Carlo Zoli

Smallcodes S.r.L.
Via del Campuccio 118, 50125 Firenze, Italy

[carlo.zoli@smallcodes.com]

'Smallcodes', a Unified computational linguistics toolbox for minority languages

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, Iannàccaro, 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 allows 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 Romantsch 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 (*fig. 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”?

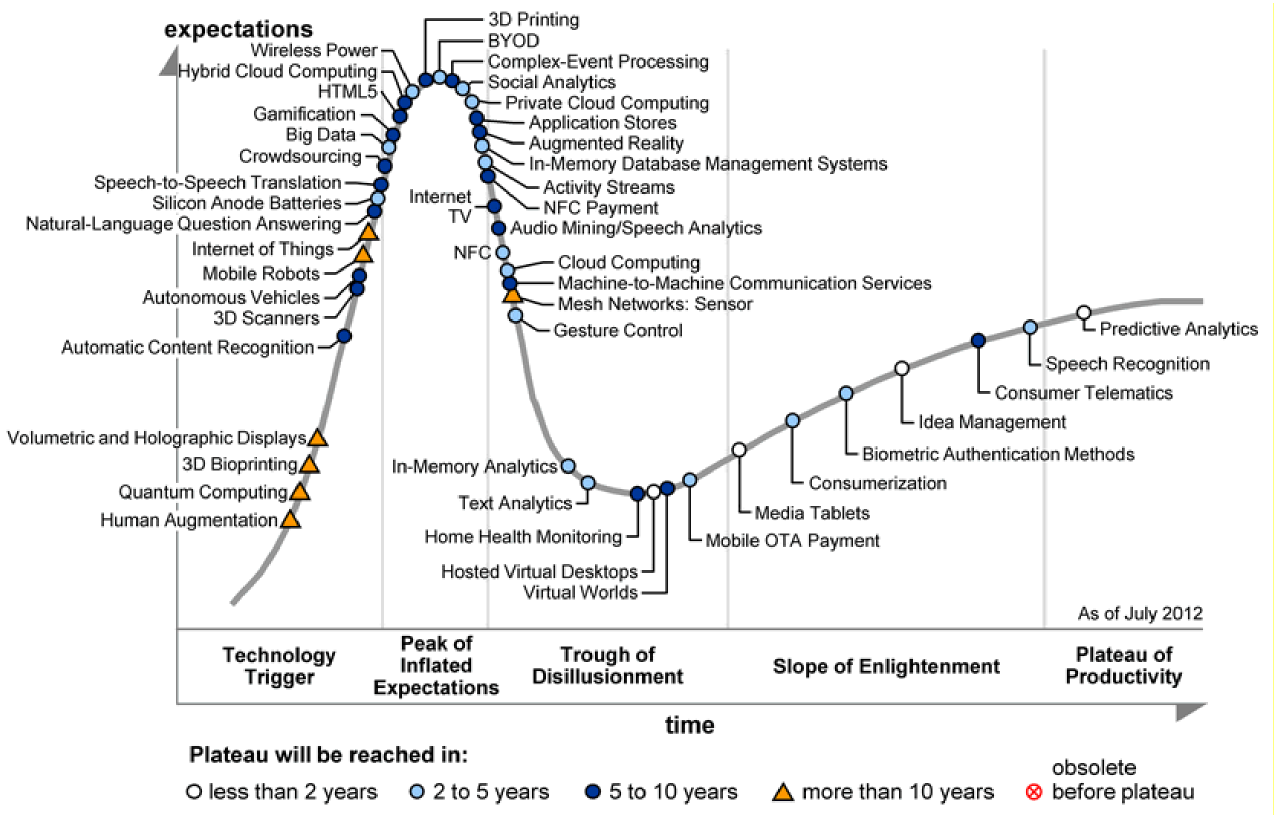


Fig. 1: Gartner’s hype cycle as of July 2012

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, Iannàccaro, 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 (*fig.2*) 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).

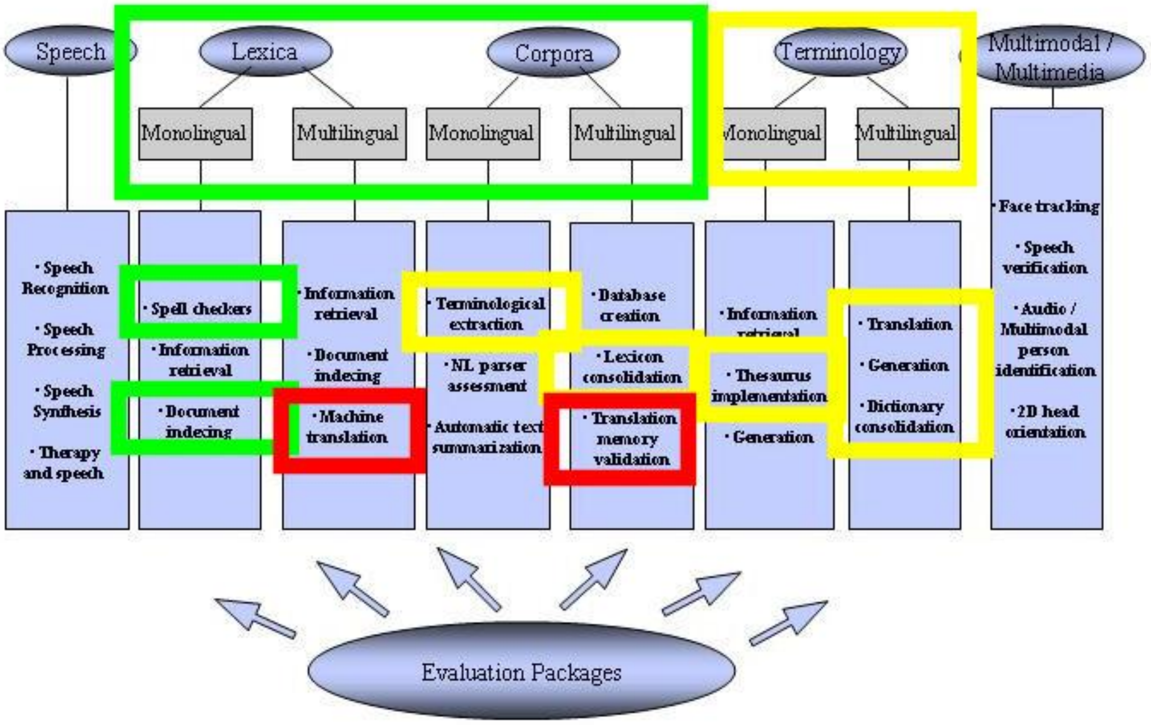
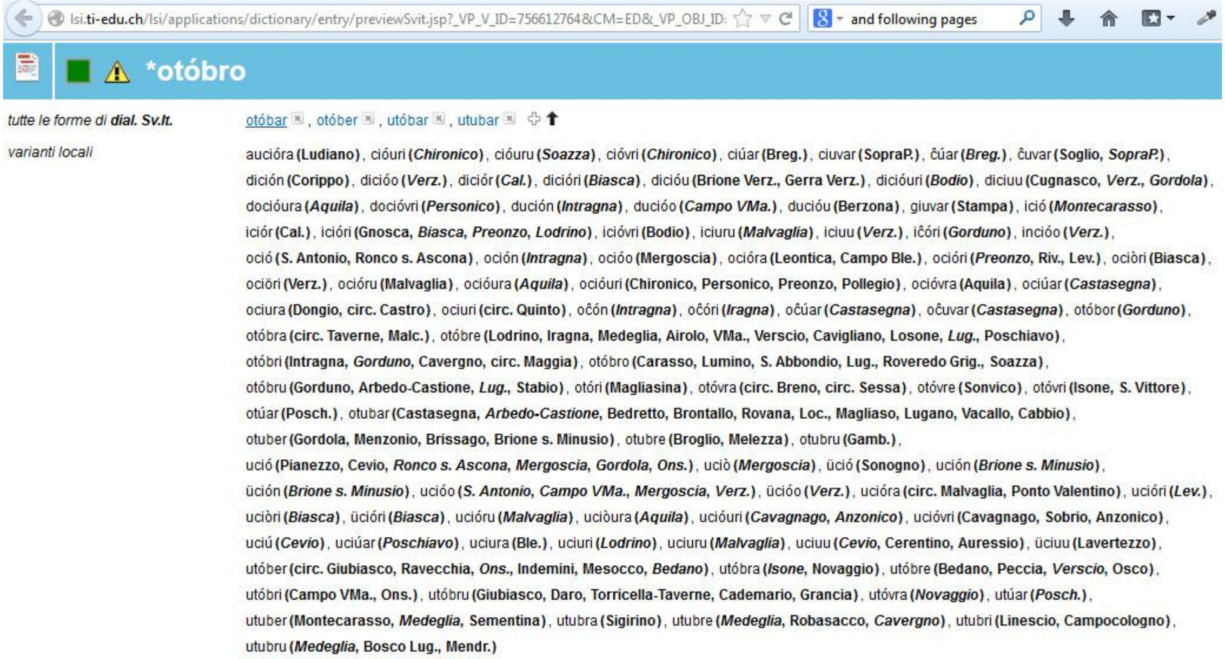


Fig. 2: Computational linguistics packages for minority languages (Scannel, 2008)

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 form informal databases and being the result of an OCR or parsing of ancient dictionaries.

The figure below (*fig. 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 (*fig. 4*), we have the same entry in a human-readable form (actually an XML + CSS wich can be easily imported in a professional publishing tool as Adobe Indesign, see *fig. 6*); *fig. 5* shows the XML of fig 4 in the classic machine-readable form.


lsi.ti-edu.ch/lsi/applications/dictionary/entry/previewSvit.jsp?_VP_V_ID=756612764&CM=ED&_VP_OBJ_ID:
and following pages

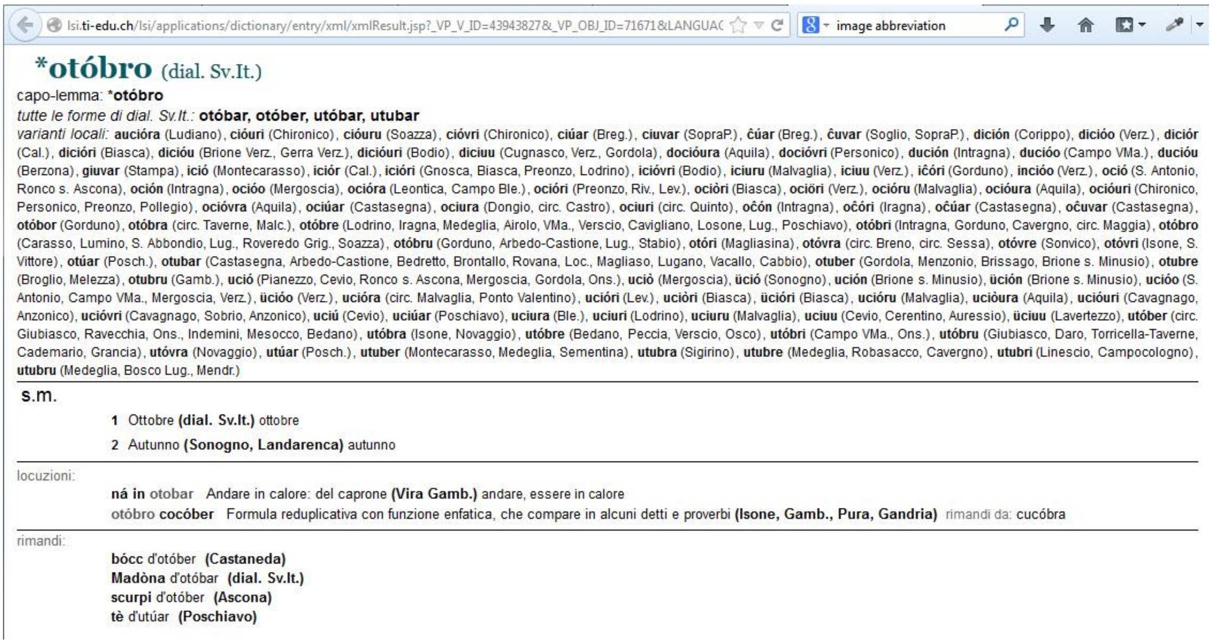
🟩 ⚠ *otóbro

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

varianti locali

auciòra **(Ludiano)**, cióuri **(*Chironico*)**, cióuru **(*Soazza*)**, cióvri **(*Chironico*)**, ciúar **(Breg.)**, ciuvar **(SopraP.)**, ĉúar **(*Breg.*)**, ĉuvar **(Soglio, *SopraP.*)**, dición **(Corippo)**, dicióo **(*Verz.*)**, diciór **(*Cal.*)**, dicióri **(*Biasca*)**, diciòu **(Brione Verz., Gerra Verz.)**, dicióuri **(*Bodio*)**, diciuu **(Cugnasco, *Verz., Gordola*)**, dociòura **(*Aquila*)**, docióvri **(*Personico*)**, dución **(*Intragna*)**, ducióo **(*Campo VMa.*)**, duciòu **(Berzona)**, giuvar **(Stampa)**, ició **(*Montecarasso*)**, iciór **(Cal.)**, icióri **(Gnosca, *Biasca, Preonzo, Lodrino*)**, icióvri **(Bodio)**, iciuru **(*Malvaglia*)**, iciuu **(*Verz.*)**, iĉóri **(*Gorduno*)**, incióo **(*Verz.*)**, oció **(S. Antonio, Ronco s. Ascona)**, oción **(*Intragna*)**, ocióo **(Mergoscia)**, ocióra **(Leontica, Campo Ble.)**, ocióri **(*Preonzo*, Riv., Lev.)**, ociòri **(Biasca)**, ociöri **(Verz.)**, ocióru **(Malvaglia)**, ocióura **(*Aquila*)**, ocióuri **(Chironico, Personico, Preonzo, Pollegio)**, ocióvra **(Aquila)**, ociúar **(*Castasegna*)**, ociura **(Dongio, circ. Castro)**, ociuri **(circ. Quinto)**, oĉón **(*Intragna*)**, oĉóri **(*Iragna*)**, oĉúar **(*Castasegna*)**, oĉuvar **(*Castasegna*)**, otóbor **(*Gorduno*)**, otóbra **(circ. Taverne, Malc.)**, otóbre **(Lodrino, Iragna, Medeglia, Airolo, VMa., Verscio, Cavigliano, Losone, *Lug.*, Poschiavo)**, otóbri **(Intragna, *Gorduno*, Cavergno, circ. Maggia)**, otóbro **(Carasso, Lumino, S. Abbondio, Lug., Roveredo Grig., Soazza)**, otóbru **(Gorduno, Arbedo-Castione, *Lug.*, Stabio)**, otóri **(Magliasina)**, otóvra **(circ. Breno, circ. Sessa)**, otóvre **(Sonvico)**, otóvri **(Isone, S. Vittore)**, otúar **(Posch.)**, otubar **(Castasegna, *Arbedo-Castione*, Bedretto, Brontallo, Rovana, Loc., Magliaso, Lugano, Vacallo, Cabbio)**, otuber **(Gordola, Menzonio, Brissago, Brione s. Minusio)**, otubre **(Broglio, Melezza)**, otubru **(Gamb.)**, ució **(Pianezzo, Cevio, *Ronco s. Ascona, Mergoscia, Gordola, Ons.*)**, uciò **(*Mergoscia*)**, üció **(Sonogno)**, ución **(*Brione s. Minusio*)**, üción **(*Brione s. Minusio*)**, ucióo **(*S. Antonio, Campo VMa., Mergoscia, Verz.*)**, ücióo **(*Verz.*)**, ucióra **(circ. Malvaglia, Ponto Valentino)**, ucióri **(*Lev.*)**, uciòri **(*Biasca*)**, ücióri **(*Biasca*)**, ucióru **(*Malvaglia*)**, uciòura **(*Aquila*)**, ucióuri **(*Cavagnago, Anzonico*)**, ucióvri **(Cavagnago, Sobrio, Anzonico)**, uciú **(*Cevio*)**, uciúar **(*Poschiavo*)**, uciura **(Ble.)**, uciuri **(*Lodrino*)**, uciuru **(*Malvaglia*)**, uciuu **(*Cevio*, Cerentino, Auressio)**, üciuu **(Lavertezzo)**, utóber **(circ. Giubiasco, Ravecchia, *Ons.*, Indemini, Mesocco, *Bedano*)**, utóbra **(*Isone*, Novaggio)**, utóbre **(Bedano, Peccia, *Verscio*, Osco)**, utóbri **(Campo VMa., Ons.)**, utóbru **(Giubiasco, Daro, Torricella-Taverne, Cademario, Grancia)**, utóvra **(*Novaggio*)**, utúar **(*Posch.*)**, utuber **(Montecarasso, *Medeglia*, Sementina)**, utubra **(Sigirino)**, utubre **(*Medeglia*, Robasacco, *Cavergno*)**, utubri **(Linescio, Campocologno)**, utubru **(*Medeglia*, Bosco Lug., Mendr.)**

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


lsi.ti-edu.ch/lsi/applications/dictionary/entry/xml/xmlResult.jsp?_VP_V_ID=43943827&_VP_OBJ_ID=71671&LANGUAG
image abbreviation

***otóbro** (dial. Sv.It.)

capo-lemma: ***otóbro**

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

varianti locali: **auciòra** (Ludiano), **cióuri** (Chironico), **cióuru** (Soazza), **cióvri** (Chironico), **ciúar** (Breg.), **ciuvar** (SopraP.), **ĉúar** (Breg.), **ĉuvar** (Soglio, SopraP.), **dición** (Corippo), **dicióo** (Verz.), **diciór** (Cal.), **dicióri** (Biasca), **diciòu** (Brione Verz., Gerra Verz.), **dicióuri** (Bodio), **diciuu** (Cugnasco, Verz., Gordola), **dociòura** (Aquila), **docióvri** (Personico), **dución** (Intragna), **ducióo** (Campo VMa.), **duciòu** (Berzona), **giuvar** (Stampa), **ició** (Montecarasso), **iciór** (Cal.), **icióri** (Gnosca, Biasca, Preonzo, Lodrino), **icióvri** (Bodio), **iciuru** (Malvaglia), **iciuu** (Verz.), **iĉóri** (Gorduno), **incióo** (Verz.), **oció** (S. Antonio, Ronco s. Ascona), **oción** (Intragna), **ocióo** (Mergoscia), **ocióra** (Leontica, Campo Ble.), **ocióri** (Preonzo, Riv., Lev.), **ociòri** (Biasca), **ociöri** (Verz.), **ocióru** (Malvaglia), **ocióura** (Aquila), **ocióuri** (Chironico, Personico, Preonzo, Pollegio), **ocióvra** (Aquila), **ociúar** (Castasegna), **ociura** (Dongio, circ. Castro), **ociuri** (circ. Quinto), **oĉón** (Intragna), **oĉóri** (Iragna), **oĉúar** (Castasegna), **oĉuvar** (Castasegna), **otóbor** (Gorduno), **otóbra** (circ. Taverne, Malc.), **otóbre** (Lodrino, Iragna, Medeglia, Airolo, VMa., Verscio, Cavigliano, Losone, Lug., Poschiavo), **otóbri** (Intragna, Gorduno, Cavergno, circ. Maggia), **otóbro** (Carasso, Lumino, S. Abbondio, Lug., Roveredo Grig., Soazza), **otóbru** (Gorduno, Arbedo-Castione, Lug., Stabio), **otóri** (Magliasina), **otóvra** (circ. Breno, circ. Sessa), **otóvre** (Sonvico), **otóvri** (Isone, S. Vittore), **otúar** (Posch.), **otubar** (Castasegna, Arbedo-Castione, Bedretto, Brontallo, Rovana, Loc., Magliaso, Lugano, Vacallo, Cabbio), **otuber** (Gordola, Menzonio, Brissago, Brione s. Minusio), **otubre** (Broglio, Melezza), **otubru** (Gamb.), **ució** (Pianezzo, Cevio, Ronco s. Ascona, Mergoscia, Gordola, Ons.), **uciò** (Mergoscia), **üció** (Sonogno), **ución** (Brione s. Minusio), **üción** (Brione s. Minusio), **ucióo** (S. Antonio, Campo VMa., Mergoscia, Verz.), **ücióo** (Verz.), **ucióra** (circ. Malvaglia, Ponto Valentino), **ucióri** (Lev.), **uciòri** (Biasca), **ücióri** (Biasca), **ucióru** (Malvaglia), **uciòura** (Aquila), **ucióuri** (Cavagnago, Anzonico), **ucióvri** (Cavagnago, Sobrio, Anzonico), **uciú** (Cevio), **uciúar** (Poschiavo), **uciura** (Ble.), **uciuri** (Lodrino), **uciuru** (Malvaglia), **uciuu** (Cevio, Cerentino, Auressio), **üciuu** (Lavertezzo), **utóber** (circ. Giubiasco, Ravecchia, Ons., Indemini, Mesocco, Bedano), **utóbra** (Isone, Novaggio), **utóbre** (Bedano, Peccia, Verscio, Osco), **utóbri** (Campo VMa., Ons.), **utóbru** (Giubiasco, Daro, Torricella-Taverne, Cademario, Grancia), **utóvra** (Novaggio), **utúar** (Posch.), **utuber** (Montecarasso, Medeglia, Sementina), **utubra** (Sigirino), **utubre** (Medeglia, Robasacco, Cavergno), **utubri** (Linescio, Campocologno), **utubru** (Medeglia, Bosco Lug., Mendr.)

s.m.

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

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

locuzioni:

ná in otobar Andare in calore: del caprone **(Vira Gamb.)** andare, essere in calore

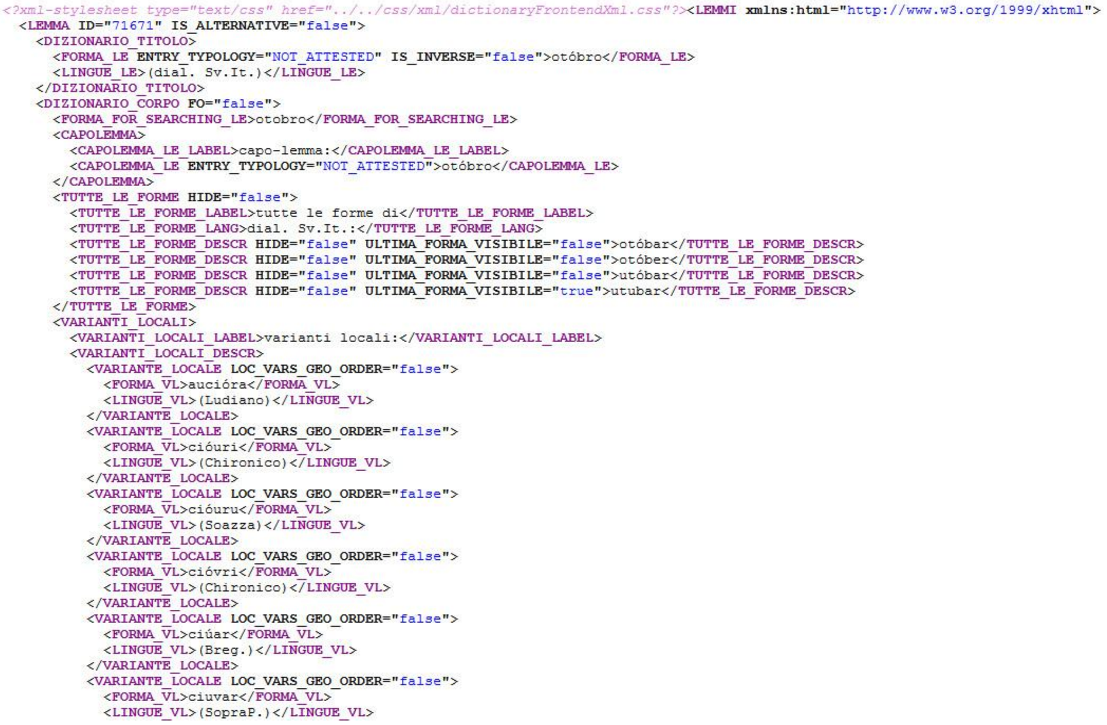
otóbro **cocóber** Formula reduplicativa con funzione enfatica, che compare in alcuni detti e proverbi **(Isone, Gamb., Pura, Gandria)** rimandi da: cucóbra

rimandi:

bócc d'otóber **(Castaneda)**
Madòna d'otóbar **(dial. Sv.It.)**
scurpi d'otóber **(Ascona)**
tè d'utúar **(Poschiavo)**

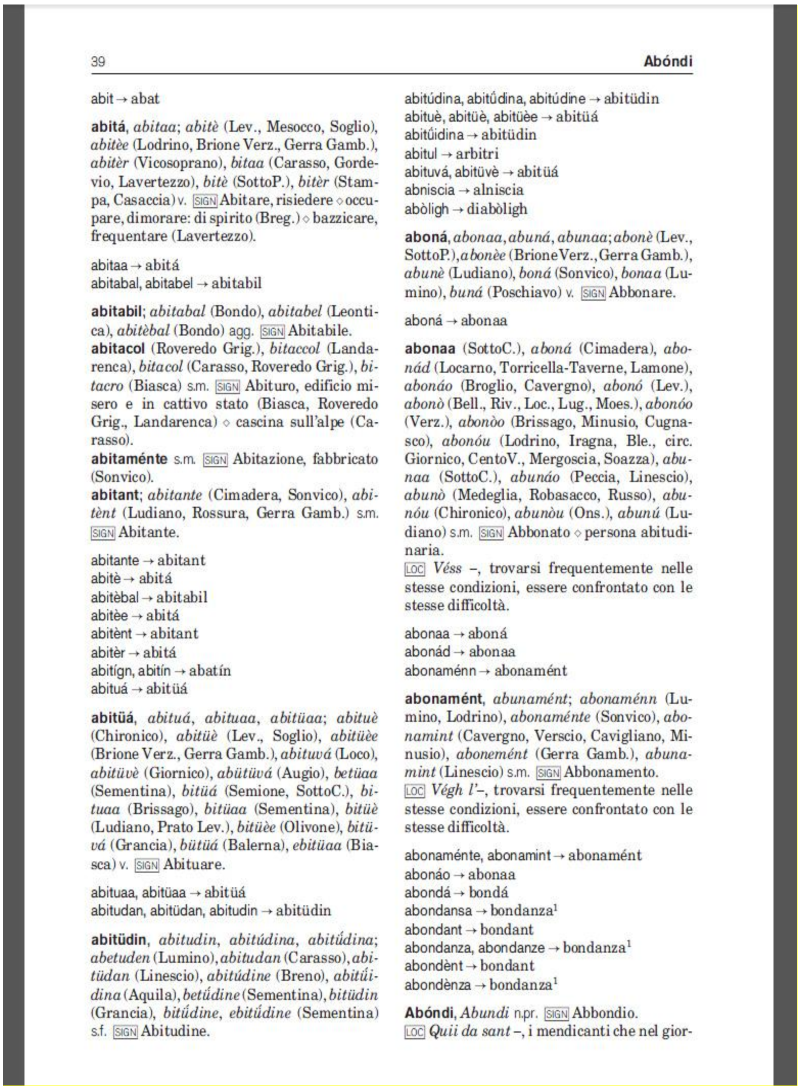
Fig. 4: The same entry in a human-readable form.[1]

[1] 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 type="text/css" href="../../css/xml/dictionaryFrontendXml.css"?><LEMMI 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>
    <DIZIONARIO_CORPO FO="false">
      <FORMA_FOR_SEARCHING_LE>otobro</FORMA_FOR_SEARCHING_LE>
      <CAPOLEMMA>
        <CAPOLEMMA_LE_LABEL>capo-lemma:</CAPOLEMMA_LE_LABEL>
        <CAPOLEMMA_LE ENTRY_TYPOLOGY="NOT_ATTESTED">otóbro</CAPOLEMMA_LE>
      </CAPOLEMMA>
      <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">utóbar</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">
            <FORMA_VL>ciuvar</FORMA_VL>
            <LINGUE_VL>(SopraP.)</LINGUE_VL>
```

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



39 **Abóndi**

abit → abat

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

abitaa → abitá
abitabal, abitabel → abitabil

abitabil; *abitabal* (Bondo), *abitabel* (Leontica), *abitèbal* (Bondo) agg. SIGN Abitabile.
abitacol (Roveredo Grig.), *bitaccol* (Landarenca), *bitacol* (Carasso, Roveredo Grig.), *bitacro* (Biasca) s.m. SIGN Abituro, edificio misero e in cattivo stato (Biasca, Roveredo Grig., Landarenca) ◇ cascina sull'alpe (Carasso).
abitaménte 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
abitígn, abitín → abatín
abituá → abitüá

abitüá, *abituá*, *abituaa*, *abitüaa*; *abituè* (Chironico), *abitüè* (Lev., Soglio), *abitüèe* (Brione Verz., Gerra Gamb.), *abituvá* (Loco), *abitüvè* (Giornico), *abütüvá* (Augio), *betüaa* (Sementina), *bitüá* (Semione, SottoC.), *bituaa* (Brissago), *bitüaa* (Sementina), *bitüè* (Ludiano, Prato Lev.), *bitüèe* (Olivone), *bitüvá* (Grancia), *bütüá* (Balerna), *ebitüaa* (Biasca) v. SIGN Abituare.

abituaa, abitüaa → abitüá
abitudan, abitüdan, abitudin → abitüdin

abitüdin, *abitudin*, *abitúdina*, *abitǘdina*; *abetuden* (Lumino), *abitudan* (Carasso), *abitüdan* (Linescio), *abitúdine* (Breno), *abitǘidina* (Aquila), *betǘdine* (Sementina), *bitüdin* (Grancia), *bitǘdine*, *ebitǘdine* (Sementina) s.f. SIGN Abitudine.

abitúdina, abitǘdina, abitúdine → abitüdin
abituè, abitüè, abitüèe → abitüá
abitǘidina → abitüdin
abitul → arbitri
abituvá, abitüvè → abitüá
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* (Broglio, Cavergno), *abonó* (Lev.), *abonò* (Bell., Riv., Loc., Lug., Moes.), *abonóo* (Verz.), *abonòo* (Brissago, Minusio, Cugnasco), *abonóu* (Lodrino, 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, Lodrino), *abonaménte* (Sonvico), *abonamint* (Cavergno, 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
abondá → bondá
abondansa → bondanza^1
abondant → bondant
abondanza, abondanze → bondanza^1
abondènt → bondant
abondènza → bondanza^1

Abóndi, *Abundi* n.pr. SIGN Abbondio.
LOC *Quii da sant* –, i mendicanti che nel gior-

Fig. 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 (*fig. 7*): the simple one detects “*soundslike* typical mistakes”; the advanced one detects “linguistic-background driven mistakes”. See fig. 8 and fig. 9 for MS Word and web interface of the “dialectal” spellchecker (Zoli, 2008).

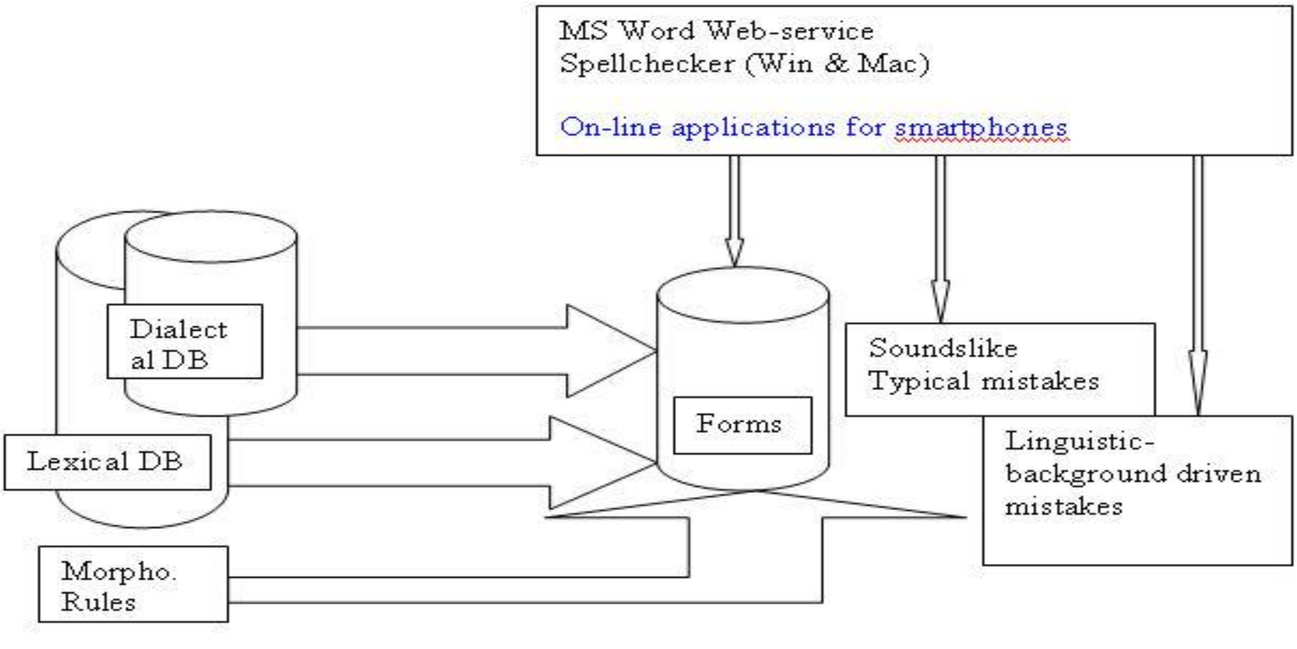
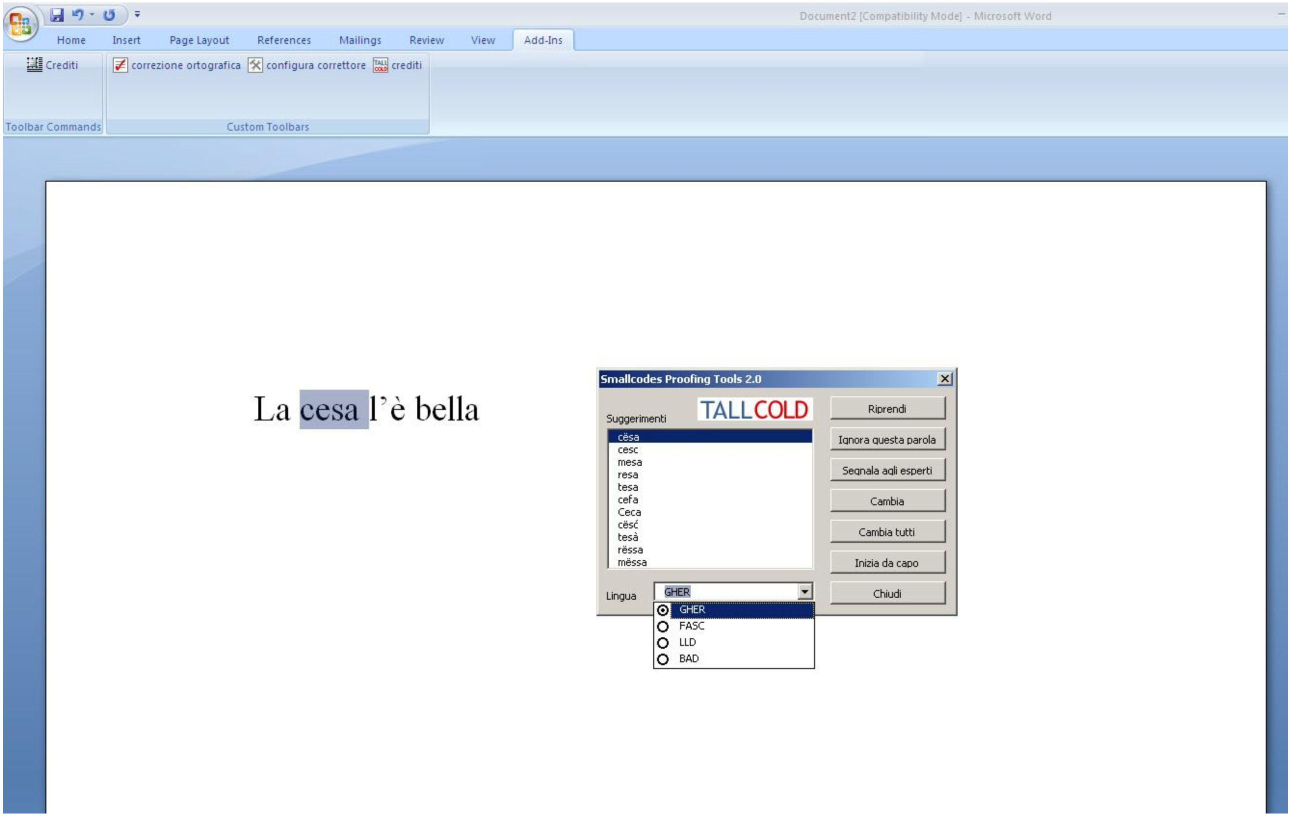


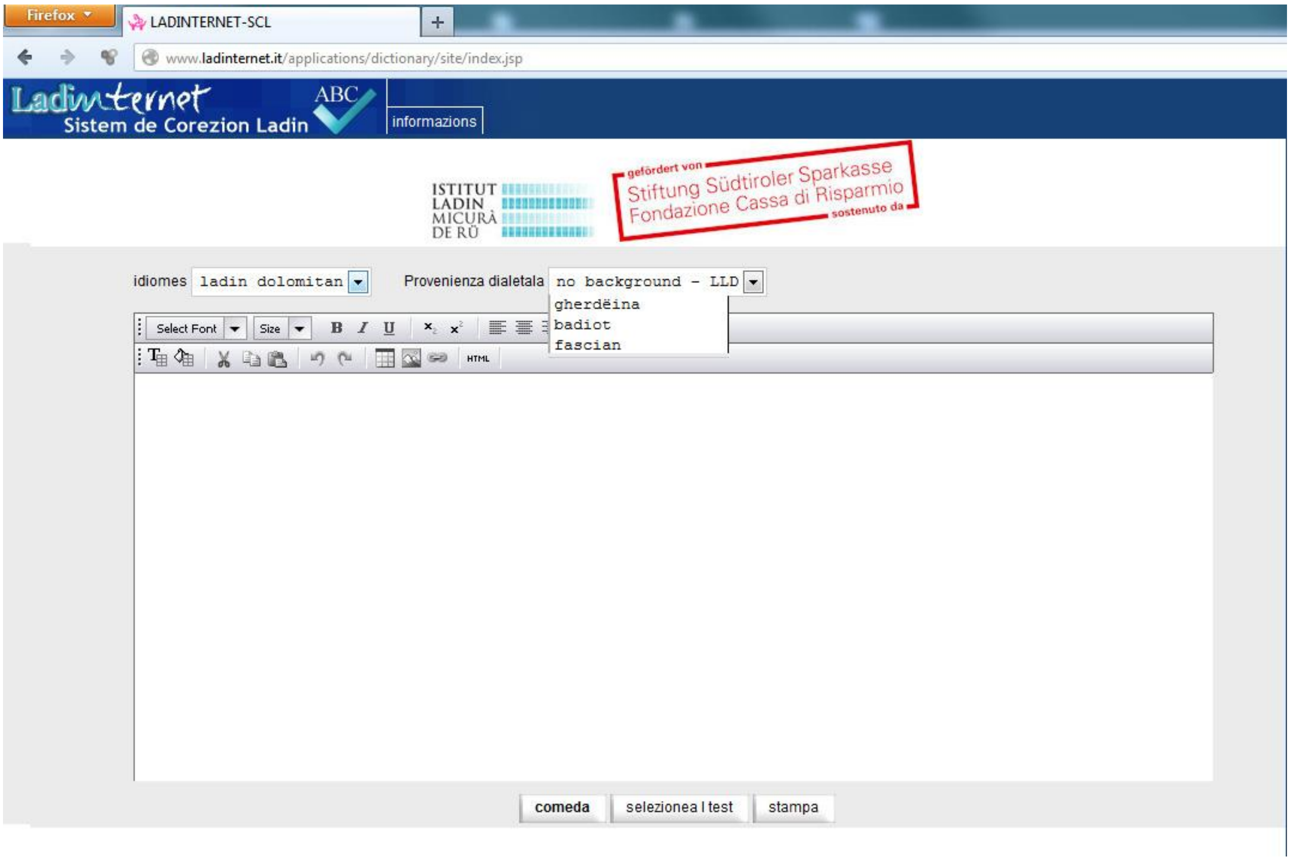
Fig. 7: functioning of an advanced spell-checking system



La cesa l'è bella

Lingua: GHER

Fig. 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.



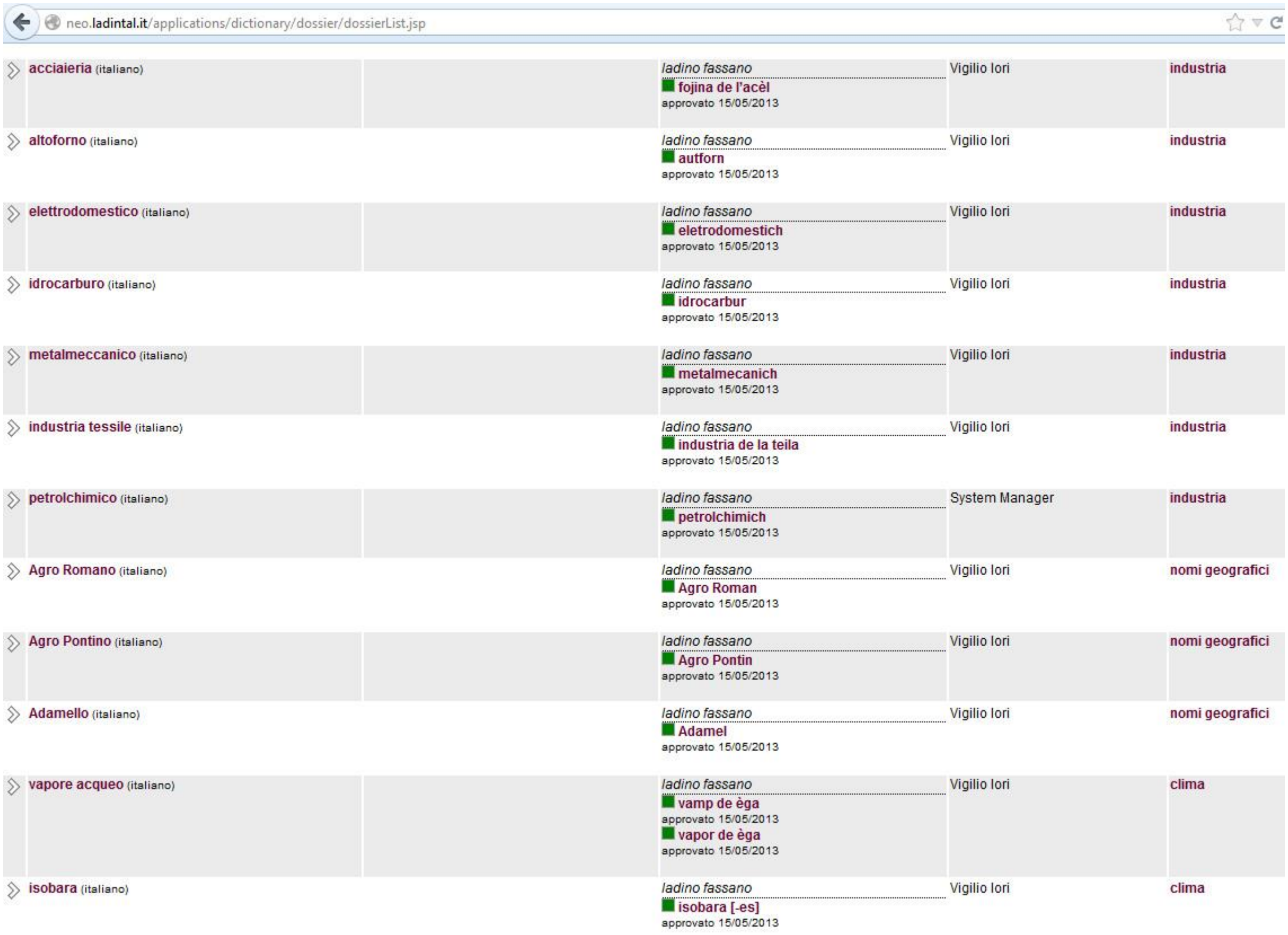
www.ladinternet.it/applications/dictionary/site/index.jsp

idiomes: ladin dolomitan

Provenienza dialetala: no background - LLD

Fig. 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.

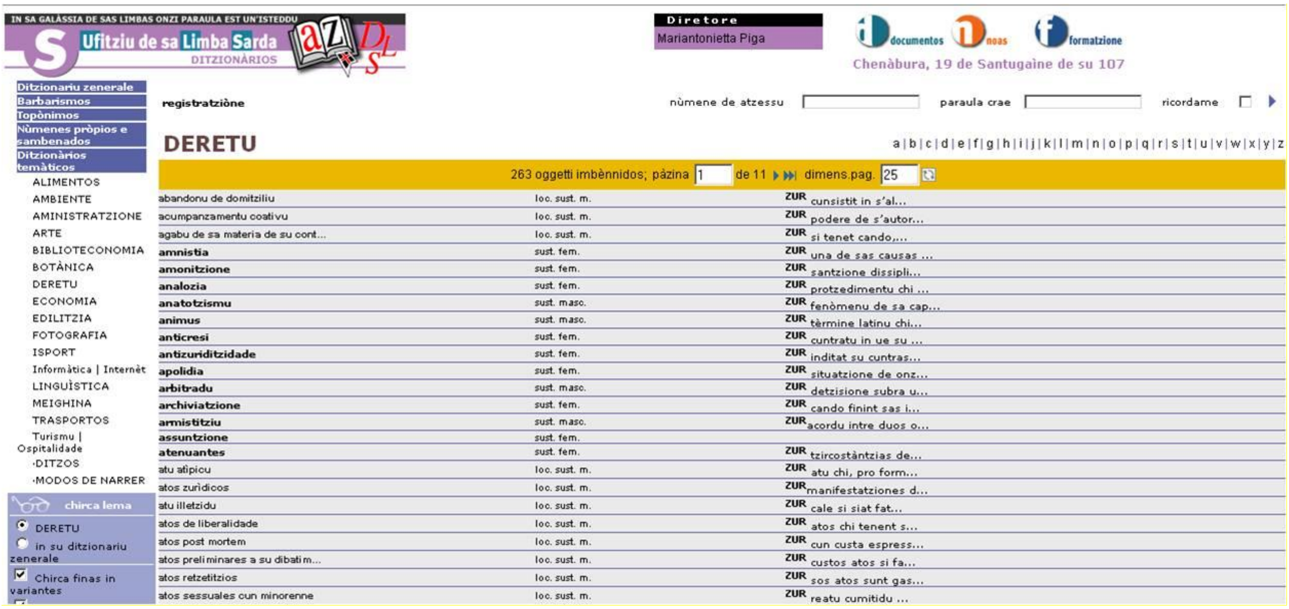
The **third step** is the terminology module. The creation of the terminology is a fundamental procedure if we want the language to be employed, for example, in school teaching (see for example fig. 10, which shows a collaborative webTool for neology, used by the authors of schoolbooks in Ladin Dolomitan), 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* 🟩 fojina de l'acèl approvato 15/05/2013	Vigilio Iori	industria
altoforno (italiano)		*ladino fassano* 🟩 autforn approvato 15/05/2013	Vigilio Iori	industria
elettrodomestico (italiano)		*ladino fassano* 🟩 eletrodomestich approvato 15/05/2013	Vigilio Iori	industria
idrocarburo (italiano)		*ladino fassano* 🟩 idrocarbur approvato 15/05/2013	Vigilio Iori	industria
metalmeccanico (italiano)		*ladino fassano* 🟩 metalmecanich approvato 15/05/2013	Vigilio Iori	industria
industria tessile (italiano)		*ladino fassano* 🟩 industria de la teila approvato 15/05/2013	Vigilio 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	Vigilio Iori	nomi geografici
Agro Pontino (italiano)		*ladino fassano* 🟩 Agro Pontin approvato 15/05/2013	Vigilio Iori	nomi geografici
Adamello (italiano)		*ladino fassano* 🟩 Adamel approvato 15/05/2013	Vigilio Iori	nomi geografici
vapore acqueo (italiano)		*ladino fassano* 🟩 vamp de èga approvato 15/05/2013 🟩 vapor de èga approvato 15/05/2013	Vigilio Iori	clima
isobara (italiano)		*ladino fassano* 🟩 isobara [-es] approvato 15/05/2013	Vigilio Iori	clima

Fig. 10: An example of the work flow (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.



Diretore
Mariantonietta Piga

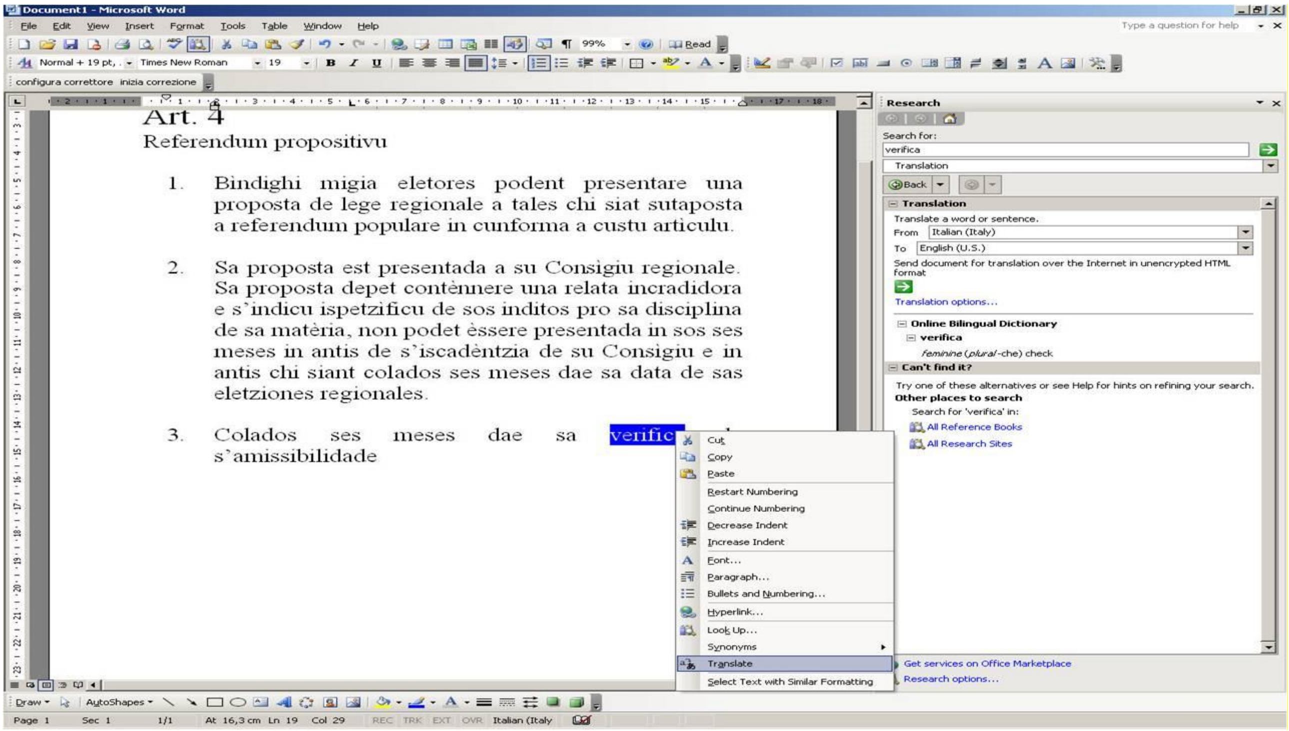
Chenàbura, 19 de Santugaine de su 107

DERETU

263 oggetti imbènnidos; pàzina 1 de 11 dimens.pag. 25

abandonu de domitziliu	loc. sust. m.	ZUR cunsistit in s'al...
acumpanzamentu coativu	loc. sust. m.	ZUR podere de s'autor...
agabu de sa materia de su cont...	loc. sust. m.	ZUR si tenet cando,...
amnistia	sust. fem.	ZUR una de sas causas ...
amonitzione	sust. fem.	ZUR santzione dissipli...
analozia	sust. fem.	ZUR protzedimentu chi ...
anatotzismu	sust. masc.	ZUR fenòmenu de sa cap...
animus	sust. masc.	ZUR tèrmine latinu chi...
anticresi	sust. fem.	ZUR cuntratu in ue su ...
antizuriditzidade	sust. fem.	ZUR inditat su cuntras...
apolidia	sust. fem.	ZUR situatzione de onz...
arbitradu	sust. masc.	ZUR detzisione subra u...
archiviatzione	sust. fem.	ZUR cando finint sas i...
armistitziu	sust. masc.	ZUR acordu intre duos o...
assuntzione	sust. fem.	
atenuantes	sust. fem.	ZUR tzircostàntzias de...
atu atìpicu	loc. sust. m.	ZUR atu chi, pro form...
atos zurìdicos	loc. sust. m.	ZUR manifestatziones d...
atu illetzidu	loc. sust. m.	ZUR cale si siat fat...
atos de liberalidade	loc. sust. m.	ZUR atos chi tenent s...
atos post mortem	loc. sust. m.	ZUR cun custa espress...
atos preliminares a su dibatim...	loc. sust. m.	ZUR custos atos si fa...
atos retzetitzios	loc. sust. m.	ZUR sos atos sunt gas...
atos sessuales cun minorenne	loc. sust. m.	ZUR reatu cumitidu ...

Fig. 11: Web page output of terminology module



Art. 4

Referendum propositivu

1. Bindighi migia eletores podent presentare una proposta de lege regionale a tales chi siat sutaposta a referendum populare in cunforma a custu articulu.

2. Sa proposta est presentada a su Consigiu regionale. Sa proposta depet contènnere una relata incradidora e s'indicu ispetzificu de sos inditos pro sa disciplina de sa matèria, non podet èssere presentada in sos ses meses in antis de s'iscadèntzia de su Consigiu e in antis chi siant colados ses meses dae sa data de sas eletziones regionales.

3. Colados ses meses dae sa verific s'amissibilidade

Research

Search for: verifica

Translation

Back

Translation

Translate a word or sentence.

From Italian (Italy)

To English (U.S.)

Send document for translation over the Internet in unencrypted HTML format

Translation options...

Online Bilingual Dictionary

verifica

feminine (*plural* -che) check

Can't find it?

Try one of these alternatives or see Help for hints on refining your search.

Other places to search

Search for 'verifica' in:

All Reference Books

All Research Sites

Get services on Office Marketplace

Research options...

Fig. 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*)[2] 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 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 enables 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 form 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 (Krauwer, 2003) and LCTL at the Linguistic Data Consortium rely on somewhat similar ideas [3]

The main difference is that this projects 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 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

[2] http://corpuslad.ladintal.it/

[3] http:// projects.ldc.upenn.edu/LCTL/index.html

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 [4], DoBeS[5]), 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. Namely

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)
- 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.[6]

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: □ / □; □ / □; □ / □; □ / □; □ / □ (Boukous, 2009).

[4] http://www.language-archives.org/ 21/07/2013

[5] http://dobes.mpi.nl/ 21/07/2013

[6] http://lucene.apache.org/ 29/05/2013.

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.

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; Iannàccaro, 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 fur Dialektologie and Linguistik, 1976.

Krauwer, 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 NAACLT 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/Macomer, 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'inventérr e d'valantaij 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
- Ufici Lenghe **Furlane** – Provincia di Udine
- Agjenzie regjonâl pe lenghe **furlane** (ArLeF)