Encouraging the presence in the cyberspace of the lesser used languages through writing and proofing tools: the case of Sardinian language

Carlo Zoli
Smallcodes / Italy
carlo.zoli@smallc<odes.com
November 21, 2012

SARDINIAN LANGUAGE

The Sardinian Language has a special place among Romance Languages: in the classical manual of Tagliavini ‹‹Il sardo ha una sua speciale fisionomia e individualità che lo rende, in certo qual modo, "il più caratteristico degli idiomi neolatini"››.

If we accept the classical, and somewhat arbitrary subidivision of romance languages into 9 families (Italic / Occitan / Catalan / Langues d'oil / Romanian / Iberic / Rheto-Romance / Sardinian / Dalmatic, which is extinct), only Sardinian and R-R (Ladin / Romantsch / Friulian) have not developed at present day a well international recognised literary, standard form. Curiously the attestations of Sardinian, clearly recognisable in early times, are the first and the most abundant ones among Latin vulgars, and they date back at least to IX-X-XIth century. Sardinian has come close to achieving this result, in medieval times, to becoming the official language of Sardinia, in the short period of self-government of the "*Giuigados*"; but then political, and of course linguistic, history has taken another path, and Catalan, and then Spanish, and at last Italian have played this role, leaving this incredible offspring of Latin language in the position of an oral vernacular, dialectally fragmented and less and less conscious of its unity, with no standard written form (a proto-standard for poetry has developed, but scarcely for prose, and nothing at all for modern official purposes until the initially timid, then more earnest attempts in the latest 20 years).

THE BASIC UNIQUENESS OF SARDINIAN

As it has been pointed out by Hagège, it is typical of threatened languages to overestimate their internal diversity (he calls it "*purisme dialectal*"): internal diversity is

present in every natural language; in highly prestigious languages this diversity is often shadowed by the power and the prestige of the standard written languages, which covers, protects we can say, the oral varieties with a solid roof (*Dachsprache*), thus limiting their natural tendency, the drift towards diachronic and diatopic differentiation. Not surprisingly, since Sardinian people have begun to use Italian for every formal situation and more and more often for ordinary conversation, especially among speakers from different villages or towns (*biddas diferentes*), this internal variation has commenced to be perceived as a barrier for mutual understanding; but this has never been true for centuries. Traditional nomadic shepherds have, for generations, moved across the island adapting their variety to that of the areas where they were moving with cattle, also because phonetic variation in Sardinian is highly regular and predictable, and syntactic and lexical variation is not a problem for intercomprehension.

STANDARD WRITING → CYBERSPACE → PRESTIGE

I have three children and I often tell them – a little bit kidding - that only three are the things that make their world really different from the one I was living in when I was their age. These three things are Ryanair, the euro, and the internet. It is crystal clear that the Internet is one of the biggest revolutions of human history, and we have the privilege to assist at it, and, maybe, to make a little contribution to it. We do not understand the significance of this revolution yet, we have not seen the next big things yet (whatever they may be), I believe.

Schematically, we may divide the cyberspace into two or three sub-spaces (following the classical distinction between internet 1.0 and internet 2.0): the formal, top-down content and the new (not-so-new now) bottom-up (so called "user generated" / blog) and collaborative content.

Now everybody speaks about Facebook, Twitter, LinkedIn, the social networks, etc.: if we measure the twits and the posts, we may find that hundreds, or thousands, of languages are represented, and we may find some consolation here: http://indigenoustweets.com/

I talk about "user generated content" (SMS / social networks / e-mails) and "collaborative content" (Wikipedia) as it they were the same thing, which they are not, but I am not sure that that we can measure the 100k articles of Wikipedia in Basque and the 20k (maybe) in Sardinian on the same scale. These language that are basically out of the school, out of the Internet 1.0 "formal publishing industry", end to show up a very poor language in social media and also in "collaborative media"

SOCIOLINGUISTIC SITUATION IN CYBERSPACE

When I listened to my friend Claudia Sorìa, or to Rehm and Mariani, this morning, with all the difficulties they point out, I was caught by a deep sorrow: the languages I work with, I struggle together with, are at a stage where it is impossible even to think about "terminology extractors", or "resumé automatique", or "question answering".

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. I am talking about sociolinguistic quality, not literary or aesthetic value. It may be curious, and I often have to defend my position, especially in Italy, that an institution 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"? I say that 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 hadn't been… English as a world language. But if we want to foster our small languages in the real world, and in the cyberspace (the two things will tend to be partially 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, where I come from, 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 Shenzhen 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. How?

In this respect, the slide of Georg Rehm this morning about the Gartner hype cycle is of great importance: many "next big things" are 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 "january"?