

## 'Smallcodes', a Unified Computational Linguistics Toolbox for Minority Languages

Carlo Zoli

Smallcodes S.r.L., Firenze, Italy

*L'articolo presenta la toolbox di Smallcodes, uno strumento web indispensabile per le lingue minoritarie e valuta la sua possibile applicazione al Tamazight. Lo strumento è stato sviluppato per permettere anche alle minoranze linguistiche di incrementare la loro presenza nel cyberspace e quindi passare da una realtà solo orale e al mondo scritto di Internet. La toolbox unificata di Smallcodes è composta da diversi moduli integrati tra loro: un dizionario che tenga conto delle esigenze speciali (ad esempio i caratteri) delle lingue di minoranza; un correttore ortografico studiato per la diversità dialettale; una sezione di terminologia che aiuti la pianificazione di neologismi. Crediamo infatti che per permettere alle lingue meno usate di sopravvivere in un mondo ultra-connesso dove la maggior parte degli input sono mediati dal web e dalla lingua scritta, sia necessario dotare queste lingue di un "kit di sopravvivenza" per fornire loro gli stessi strumenti e risorse delle maggiori lingue nazionali.*

### Introduction

Living in a hyper-connected world means that most of the inputs we receive daily are mediated by the Web and they are thus mainly written and not orally transmitted. Therefore only those inputs with the right features, such as a good visibility, the right patterns and an understandable language are winners in the vast world of the written media. Contents in English language are therefore the most widely spread because they most often meet these parameters. It is clear that, if we want to save and preserve minority languages, we must necessarily let these lesser-used languages have access to the tools and resources of the same technological level as those of "bigger" languages. This can be done only sharing experience, expertise and costs between minorities.

We believe that a computational linguistics toolbox can offer a unique solution for minority languages to increase their presence in the written media, among which the cyberspace is and will be the most pervasive. Basically, a first set of resources that are needed to undertake the path to a complete NLP toolbox are, not necessarily in this order, lexica, morphological analyzers / synthesizers, phonetic

similitude patterns, neology / terminology thesauri, corpora and parsers (Scannel, 2011).

The aim is to create a **single** and **integrated** platform for language technology dedicated to minority languages. The big novelty in the design of this instrument is that we put all the tools together in one single highly interoperable box. This unified and comprehensive toolbox is designed for the creation and management of electronic language resources, to respond to the following needs (here we use the classic tripartition of corpus, status and acquisition planning introduced by Heinz Kloss (1976):

- Corpus planning: such a toolbox is necessary to study minority languages both in their internal variability and from a standardized point of view.
- Status planning: the tools for neology provide a rapid introduction in the world of administration and education whereas the orthographical and auto-completion tools are intended to give an easy means in order to move from the local oral variety to the standard written form.
- Acquisition planning: the toolbox aims at providing language students with a comprehensive tool made for acquiring the language and practicing its use in everyday life.

Designing such a tool for minority languages is in some ways more difficult than making it for an official national language, because only the latter has an ancient and well-established written tradition. And yet this action is even more necessary, because computational linguistics for minority languages is not an accessory “luxury”, but it is a necessary (unfortunately not sufficient) condition to survive in a globalized world (Dell’Aquila, Iannàcaro, 2011).

Smallcodes platform has an explicit eco-linguistic intent because it wants to create interest around poorly investigated topics by mainstream universities. In fact, Smallcodes works as a commercial firm when working with industrial, commercial and government partners, but we also work as a non-profit organization when collaborating with non-profit, volunteer, ONG partners or when participating in co-funding of national or international projects.

## **A first-level toolbox**

The bare minimum to ensure any language a scientific and systematic presence in the written world is made of:

- A lexicon.
- A spell-checker tool.
- A terminology module.

The final aim is the maintenance and/or re-integration of language in society and these tools are the necessary means to develop the chain of corpus planning → status planning → acquisition planning. It must be clear that these technologies are just means: the main purpose is in fact the maintenance of the language in social

life. But in the contemporary world the social use passes through the written form, and the written form passes through technology.

Minority languages lack of those fundamental IT tools that allow scientists to study other bigger languages (i.e. “terminology extractors”, or “resumé automatique”, or “question answering”). In fact, when we talk about minority, small, lesser used languages, we have to face not only their (relatively) scarce presence in the cyberspace, but also the quality of this presence. We are talking in terms of sociolinguistic quality, not about the literary or the aesthetic value.

It may be curious that an institution like ours that has devoted its life to preserve linguistic diversity is such a strong defender of standardisation. Is not standardisation an enemy of natural autochthonous languages as much as colonialism or “English glottofagy”? On the contrary, we must be realistic: there are very powerful tools that have been developed for standardised languages which have meant years of development and millions of investments. It would be crazy not to use them; it would be fool to think that Google Translator, or the incredible results of the search engines or of the semantic Web would have been achieved if English had not been... English as a world language [Zoli, 2012 (1)]. If we want to foster our small languages in the real world, and in the cyberspace (the two things will tend to be asymptotically the same) we must be as “dwarfs sitting on the shoulders of giants”, as we say in Italian. And to do this we have to pay a little price: giving the language a common standard written form. This is absolutely not sufficient, but it is terribly necessary and, in most of the contexts where we work, it is not obvious at all.

Would it be sensible to go to Microsoft and ask them to localize Windows in 3 different Sardinian languages? Would it be conceivable to go to Shēnzhèn at Apple Developers meeting and ask for 5 different Romansch forms of iOS, or of Siri? We must make all the possible profit from these global instruments as Google or Siri and sit on the shoulders of these giants.

In this respect, having a look at Gartner’s hype cycle as of July 2012 (*Figure 1*) is of great importance. Many expectations concern technology languages, but can we think about information extraction, or integration with calendars or smart phones, if people do not agree on how to write “Thursday”.



In order to meet technological expectations is therefore necessary to promote the written use of the language (which is necessarily electronic and not handwritten). The writing is - just in terms of status planning - often in the domains of administration, bureaucracy and schools (Dell'Aquila, Iannàcaro, 2011:98ff). These are the more permeable areas to language policy, while those of literary creativity are often oriented towards localisms and are more reluctant to accept a standardized script.

The written use must be then promoted and facilitated. In this sense, the advantage of minority languages is that very often the official institute for the defence of the language is unique and known by many: the Institut Royal de la Culture Amazighe for Tamazight or the Istitut Cultural Ladin for Ladin language are authority whose prestige is recognized by most of the speakers of the target minority language.

## The involved fields

Not every research field of computational linguistics can be involved at the beginning of the process. At least in the first stage, it is necessary to focus on fewer and simpler areas of interest, having clear in mind that, especially for normal users, for school pupils and teachers, for a non-specialist audience *a fairly-good 'something' is much better than a perfect 'nothing'* (Scannell, 2011). Preliminarily, it is necessary to have a unifying - better than "unified" writing system (field: Writing). Then, the following step is represented by the creation of a common-use dictionary and (if this is possible with budget and workforce available) a dialectal dictionary of local varieties, plus the retrieval of studies and corpora on terminology, neologism and modernization of the lexicon (field: Dictionaries). It is then very useful to have spell-checking instruments such as online spell-checkers (available online and for Microsoft Word and/or Open Office) and automatic correction systems in all these cases (field: Writing aid). A further effort is the creation of corpora and archives of ancient texts, the production of e-books, audio books and didactic material online with downloadable and printable files off-line (fields: Digital libraries / School teaching). An optional subsequent action is the creation of a Web-TV and of free-press magazines and newspaper which is mediated and generated by the Web (field: Digital instruments of mass communication).

The chart below (*Figure2*) shows the fields of interest to be exploited for minority languages according to the urgency of the action to be taken (ranging from green: very urgent – yellow: possible – red: optional).

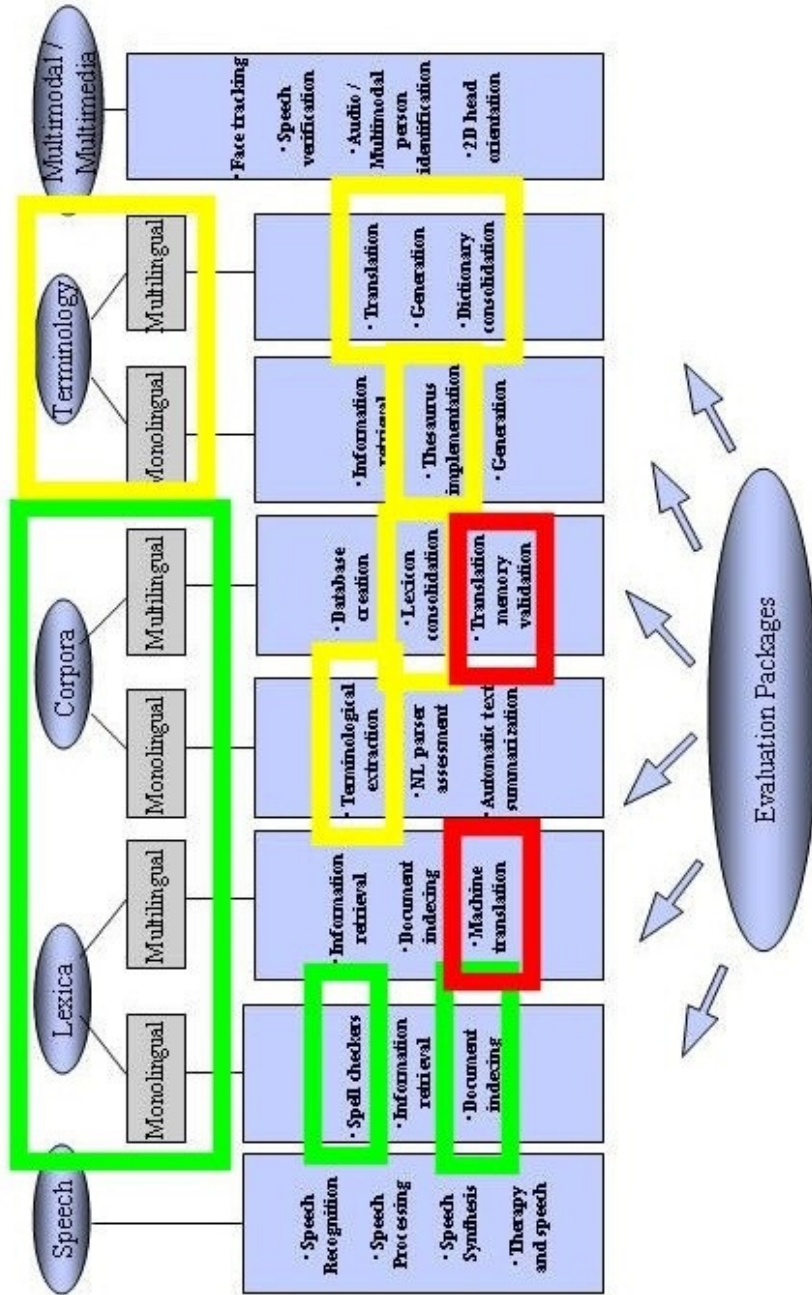


Figure 2: Computational linguistics packages for minority languages (Scamnel, 2008)

After having collected enough lexical material (**preliminary step**), it is then possible to plan the dictionary (**first step**). In fact, lexical lists of various kinds are the necessary condition in order to set up the dictionary. They can be wordlists of local or global language (i.e. conforming to local varieties of the language or to the standardized spelling); they can also be imported from informal databases and being the result of an OCR or parsing of ancient dictionaries.

The figure below (*Figure 3*) shows an example of 'standardizing' dictionary with registration of local varieties. Here is the extreme case of the entry otóbro ('October') which has around 150 different phonetic realizations ascribable to three consonantal macro-phenomena (1. maintenance of etymological t; 2. palatalization of t > c. 3. loss of b). As it can be seen, the standard forms have been chosen among those forms which are more "etymologically regular" (Lurà et al., 2009). Then (*Figure 4*), we have the same entry in a human-readable form (actually an XML + CSS which can be easily imported in a professional publishing tool as Adobe Indesign, see *Figure 6*); *Figure 5* shows the XML of fig 4 in the classic machine-readable form.

th/isi/applications/dictionary/entry/previewSvIt.jsp?\_VP\_V\_ID=756612764&CM=ED&\_VP\_OBJ\_ID=...

otóbar, otóber, utóbar, utubar

**\*otóbro**

**I. Sv.it.**

auciára (**Ludiano**), cióuri (**Chironico**), cióuru (**Soazza**), cióuri (**Chironico**), ciár (**Breg.**), ciuar (**Soprap.**), ciár (**Breg.**), ciuar (**Soglio, Soprap.**),  
dición (**Corippo**), dició (**Verz.**), diciór (**Cal.**), dición (**Biasca**), diciú (**Brione Verz.**), gerra (**Verz.**), diciú (**Bodío**), diciú (**Cugnasco, Verz., Gordola**),  
docióra (**Aquila**), docióri (**Personico**), dución (**Intragna**), dució (**Campo VMa.**), duciú (**Berzona**), giubar (**Stampa**), ió (**Montecarasso**),  
iciór (**Cal.**), ió (**Gnosca, Biasca, Preonzo, Lodrino**), ió (**Bodio**), ió (**Malvaglia**), ió (**Verz.**), ió (**Gorduno**), ió (**Verz.**),  
oció (**S. Antonio, Ronco s. Ascona**), oción (**Intragna**), oció (**Mergoscia**), ocióra (**Leontica, Campo Ble.**), ocióri (**Preonzo, Riv., Lev.**), ocióri (**Biasca**),  
ocióri (**Verz.**), ocióru (**Malvaglia**), ocióra (**Aquila**), ocióri (**Chironico, Personico, Preonzo, Pollegio**), ocióra (**Aquila**), ociuar (**Castasegna**),  
ociura (**Dongio, circ. Castro**), ociuri (**circ. Quinto**), ocón (**Intragna**), ocóri (**Iragna**), ocúar (**Castasegna**), ocúar (**Gorduno**),  
otóbra (**circ. Taverne, Malc.**), otóbre (**Lodrino, Iragna, Medeglia, Airola, VMa., Verscio, Cavigliano, Losone, Lug., Poschiavo**),  
otóri (**Intragna, Gorduno, Caveragno, circ. Maggia**), otóbro (**Carasso, Lumino, S. Abbondio, Lug., Roveredo Grig., Soazza**),  
otóbru (**Gorduno, Arbedo-Castione, Lug., Stabio**), otóri (**Magliasina**), otóra (**circ. Breno, circ. Sessa**), otóre (**Sonvico**), otóri (**Isona, S. Vittore**),  
otúar (**Posch.**), otubar (**Castasegna, Arbedo-Castione, Bedretto, Brontallo, Rovana, Loc., Magliaso, Lugano, Vacallo, Cabbio**),  
otuber (**Gordola, Menzoni, Brissago, Brione s. Minusio**), otubre (**Broglio, Melezza**), otubru (**Gamb.**),  
ució (**Planezzo, Cevio, Ronco s. Ascona, Mergoscia, Gordola, Ons.**), ució (**Mergoscia**), ució (**Sonogno**), ución (**Brione s. Minusio**),  
ución (**Brione s. Minusio**), ució (**S. Antonio, Campo VMa., Mergoscia, Verz.**), ució (**Verz.**), ucióra (**circ. Malvaglia, Ponto Valentino**), ucióri (**Lev.**),  
ucióri (**Biasca**), ucióru (**Biasca**), ucióra (**Malvaglia**), ucióra (**Aquila**), ucióri (**Cavagnago, Anzonico**), ucióri (**Cavagnago, Sobrio, Anzonico**),  
uciú (**Cevio**), uciuar (**Poschiavo**), uciura (**Ble.**), uciuri (**Lodrino**), uciuru (**Malvaglia**), uciú (**Cevio, Cerentino, Auressio**), uciú (**Lavertezzo**),  
utóber (**circ. Giubiasco, Ravecchia, Ons., Indemini, Mesocco, Bedano**), utóbra (**Isona, Novaggio**), utóbre (**Bedano, Peccia, Verscio, Osco**),  
utóbru (**Campo VMa., Ons.**), utóbru (**Giubiasco, Daro, Torricella-Taverne, Cademario, Grancia**), utóra (**Novaggio**), utuar (**Posch.**),  
utuber (**Montecarasso, Medeglia, Sementina**), utubre (**Sigirino**), utubre (**Medeglia, Robasacco, Caveragno**), utubri (**Linescio, Campocologno**),  
utubru (**Medeglia, Bosco Lug., Mendr.**)

Figure 3: An example of a standardizing dictionary with registration of local varieties



← [siti-edu.ch/siti/applications/dictionary/entry/xml/xmlReult.jsp\\_VP\\_V\\_ID=4394927&\\_VP\\_OBR\\_ID=71671&LANGUAK](https://siti-edu.ch/siti/applications/dictionary/entry/xml/xmlReult.jsp_VP_V_ID=4394927&_VP_OBR_ID=71671&LANGUAK)

## \*otóbro (dial. Sv.It.)

capo-lemma: **otóbro**

*tutte le forme di dial. Sv.It.:* **otóbar, otóber, utóbar, utubar**

*varianti locali:* **aucóira** (Ludiano), **cióuri** (Chironico), **cióuru** (Soazza), **cióuri** (Chironico), **ciúar** (Breg), **ciúar** (Soprap), **ciúar** (Soglio, Soprap), **dicón** (Carpino), **dicóio** (Verz), **dicíor** (Cal), **dicíon** (Bassca), **dicíou** (Briore Verz, Gera Verz), **dicíou** (Bodio), **dicíou** (Cugnasco, Verz, Gordola), **docóura** (Aquila), **docóuri** (Personico), **dicúon** (Intragna), **dicúio** (Campo Vlla), **dicúiu** (Bercina), **giubar** (Stampa), **icóio** (Montecarasso), **icóir** (Cal), **icóiri** (Grosca, Bassca, Preanzo, Lodrino), **icóirou** (Bodio), **icúuru** (Malavaglia), **icúuu** (Verz), **icóiri** (Gorduno), **incóio** (Verz), **ocóio** (S. Antonio, Ronco s. Ascona), **ocóion** (Intragna), **ocóio** (Mergosia), **ocóira** (Leontica, Campo Ele), **ocóiri** (Preanzo, Riv, Lev), **ocóiri** (Bassca), **ocóiri** (Verz), **ocóiru** (Malavaglia), **ocóura** (Aquila), **ocóuri** (Chironico, Personico, Preanzo, Pollegio), **ocóvra** (Aquila), **ocúiar** (Castasegna), **ocúra** (Dongo, circ. Castro), **ocúiri** (circ. Quinto), **ocón** (Intragna), **ocón** (Iagna), **ocúiar** (Castasegna), **ocúvar** (Castasegna), **otóbor** (Gorduno), **otóbra** (circ. Taveme, Malc), **otóbre** (Lodrino, Iagna, Medeglia, Airolo, Vlla, Verscio, Carigliano, Losone, Lug, Poschiano), **otóbrri** (Intragna, Gorduno, Caregno, circ. Maggia), **otóbro** (Carasso, Lumino, S. Abbondio, Lug, Roveredo Grig, Soazza), **otóbru** (Gorduno, Abedo-Castione, Lug, Stabio), **otóri** (Maglissina), **otóvra** (circ. Brieg, circ. Sessa), **otóvre** (Somico), **otóvri** (Isone, S. Vittore), **otúar** (Posch), **otúbar** (Castasegna, Abedo-Castione, Bedretto, Bronálio, Rovana, Loc, Magliaso, Lugano, Cabbio), **otúber** (Gordola, Merzono, Bissago, Briore s. Minusio), **otúbre** (Broglio, Melezza), **otúbruu** (Gamb), **ucóio** (Pianezza, Celio, Ronco s. Ascona, Mergosia, Gordola, Ono), **ucóio** (Mergosia), **ucóio** (Sonogno), **ucóion** (Briore s. Minusio), **ucóion** (Briore s. Minusio), **ucóio** (S. Antonio, Campo Vlla, Mergosia, Verz), **ucóio** (Verz), **ucúia** (circ. Malavaglia, Pomo Valerino), **ucóiri** (Lev), **ucóiri** (Bassca), **ucóiri** (Bassca), **ucóiru** (Malavaglia), **ucóúra** (Aquila), **ucóuri** (Caragnago, Anzonico), **ucóivri** (Caragnago, Sobrio, Anzonico), **ucúci** (Cevio), **ucúia** (Poschiano), **ucúia** (Ele), **ucúiri** (Lodrino), **ucúiru** (Malavaglia), **ucúiu** (Cevio, Carlenho, Aressio), **ucúiu** (Lavenazzo), **utóber** (circ. Giubiasco, Ravechia, Ono, Indemini, Mesocco, Bedano), **utóbra** (Isone, Novaggio), **utóbre** (Bedano, Peccia, Verscio, Oso), **utóbrri** (Campo Vlla, Ono), **utóbruu** (Giubiasco, Dara, Torricella-Taveme, Cademario, Grancia), **utóvra** (Novaggio), **utúar** (Posch), **utúber** (Montecarasso, Medeglia, Semolina), **utúbra** (Sigrino), **utúbre** (Medeglia, Robasacco, Caregno), **utúbrri** (Inverso, Campocologno), **utúbruu** (Medeglia, Bosco Lug, Mendr)

**S. It.**

**1** Ottobre (dial. Sv.It.) ottobre

**2** Autunno (Sonogno, Landarencia) autunno

*locuzioni:*

**ná in otóbar** Andare in calore, del caprone (**Vira Gamb**) andare, essere in calore

**otóbro cocóber** Formula reduplicativa con funzione enfatica, che compare in alcuni detti e proverbi (**Isone, Gamb, Pura, Gandria**) *rimandi da cucóbra*

*rimandi:*

**bocó** dotóber (Castaneda)

**Madóna** dotóbar (dial. Sv.It.)

**scurpi** dotóber (Ascona)

**te** dútuar (Poschiano)

Figure 4: The same entry in a human-readable form<sup>1</sup>

<sup>1</sup>Please note that the current tendency in normalization is to suggest a single graphic form but to allow free choices in local meanings and lexical types. The image shows the lexical type *otóbro* ('October') which in some places means 'autumn, fall'. Symmetrically, for the concept of "October", we could have many other lexical types, such as 'Month of St. Martin' or 'Month of chestnuts'.

```

<?xml-stylesheet types="text/css" href=".../css/xml/dictionaryFrontend.xml.css"?><LEMMA xmlns:html="http://www.w3.org/1999/xhtml" >
<LEMMA ID="71671" IS ALTERNATIVE="false">
<DIZIONARIO TITOLO>
<FORMA LE ENTRY TYPOLOGY="NOT ATTESTED" IS_INVERSE="false">otóbro</FORMA_LE>
<LINGUE_LE>(dial. Sv.It.)</LINGUE_LE>
</DIZIONARIO TITOLO>
<FORMA FOR_SEARCHING_LE>otobro</FORMA_FOR_SEARCHING_LE>
<CAPOLEMA>
<CAPOLEMA LE LABEL>capo-lemma:</CAPOLEMA LE LABEL>
<CAPOLEMA LE ENTRY TYPOLOGY="NOT ATTESTED">otóbro</CAPOLEMA_LE>
</CAPOLEMA>
<TUTTE LE FORME HIDE="false">
<TUTTE LE FORME LABEL>tutte le forme di</TUTTE LE FORME LABEL>
<TUTTE LE FORME LANG>dial. Sv.It.:</TUTTE LE FORME LANG>
<TUTTE LE FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">otóbar</TUTTE LE FORME_DESCR>
<TUTTE LE FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">otóber</TUTTE LE FORME_DESCR>
<TUTTE LE FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="false">otóbar</TUTTE LE FORME_DESCR>
<TUTTE LE FORME_DESCR HIDE="false" ULTIMA_FORMA_VISIBILE="true">utubar</TUTTE LE FORME_DESCR>
</TUTTE LE FORME>
<VARIANTI_LOCALI>
<VARIANTI_LOCALI LABEL>varianti locali:</VARIANTI_LOCALI_LABEL>
<VARIANTI_LOCALI_DESCR>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">
<FORMA_VL>aucióra</FORMA_VL>
<LINGUE_VL>(Ludiano)</LINGUE_VL>
</VARIANTE_LOCALE>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">
<FORMA_VL>cióuri</FORMA_VL>
<LINGUE_VL>(Chironico)</LINGUE_VL>
</VARIANTE_LOCALE>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">
<FORMA_VL>cióuru</FORMA_VL>
<LINGUE_VL>(Soazza)</LINGUE_VL>
</VARIANTE_LOCALE>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">
<FORMA_VL>cióvri</FORMA_VL>
<LINGUE_VL>(Chironico)</LINGUE_VL>
</VARIANTE_LOCALE>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">
<FORMA_VL>ciúar</FORMA_VL>
<LINGUE_VL>(Breg.)</LINGUE_VL>
</VARIANTE_LOCALE>
<VARIANTE_LOCALE LOC_VARS GEO_ORDER="false">

```

Figure 5: Machine-readable output of the same entry in a LMF (Francopoulo et al., 2006), compliant XML-schema

abit → abat

**abitá**, *abitaá*; *abitè* (Lev., Mesocco, Soglio), *abitèe* (Lodrina, Brione Verz., Gerra Gamb.), *abitèr* (Vicosoprano), *bitaa* (Carasso, Gordio, Lavertezzo), *bitè* (SottoP.), *bitèr* (Stampa, Casaccia) v. **SIGN** Abitare, risiedere ◊ occupare, dimorare: di spirito (Breg.) ◊ bazzicare, frequentare (Lavertezzo).

abitaá → abitá

abitabal, abitabel → abitabil

**abitabil**; *abitabal* (Bondo), *abitabel* (Leontica), *abitèbal* (Bondo) agg. **SIGN** Abitabile.

**abitacol** (Roveredo Grig.), *bitacol* (Landarenca), *bitacol* (Carasso, Roveredo Grig.), *bitaero* (Biasca) s.m. **SIGN** Abituro, edificio misero e in cattivo stato (Biasca, Roveredo Grig., Landarenca) ◊ cascina sull'alpe (Carasso).

**abitamènte** s.m. **SIGN** Abitazione, fabbricato (Sonvico).

**abitant**; *abitante* (Cimadera, Sonvico), *abitènt* (Ludiano, Rossura, Gerra Gamb.) s.m. **SIGN** Abitante.

abitante → abitant

abitè → abitá

abitèbal → abitabil

abitèe → abitá

abitènt → abitant

abitèr → abitá

abitìgn, abitín → abatín

abitúá → abitúá

**abitúá**, *abitúá*, *abituáa*, *abitúaa*; *abitùe* (Chironico), *abitùe* (Lev., Soglio), *abitùee* (Brione Verz., Gerra Gamb.), *abituvá* (Loco), *abitùvè* (Giornico), *abütüvá* (Augio), *betüaa* (Sementina), *bitüá* (Semione, SottoC.), *bituaa* (Brissago), *bitüaa* (Sementina), *bitùe* (Ludiano, Prato Lev.), *bitùee* (Olivone), *bitüvá* (Grancia), *bütüá* (Balerna), *ebitüaa* (Biasca) v. **SIGN** Abituare.

abituáa, abitüaa → abitúá

abitudan, abitüdan, abitudin → abitüdin

**abitüdin**, *abitudin*, *abitüdüna*, *abitüdüna*; *abetuden* (Lumino), *abitudan* (Carasso), *abitüdan* (Linescio), *abitüdüne* (Breno), *abitüdüna* (Aquila), *betüdüne* (Sementina), *bitüdin* (Grancia), *bitüdüne*, *ebitüdüne* (Sementina) s.f. **SIGN** Abitudine.

abitüdüna, abitüdüna, abitüdüne → abitüdin

abitùe, abitùe, abitùee → abitúá

abitüdüna → abitüdin

abitul → arbitri

abituvá, abitùvè → abitúá

abniscia → alniscia

abòligh → diabòligh

**aboná**, *abonaa*, *abuná*, *abunaa*; *abonè* (Lev., SottoP.), *abonèe* (Brione Verz., Gerra Gamb.), *abunè* (Ludiano), *boná* (Sonvico), *bonaa* (Lumino), *buná* (Poschiavo) v. **SIGN** Abbonare.

aboná → abonaa

**abonaa** (SottoC.), *aboná* (Cimadera), *abonád* (Locarno, Torricella-Taverne, Lamone), *abonáo* (Broglia, Caveragno), *abonó* (Lev.), *abonò* (Bell., Riv., Loc., Lug., Moes.), *abonóo* (Verz.), *abonòo* (Brissago, Minusio, Cugnasco), *abonóu* (Lodrina, Iragna, Ble., circ. Giornico, CentoV., Mergoscia, Soazza), *abunaa* (SottoC.), *abunáo* (Peccia, Linescio), *abunò* (Medeglia, Robasacco, Russo), *abunóu* (Chironico), *abunòu* (Ons.), *abunú* (Ludiano) s.m. **SIGN** Abbonato ◊ persona abitudinaria.

**LOC** *Véss* -, trovarsi frequentemente nelle stesse condizioni, essere confrontato con le stesse difficoltà.

abonaa → aboná

abonád → abonaa

abonaménn → abonamént

**abonamént**, *abunamént*; *abonaménn* (Lumino, Lodrina), *abonaménte* (Sonvico), *abonamint* (Caveragno, Verscio, Cavigliano, Minusio), *abonemént* (Gerra Gamb.), *abunamint* (Linescio) s.m. **SIGN** Abbonamento.

**LOC** *Végh l'*-, trovarsi frequentemente nelle stesse condizioni, essere confrontato con le stesse difficoltà.

abonaménte, abonamint → abonamént

abonáo → abonaa

abonád → bondá

abondansa → bondanza<sup>1</sup>

abondant → bondant

abondanza, abondanze → bondanza<sup>1</sup>

abondènt → bondant

abondèzza → bondanza<sup>1</sup>

**Abóndi**, *Abundi* n.p. **SIGN** Abbondio.

**LOC** *Quii da sant* -, i mendicanti che nel gior-

Figure 6: XML above imported automatically into Adobe InDesign for automatic layout for printing.

The **second step** should be the integration of a morphological analyzer-synthesizer within the dictionary, in order to develop a fully integrated spell-checker for the minority language. The majority of spell-checking systems (e.g. HunSpell which is the base of LibreOffice, Firefox, Chrome, etc. proofing tools) are fed with wordlists which are not integrated and often not even exported from a coherent dictionary authoring system (Németh 2011); the same can be said for morphological engines or corpus analysis software, such as NOOJ (Ben Hamadou, Mesfar, Silberztein, 2010): they may provide powerful tools, but they are never integrated with a dictionary authoring and publishing system, and their use is normally confined to NLP specialists, and often well beyond the reach of traditional linguists not to say general public, school teachers or public administration staff. In fact, having an integrated system means that every change is reported automatically in both modules of the system and that the spell-checker is always up to date, and so is authoring, Web publication, Smartphone app generation, and even traditional paper publishing are all steps of a highly integrated procedure. This is especially useful in treating minority or lesser-used language, where the fieldwork is always active and new additions, changes, creation of neology and terminology, and even spell reforms are frequent events. As modern spell-checkers, our module works with a “best-guess” pattern of the rule, based on statistic algorithms, on *Levenshtein distance* (Levenshtein, 1966) and on *double metaphone* (Philips, 1990).

In addition, it includes dialectal-driven error patterns, which are fundamental for minority languages. In fact, every correction system sets up its guesses upon similarities of words. Our system adds to this method the awareness that, for semi- or recently standardized languages where the overwhelming majority of writers are *de facto* illiterate in their language, most errors can be caused by the knowledge of a word in one particular language variety that is not the standard form: in minority languages people do not only misspell: they simply can't write, even if they can perfectly speak (and write in the dominant language). The two word forms (standard and non-standard) may differ a lot sometimes: the non-standard word can be, for example, more similar to a word with a complete different meaning than to its standard equivalent; or it can also be so graphically far from the standard form that the system is not able to find the equivalence using the statistic algorithm or the standard pattern matching. The system must then know that there can be odd correspondences. We can offer a typical example from Sardinian language (the first language for which we developed the spell-checker): the word *berbeghe* (sheep) is pronounced /brebei/ in South Sardinia. If we analyze the differences among the two words, we can understand that a simple system would not be able to guess the standard form (*berbeghe*) starting from the non-standard one (*brebei*) (Corongiu, 2013). Conversely, our dialect-oriented spell-checker knows these odd correspondences and the rules that allow to guess them. Our system uses therefore two guess pattern, shown in the table below (*Figure 7*): the simple one detects “*soundslike* typical mistakes”; the advanced one detects “*linguistic-background* driven mistakes”. See *Figure 8* and *Figure 9* for MS Word and web interface of the “*dialectal*” spellchecker (Zoli, 2008).



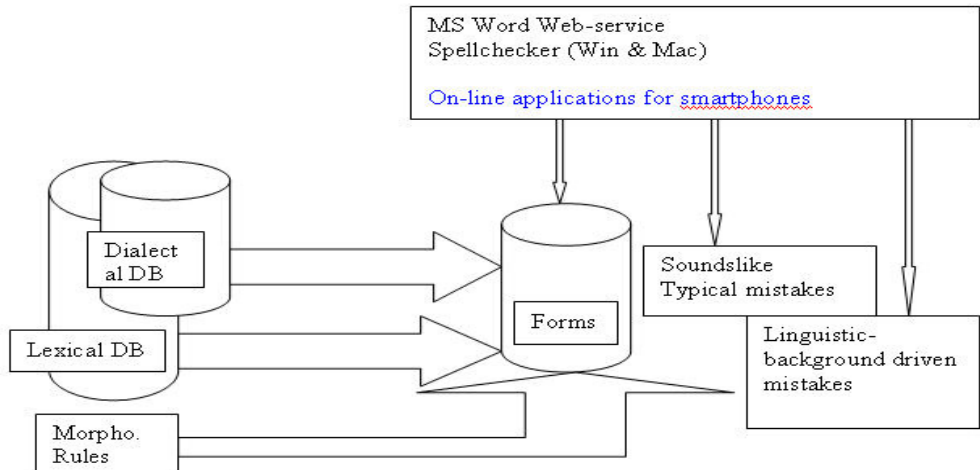


Figure 7: Functioning of an advanced spell-checking system

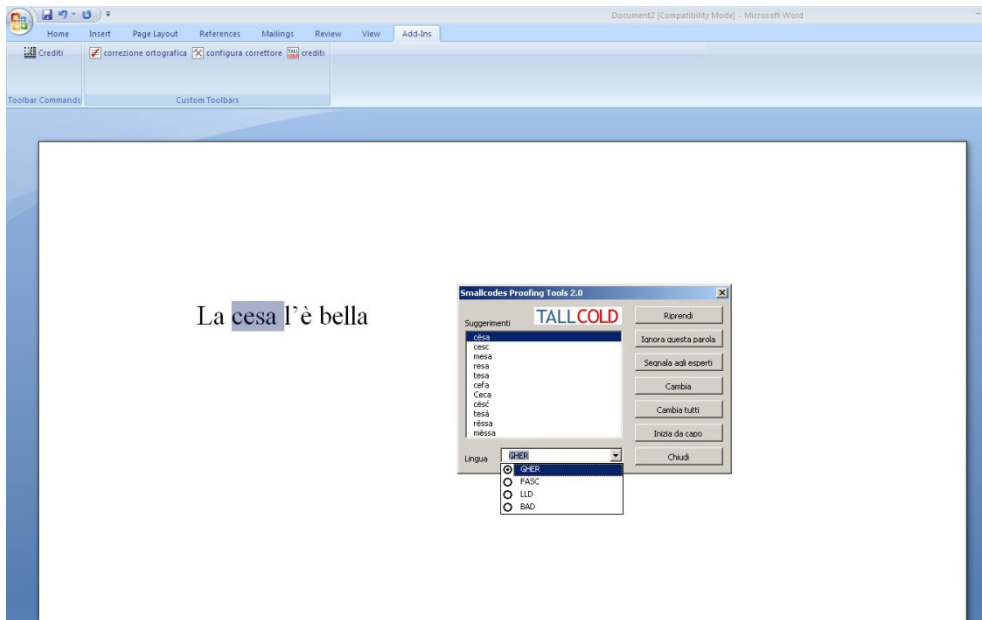
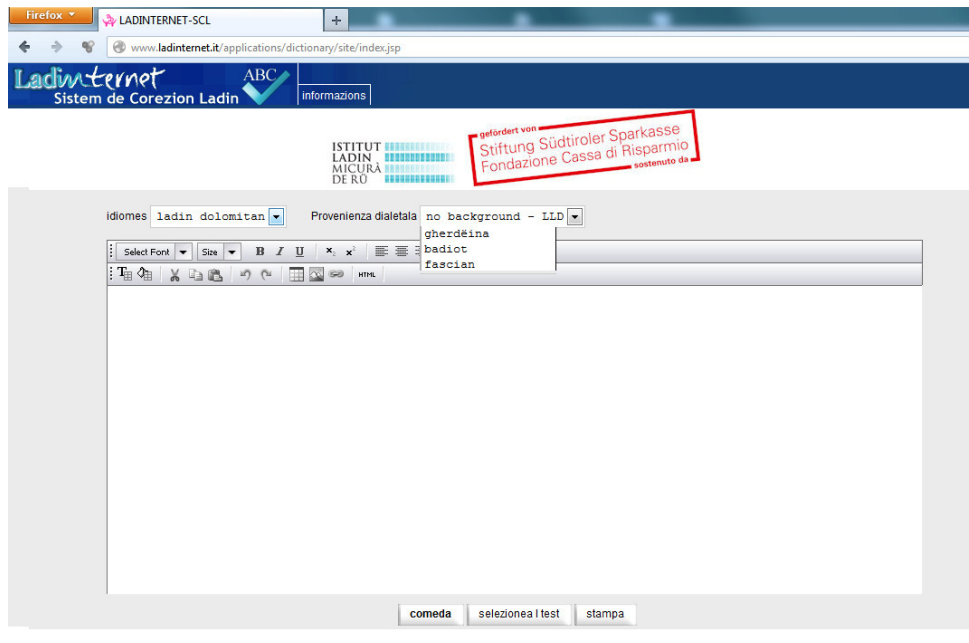


Figure 8: Spell-checking of standard Ladin language with correction based on the typical errors caused by the three main dialectal backgrounds (corresponding to the three major oral dialects spoken in the respective alpine valleys: Gherdëina, Badiot, Fascian)




*Figure 9: Same system as above accessing exactly the same data via the same SOAP Web-service but for use in a text area within a Web browser.*

The **third step** is the terminology module. The creation of the terminology is a fundamental procedure if we want the language to be employed, for example, in school teaching (see for example Figure 10, which shows a collaborative webTool for neology, used by the authors of schoolbooks in Ladin Dolomitan), and administrative / official translation (see fig 11 & 12 for a tool of computer-aided technical translation for Sardinian languages, used by various public bodies). Languages which do not have a written tradition normally lack of technical lexicon. These new words need therefore to be created and the method for their creation already exists: the sources are the other international languages that have made this procedure before and the other minority languages that have already solved these issues. Another possibility is to re-use old words whose original meaning is losing importance in today's life and make these words express new meanings. A typical example is the vocabulary used for cars nowadays in Italian: this is nothing more than the recovered lexicon for horse carriages; similarly, the lexicon of Air Navigation is directly taken from Maritime Navigation vocabulary. English typically uses this strategy for neologisms, exploiting metaphors and meaning extensions of pre-existing words. Romance languages, on the other hand, favour the use of loan words, drawing inspiration from present or past prestigious languages.

neo.ladintal.it/applications/dictionary/dossier/dossierList.jsp				
⌵	acciaieria (italiano)	ladino fassano folina de facel approvato 15/05/2013	Vigiljo Iori	industria
⌵	altolorno (italiano)	ladino fassano autlon approvato 15/05/2013	Vigiljo Iori	industria
⌵	elettrodomestico (italiano)	ladino fassano elettrodomestich approvato 15/05/2013	Vigiljo Iori	industria
⌵	idrocarburo (italiano)	ladino fassano idrocarburi approvato 15/05/2013	Vigiljo Iori	industria
⌵	metalmecanico (italiano)	ladino fassano metalmecanich approvato 15/05/2013	Vigiljo Iori	industria
⌵	industria tessile (italiano)	ladino fassano industria de la tella approvato 15/05/2013	Vigiljo Iori	industria
⌵	petrolchimico (italiano)	ladino fassano petrolchimich approvato 15/05/2013	System Manager	industria
⌵	Agro Romano (italiano)	ladino fassano Agro Roman approvato 15/05/2013	Vigiljo Iori	nomi geografici
⌵	Agro Pontino (italiano)	ladino fassano Agro Pontin approvato 15/05/2013	Vigiljo Iori	nomi geografici
⌵	Adamello (italiano)	ladino fassano Adamel approvato 15/05/2013	Vigiljo Iori	nomi geografici
⌵	vapore acqueo (italiano)	ladino fassano vamp de ega approvato 15/05/2013 vapor de ega approvato 15/05/2013	Vigiljo Iori	clima
⌵	isobara (italiano)	ladino fassano isobara [asj] approvato 15/05/2013	Vigiljo Iori	clima


Figure 10: An example of the workflow (with various status of approval) for the creation and consolidation of terminology in Ladin language: please note that the system is fully integrated with the dictionary module so that specific word-lists can be included or excluded from the general dictionary, exported for the Web or via Web-service for use within other applications

IN SA GALASSIA DE SAS LIMBAS ONZI PARANLA EST UN'ISTEDDU



**Uffiziu de sa Limba Sarda**  
DITZIONARIOS

**Direttore**  
Marionetta Piga



documentos news formazione

Chenàbura, 19 de Santugaine de su 107

registraztzione

**DERETU**

263 oggetti imbedduidos; pàgina 1 de 11 >>> dimens pag. 25

numene de atessu  paraula crae  ricordame

a|b|c|d|e|f|g|h|i|j|k|l|m|n|o|p|q|r|s|t|u|v|w|x|y|z

abandonu de domitiliu	loc. sust. m.	ZUR	cunsistit in s'ai...
acompagnamentu coativu	loc. sust. m.	ZUR	podere de s'autor...
agabu de sa mallera de su cont...	loc. sust. m.	ZUR	si tenet cando...
amnistia	sut. fem.	ZUR	una de sas causas...
amontizione	sut. fem.	ZUR	antioneu dissipli...
analozia	sut. fem.	ZUR	protzedimentu chi...
anatolizismu	sut. masc.	ZUR	fenòmenu de sa cap...
animus	sut. masc.	ZUR	tèrmine latinu chi...
anticerei	sut. fem.	ZUR	contratu in ue su...
antizuriditizidade	sut. fem.	ZUR	indikat su contras...
apolida	sut. fem.	ZUR	situatzione de ont...
arbitradu	sut. masc.	ZUR	detzisione subra u...
archiviatzione	sut. fem.	ZUR	cando finint sas l...
armistitziu	sut. masc.	ZUR	acordu intre duos o...
assuntzione	sut. fem.	ZUR	tircostàntzias de...
atenuantes	sut. fem.	ZUR	atu chi, pro form...
atu alpicu	loc. sust. m.	ZUR	manifestatziones d...
atos zuridicos	loc. sust. m.	ZUR	cale si siat fat...
atu illetziu	loc. sust. m.	ZUR	atos chi tenent s...
atos de liberalidade	loc. sust. m.	ZUR	atos chi tenent s...
atos post mortem	loc. sust. m.	ZUR	atos chi tenent s...
atos preliminaries a su dibalim...	loc. sust. m.	ZUR	atos chi tenent s...
atos retetzitos	loc. sust. m.	ZUR	atos chi tenent s...
atos sexuales cun minorene	loc. sust. m.	ZUR	atos chi tenent s...

chircu lema

DERETU

in su ditzionariu generale

Chircas finas in variantes

Figure 11: Web page output of terminology module



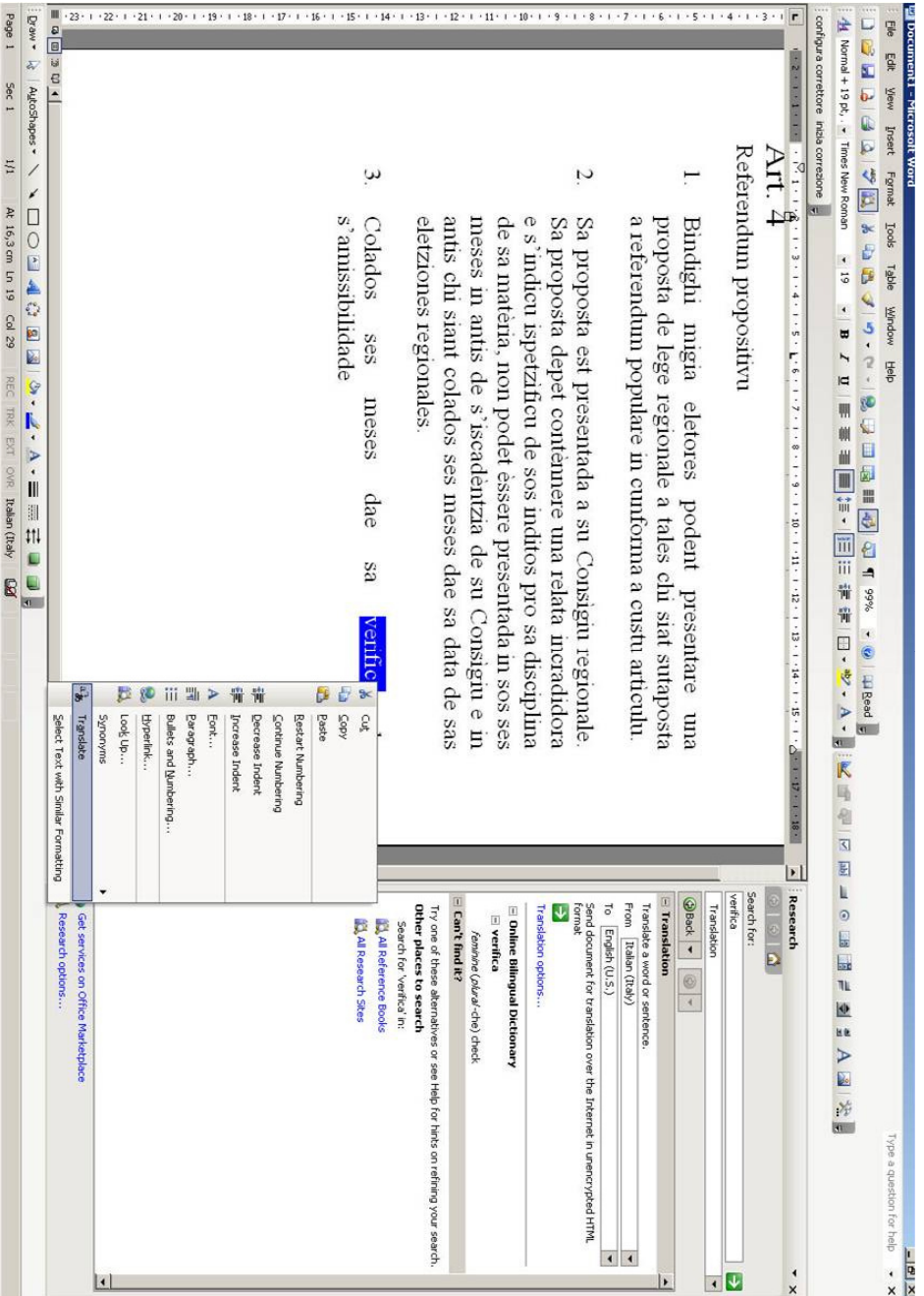


Figure 12: Web-service output for word-to-word terminology translation

The **fourth** (optional) **step** is the creation of a reference thesaurus, namely the collection of written material of any literary kind and historical period which offers a precious help in the consolidation of lexicon. A first-level thesaurus does not need to be exaggeratedly ambitious: actions such as pos tagging, machine-translation or stylistic analysis can be left out of the first phase of our project, not because they are not interesting or important but because they do not belong to the very first set of tools that every language needs to start its digital preservation (Soria, Zoli, 2012). What is very useful, at the beginning, is the possibility of having phrase quotations for literary dictionaries, the lemmatization and the *aligning of old orthography with new orthography*.

The material collected and organized in the corpus allows to compare old and new texts and wordlists and represents an authoritative support to be consulted at any time. The thesaurus, thanks to frequency parameters, can offer guarantees on the effective use of a word or on its register / linguistic style.

The choice of literary texts is done in order to let the speakers' community recognize them as "properly written" and trustable. Old literary texts often have, in fact, a special prestige and are regarded as models in the lexical and semantic field (Videsott, 2011). But very often, if not always, literary texts in minority languages are written in different, incoherent spellings. We must preserve the original script and at the same time re-publish the text in modern / standardized scripts as far as it is possible. The automatic statistical alignment of the two version of the same text (old and modern) allows the users and the researches to quote literature both "as it was written" and "as it would be written today".

Our Ladin thesaurus (*Corpus dl Laden leterar / Wörterbuch des literarischen Ladinisch / Corpus letterario del ladino*)<sup>2</sup> is a electronic corpus of literary works written in Ladin language. It currently stores more than 1,200 texts from Ladin valleys (Val Badia, Val Gardena, Val di Fassa, Fodom, Ampezzo): the material has been completely scanned and digitalized and it consists of an archive of more than 250,000 different words.

Such a thesaurus represents the last step of the complete system and its strong connection with the dictionary module contributes to show the importance of an integrated system for the study and filing of lexical material.

## Some possible objections

In most contexts where a minority language is struggling to be recognized and protected, standardization is feared by many people. These people, especially when they master the lesser-used language, are afraid that a major standard form might hide away their native varieties, which have a strong identity value for them. At the same time, the speakers who fear standardization, also reject the use of tools such as electronic instruments for spell-checking (according to the belief that everyone writes in his or her own way, or in a way which is totally respectful of local pronunciations cfr. Vitali 2008). This attitude contribute to relegate minority

---

<sup>2</sup> <http://corpuslad.ladintal.it/>

languages such as Tamazight to the status of dialects and prevent them to evolve and flourish.

Instead, it must be clear that standardized spelling only makes sense for a written language. If there were, for example, a talk show in one minority language, the titles and explanatory signs would be in standard, but the presenter and the guests would talk in their own dialects (as it happens in German Switzerland or in Norway) [Zoli, 2012 (2)]. The spell-checker we developed is aimed at appointing languages such as Tamazight a written authority in which every rule is fixed and scientifically established. Only in this way, we believe, small local languages can be protected by the unified bigger standard form (which is, again, only a standard script that tries to enable everyone to read in his or her own dialect, as long as a small effort is made to find regularities and correspondences between every vernacular variety and the standard form).

Moreover, some speakers might challenge the effectiveness of the terminological research we support with our third module of the integrated system. In fact, some people do not accept the creation of neologisms because they are alien to the traditional language these speakers learned as children (“*my grandma would have never said that!*”). As a matter of fact, no language, at an early stage, has the words to express novelties or brand new concepts, but the school has made us believe that certain languages are rich for some sort of divine predestination (Pellegrini, 1977). If we take any Italian or French vocabulary, we can see that about 4000-5000 words are derived directly from Latin: these words are the most frequent and they concern concepts or things which are the backbone of the language. Another 20,000 are also derived from Latin, but these other words were created later, invented by the humanist scholars and writers. When a new concept was needed, scholars used to draw it from the inexhaustible mines of Latin or Greek and superficially make the word fit the graphics system and the phonetics of the target language. This explains why French and Italian basic words directly derived from Latin only remotely resemble each other (*occhio / œil, bocca / bouche, casa / chez*), while the scientific terminology is virtually identical (*oculare / oculaire, orale / oral, domestico / domestique*). The former have come straightforwardly from Latin and their form is the result of phonetic modification. The latter have been reinserted later and belong to scientific or academic (i.e. terminological) vocabulary. The creation of new terminology is therefore at the basis of the rehabilitation of a language and it is a necessary step for this language to acquire prestige and be adopted in the public sphere (e.g. school or public administration).

## **A quick comparison with possibly similar tools**

It has to be said that similar ideas have been around for a while: efforts like BLARK (Krauer, 2003) and LCTL at the Linguistic Data Consortium rely on somewhat similar ideas<sup>3</sup>.

The main difference is that this project aims to be industry-standard: the idea, as it is expressed in the manifesto below, is to give long digital life not only to data, but

---

<sup>3</sup> [http:// projects ldc.upenn.edu/LCTL/index.html](http://projects ldc.upenn.edu/LCTL/index.html)

also to applications, source-code, etc. A limit of software tools that come from the academic and pure-research world is that they often cannot be maintained by “big” teams of professional software developers, but often are either quickly abandoned (not by the users, by the developers when the research project, and consequently the funding, is over) or suffer an inevitable technical obsolescence (the case of E-Meld is paradigmatic).

We could say that the Smallcodes project, stemming from the private industry sector and approaching the research world (rather than vice versa) has a, so to say, *different business model*.

The business model is not that the language experts or researchers adopt the system as users, basically using it “at their own risk” or contributing to the development, in a classical open-source fashion.

On the contrary, the Smallcodes business model is that the software is centrally developed, and partnerships and funding opportunities are established every time a new language group enters the “community”. Every new language expert group adds new expertise, new funding, requests new features, but development is pursued in an industrial fashion, with attention to the latest web technologies, with highly resourced staff in an a “web 2.0 commerciale way”; then, the business itself is basically non-profit, but all the same this is different from software development done inside the linguistic academic world, which cannot have the structure and the attitude of a commercial software house.

Finally it is more common to find a commitment for sharing language resources (see for example OLAC<sup>4</sup>, DoBeS<sup>5</sup>), whereas Smallcodes focuses more on the sharing of *software tools*.

## **A possible employment of the toolbox for Tamazight**

Our aim is to have one integrated tool which will be multi-accessible and will give multiple simultaneous outputs. These are as follows:

A) “human readable data”: a web-app which will provide (not necessarily all, and not necessarily at the same time):

- Online authoring of dictionary of standard language (with synonyms, antonyms, WordNet-like synset relationships (Fellbaum 1998).
- Online authoring of dictionary of dialectal variation
- Online authoring (with collaborative discussion and workflow) of neology/terminology
- Web publishing for public consultation of dictionary
- Web publishing of conjugation / declination tables (paradigms / schemas)

---

<sup>4</sup> <http://www.language-archives.org/> 21/07/2013

<sup>5</sup> <http://dobes.mpi.nl/> 21/07/2013

- Integrated output of XML files for paper publishing (Adobe InDesign format)
- Integrated output of e-books (ePub format)
- Integrated output of XML files for Android / iOS dictionary apps.

B) “machine-readable data”: a Web-service which will provide (not necessarily all, and not necessarily at the same time):

- spell-checking
- dictionary look-up
- thesaurus look-up
- glossary look-up → encyclopaedic information
- terminological word-to-word translation
- morphological analysis and synthesis.

The Web-service will provide data to many different applications: for use in a browser, or integrated in a word-processor (via XML SOAP web-service) or, again, integrated in e-Books for dictionary / terminology lookup.

Every language has its own peculiarities in terms of phonology and morphology. A comprehensive tool must take account of a very large number of possible differences among languages and anticipate the changes that each language will require to the system.

The case of Tamazight is more complicated: we must be able to search for an entry using the Tamazight script but also with the corresponding Latin characters. We must therefore create the correspondences between graphemes and insert them into the system. We must also take into account all possible variations in transliteration and design a list of interchangeable graphemes. All this will be accomplished with multiple Lucene indexes.<sup>6</sup>

We also know that the Tamazight, as every language of recent standardisation can have oscillation in writing, and event different realisations of the same phoneme:  $\ominus / \oplus$ ;  $+$  /  $\times$ ;  $\wedge$  /  $\vee$ ;  $\Re$  /  $\mathfrak{R}$ ;  $\mathfrak{X}$  /  $\mathfrak{X}$  (Boukous, 2009). The dictionary module and the attached spell-checker must take account of all these possibilities.

In addition to these mutation processes, Tamazight language also possesses many assimilation processes, such as the propagation of emphasis or the assimilation of voiced and unvoiced consonants. These aspects concern instead the spell-checker, which must then conceive special rules for words formation which reckon with these phenomena. Only a strong language-aware spellchecker and metaphone algorithm can achieve good results: in the situation of non-latin, recently-standardized and highly diatopically variable languages standard spell-checking simply does not work.

---

<sup>6</sup> <http://lucene.apache.org/> 29/05/2013.

Again, with regard to the spell-checking module, we must have a look at morphology: there are typical functions of Tamazight language which do not concern, for example, Romance languages, such as the discontinuous affixation. Therefore, we have to formulate a set of rules in order to automate the process of spelling correction. In this particular case, we must take account, for example, of the incredible variety of patterns in plural formation. Moreover, we must add the categories of grammatical cases and consider the morphological changes that words undergo in this inflection (in addition to gender and number inflection). Yet again, there are several morphological changes in verbal inflection, such as personal endings, aspect, derivative morphemes (causative, reciprocal and passive), noun agreement (for the participle form) (Boukhris, 2008).

These examples are useful to show that an effective comprehensive system must be able to adapt to the needs of every language. Tamazight has a complex grammar, even if, when compared to other distant languages (such as Mexican languages, which whom we work with), it has some sort of similarities with European languages. We are struggling in order to reach the best results in the consideration of the largest amount of lexical and grammatical possibilities. Every new language we introduce in our system is an important piece of the puzzle that allow us to test the capabilities of our system, add new concepts, discard old beliefs. This expectation, we think, is our way to put into practice the principle of cooperation among minority languages of the world.

## **Our manifesto (Zoli, 2008)**

As we have seen, language technology offers significant opportunities for minority languages and can be a major force in addressing and alleviating some of the difficulties they face. Speech and language technologies are in fact a powerful means to bring together speakers' communities, to have a major impact on language learning support, to promote inclusion of elderly or impaired people and to foster widespread use of a language through digital means (Soria, Zoli 2012).

In developing the integrated system we describe here we have been inspired by some beliefs. First of all, we believe that any serious project of cultural defence should start from the defence of the language, and that modernization is to be achieved through a written form of the language, as coherent and as widely accepted as possible. We firmly think that digital technologies can play a crucial role in this process of language modernization and in that of promotion and diffusion of the language among younger generations. Finally, speaking of technology, we believe that the highest possible degree of standardization (in file formats, in communication protocols, in programming languages, in DBMS's) is mandatory. Only so it is possible to guarantee "long digital life" to language resources and only so we can allow a real exchange of information, data and technologies.

Several years of experiences have allowed us to reach different results:

- Our platform supports the standard Unicode (diacritics and all sorts of characters are accepted).
- The interface language can be changed very simply at any desired moment.
- There is a high-parameterization (nothing is hard-coded).
- Our software meets industrial standards.
- All modules have achieved a real interoperability.

The final aim was to develop a unique tool which can be integrated in the main writing systems (Word, Libre Office, Web browser, etc.) and which can operate at all the different levels (or modules) of the toolbox. This, we believe, has shown to be one of the most complete and effective "survival kits" for all endangered minority languages such as Tamazight.

## Bibliography

- Ben Hamadou, Abdelmajid; Mesfar, Slim; Silberztein, Max. “Finite State Language Engineering: NooJ 2009”. International Conference and Workshop. Touzeur: Centre de Publication Universitaire, 2010.
- Boukous, Ahmed. *Phonologie de l'Amazighe*. Rabat: Institut Royal de la Culture Amazighe, 2009.
- Boukhris, Fatima. *La Nouvelle Grammaire de l'Amazighe*. Rabat: Institut Royal de la Culture Amazighe, 2008.
- Corongiu, Giuseppe. *Il sardo: una lingua normale*. Cagliari: Condaghes, 2013.
- Dell'Aquila, Vittorio; Iannàcaro, Gabriele. *La pianificazione linguistica*. Roma: Carocci Editore, 2011.
- Kloss, Heinz “Abstandsprachen und Ausbausprachen”. In Göschel, Joachim; Nail, Norbert; Van der Els, Gaston. *Zur Theorie des Dialekts: Aufsätze aus 100 Jahren Forschung*. Zeitschrift für Dialektologie und Linguistik, 1976.
- Krauer, Stevem. The Basic Language Resource Kit (BLARK) as the First Milestone for the Language Resources Roadmap *Proceedings of SPECOM 2003*, Moscow, 2013.
- Fellbaum, Christiane, ed. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.
- Francopoulo, Gil et al. *Lexical markup framework (LMK) Genoa*: LREC, 2006.
- Levenshtein, Vladimir I. *Binary codes capable of correcting deletions, insertions, and reversals*. Soviet Physics Doklady, 1966.
- Lurà Franco et al. “Dalla carta al web: la versione informatica del lessico dialettale della Svizzera italiana”, In: Ruffino G., D'Agostino M. *Storia della lingua italiana e dialettologia*, atti del VIII Convegno Internazionale dell'Associazione per la Storia della Lingua Italiana, Palermo, 2009.
- Németh, László <http://hunspell.sourceforge.net/> (27/05/2013).
- Pellegrini, Giovan Battista. *Carta dei dialetti d'Italia*. Pisa: Pacini editore, 1977.
- Philips Lawrence. “Hanging on the Metaphone”, In *Computer Language*, Vol. 7, No. 12 (December), 1990.
- Scannel, Kevin. “New computational resources for indigenous and minority languages”, 17th annual NAACL conference. Isle of Man, 2011.
- Scannel, Kevin. “Semi-automated construction of semantic networks using web corpora”, Words, Texts and Dictionaries conference. University of Wales Centre for Advanced Welsh and Celtic Studies, Aberystwyth, 2008.
- Vitali, Daniele. “Appello ai romagnoli per studiare la diversità dialettale” *La Ludla XII*, 2008.



Soria, Claudia, Zoli, Carlo. “New markets for Language Technology for minority languages”, Maaya Conference. Paris, 2012.

Videsott, Paul. Vocabolar dl Ladin Leterar / Wörterbuch des literarischen Ladinisch / Vocabolario del Ladino letterario (VLL). Projektbeschreibung, 2011.

Zoli, Carlo. “Encouraging the presence in the cyberspace of the lesser used languages through writing and proofing tools: the case of Sardinian language”, Maaya Conference. Paris, 2012.

Zoli, Carlo. “La scrittura standard del romagnolo: un’urgenza non rimandabile” La Ludla IX, 2012.

Zoli, Carlo. “Trattamento digitale delle lingue al servizio delle lingue meno usate”, Corongiu G., Romagnino C. Sa Diversidade de sas Limbas in Europa, Itàlia e Sardigna. Atos de sa cunferèntzia regionale de sa limba sarda, Macumere/Macommer, 2008.

## Appendix: Institutions whom which we work

- Institut national des langues et civilisations orientales (**INALCO** - Paris)
- **Rromani Baxt** - Paris
- PARIS 3 (prof. J.-L. Léonard – **Meso-American** languages)
- Chubri, institu d'inventèrr e d'valantaj du **Galo**
- Università Orientale di Napoli (prof. M. Gnerre - **Meso-American** languages)
- Chambra d'Òc – **Occitan, Francoprovençal**
- Regione Piemonte – Minority Department (**Walser, Occitan**)
- Bureau Régional Ethnographie et Linguistique – Val D'Aosta (**Francoprovençal language**)
- Istituto di Dialettologia ed Etnografia della Svizzera Italiana (**Lombard = north Italian dialects of Italian Switzerland**)
- Ufitziu pro sa **Limba Sarda** – Regione Autònoma de sa Sardigna
- Istitut **ladin** “Micurà de Rü” – Val Gardena-Val Badia
- Istituto culturale **ladino** “Majon di Fascegn” – Val di Fassa
- Union Generèla di **Ladins** dla Dolomites - SPELL
- Istituto Culturale **Mòcheno** Palù TN
- Istituto Culturale **Cimbro** Luserna TN
- Uffici Lenghe **Furlane** – Provincia di Udine
- Agenzie regionâl pe lenghe **furlane** (ArLeF)