way these concepts are expressed across tiers. Using GrAF, via the software library Poio API[1], we would like to show that Advanced Convertibility allows us to stay within a given "semantics" (nodes, annotations, edges) when converting from one format into another. "Graph semantics" is learned and can be applied to any transformation. This makes our approach sustainable.

IGT normally consists of 3-5 lines, which are also called "tiers"[2]. Originally, researchers used IGT in descriptive linguistics and related disciplines to discuss features of languages in articles and books. An IGT was, at least in the more empirically oriented fields of linguistics, regarded as evidence in the evaluation of a hypothesis. Example (1) shows a typical IGT:

(1) Example from Kaguru (ISO-693-3 kki)

Kamei	*howoluta*	*kunyumbangwa*
kamei	*ha-wa-lut-a*	*ku-nyumba-ngwa*
then	*PAST-2-go-FV*	*17-house:9/10-somebody's*
adv	tm-sm-v-fv	sm-n-prn
imwe,		
di-mwe		
5-one		
ncp-num		

Then they went to one house

In this case the example conists of tiers for "words", "morphemes", "(morpho-syntactic) glosses", "part-of-speech" and "free translation" that are partly aligned vertically. There were several attempts to standardize IGT, as for example in the "Leipzig Glossing Rules"[3] as the most prominent example. But none of those attempts where accepted by the community and today a diversity of tier names, tier structures and annotation schemes co-exist in published data. This is one of the most important reasons why linguists struggle to analyse and compare data from different projects. In our paper we want to demonstrate how GrAF as a pivot data model can support researchers to exchange and analyse data in different file formats and with different tools.

2. Annotation graphs as pivot model

Our approach depends on the use of annotation graphs, i.e. the recently standardized implementation of the Linguistic Annotation Framework (LAF) as described in ISO 24612[4]. LAF was developed as an underlying data model for linguistic annotations designed to allow a better insight into

[1] `http://media.cidles.eu/poio/poio-api/`, accessed 19.2.2014

[2] For a discussion of this data format its variants and its problems see for example (Bow Catherine and Bird, 2003), (Palmer and Erk, 2007) and (Beermann and Mihaylov, 2013).

[3] `http://www.eva.mpg.de/lingua/resources/glossing-rules.php`, accessed 26.3.2014

[4] "Language resource management - Linguistic annotation framework", `http://www.iso.org/iso/catalogue_detail.htm?csnumber=37326`, accessed 3.2.2014

commonality as well as divergence of annotations from different sources, while GrAF is an implementation of LAF. The goal of LAF, as formulated by (Ide and Romary, 2004) is to combine flexibility for the annotators with data mobility, a goal that we share.
Since natural occurring language is a multi-layered information structure, linguistic annotations tend to be multidimensional and linguistic resources not rarely consist of audio- video- and text material and their annotations. Annotations in turn have their own complexity, they not only reflect resource internal properties but also scientific development and analytic stances. For the well researched languages, the first generation of annotated corpora featured part-of-speech and syntactic annotations, while recent corpora are annotated (in addition) for semantic and discourse information (Dipper, 2005). A similar development can also be observed for the annotation of lesser described languages where structured resources, mostly IGT, consist for the most part of small heterogeneous corpora[5].
In this section we want to demonstrate IGT data conversions. The real challenge lies in showing that our data model, the GrAF, is sufficiently expressive to hold the existing tier based information structure as well as its heterogeneity. GrAF is an implementation of LAF, originally developed to publish the "Manually Annotated Subcorpus" (MASC) of the American National Corpus [6], and consists of three parts:

- an abstract data model;
- an API for manipulating the data model;
- a straightforward XML serialization of the data model.

In the following we will present the challenges that we encountered in mapping IGT data from various file formats onto the GrAF model. Section 3 will then focus on an architecture for Advanced Convertibility based on the Python library Poio API. We explicitly do not want to introduce a new file format like GrAF-XML in our research area, as we think that there are currently enough de dicto standards available that cover the linguistic use cases. We will also not cover details about the layout of IGT GrAF graphs in Poio API (see (Bouda et al., 2012), (Blumtritt et al., 2013) for more information on how tiers and annotations are represented in Poio API).

2.1. IGT within the Linguistic Annotation Framework

Although LAF and annotation graphs in general are an underlying model for linguistic annotation, it does not mean that differences in semantics between applications and their file formats and data models do disappear when we load IGT data into GrAF. As (Bird and Liberman, 2001) note in their influential paper about annotation graphs,

> "It is important to recognize that translation into AGs does not magically create compatibility among systems whose semantics are different."

We expect that this is the case between systems in different research areas like corpus linguistics and language documentation, but also within the "IGT world" we have a high heterogeneity among data models even in cases where the data was created with only one and the same application. As already mentioned, one reason for this are differences in the theoretical backgrounds of linguists, so that the data finally uploaded into an archive can be regarded as an "island" with only few connections to other corpora. The more abstract the annotation format becomes, or the more theory inspired the annotations appear, the more controversial they tend to become. Also different annotation models where annotations from one tier, for example in Toolbox, appear on different tiers,for example in TypeCraft, may lead to less acceptability of the annotation itself. We will discuss such cases in more detail in section 3.
Here we would like to mention a further challenge coming from the need of under-specifying data, and to give enough room for underspecified data, or data still subject to linguistic investigation. Most field linguists see it as part of their research ethics to start with as few theoretical assumptions as possible when they study a language. In practice this approach has the advantage that "unseen" linguistic structures can be discovered and described and that the annotation of the data stays close to the spirit of the specific language, but generally this makes it hard to process and analyse the data later. Our proposed solution to assure IGT Convertibility consists of a two-step process: first, we convert the data model of any IGT input format into an annotation graph; then, we apply a rule-based transformation on the graph, in order to be able to compare the graphs and to convert between data models of file formats. In order to do so, we first defined a subset of GrAF graphs that adhere to the "tier model" characteristic for IGT data. An "annotation tier" is a collection of annotations that are either directly linked to the primary data (via string offsets or timing information), or to annotations on another tier. Our goal is then to map the existing relations between elements of different tiers to a graph with "nodes" and "edges". We regard this process as a conversion from the semantics of IGT to the semantics of annotation graphs. Judging from prior work with ELAN files in (Blumtritt et al., 2013), the mapping process consists of adding edges with appropriate labels between nodes that we created from the annotations, and we like to assume that our approach can capture any data model that exists in the "IGT world", but of course the set of annotation graphs in LAF is much larger. In a general graph there might be edges between any two nodes, while in our case edges are limited to annotations on tiers that are in a parent-child relationship in the original data model. This poses the question of validation, i.e. how to restrict the set of annotation graphs to a subset that encodes tier-based data. Part of Poio API is a solution for exactly this problem: we encode the original tier hierarchy in a separate entity, the "data structure type" to make sure that the data in the AG is compatible with any IGT model. Any ap-

[5] In Western Europe, The Language Archive at the Max Planck Institute in Nijmegen and the Endangered Languages Archive, ELAR at SOAS are probably the two best known archives for these corpora.

[6] `http://www.anc.org/data/masc/`, accessed 3.2.2014

plication based on Poio API can thus validate the data and always check which tier hierarchies the annotation graphs contains. The second step of our workflow, the transformation, will use this notion of "data structure type" to support transformations of annotation graphs without violating the tier model. It makes it easy to let users work with the semantics they know (i.e. "tiers" and "annotations" with a tier hierarchy encoded in the "data structure type") while the internal model of "nodes" and "annotations" makes the implementation of the transformation in Poio API as general as possible. There is a trade-off of course, as we lose the full power of annotation graphs and the option to manage random edges within Poio API. We still think that our use cases benefit massively from a uniform data model and a general conversion mechanism.
The next section will cover the implementation of the second step in our workflow, the transformation between semantics, in more detail.

2.2. Implementation in Poio API

Poio API was originally developed as part of a curation project of the CLARIN-D[7] working group 3 "Linguistic Fieldwork, Anthropology, Language Typology". It is a free and open source Python library to manage and analyse linguistic data from language documentation projects. For this purpose it uses the GrAF model as internal representation of the data. As already mentioned, one goal of Poio API is to make sure that the annotation graphs adhere to a certain layout, so that the data can be accessed via tier labels and their annotations at any point. Another benefit is that Poio API contains rudimentary functions to search and analyse sets of files in different formats. Think of GrAF as an assembly language for linguistic annotation, then Poio API is a library to map from and to higher-level languages. What we describe in this section is work in progress, especially the general, rule-based mapping mechanism that we propose here. We will thus focus on our experience from initial prototypes and sketch the next steps in our work.
In the present project we use Poio API to achieve what we call "Advanced Convertibility" which is greatly needed. We mentioned already that corpus work has become more interesting for scholars in the theoretical and applied fields of linguistics. In addition the computational fields also turn to data from non-mainstream languages, and with a growing "market", corpus tools appear that cater to higher-level annotations requiring advanced facilities for example for semantic or discourse annotation. These new tools in turn generate interest among new group of scholars. With new players in the game and a growing interest in more granular annotation, the need to integrate layers of analysis is growing. So far it has been common that annotating linguists use one linguistic tool to the exclusion of others, the trend however is towards combing tools. Linguistic tool providers are excited about the emerging of linguistic services which combine several linguistic tools in order to allow a more comprehensive analysis of language data; Weblicht[8], a CLARIN-D development, is such an environment and also WebANNO[9].
The success of this development crucially depends on our ability to migrate data efficiently between tools, and the Poio API development, and also the present collaborative project between TypeCraft[10] and the "Centro Interdisciplinar de Documentação Linguística e Social"[11], who developed the Poio API, needs to be seen in this context. Our first use case, as described in section 3, was to import Toolbox files from several projects into the TypeCraft web application. This project requires conversions between the Toolbox and the TypeCraft data model. In a first prototype, described below, we hard-coded a set of rules into a Typecraft XML writer, so that we could make the result visible to the human eye to learn about diversions. Conceptual differences had to be distinguished from errors and random idiosyncrasies. It is the conceptual differences leading to differences in the semantics that we will describe below. The conversion itself is rather straightforward. A writer traversed the graph while writing the Typecraft XML file. In addition we used regular expression to map specific annotations (e.g. "the POS tag 'prn' will be written as 'PN'"). In an iterative process we refined those rules and are currently implementing the transformation based on GrAF graphs to be independent of any input and output format.

3. Toolbox-to-TypeCraft conversions A case study

Annotated language data tends to be a hydra. Its heads are the inherent multi-layered nature of language that wants to be expressed, the manifold of the linguistic fields engaging in annotating with their specific conventions and demands, and the multitude of scientific ventures in which this data serves. Also in the IGT world when merging and comparing data, the goal is to maintain its semantic coherence while giving room for the representation of alternative and conflicting annotations.
In our project we started simple and focused on core annotations. During the prototyping phase we developed the following requirements for our solution:

1. We start with a small set of unrelated Toolbox projects representing different languages from a random sample of annotators working on unrelated projects.

2. We focus on core tiers: the text, the morpheme, the gloss and, if available, the part-of-speech tier.

3. We isolate the conceptual distinctions and separate them from the denotational variants.

4. Denotational variants are treated by string mapping from tags into tags.

[7] http://de.clarin.eu/en/, accessed 26.3.2014
[8] http://clarin-d.de/en/language-resources/weblicht-en, accessed, 26.3.2014
[9] http://www.ukp.tu-darmstadt.de/software/webanno/, accessed 26.3.2014
[10] http://typecraft.org/tc2wiki/Main_Page, accessed 20.02.2014
[11] http://www.cidles.eu/, accessed 26.3.2014

Our conversion project is part of a software development project which allows us to make additional design decisions, and to implement new types of functionality to facilitate complex conversion issues. We will not further describe that part of our project here, but in a nutshell, the project aims for the following: Since the definition of rules is never ”complete”, we will have a user interface to let users define mappings for those annotations that we could not map automatically. We expect that users will want the system to learn from their decisions.

Working with selected Toolbox projects confirmed what we expected, namely that the data is heterogeneous within any one project and even more diverse in comparison. Yet in some respect recommendations such as those formulated in the Leipzig Glossing Rules seem to function as conventional standards, which is of course helpful.

Table 1 below gives an overview over some of the data we have worked with in the first phase of the project, while examples (2) - (5) are representative for our Toolbox data.

Table 1 Overview over some of our first phase data sets

LanguageName	ISO-693-3	Source	Data carrier/Tool
Kaguru	kki	Endangered Language Archive http://elar.soas.ac.uk/deposit/0093	Toolbox
Paunaka	pnk	The Paunaka Documentation Project, University of Leipzig, Funded by ELDP, SOAS, University of London, Duration February 2011 to January 2013. courtesy of Lena Terhart	Toolbox
Baram	brd	Endangered Language Archive http://elar.soas.ac.uk/deposit/0007	Toolbox
Sri Lanka Malay	sci	https://corpus1.mpi.nl/ds/imdi_browser/?openpath=MPI515582%23 courtesy of Sebastian Nordhoff	Toolbox
Mandinka	mnk	Denis Creissels et Pierre Sambou, *Le mandinka : phonologie, grammaire, textes*, Paris : Karthala, 639 pages. courtesy of Dennis Creissels	WORD file-based

(2) Paunaka

\tx *kuinabu pueroinabu*
\mb *kuina-bu puero-ina-bu*
\ge *NEG-?IPFV be.able-IRR.NOM-?IPFV*
\ps adv-suff n-suff-suff

(free translation of this phrase not available)

(3) Kaguru

\t *Kamei howoluta kunyumbangwa*
\m *kamei ha-wa-lut-a ku-nyumba-ngwa*
\g *then PAST-2-go-FV 17-house:9/10*
\p adv tm-sm-v-fv sm-n-prn num

\f Then they went to one house

(4) Baram

\tx *jen nepna cə əbə*
\mb *jen ŋi-pəna cə əbə*
\ge *take=away NPST-must EPIS now*

\ft Now it should be brought.

(5) Sri Lanka Malay

\tx *saudara saudari*
\mb *su-*udara su-*u=nang=dheri*
\ge *PAST-**** PAST-****=DAT=ABL*
dan Muhammad
**d-an Muhammad*
*****-NMLZR ****

(free translation of this phrase not available)

Our conversion core can spread over 5 tiers, but not all tiers might be available: the text, the morph, the gloss, the part of speech and the free translation tier. Conversions of non-Latin scripts do not represent a problem when converting to TypeCraft, they map directly to the original script tier. IGT is broken up in phrases, but in our IGT test-set not for all of the phrases free translations are available. Generally, the part-of-speech tier is less frequent for IGT data than any of the other tiers. This is due to the fact that IGT is a conventional data type for which a text tier and a gloss tier plus a free translation was all that was required. Text tiers may indicate morpheme breaks. Note that IGT data reflects constraints which are not part of linguistics but result from the fact that annotating were done by hand on paper, or in standard word-processing programs. This made it difficult to keep morphemes and their annotations vertically aligned with the words in the text tier of which they were a part. Therefore extra symbols were introduced that distinguish the bound from the free morphemes. A hyphen attached to the last letter of a string indicated a prefix, and a hyphen attached to the first letter a suffix. More extra symbols were used too for example distinguishing clitics from affixes or annotations of inflectional from annotations of lexical properties. As conventional standards the use of these extra characters on the gloss tier is rather systematic, according to our observations, and therefore lends itself easily to a mapping algorithms. Important is that especially the gloss tier for standard IGT does not host elements of the same type - semantically or other. Translational glosses are mixed with symbolic glosses and the just mentioned extra set of symbols add even more linguistic distinctions. The same is in fact true for the part of speech tiers. Here one finds next to part of speech categories which are always assigned to words, also elements indicating bound morphemes of all sorts. The part of speech tiers of (2) and (3) exemplify this fact. The de facto use of annotation tiers thus reveals that annotators adhere to conventional standards to a certain extent, but that tiers nevertheless contain elements of quite heterogeneous categories.

For our project we map from Toolbox to TypeCraft. While for Toolbox the ”tier” is a basic concept, this is not the

case for TypeCraft where it is a unit of display. This point is probably best illustrated by going through a TypeCraft XML serialisation for the first two words of (3) above, here repeated as (6):

(6) Kaguru

\t *Kamei howoluta kunyumbangwa*
\m *kamei ha-wa-lut-a ku-nyumba-ngwa*
\g *then PAST-2-go-FV 17-house:9/10*
\p adv tm-sm-v-fv sm-n-prn num

\f Then they went to one house

(6) is mapped into the TypeCraft XML, as shown in the following listing:

```
<phrase id="41437" valid="VALID">
<original>
  Kamei howoluta kunyumbangwa ...
</original>
<translation>
  Then they went to one house ...
</translation>
<description/>
<word head="false" text="Kamei">
<pos>ADV</pos>
<morpheme
  baseform="kamei" meaning="then"
  text="kamei"/>
</word>
<word head="false" text="howoluta">
<pos>V</pos>
<morpheme baseform="ha" text="ha">
<gloss>PAST</gloss>
</morpheme>
<morpheme baseform="wa" text="wa">
<gloss>CL2</gloss>
</morpheme>
<morpheme
  baseform="lut" meaning="go"
  text="lut"/>
<morpheme baseform="a" text="a">
<gloss>FV</gloss>
</morpheme>
</word>
```

The TypeCraft XML serialisation preserves word as well as morpheme order. Word elements can only have one POS element but may have one to many morpheme elements. In a word element, morphemes that precede the morpheme that has its meaning attribute specified (thus counting as stem), are prefixes, the morphemes following the stem are suffixes. To preserve the linear order of morphemes under mapping is no problem, yet we need to distinguish stems from other morphemes, since as stems they have a value assigned to their meaning attribute. Under conversion we need to map all translational glosses from the TB gloss tier to this meaning attribute.

Now looking at the \mb tier in Toolbox, except for the Paunaka data (2), where the elements on the \mb tier seem to be morphs rather than morphemes and therefore map into the text attribute of a morpheme in TypeCraft XML, all the other IGT samples store morphemic forms in this tier which map to the baseform attribute rather than to the text attribute of the TypeCraft morpheme element.

Under the mapping from Toolbox to TypeCraft, elements from one tier can go to different places in the XML serialisation. Alternatively some elements of a tier might map into the values of one specific attribute while others need to be suppressed. This is the case for the part of speech tiers that we have encountered so far. To have a fix point for the mapping between different conceptual models in file formats, we therefore propose the introduction of "pivot tiers" in parallel to the GrAF pivot data model. This would allow users to decide which of the annotations from their tiers are mapped onto which of our pivot tier. Again, users stay within their conceptual universe, while the definition of a fix point allows us to set an anchor for the diversity of tier and annotation schemes. This definition is also compatible with the internal representation of GrAF graphs and the data structure types, as described above. We also propose an application-driven approach here, and only add new tier types to our set of pivot tiers as needed. In the context of prior IGT conversion projects a set of six pivot tier types (utterance, words, morphemes, part of speech, gloss and translation) was implemented. Discourse annotations and in depth semantic analysis might require additional tiers. We are currently looking for new applications that allow us to extend or stablise the set of pivot tiers in Poio API.

Another advantage of such a set of pivot tiers is that their definition would make it possible to link data sets via ISOcat entries. This crucial step to link data from language documentation projects in the Semantic Web is still in progress, partly because there was no agreement between the projects about tier and annotation labels, for example. We propose here an application-driven approach to create the appropriate categories, and hope to be able to derive a set of pivot tiers from de facto standards and thus clearly from within the wider community. We think that such a pragmatic step would significantly speed up the integration of archived language documentation data within the Semantic Web.

4. References

Dorothee Beermann and Pavel Mihaylov. 2013. Typecraft collaborative databasing and resource sharing for linguists. *Language Resources and Evaluation*, pages 1–23.

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, 33:23–60.

Jonthan Blumtritt, Peter Bouda, and Felix Rau. 2013. Poio API and GraF-XML: A radical stand-off approach in language documentation and language typology. In *Proceedings of Balisage: The Markup Conference 2013*, Montréal.

Peter Bouda, Vera Ferreira, and António Lopes. 2012. Poio API - an annotation framework to bridge language documentation and natural language processing. In *Proceedings of the Second Workshop on Annotation of Corpora for Research in Humanities*, Lisbon. Edies Colibri.

Baden Hughes Bow Catherine and Steven Bird. 2003. Towards a general model of interlinear text. In *Proceedings EMELD Conference 2003: Digitizing and Annotating Texts and Field Recordings*.

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

Sebastian Drude. 2002. Advanced glossing: A language documentation format and its implementation with shoebox. In *Proceedings of the 2002 International Conference on Language Resources and Evaluation*, Paris. ELRA.

Nancy Ide and Laurent Romary. 2004. International Standard for a Linguistic Annotation Framework. *Nat. Lang. Eng.*, 10(3–4):211–225, sep.

Nancy Ide and Keith Suderman. 2007. GrAF: A graph-based format for linguistic annotations. In *Proceedings of the Linguistic Annotation Workshop*, Prague, June. Association for Computational Linguistics.

Alexis Palmer and Katrin Erk. 2007. An xml format for interlinearized glossed texts. In *Proceedings of the Linguistics Annotation Workshop (LAW-07)*, Prague, Czech Republic.

Crowd-Sourced, Automatic Speech-Corpora Collection – Building the Romanian Anonymous Speech Corpus

Stefan Daniel Dumitrescu, Tiberiu Boroș, Radu Ion
Research Institute for Artificial Intelligence
Calea 13 Septembrie, nr. 13, Bucharest, Romania
E-mail: sdumitrescu@racai.ro, tibi@racai.ro, radu@racai.ro

Abstract

Taking the example of other successful initiatives such as VoxForge, we applied the concept of crowd-sourcing to respond to a particular need: the lack of free-speech, time-aligned, multi-user corpora for the Romanian language. Such speech corpora represent a valuable asset for spoken language processing application because they offer the means to (1) train and test acoustic models and (2) develop and validate various methods and techniques that are intended to enhance today's ASR and TTS technologies.

Keywords: crown-sourced, corpora collection, speech processing

1. Introduction

Dialog-enabled interfaces based on spoken language processing applications offer well-known benefits such as (1) enabling accessibility for the visually impaired or dyslexic people (TTS), (2) offering assistance to disabled people by enabling them to control various appliances through the use of voice (ASR) and (3) providing assistive technologies that help improve the general quality and comfort of life, because dialogue is the most common way of interaction between people and it is also a preferred alternative to the classic interfaces that require text input and physical interaction with touchscreens, keyboard and mouse. Having dealt with numerous issues regarding these types of technologies in the past, one major bottleneck still plagues research and the progress of multilingual ASR and TTS systems: the lack of freely available resources. According to the MetaNet White Paper Series (Trandabăț et al., 2011), through the analysis of the current state of resources of tools available for speech processing, Romanian was classified into the fragmentary support class along with 14 other European languages (fragmentary being the second lowest grade out of five).

Taking the example of other successful initiatives such as VoxForge[1], we applied the concept of crowd-sourcing to respond to a particular need: the lack of free-speech, time-aligned, multi-user corpora for the Romanian language. Such speech corpora represent a valuable asset for spoken language processing application because they offer the means to (1) train and test acoustic models and (2) develop and validate various methods and techniques that are intended to enhance today's ASR and TTS technologies.

The current paper presents an approach to harnessing the power of crowdsourcing in an attempt to solve a lack of specific Romanian speech resources. The authors detail the current level of the online platform's implementation, results, and further present their roadmap to significantly improve user participation and directly increase the collected corpus size.

More specifically, we will describe an online interactive platform that, under the umbrella of entertainment and through computer-user interaction will automatically collect such corpora. The concept of crowd-sourcing is not new, but to our knowledge an approach such as ours has not been proposed or done before in the field of spoken language processing.

2. Approach and general system architecture

TTS and especially ASR systems require large amounts of preprocessed and time-aligned speech corpora, the creation of which is a highly time and resource consuming process. We are addressing this problem by using crowd-sourcing with virtually zero labor maintenance costs in comparison to alternative methods. This approach has been successfully applied to other demanding tasks like image annotation. Furthermore, a free-speech, time-aligned, multi-user corpus is difficult to be found even for the English language (which is the best resourced amongst languages) and practically non-existent for Romanian.

The corpus creation process is based on many users, each speaking a few different sentences, thus obtaining many sentence/speech pairs. The speech segments will be automatically aligned to their corresponding texts using HTK (Young et al., 2002). The resulting speech corpus is directly usable, as HTK (http://htk.eng.cam.ac.uk/) provides good alignments (Woodland et al., 1999); there are several research papers that describe methods and techniques for fine-tuning the segmentation performed with HTK (Meen et al., 2005; Sethy and Narayanan, 2002, etc.) that we will use in our speech alignment process in order to obtain high quality corpora.

Furthermore, the platform itself is autonomous. It will continue to function indefinitely, improving itself with the corpora it collects as well as delivering progressively larger time-aligned and unaligned corpora.

The platform will provide the following:

1. **Time-aligned speech corpus**: used to train better ASR and TTS systems; also used to

[1] http://www.voxforge.org

automatically improve the platform itself.

2. **Free-speech un-annotated corpus**: a resource of unquestionable scientific value, can be further processed to create test sets.
3. **Improved ASR and TTS algorithms**: this development is based on our experiments with the proposed platform and the size of the gathered corpora;

3. A first step towards RASC – the results of a proof of concept platform

Before attempting to build a dedicated system, we needed a proof of concept that our idea is valid. As such, we constructed a relatively simple website that integrates the major components needed to create a speech corpus: a simple user-interface, a recording module, a storage module and a database of sentences. The website is freely accessible at http://rasc.racai.ro/ . So far, the platform offers more than 10000 sentences that can be read by the interested user.

The 10000+ available sentences have been automatically chosen to provide a balanced choice of triphones. The corpus from which the sentences were extracted was the sentence-split, full dump of the Romanian Wikipedia as of June 2012, because, belonging to the encyclopedic genre, it contains a wide range of domains and thus, of many different word types from specific terminologies. To make it easier to read, we have selected only short sentences (up to at most 20 words). This also helps with the prosody which can get somewhat unnatural when the reader reads out a long sentence due to unforeseen punctuation or even words, given that we must assume the reader will not take the time to read the entire sentence first and then voice it. The sentence also had to be properly terminated (with a full stop '.' – in Wikipedia sentences rarely end with other punctuation like question marks) and to begin with an uppercase letter. This restriction was also meant to ensure better prosody, and for us to be sure that we had full sentences and not sentence chunks. Additionally, we avoided any sentences that contained numbers because they have ambiguous normalized (expanded) forms and person names, because it is likely that they follow different letter-to-sound rules depending on their etymology. On this set of sentences we applied the triphones balancing algorithm described next.

To keep the number of triphones from each type as balanced as possible (a perfect balancing is not possible because there are triphones that are intrinsically rare) we have applied the following algorithm:

1. Compile an initial frequency of triphones from the whole corpus;
2. If a sentence contained a rare triphone (with a frequency below 100), keep it;
3. If a sentence contained only very frequent triphones (with frequencies over the H index of the initial distribution), discard it.
4. Default action: keep the sentence.

At the time of writing, the platform collected speech for 3242 sentences during the four month time-frame. As the platform does not require any user login to facilitate participation, we cannot uniquely identify users and as such we cannot compute the exact number of contributing users. We can approximate the current number of contributors to around 65 persons based on the selectable information they have provided: gender, age group, dialect and microphone type (though it may happen that two contributors used the same computer in sequence and are counted only once, or, a single user, after cleaning the browser's cookies is counted twice).

The following tables present the current statistics at the time of writing (late March 2014):

Male	**Female**
35.5%	64.5%

Table 1: Gender distribution.

18-35 years	**35-60 years**
71.7%	28.3%

Table 2: Age distribution.

Dialect	**Percent**
Bucovina	1.3%
Muntenia	67.7%
Moldavia	28.8%
Oltenia	2.2%

Table 3: Dialectal distribution.

Microphone type	**Percent**
Standard desktop	58.5%
Webcam microphone	0.7%
Laptop (embedded)	19.3%
Headsets with microphone	21.5%

Table 4: Microphone type distribution.

These statistics (and more) are generated live and can be found at the RASC Website[2]. On the same site the entire corpus is freely available for download.

We have drawn the following conclusions:

1. Without any online promotion, our platform attracted a fair number of users (that reached the platform by word-of-mouth mostly, and a few by web searches as the site was automatically indexed)
2. Online promotion would attract a significantly increased number of users; however, for a user to record more than a few sentences, we need to offer something in return → we need to make an interactive platform (proposed in step 2)
3. We have validated the recording and storage components; we have implemented and tested the time-aligning component and have successfully obtained a processed speech corpus from our available recorded sentences.

4. The second step towards RASC – a full-fledged, interactive speech collection platform

By further exploring experience of all crowd-sourced applications available on-line, one can easily see that users become more cooperative when they are presented

[2] http://rasc.racai.ro

with a reward for their efforts. The user feedback is highly dependent on the pragmatism of the reward. The basic version of the RASC platform only permits the user to listen to his voice while he is recording and to see small statistics regarding his contribution to the project. In order to improve our data collection rate, we propose to significantly extend our platform by offering its users a reward for their recording effort in the form of ludic activities. Thus, our target is two-fold: (1) exploit the fact that leisure is an important proficiency factor of online platforms and (2) create the "viral" effect (make our platform known to as many users as possible, through the use of social networks, portals and blogs).

Probably one of the attractive means to convince people to spend their time on an interactive platform is through the umbrella of games. The two types of collecting data in the game scenario are:

(1) Through the introduction of user adaptation before playing: the system will require the user to clearly read a few sentences containing as many varied diphones and triphones as possible. Even at this initial point we have a few welcomed outcomes: (1) the system learns the user's voice and (2) for each read sentence, we obtain a pair of text/speech segments that we further automatically process into our growing corpus.

(2) Though directly asking the user to speak out loud a given sentence: reading a text is a primary component of the game.

Thus, we propose three interactive games:

Game 1 - Voice mimicking: after voice adaptation, the system will allow the user to input text and play it back using the user's own voice, further allowing him to distort it using voice effects like pitch shift. The user can save or share the results.

Game 2 - Voice-morphing karaoke: the user will read lyrics (without singing them!) from various karaoke songs. We will modify his/her voice parameters to match that from the song, generating his voice on the recordings (just like a normal karaoke system). Another side-effect of this game is that the user will read lyrics, further growing the speech corpus.

Game 3 - Voice-chat with a computer robot (bot): this game will be a prototype bidirectional speech-to-speech system between the user and a computer bot. Based on arguably simple algorithms, there currently are hundreds of text chat bots online (e.g. cleverbot.com) that can sustain a mildly reasonable "conversation" with a user. We intend to package such a bot with a speech-to-speech interface. The result will simply make for fun experience, attracting users through its interface. We will also obtain a lot of speech samples used in the un-annotated speech corpus.

We must note that all the game ideas are not new, the methods and technologies used to power them (Natural Language Processing and Sound Processing) have been tried and successfully tested before in different scenarios, making a strong argument about the feasibility of such an approach.

5. Technical solutions for developing the platform

In this section we cover the current state of art technologies involved in the connected processes in the game platform we are developing. As previously mentioned, from the scientific point of view, the challenges come from speaker adaptation, automatic speech alignment and voice morphing.

Speaker adaptation is a technique used by speech recognition and speech synthesis application. There are several techniques used for this task, among which three model-based adaptation techniques have become prevalent over other methods: the Maximum A-Posteriori (MAP) adaptation of GMM-HMM parameters, the Maximum Likelihood Linear Regression (MLLR) (Leggetter and Woodland, 1995; Gales, 1998) of GMM parameters and multiple the Speaker Space (SS) (Kosaka and Sagayama, 1994) technique – which is not useful in our case. Currently, our system relies on the MLLR speaker adaptation technique implemented by the HMM Speech Synthesis System (HTS) (Zen et al., 2007).

As previously presented in the article, **automatic speech alignment** is used to determine the time relationship between a spoken utterance and its text equivalent. Based on the HMM parameters determined in the monophone

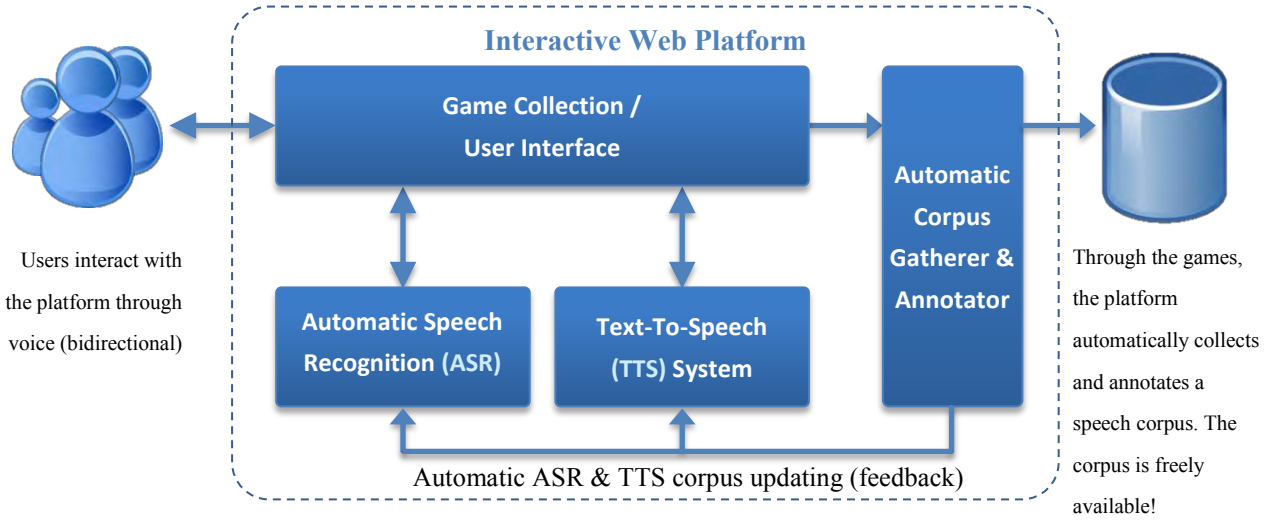

Figure 1 – System architecture

estimation phase of HTK we use dynamic programming to obtain the time-alignments that yield the lowest global cost for the given utterance.

Voice morphing is a process through which the parameters of a person's voice (in our case pitch and phoneme duration) are altered according to a pre-defined goal. We currently have two implementations: a time-domain Pitch Synchronous Overlap and Add (PSOLA) algorithm backed up by a dynamic pitch tracking algorithm and a re-synthesis method based on spectral approximation of Mel Generalized Cepstral Analisys (MGCEP), implemented by the Speech Signal Processing Toolkit (SPTK) (Imai et al., 2009)

6. Current state of the project

The platform implementation is underway and it will be made available once all the games are available, in order to avoid user bouncing because of an unfriendly experience with unfinished software.

1. Voice mimicking. Currently the mimicking module is completed and is undergoing tests to see how it scales to multiple users. We are trying to optimize the speed of the voice adaptation system, but it currently provides satisfactory performance for the production environment.

2. Voice morphing karaoke. Most of the basic components of the system are operational. We are currently working on integration of the modules and are checking karaoke licensing for some Romanian songs. Additionally work has been done on the voice tracing module that is designed to help us in extracting prosody related information from the original singer's voice. We estimate that this system would be operational within 3 months.

3. Voice chat. This last and most technologically challenging game is also under development. Structurally, the voice chat bot is split into three distinct systems: the Automatic Speech Recognition input module, the Natural Language Processing module and the Text-to-Speech output module. While not going into details regarding the ASR and TTS modules, (ASR can even be replaced by Google's ASR plugin for the Romanian language, if necessary), the NLP intermediary module has raised a number of issues. To begin with, we had two distinct options: either build it ourselves or use an already existing implementation. To build it ourselves is possible, but would require too much time that would be better spent on more important issues. However, using an already built chat engine is also difficult, the major challenge being that we need it to work for the Romanian language, while almost all current implementations are in English. (As a side note, it is surprising that even today, the Unicode standard needed to represent a few special diacritical characters in Romanian, is not widely supported in these chat bots).

We did find a few usable chat bots that are trainable and promise decent results (e.g. CleverBot trainable by www.cleverscript.com, which, apparently, has been conversed with more than 150 million times since its launch in 1997). These systems are rule-based and rely on basic string similarity measures to decide what option matches best the user's input in a given state. So, because we are not able to directly use a general domain chat bot, we have to create the rules table first, a process which we think can only be feasibly done automatically. CleverBot itself is supposed to learn from its conversations to expand its database. This is still an ongoing research item for the RASC platform.

7. Conclusions

Currently, Romanian speech research suffers from a lack of resources. In response, we designed and implemented an online, self-sustainable, self-improving platform. This crowd sourced platform is used to automatically obtain time-aligned free- and restricted-domain speech corpora (depending on the chosen sentences), with the long-term goal of improved ASR&TTS systems for the Romanian language.

The initial proof of concept platform has shown that such an approach is feasible. Within four months since launch, we built the necessary platform components and gathered over 3000 spoken sentences from volunteers.

The final platform will further involve its users, attempting to increase the time spent recording sentences by offering multimedia feedback through the use of a series of interactive games like Voice mimicking or Voice-morphing karaoke. Furthermore we intend to develop a Speech-to-Speech voice-chat with a computer bot (e.g. similar to cleverbot.com, but with ASR&TTS) as another interaction option with the platform.

At present, we do not know of such a game-enabled approach to collecting speech corpora. Our crowd-sourced platform will be automatic, sustainable (self-improving based on the collected corpus), attractive for its users through its interactive games, free and fast (login is optional, not required, for user convenience). Finally, our goal is to deliver to the Romanian speech community a free time-aligned speech corpus, a valuable resource for further research.

8. References

Gales, M. J. (1998). Maximum likelihood linear transformations for HMM-based speech recognition. Computer speech & language, 12(2), pp. 75-98.

S. Imai, T. Kobayashi, K. Tokuda, T. Masuko, K. Koishida, S. Sako, and H. Zen. 2009. Speech signal processing toolkit (SPTK), Version 3.3. http://sp-tk.sourceforge.net. (accessed March 2014).

Kosaka, T., & Sagayama, S. (1994, April). Tree-structured speaker clustering for fast speaker adaptation. In Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on (Vol. 1, pp. I-245). IEEE.

Leggetter, C. J., & Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech & Language, 9(2), pp. 171-185.

Trandabăț, D., Irimia, E., Barbu Mititelu, V., Cristea, D., Tufiș, D. (2012). The Romanian Language in the Digital Age / Limba română în era digitala. White Paper Series, Vol 16, Springer.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P. (2009). The HTK book (for HTK version 3.2) Cambridge university engineering department.

Woodland, P. C., Hain, T., Moore, G. L., Niesler, T. R., Povey, D., Tuerk, A., & Whittaker, E. W. D. (1999). The 1998 HTK broadcast news transcription system: Development and results. In *Proc. DARPA Broadcast News Workshop*, pp. 265-270.

Meen, D., Svendsen, T., & Natvig, J. E. (2005). Improving phone label alignment accuracy by utilizing voicing information. *SPECOM 2005 Proceedings*, pp. 683-686.

Sethy, A., and Shrikanth S. Narayanan. (2002). "Refined speech segmentation for concatenative speech synthesis." *INTERSPEECH*.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A., & Tokuda, K. (2007, August). The HMM-based speech synthesis system (HTS) version 2.0. In *Proc. of Sixth ISCA Workshop on Speech Synthesis*, pp. 294-299.

Crowdsourcing for the Development of a Hierarchical Ontology in Modern Greek for Creative Advertising

Katia Kermanidis[1], Manolis Maragoudakis[2], Spyros Vosinakis[3]
[1]Department of Informatics, Ionian University, 49100 Corfu, Greece
[2]Department of Information and Communication Systems Engineering, University of the Aegean, 83200 Karlovasi, Samos, Greece
[3]Department of Product and Systems Design Engineering, University of the Aegean, 84100 Ermoupoli, Syros, Greece
E-mail: kerman@ionio.gr, mmarag@aegean.gr, spyrosv@aegean.gr

Abstract

This paper describes the collaborative development of a hierarchical ontology in the domain of television advertisement. The language addressed is Modern Greek, one of the not widely spoken and not-richly-equipped-with-resources languages. The population of the ontology is achieved through collaborative crowdsourcing, i.e. players annotate ad video content through a multi-player videogame, implemented especially for this purpose. The provided annotations concern the ad content, its production values, its impact, and they constitute the ontology terms and concepts. Dependencies, correlations, statistical information and knowledge governing the ontology terms and concepts are to be revealed through data mining and machine learning techniques. The extracted knowledge constitutes the core of a support tool, i.e. a semantic thesaurus, which will help ad designers in the brainstorming process of creating a new ad campaign. Unlike existing creativity support models, that are static and depend on expert knowledge, thereby hurting creativity, the proposed support tool is generic in nature (as it is based on a collaborative crowdsourcing-based semantic thesaurus), dynamic and minimally restricting the brainstorming process.

Keywords: collaborative tagging, hierarchical advertisement ontology, Modern Greek

1. Introduction

The design of a new advertisement campaign is a creative process that relies to a large extent on brainstorming. The impact of advertising, as well as the creative processes it involves, has been studied thoroughly (Amos et al., 2008; Aitken et al., 2008; Hill and Johnson, 2004).

Several tools have been proposed and implemented for creativity support (Opas, 2008), and they are usually based on creativity templates (Goldenberg et al., 1999), decision making systems (Burke et al., 1990), wording schemata (Blasko and Mokwa, 1986), predefined concept associations (Chen, 1999; MacGrimmon and Wagner, 1994), or conversation-sensitive, picture-triggered brainstorming (Wang et al., 2010). Most of these tools use static non-expandable databases and hand-crafted associations, term-relations and transformations. Such static, passive, expert-dependent knowledge models can hurt creativity (Opas, 2008).

The present work describes part of the process of the development of a support tool for creative advertising. A hierarchical ontology that consists of concepts and terms related to television advertisement constitutes the core of the support tool.

The language of the ontology is Modern Greek, the only Indo-european language of the Hellenic language family that has survived. Modern Greek is one of the less widely spoken languages and not as richly equipped with linguistic resources (Gavrilidou et al., 2012).

While the structural backbone of the ontology (its concepts-categories and their hierarchy) is static and expert-dependent (provided by ad specialists), its content (the terms populating the categories) is provided dynamically through crowdsourcing. As crowd wisdom relies heavily on crowd size (Surowiecki, 2005), an annotation tool that is web-based, accessible to all, and manages to attract the annotator's attention and keep him/her entertained for a long time is implemented for the task. Namely, a novel multi-player action and strategy videogame has been implemented for collaboratively annotating the content of the ad videos. Serious games have been developed and used extensively as annotation tools due to their popularity and entertaining nature.

Furthermore, machine learning and data mining techniques will be employed to extract relations and dependencies among concepts and terms. The resulting ad support tool will be based on this semantic thesaurus of concepts, terms and relations and will provide ad designers with access to a rich library of video ads, with the ability to search the videos by content based on a query of keywords, to retrieve statistical and co occurrence data regarding the ads, and to access the consumers' (players') evaluation on the impact the advertisement had on them.

Unlike previous approaches to creativity support tool design, the resulting tool will not rely on hand-crafted rules and relations. The knowledge it relies on is data-driven, automatically derived, generic, dynamic, scalable, expandable, robust and therefore minimally restricting in the creative process and imposing minimal limitations to ideation or brainstorming.

Level 0	Level 1 Concepts	Level 2 Sub-concepts	Level 3 Sub-concepts	Level 4 Sub-concepts	Level 5 Sub-concepts	Concept Terms/Values
Root	Cinematography	Sound	Music/Song Recognisability			Yes/No/NA
			Song/Music Type			rock/classical/ethnic/jazz/soundtrack/dance/pop/NA/other
		Filming	Photography			picturesque/landscape/ airphoto/other/NA
			Style			realistic/cartoon/fairy tale/animated/fiction/other
	Location	indoors				home/office/work/store/other/NA
		outdoors				urban/rural/space/other/NA
	Ad Impact	Convincing power				very convincing/a little convincing/not convincing
		Opinion				positive/negative/neutral/NA
		Improvement suggestions				content/style/characters/location/music/photography/story/other/NA
	Production	Producer				producer name
		Director				director name
		Production value				amateur-like/professional/high-budget/other/NA
	Participating elements	Main character	Recognisability			famous/movie character/unknown/other/NA
			Type	Human	Gender	male/female/other/NA
					Age	baby/child/teenager/Youth/Middle-aged/Senior/Other/NA
					Occupation	employee/housework/farmer/scientist/public servant/other/NA
				Animal		pet/wild/exotic/small/big/other
				Inanimate		animated/cartoon/inanimate/other/NA
		Key-participants				tool/furniture/vehicle/appliance/gadget/other/NA
	Message Communication	Structure				single stand-alone episode/multiple stand-alone episodes/sequel/other
		Indirect critique on competition				yes/no
		Linguistic schemata				word game/metaphor/paraphrase/figure of speech/other/NA
		Humoristic elements				humoristic word/humoristic phrase/humoristic scenario/other/NA
		tag lines				line from movie/from song/proverb/historical saying/other/NA
		brand name				company brand name/region brand name/other/NA
	Product Type	Novelty				known/new/original/other/NA
		Product	Product	Food		bio/dietary/health poduct/other/NA
				Beverage		alcoholic/non-alcoholic/soft drink/juice/other/NA
				Electrical device	Device type	home appliance/electrical tool/widget/other/NA
					Energy class	A/B/C/D/E/other/NA
				Electronic device		phone/computer/laptop/tablet/GPS/camera/i-pod/i-pad

				Store		clothes/grocery/cosmetics/houseware/electronics/electrical/bookstore/e-shop/other/NA
				vehicle	Type	family/SUV/sports/motorcycle/small/other/NA
					Value	luxury/expensive/affordable /other/NA
				household		kitchenware/detergent/cosmetics/furniture/decorative/bathroom/linens/other/NA
			Service	telecommunications		mobile/internet/double play/triple play/other/NA
				TV		
				Banking		loan/investment/insurance/other
				Insurance		life/investment/other/NA
				Healthcare		yes/no
			Product and Service			yes/no
			Other			yes/no
		Target group				youth/housewifes/professionals/seniors/hobby/entertainment/men/women/parents/children/teenagers/other/NA
		Product Origin				mentioned/implied/not mentioned/other/NA

Table 1: The ontological backbone

The focus of the present paper is, first, the description of the advertisement ontology backbone structure, and, second, the presentation of *House of Ads*, the multi-player video game used as the annotation tool for collecting the ad video terms that will populate the ontology. Thereby a novel pathway to the collaborative development of a semantic thesaurus for a less widely spoken language is proposed.

In the remainder of this paper, Section 2 describes the structure of the ontological backbone and Section 3 presents the basic design features of the *House of Ads* videogame. Section 4 mentions the upcoming research challenges involved in the ad support tool development process, while Section 5 concludes the present work.

2. The Ontology Hierarchy

Certain widely-advertised product and service types were selected by marketing experts. Product types include food/beverage, electrical and electronic devices, stores, vehicles and household items, while service types include telecommunications, television, investing, healthcare and banking. After watching approximately three hundred ad spots of these particular product and services types, the experts then sketched the hierarchical structure of the advertisement ontology. It is designed to be scalable, so it can constantly be enriched and updated.

The ontology hierarchy organizes this information into five levels, excluding the root node, shown in Table 1. The last column is the list of possible terms populating the particular sub concept. The structure includes concepts/categories that are related to the ad content (product/service type, main character(s), other participants, location), its production values (quality, producer/director information), cinematography (sound, filming), message communication techniques (humor, tag lines, linguistic schemata, critique on competition), its target group and consumer impact (convincing power, opinion about the ad, improvement suggestions).

3. Crowdsourcing

Except for information regarding the producer and director of an ad video, which will be provided by ad experts (as this information is not known to the everyday consumer), the remaining concepts in the hierarchy are populated by everyday consumers through crowdsourcing techniques, i.e. collaborative annotation of the ad videos. Thereby the generic, minimal-human-expertise demanding and data-driven nature of the proposed support tool is ensured. It is evident that the performance of the tool relies heavily on the plethora of provided annotations; therefore the annotation tool needs to be attractive, engaging, fun and addictive. To this end, *House of Ads*, a browser-based game, has been designed and implemented especially for the task at hand.

Several toolkits exist for annotating text (Wang et al., 2010; Chamberlain et al., 2008), images (Catmaid, Flickr, Riya, ImageLabeler and Imagenotion (Walter and Nagypal, 2007)), or video, like VideoAnnex and YouTube Collaborative Annotations.

Von Ahn (2006), was the first to acknowledge that the high popularity degree of video games can be channelled towards more "serious" applications (e.g. educational or crowdsourcing). Several games for annotating video content (Diakopoulos, 2009) and for ontology population

(Kallergi and Verbeek, 2009; Siorpaes and Hepp, 2008) have been proposed.

The nature of textual data has not allowed for the design of genuinely entertaining gaming annotation software. The annotation of ad videos, however, inspires the design of software that can keep the player's interest and engagement level active for a very long time. The ontologies aimed at usually (e.g. Imagenotion) have a less intricate structure than the one described herein, i.e. they are not as deep, they contain fewer categories, and most categories are equally easy/hard to annotate. *House of Ads* is a more elaborate game platform, suitable for addressing the hierarchy of Table 1, which is more intricate and contains annotations of varying difficulty (e.g. identifying the linguistic schema that the ad uses as a message communication tool is harder than identifying the main character). The design of engaging game scenarios with usable and attractive interfaces has been recognized as one of the key challenges in the design of Games with a Purpose for content annotation (Siorpaes and Hepp, 2008).

3.1 House of Ads

House of Ads is an arcade-style, top-down and puzzle-like game, accessible to anyone. The fun elements of interaction and competition (Prensky, 2001) are ensured by including typical action-game challenges rather than simply adopting a quiz-like gameplay. *House of Ads* supports one to four players and includes two gameplay modes: the combat mode and the quiz mode (Figures 1 and 2 show some indicative screenshots of the two gameplay modes. The figures have been translated into English for comprehensibility purposes).

In the combat mode the player moves through rooms on a house floor. Rooms represent concepts and sub-concepts, while collectible items within a room stand for concept terms. On a TV screen the ad video is reproduced (the player can pause, rewind or fast-forward the video through a slider). The ad and the corresponding questions are selected from a database so that one player does not annotate the same ad twice, and so a sufficient number of answers is accumulated for all ads. Thus, the floor is dynamically created. When entering a room, the concept represented by the room appears on screen, as well as the term associated with each item in the room. The goal for the player is to collect as many items (provide as many terms) as possible that characterize the ad as quickly as possible and exit the house. Every player can block other players from reaching their goal by using a set of available weapons (fence, bombs etc.). The house platform supports different levels of difficulty; as the game evolves the house becomes more complicated, different floors appear with different rooms and items that demand more fine-grained, more difficult annotations and more complex attack/defence strategies. The player is free to switch to a different ad video at any time.

One significant challenge in designing annotation games is the real-time evaluation of the provided annotations and the scoring of the players. In *House of Ads* players are credited with a monetary amount with every item they choose. The harder the annotation term, the higher is the credited amount. In the case that multiple (more than one) players play the same ad simultaneously, or the ad has been played in the past, choosing an item that has already been chosen for the same ad means that the item is correct and the respective credited amount is free for the player to spend in real time (immediately after choosing the item). Otherwise, the correctness of the selection cannot be determined in real time and the credited amount is temporarily blocked until correctness is decided in a future game (a future player chooses the same item for the same ad). The respective blocked amount is then freed and the player, by logging in the game later, may see that his amount (score) has increased. The earned unblocked money can be used to buy new weapons or to improve the abilities of a player. If a player systematically collects incorrect items, his/her credibility drops and, if it falls below a certain threshold, the game ends and the player has to start over. Thereby cheaters can also be easily spotted.

Contradictive item selection (i.e. while playing simultaneously, players have selected contradicting items from a room for the same ad) is addressed in the quiz mode. The contradicted items are presented again to all the players simultaneously, and the player who first selects the correct answer receives the money that was blocked so far.

At the end of each stage, the player will be asked to comment on the convincing ability and the impact of the ad, providing his personal opinion. There is a large number of available questionnaires for the evaluation of advertising campaigns[1] [2], based on which this type of information may be provided.

3.2 Initialization Phase

In order to bootstrap the annotation process, and develop a ground truth benchmark that will enable game scoring and help establish the credibility of players, an initial annotation phase is carried out that does not involve the *House of Ads* game. In more detail, marketing students will be asked to watch five to six ad videos (different videos will be shown to each participant). After each video, they will fill out a questionnaire and the answers constitute the annotation tags. Each ad will correspond to one questionnaire, and each ad questionnaire will be filled out by more than one participant to allow for cross-checking and validation of the answers. The answers, once cross-checked, will form an initial set of correct annotations (a ground truth) that will facilitate the following gameplay phase.

[1] www.surveyshare.com/templates/televisionadvertisementevaluation.html

[2] www-sea.questionpro.com/akira/showSurveyLibrary.do?surveyID=119&mode=1

Figure 1: The 'House of Ads' combat mode

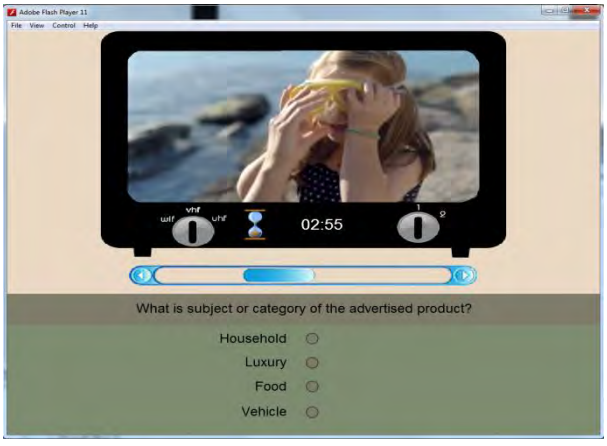

Figure 2: The 'House of Ads' quiz mode

4. Upcoming Challenges

Statistical analysis, data mining and machine learning schemata will be applied to the collected annotations. Ad concepts will constitute learning features and their annotation terms will constitute feature values. The goal is to

- retrieve correlation information. Co occurrence analysis between terms will help identify correlation values between various aspects of an advertisement.
- perform feature selection. Setting the product/service type as a classification label, the ad features that optimally characterize it and distinguish it from the remaining product/service types can be identified.
- extract rules from the data. Data-driven rules (e.g. association rules) may be extracted that will form patterns among the ad features.
- extract features that affect most the ad impact on the players/consumers. Given a specific product/service type, the identification of video features that play the most significant role in forming the consumers' opinion about the ad is of significant interest.
- perform prediction of consumer impact. Training a learning algorithm with already annotated videos, regarding their impact on the consumers, enables the prediction of the impact that new ad videos will have on the public.

The sparseness in the data, the large number of features and their heterogeneous nature will factor in deciding upon the learning techniques to be employed. Dimensionality reduction will be performed (Lee et al., 2010), and Support Vector Machines classifiers (Vapnik, 1995) will be taken into consideration as they are suitable for dealing with large feature spaces.

Visualization of instances and attributes on a new feature space, according to the most significant vectors will reveal clusters of similar ontology mappings together with advertised products. Support Vectors will be used to weight each attribute according to a certain criterion, such as correctness of user responses and could be used to

evaluate the most significant ontology nodes.

The aforementioned knowledge will be made accessible to professionals in the advertising domain through a user-friendly interface. The innovative generic nature of the tool will allow it to be flexible, scalable and adjustable to the end user's needs. Most importantly, unlike creativity templates, the generic nature does not impose any sort of 'mold' or template to the creative advertiser's way of thinking.

5. Conclusion

The present work describes how entertaining crowdsourcing may be employed for the collaborative development of a hierarchical advertisement ontology for a less widely spoken language, i.e. Modern Greek.

In a following research step, the provided annotations will be processed with data mining and machine learning techniques in order to reveal correlations between ad filming, products and consumer impact. The resulting semantic thesaurus and extracted knowledge will form the core of a support tool that will help ad designers during the brainstorming process of creating a new ad campaign.

6. Acknowledgements

This Project is funded by the National Strategic Reference Framework (NSRF) 2007-2013: ARCHIMEDES III – Enhancement of research groups in the Technological Education Institutes.

7. References

Aitken, R., Gray, B., and Lawson R. (2008). Advertising Effectiveness from a Consumer Perspective. *International Journal of Advertising*, 27 (2), pp. 279–297.

Amos, C., Holmes, G., and Strutton, D. (2008). Exploring the Relationship between Celebrity Endorser Effects and Advertising Effectiveness. *International Journal of Advertising*, 27 (2), pp. 209–234.

Blasko, V., Mokwa, M. (1986). Creativity in Advertising: A Janusian Perspective. *Journal of Advertising*, 15(4), pp. 43–50.

Burke, R., Rangaswamy, A., Wind, J. Eliashberg, J. (1990). A Knowledge-based System for Advertising Design. *Marketing Science*, 9(3), pp. 212–229.

Chamberlain, J., Poesio, M. and Kruschwitz, U. (2008). Phrase Detectives: A Web-based Collaborative Annotation Game. *Proceedings of I-Semantics*.

Chen, Z. (1999). Computational Intelligence for Decision Support. CRC Press, Florida.

Diakopoulos, N. (2009). Collaborative Annotation, Analysis and Presentation Interfaces for Digital Video. Ph.D. Thesis. Georgia Institute of Technology.

Ericsson, K., Simon, H. (1993). Protocol Analysis: Verbal Reports as Data (2nd ed.). MIT Press, Boston.

Gavrilidou, M., Koutsombogera, M., Patrikakos, A., and Piperidis, S. (2012). Language Technology Support for Greek. The Greek Language in the Digital Age, White Paper Series, pp.54-70. Springer.

Goldenberg, J., Mazursky, D. and Solomon, S. (1999). The Fundamental Templates of Quality Ads. *Marketing Science*, 18 (3), pp. 333–351.

Hill, R., Johnson, L. (2004). Understanding Creative Service: A Qualitative Study of the Advertis-ing Problem Delineation, Communication and Response (APDCR) Process. *International Journal of Advertising*, 23(3), pp. 285–308.

Kallergi, A., Verbeek, F. J. (2009). Onto-Frogger: Playing with Semantic Structure. *SWAT4LS, CEUR Workshop Proceedings*, Vol. 559, CEUR-WS.org

Lee, M., Shen, H., Huang, J. Z. and Marron, J.S. (2010). Biclustering via sparse singular value de-composition. *Biometrics*, Vol. 66, pp. 1087-1095.

MacCrimmon, K., Wagner, C. (1994). Stimulating Ideas Through Creativity Software. *Management Science*, 40 (11), pp. 1514–1532.

Opas, T. (2008). An Investigation into the Development of a Creativity Support Tool for Advertising. PhD Thesis. Auckland University of Technology.

Prensky, M. (2001). Fun, play and games: What makes games engaging? *Digital Game-based Learning*, pp. 1-31.

Siorpaes, K., Hepp, M. (2008). Games with a Purpose for the Semantic Web. *IEEE Intelligent Systems*, pp. 1541-1672.

Surowiecki, J. (2005). The Wisdom of Crowds. Anchor Books. pp. xv. ISBN 0-385-72170-6.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. Springer-Verlag. ISBN 0-387-98780-0.

von Ahn, L. (2006). Games with a Purpose. *IEEE Computer*, 39 (6), pp. 92-94.

Walter, A., Nagypal, G. (2007). IMAGENOTION – Collaborative Semantic Annotation of Images and Image Parts and Work Integrated Creation of Ontologies. In *Proceedings of the 1st Conference on Social Semantic Web (CSSW*). LNCS, pp. 161-166, Springer.

Wang, A., Hoang, C. D. V., and Kan, M. Y. (2010). Perspectives on Crowdsourcing Annotations for Natural Language Processing. Technical Report (TRB7/10). The National University of Singapore, School of Computing.

Wang, H-C., Cosley, D. and Fussell, S. R. (2010). Idea Expander: Supporting Group Brainstorming with Conversationally Triggered Visual Thinking Stimuli. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW). Georgia, USA (2010)