Sardinian on Facebook: Analysing Diatopic Varieties through Translated Lexical Lists

Irene Russo
ILC CNR
Pisa
irene.russo@ilc.cnr.it

Simone Pisano
Università Guglielmo Marconi
Roma
s.pisano@unimarconi.it

Claudia Soria
ILC CNR
Pisa
claudia.soria@ilc.cnr.it

Abstract

English. Presence of regional and minority languages over digital media is an indicator of their vitality. In this paper, we want to investigate quantitative aspects of the use on Facebook of the Sardinian language. In particular, we want to focus on the co-existence of diatopic varieties. We extracted linguistic data from public pages and, through the translation of the most frequent words, we find out similarities and differences between varieties.

Italiano. *La presenza e l' uso delle lingue regionali e minoritarie sui mezzi digitali è un indicatore della loro vitalità. In questo lavoro vogliamo concentrarci sugli aspetti quantitativi del sardo usato su Facebook. In particolare, vogliamo analizzare le varietà diatopiche estraendo i dati linguistici dalle pagine pubbliche. Mediante la traduzione delle parole più frequenti abbiamo trovato similarità e differenze tra le varietà.*

1 Introduction

Everyday life makes an increasingly extensive use of digital devices that involve language use; for this reason, usability of a language over digital devices is a sign for that language of being modern, relevant to current lifestyles and capable of facing the needs of the XXI century. A positive correlation between presence in new technologies and better appreciation of a language has been repeatedly observed in the literature, see for instance (Eisenlohr, 2004) and (Crystal, 2010). Regional and minority languages (RMLs henceforth) are usually very poorly represented digitally (Soria, 2016).

Since poor digital representation of regional and minority languages further prevents their usability on digital media and devices, it is extremely important to enhance every bottom-up effort that can boost the quantity of available digital content. In fact, if the perception of the marginal role and limited applicability of RMLs persists, their attractiveness diminishes.

An increase in quantity of digital content available online represents today an opportunity for regional and minority languages. Online speakers can make visible the existence of a community that uses the language to interact; they can use online communication to converge toward a standard and they can instruct less skilled speakers toward better mastering of the rules of the language, especially when the language is not formally included in education. From the perspective of computational linguistics, the presence of digital content written in RMLs means that corpora can be built for them and basic tools (lemmatizers, spell checkers, lexicons etc.) can be developed.

The presence of RMLs over digital media and their usability through digital devices is often limited to instances of digital activism and/or by means of cultural initiatives focused on the preservation of cultural heritage.

In this paper we promote the first study we are aware of about the use on social networks (more specifically, Facebook) of Sardinian, an Italian minority language characterised by the coexistence of varieties and the difficulties for the promoted standard to emerge as unifying factor. Our starting hypothesis concerned the vitality on social networks of a language that is mainly spoken. With the help of a Sardinian linguist, we identi-

fied a small set of FB public groups where specific varieties of Sardinian are chosen as their main language plus groups where generic, not further defined Sardinian is used to communicate. We extracted messages from these pages and created a frequency lexicon for each variety. The most frequent 150 words have been translated by a Sardophone expert linguist with the aim of finding differences and commonalities between varieties. This preliminary analysis is the first step toward the use of computational linguistics methodologies in the promotion of a standard for Sardinian based on quantitative data.

2 Sardinian today: Main Varieties and Standardization Efforts

Sardinian is an autonomous Romance language spoken in the island of Sardinia. According to (Lupinu, 2007) it is known by approximately 68,4% of the population of the island. Ethnologue[1] lists four varieties for Sardinian: Northwestern Sardinian or Sassarese (100,000 speakers ca.), Campidanese (500,000 speakers ca.), Central Sardinian or Logudorese (500,000 speakers ca.) and Gallurese (100.000 speakers ca.)

The most important differences from a lexical, phonological and morphological point of view within Sardinian can be found between Central-Southern and Central-Northern dialects.

Scholars use to divide Sardinian in two main varieties: Logudorese and Campidanese, the first one spoken in the North and in the center of the island and the second one spoken in the South.

Logudorese and Campidanese can be related to two different pre-existing written standards: the so-called *Logudorese* (or *Logudorese illustre*) was used for the first time in a short poem at the end of the XV century (Manca, 2002), whereas what is known as Campidanese was the language of some religious plays at the end of the XVII Century (De Martini Abdullah Luca, 2006).

Today, Sardinian lacks of a generally agreed standard variety, although standardization efforts characterised the recent history of the Region.

The first attempt to introduce a written system based on an integration of phonetic, lexical and morphological features of modern Sardinian varieties was made in 2001, when the basic rules of LSU (*Limba Sarda Unificada*,Unified Sardinian Language) were presented (Blasco Ferrer, 2001). This proposal was sharply criticised by some sectors of the public opinion and strong disapproval came even from a part of native speakers, especially from the South, who considered this standard too much different from the language they spoke. It is a fact that it never became a model of official Sardinian.

In 2006, another model of written language was made official by the Regional Committee resolution n°16/14. This standard, called LSC (*Limba Sarda Comuna*, Common Sardinian Language)[2] made the effort of taking into account also the dialects of the transition region of the center mentioned earlier. Although regional administration recommended its use for written public documents it is still reluctantly accepted by some speakers, who perceive it as too distant from the varieties they speak.

In 2010, the Provincial Council of Cagliari took a different course choosing with the Provincial Committee resolution n°17 a linguistic norm[3] based on literary language of Southern poets and writers, in order to draw up acts, documents and even textbooks for primary children.

All these standardization efforts, politically guided or emerged bottom-up, clearly show that Sardinian speakers are aware of the role of standard orthography and grammar for the vitality and the survival of their language. On the one hand, they want to promote the idea of a unique language as a matter of identity; on the other, they dont want to lose local peculiarities by adopting standard rules that inevitably hide some local differences.

Social media are widely used by Sardinian speakers and they represent an interesting scenario for written but informal use of the language. An in-depth analysis of the type of language used by Sardinian speakers on social media is still missing. Certainly, use of everyday Sardinian in spoken and written (online) informal communication, is a sign of vitality of the language. Interaction is a powerful instrument for standardization, and the interactive modality offered by social media could reveal the emergence of coordination strate-

[1] www.ethnologue.com

[2] Regione Autonoma della Sardegna (2006), Limba Sarda Comuna. Norme linguistiche di riferimento a carattere sperimentale per la lingua scritta dellAmministrazione regionale, Cagliari, Regione Autonoma della Sardegna.

[3] Arrègulas po sortografia, sa fonètica, sa morfologia e su fueddàriu de sa bariedadi Campidanesa de sa lìngua sarda (Rules for orthography, phonetic, morphology and the vocabulary of Campidanese variety of Sardinian language)

gies toward a standard in speakers community as a natural need (Burghardt, 2016). To check this hypothesis, we started to analyse the use of different varieties of Sardinian that is being made on Facebook. According to the preliminary data of a recent survey, Facebook is the social media that is most used by Sardinian speakers, and where Sardinian is actively and extensively used[4].

3 Data Extraction and Analysis

We selected public pages and communities on Facebook that are rich in content and interactions between users. With the help of a Sardinian linguist we identified four mutually exclusive sets:

- pages where people communicate in LSC;
- pages where people communicate in Sardinian without further specification of the chosen variety;
- pages where people communicate in Campidanese;
- pages where people communicate choosing a local variety (in our case Nugoresu, local variety of Logudorese).

All the messages have been extracted from the json of the pages obtained through Facebook API. Lowercase texts have been tokenized splitting on whitespaces. Four frequency lists have been created, emoticons and symbols have been deleted. The 150 most frequent words have been translated in Italian by a Sardinian linguist that provided also PoS and morphological annotation plus all the available translations in case of polysemous words. We left in these lists Italian words because every cleaning procedure (lists of Italian words, PoS for Italian etc.) was risky: very frequent words in Sardinian can be found in Italian too (e.g. *a, chi, bonus, cosa*) with a different meaning.
Table 1 reports basic statistics about public pages and communities in the four sets listed above. Active users are the ones who wrote at least one message on the page. Number of active users and messages varies for each set but it was not possible to get a balanced sample.
In Table 2 the number of tokens and types for the four sets of Facebook groups analysed are reported. In Table 3 each possible pair of varieties is compared by checking the overlapping of translations into Italian. The second column reports how many Italian types are in common between two varieties. For example, among the most frequent 150 LSC word forms and the 150 most frequent *Sardu* word forms, 61 words have the same Italian translation. The third column contains the number of words with the same word forms in the two varieties compared, e.g. the Italian adjective *grande* has the same word form (*mannu*) in Nugoresu and Campidanese. This is a first attempt to understand if two varieties are close orthographically, considering the orthographic forms of the analysed words. We also report the number of content words found in each pair because we believe that in the future the overlapping at orthographic level should be analysed taking into account the distinction between content and function words.
The fourth column contains the number of the word forms related to the types in common which are different in the two varieties e.g. for the Italian word *è*, third singular person of verb *to be* in the present form, LSC has just one word *est*, while Campidanese has *est* and *esti*. In this case *esti* is counted as a different form and is included in the table under the fourth column.
Table 4 summarises for each pair how variability patterns are distributed, where pattern 1 to 2 means that there is one word form for variety *a* that correspond to two word forms for variety *b*. We know that the group *Sardu* contains data from more than one variety and we plan as future work a more detailed analysis. For the moment we note a clear overlapping because speakers of LSC contribute with posts and comments on pages where people communicate in Sardinian. For the same reason, when *Sardu* is one of the item in the pair we notice more variability patterns (see Table 4).
Concerning the comparisons between LSC and the two main varieties Campidanese and Logudorese, represented in our data by the local variety Nugoresu, we found evidence of the distance between the two main varieties with an overlapping of 41,5% in terms of word forms. LSC and Campidanese have an overlapping of 64,2% while LSC and Nugoresu have an overlapping of 83%. LSC emerges as a variety that tried to set a linguistic common ground and achieved this result, even if there is a bias toward Logudorese variety, one of

[4]Preliminary data of the DLDP Survey (www.dldp.eu) "Su Sardu: una limba digitale?". In July 2016, Facebook appears to be used by 98,1% of the respondents. Of those, 44% use Sardinian for writing and reading posts and messages, and 32,5% only for reading.

page name	type	#users	#active users	#messages	#variety
LSC, Limba Sarda Comuna: Sotziedade pro sa limba sarda comuna	Community	590	27	160	LSC
Iscritores in limba sarda	Public Group	331	49	916	LSC
Amigosde-sa-Limba-Sarda-Comuna	Community	1673	13	40	LSC
Solu in sardu	Public Group	15890	5701	373430	generic Sardinian
Solu poesias	Public Group	2018	158	1679	generic Sardinian
Scrieusu in campidanesu	Public Group	1984	576	17960	Campidanese
Cabuderra lngua e cultura	Public Group	116	1	18	Campidanese
Sos chi li piacheta faveddare e a iscrivere nugoresu	Public Group	984	438	1157	Nugoresu

Table 1: Basic statistics about data extracted.

FBgroup	tokens	types fr >10
LSC	71018	847
Sardu	3300408	18248
Campidanese	257110	2285
Nugoresu	379802	3412

Table 2: Basic statistics about token and types.

the complaints of Campidanese speakers (see par. 2).

4 Conclusion and Future Work

In this paper we address the following open question: could quantitative analysis of written data help Sardinian community to find out a common core (not specific of a variety) that could reinvigorate the idea of a standard? We plan future work on this issue, with the awareness that digital content on social media is both an opportunity and a challenge for this kind of analyses.

This paper is a first analysis of diatopic varieties of Sardinian through orthographical comparisons of word forms with the same meaning. Thanks to translated lists it was possible to look at commonalities and differences between varieties. Social media are a source of real data about language uses and the best observatory for regional and minority languages. Concerning Sardinian Facebook offers the possibility to test the distance between the proposed orthographic standard and the existing varieties. We will test the interplay between varieties with other methodologies to measure the distance and to find out usage patterns (e.g. Levenshtein distance for similar words).

This work is being carried out in the framework of the project DLDP (Digital Language Diversity Project, http://www.dldp.eu). DLDP is a three year project funded under the Erasmus+ programme. It aims at addressing the problem of low digital representation of EU regional and minority languages by giving their speakers the intellectual and practical skills to create, share, and reuse online digital content. DLDP fully embraces a bottom-up approach to language revitalization by addressing the speakers cognitive and practical skills as the cornerstone of effective revitalization initiatives.

Acknowledgments

This work is partially funded by the Erasmus + DLDP Project (Grant Agreement no. 2015-1-IT02-KA204-015090). The opinions expressed reflect only the authors view and the Erasmus+ National Agency and the Commission are not responsible for any use that may be made of the information contained.

References

Blasco Ferrer E., Bolognesi, R. et al. 2001. *Limba Sarda Unificada. Sintesi delle norme di base: ortografia, fonetica, morfologia, lessico*. Cagliari, Regione Autonoma della Sardegna.

Burghardt, M., Granvogl, D. and Wolff, C. 2016. Creating a Lexicon of Bavarian Dialect by Means of Facebook Language Data and Crowdsourcing. *Proceedings of LREC-2016*. Portoroz, Slovenia.

Crystal, D. 2010. *Language Death*. Cambridge University Press.

De Martini Abdullah Luca (ed.) 2001. *Libro de Comedias (by Antonio Maria da Esterzili)*. Cagliari, Cuec.

Eisenlohr, P. 2004. Language revitalization and new technologies: Cultures and electronic mediation and the refiguring of communities. *Annual Review of Anthropology*. 18(3):339361.

Lupinu, G., Mongili, A. , Oppo, A. , Spiga, R. , Perra, S. , Valdes, M. 2007. *Le lingue dei sardi: una ricerca sociolinguistica*. Assessorato alla Pubblica istruzione, beni culturali, informazione, spettacolo e sport, Regione Autonoma della Sardegna.

De Martini Abdullah Luca (ed.) 2002. *Sa Vitta et sa Morte, et Passione de sanctu Gavinu, Prothu et Januariu (by Antonio Cano)*. Cagliari, Cuec.

	common_types	types_with_same_word_forms	types_with_different_word_forms
LSC-Sardu	61	60 (21 content words)	32 (11 content words)
LSC-Campidanese	67	43 (14 content words)	44 (17 content words)
LSC-Nugoresu	65	54 (14 content words)	39 (17 content words)
Sardu-Campidanese	65	47 (15 content words)	64 (26 content words)
Sardu-Nugoresu	70	64 (16 content words)	65 (33 content words)
Campidanese-Nugoresu	81	34 (12 content words)	82 (27 content words)

Table 3: Comparison between Sardinian varieties.

	1 to 2	1 to 3	1 to 4
LSC - Sardu	8	4	1
LSC - Campidanese	3	0	0
LSC - Nugoresu	7	1	1
Sardu - Campidanese	5	1	5
Sardu - Noguruse	8	3	5
Campidanese - Nugoresu	2	0	1

Table 4: Comparison between Sardinian varieties.

Soria, C., Russo, I. , Quochi, V., Hicks, D., Gurrutxaga, A., Sarhimaa, A. and Tuomisto, M. 2016. Fostering digital representation of EU regional and minority languages: the Digital Language Diversity Project. *Proceedings of LREC-2016*. Portoroz, Slovenia.

Virdis, M. 1988. Sardisch: Areallinguistik / Aree linguistiche. Holtus G., Metzeltin, M., Schmitt, C. (eds.), *Lexicon der Romanistischen Linguistik* 4, Tubingen, Max Niemeyer, pp. 897-913.

Wagner, M. L. 1997. *La lingua Sarda. Storia, spirito e forma*. Nuoro, Ilisso.