Linguistic diversity in the information society*

Lars Borin
Språkbanken, Dept. of Swedish Language, University of Gothenburg
Gothenburg, Sweden
lars.borin@svenska.gu.se

Abstract: This presentation is intended to provide some background information as well as a broader picture of some of the issues involved in developing language technology – especially information extraction – for lower-density languages, against which to set the work presented in the other papers in this volume.

Keywords: Language resources, low-density languages, language technology, linguistic diversity

1 *Linguistic demography and language technology*

There are 5–7,000 languages spoken in the world today. The latest edition of the *Ethnologue* (Lewis, 2009) lists almost 7,000 living languages, but the actual number is difficult – arguably impossible – to determine, because of such factors as the arbitrary distinction between languages and dialects.

Their size in number of first-language speakers is very unevenly distributed. The top 30 languages in the world account for more than 60% of its population. At the other end of the scale, we find that most languages are spoken by quite small communities:

> There are close to 7,000 languages in the world, and half of them have fewer than 7,000 speakers each, less than a village. What is more, 80% of the world's languages have fewer than 100,000 speakers, the size of a small town. (Ostler, 2008, 2)

Linguists are concerned about the fact that many languages are threatened. According to one estimate (Krauss, 1992), half of the languages spoken in the world today will have gone extinct by the end of this century. This means that, on average, the last speaker of some language dies every two weeks.

To some students of language death, globalization is the number one culprit behind this development. The modern information and communication technologies so intimately connected with globalization have consequently sometimes been seen as accelerating global language extinction – as when television is referred to as "cultural nerve gas" by Krauss (1992, 6) – but sometimes also as carrying the potential to reverse it, or at least slow it down (Cunliffe and Herring, 2005; Nichols et al., 2005; Saxena, 2006).

1.1 Spoken, signed and written languages

The primary modalities of naturally occurring language are speech and sign.[1] Out of the 6,909 living languages listed in the *Ethnologue*, 126 are sign languages. I will have nothing further to say about these here, apart from noting that occasionally, there are contributions at language technology conferences dealing with sign language.

When I say that the primary modalities of language are speech and sign, I mean that there are numerous examples of languages that are spoken or signed only, i.e., with no writing, whereas the reverse seems not to occur. Historically, we assume that spoken (and signed) language had been in existence long before the first instances of writing appeared approximately 6,000 years ago. There are situations where a written language has survived beyond its spoken origins, but this is a different matter.

It is difficult to find solid estimates of how many written languages there are in the world today. The *Ethnologue* has a 'Script' entry – i.e., "Roman script", "Arabic script", etc. – for 2844 of its non-signed living lan-

* I am grateful to the language technology group at the University of the Basque Country for inviting me to present this talk at the SALTMIL workshop in Donostia.

[1] The hedge "naturally occurring" is motivated by the existence of deliberately invented languages, e.g., international auxiliary languages such as Esperanto or Volapük, or fictional languages such as Klingon, which at least in some cases are arguably first written and only later – if ever – used in the spoken mode.

guages. In addition to this there are 372 languages without script information, but where the *Ethnologue* states (under the heading of "Language development") that there are translations of (portions of) the Bible or the New Testament in this language, information that is also given for languages with a script entry entry ("[portions of the] Bible": 1090; "NT": 1080).

Thus, on the surface of it, according to the *Ethnologue*, it appears that more than half the world's languages are written. Here, it may be useful to distinguish between the mere existence of a writing system or script for a language, on the one hand, and whether there is a *tradition of writing* in the language, on the other, i.e., whether people write texts in this language on a regular basis, and in today's world, whether they communicate electronically in the language, e.g., by email or texting. For several centuries, a central activity of linguists and missionaries (often the same people) has been to devise orthographies for formerly unwritten languages, in order to translate the Bible and other religious works into these languages. However, in practice this means that the mere existence of an orthography for a particular language does not automatically mean that the speakers of the language use the orthography on a regular basis, or even that they are literate in this language. The role of the few religious writings will in this situation rather be similar to that of the Latin Bible in medieval Europe or the Quran in non-Arabic-speaking Muslim communities: a way of ensuring that the words of the Scripture do not become corrupted – a kind of linguistic freezer, as it were, used as a crutch for memory in oral presentation, rather than a means of communication.

For some languages the *Ethnologue* will tell us that they are "fully developed", meaning that "extensive literature and media exist" (according to the introduction of the *Ethnologue*). Only 62 languages are thus identified. This list is obviously too short, e.g., Basque, Faroese, Macedonian and Welsh are missing from it, to name a few conspicuous European cases. The languages identified as "fully developed" can surely be said to have a tradition of writing, and there are more such languages than those listed in the *Ethnologue*. On the other hand, it is probably true that the majority of those languages identified as either having a script or a Bible/NT translation, do not in fact have a genuine tradition of writing, in most cases because the language in question is a minority language, and literacy – if at all present – will be in another, majority or national language. A generous ballpark estimate would be that no more than 15–20% of the world's languages have a tradition of writing, i.e., on the order of a thousand languages, give or take a few hundred.

All this is relevant in the present context, because the most mature and sophisticated language technology is in effect written language technology; we work with texts, rather than speech, and with few exceptions, the applications that are discussed in this context presuppose a written language, and a standardized written language to boot. This is not to say that developing language technology aimed at primarily spoken languages would not be a worthy pursuit; on the contrary: I believe such a development could provide strong support for endangered languages. Here, I will limit myself to a discussion of written language technology, however, bexause this is where my competence lies.[2]

1.2 Lower-density languages

A related but at least partly separate issue from those discussed above is the matter of how well endowed a language is with the resources and tools necessary for the development of sophisticated language technology applications. The terminology in this area is motley, to say the least. I will be using a fairly neutral terminology which I believe was first introduced to the language technology community – the term itself is older – in the work of the Linguistic Data Consortium (LDC) in connection with the *surprise language exercise* arranged by the DARPA TIDES program[3] in 2003 (Oard, 2003; Strassel, Maxwell, and Cieri, 2003). They use the expression "density" to refer to the amount of digital language resources available in a language. Consequently, using this terminology, you can talk about high-, medium- and low-density (or lower-density)

[2] Note that even much of the speech technology that is being developed is geared toward the written language, i.e., speech-to-text and text-to-speech systems, although there are pure speech applications as well, such as spoken dialogue systems.

[3] The US Defence Advance Research Projects Agency Translingual Information Detection, Extraction and Summarization program.

languages, and even make some attempts to quantify what these terms could mean.[4]

Looking at the criteria used by LDC, reproduced below in appendix A, we see that they are applicable only to written languages. In this trivial sense, there is a one-way dependence between written languages and the density scale: The scale is not applicable to non-written languages. On the other hand, there is no correlation whatsoever with the size of a language. Large standard languages – those with numbers of native speakers in the hundreds and tens of millions and having a long tradition of writing – are not necessarily high- or even medium-density languages. This is often true for indigenous languages in former European colonies in all parts of the world.

A language that is widely used in all spheres of life will also tend to be used in those activities which give rise to linguistic resources. Conversely, if a language is confined to a few – perhaps mainly oral – situations where it is used, it will tend to lack such resources. It turns out that we are dealing here with a special case of the kind of linguistic power games that are studied in the linguistic subdiscipline known as sociolinguistics, which deals with the sociology of language and language use.

2 Sociology of language and language technology

It has been observed over and over again that the use or non-use of a language in a particular situation – where the language could in principle be used, but where there is a choice available between two or more languages – is intimately connected with the attitudes towards the language among the participants. This is perhaps the most reliable determiner of language use, and not factors such as effort, lack of vocabulary, etc., which in many cases seem to be post-hoc rationalizations motivating a choice made on attitudinal grounds. Another way of expressing this is that languages are more or less prestigious in the eyes of their speakers, and that linguistic inferiority complexes seem to be common in the world. As people are on the whole rational creatures, we may suspect that they have good reasons for eschewing their mother tongue in favor of another language. In the case of language shift, we often observe a pattern of parents speaking a more prestigious language to their children at home, rather than their first, less prestigious, language, even while paying lip service to the need for preserving the lower-status language, because they are grappling with

> [...] a conflict between wanting to do something for the language and wanting to improve the chances of the children to succeed in the macrosociety of which they are, and always will be, part. The linguist observing this state of affairs may feel regret at what is happening here; but if it is a fact that maintaining a small language at the expense of a major or national one means severely reducing prospects of an economically satisfactory life for one's children, does one have a right to blame the parents? (Winter, 1993, 311)

However, rather than taking status as an inherent and immutable characteristic of a language, we should see it for what it is, i.e., a perceived characteristic, something that lies in the eye of the beholder. As such, it can be influenced by human action. Important for our purposes here, is that it has been suggested that the creation of linguistic resources and language technology for a language may serve to raise its status.

Keeping this in mind, and also that, once we have started building language resources and language technology tools, we have set in motion a positive feedback loop. This is because the resources and tools are not independent entities; rather, as argued by Sarasola (2000), Borin (2006) and others, they can – somewhat idealized – be thought of as making up a multistoried edifice, where the lower levels form the prerequisites for those above them, all the way to the top. Researchers may differ in exactly on which level a particular linguistic resource should be located, but there seems to be a general consensus about this picture of things. The symbolic and statistical language technology communities certainly will have different opinions about how much human effort should be necessary at each step, although most machine-learning approaches used in language technology are supervised and consequently at some stage rely on human linguistic judgements, often in the form of linguistic resources manually

[4] "Lower-density language" as used here is thus to be understood as meaning the same thing as "less resourced language", used elsewhere in this volume.

created or annotated by human experts, or else automatically created or annotated, but manually checked and corrected. There is also a belief among many statistical language technology researchers that some aspects of linguistic analysis can be ignored with impunity in developing certain, even fairly sophisticated applications, the classical example being morphological analysis in information retrieval (Smeaton, 1997).

In considering how to accomplish linguistic resources and language technology tools for any language, at least two kinds of considerations will enter the picture at some stage. The most obvious one is to do with the mechanics of the whole enterprise; a tactical question, as it were: How can we in the quickest way, and spending the least amount of (human) effort, accomplish a particular set of resources for a language within a reasonable time frame, given a particular set of existing prerequisites? The other question is more strategic, and perhaps nore important in the long run: Given that we have limited resources – in terms of money, manpower and expertise – and that there is a choice of which resources we could realize within these limitations, how should we set our priorities? I will try to address both of these issues in turn in the following sections.

3 *Thrifty linguistic resource building for lower-density languages*

In the last few years, there has been an increased interest among the language technology research community in developing methodologies that would minimize both the data requirements and the human linguistic expertise needed for the creation of linguistic resources and language technology tools. Useful overviews have been presented by Maxwell and Hughes (2006), Borin (2006) and Streiter, Scannell, and Stuflesser (2006), among others.

However, looking at the literature, it seems that the only approaches that have so far produced substantial results are the non-statistical, grammar-based ones, such as the work described by Trosterud (2004), where finite-state morphological processors and constraint grammar-based disambiguation components are developed for a number of related languages. The fact that the languages are related is of great help when dealing with successive languages after the first one. The morphological component for the first language, North Sámi, required approximately 2.5 person-years of highly qualified linguistic expert work to reach the prototype stage, whereas the analogous module for the closely related Lule Sámi was completed in an additional six months (Trosterud, 2006). This and other work in the same vein reported in the literature – e.g., by Artola-Zubillaga (2004) and Maxwell and David (2008), to pick a couple at random – is characterized by deep and long-lasting involvement by linguistic expertise and further often by the creative use of digitized versions of conventional printed linguistic resources, especially dictionaries. The following observation is perhaps trivial, but bears stressing, since it is in fact often not heeded in practice: For this kind of approach to work, it is necessary that tools for providing systems with linguistic knowledge use a conceptual apparatus and notation familiar to the linguists who are supposed to be working with them.

On the other hand, most pure data-driven approaches reported in the literature are mainly small proof-of-concept experiments, which generally founder on the lack of evaluation data. Further, these approaches are data-hungry, which precludes their use with most low-density languages. There is much ongoing work addressing these issues, however, so we can probably expect some progress in this area.

In the surprise language exercise mentioned above (section 1.2), many of the teams achieved remarkable results in a very short time. For instance, the Sheffield team created a named-entity recognition (NER) system for Hindi in about one person-month, achieving an F-measure of slightly over 62% on news texts (Maynard et al., 2003).[5] This work was characterized by an eclectic, goal-driven approach to the problem at hand; all available data sources were utilized, and human volunteers were engaged to create, analyze or annotate data. Regarding the last point, one proposed way of enriching raw text resources and also of bootstrapping lex-

[5] Note, however, that the state of the art in NER is well over 90%, but that the performance of an NER system seems to be correlated mainly to the size and quality of its gazetteers, rather than to the kind of processing approach chosen (data-driven or grammar-based) (Johannessen et al., 2005).

ical resources is the "Wikipedia way", i.e., pooling voluntary efforts by many contributors into an open content resource (Streiter, Scannell, and Stuflesser, 2006). One well-known example is the Wiktionary project ⟨http://www.wiktionary.org/⟩; another example, more interesting as a language technology resource, is the free Swedish synonym dictionary project (Kann and Rosell, 2005).

In conclusion, if we want guaranteed results, there is still no way of avoiding good old-fashioned linguistics entirely. Some tasks are less linguistics-dependent than others, however – e.g., NER – and in some cases one may get away with more naive approaches provided that the interaction with the user is arranged in a suitable way that compensates for the lack of linguistic knowledge in the system, such as in typical web search engines or the cross-lingual NER system described by Steinberger and Pouliquen (2007).

4 Strategic considerations

It is often claimed that in order to survive as modern languages, low-density languages need to establish a presence in the information society. The sociopolitical situation of these languages varies enormously, however. There are large languages with a long literary tradition, which nevertheless live under the shadow of a former colonial language. Prototypical examples are South Asian languages such as Hindi or Tamil. Because English is a second, high-status, language among the technology-aware middle classes in South Asia, even family members will communicate among themselves by email or texting in English rather than in Hindi, their language of everyday oral communication (personal observation). In such a situation, will it make sense to offer language tools in support of Hindi?

The internet – the WWW and email – is becoming the central component of the information society. Increasingly, people use the internet as their main or only source of information and means of communication. In my view, there is an opportunity here for promoting language resources and language technology tools for low-density languages, for concrete practical aims as well as a means of raising the status of these languages.

The next generation of the World Wide Web has been touted as "the Semantic Web", where all the available information will be interlinked using logical representations and formal reasoning over these representations. It has been pointed out, perhaps most consistently by Yorick Wilks (Wilks, 2008; Wilks and Brewster, 2009), that the content of the Web which by some magical means will be turned into the logical representations of the Semantic Web, in fact is predominantly textual, and, we may add, increasingly multilingual. Wilks's conclusion is that the "magical means" will be nothing other than natural language processing, i.e., language technology, and that the key language technology for turning the textual web into the semantic web will be information extraction. To this we may add that technologies for interacting in natural language with the Semantic Web are likely to become increasingly important, e.g., Q&A and dialogue systems.

Consequently, those languages for which information extraction resources and tools will be available – either monolingual or as part of multi- and cross-lingual applications, will probably exhibit a more secure and prominent presence on the Semantic Web than those lacking such resources, and as a consequence, acquire the status in the eyes of their speakers that such a presence confers.

5 Conclusion

Strategically, then, it would make good sense to focus on those aspects of language resource and technology creation for a low-density language, which could be judged to facilitate the (rapid) development of suitable information extraction applications for it.[6]

In this way, they hopefully stand a good chance to carve a niche for themselves and the cultures of their language communities in the information society of the future, ensuring that the world of the Semantic Web remains a linguistically and culturally rich and diverse place.

[6] Suitable in the sense that they should be adapted to the kinds of information and genres available online in this language – mythological texts, traditional medicine, newspapers, and what have you.

References

Artola-Zubillaga, Xabier. 2004. Laying lexical foundations for NLP: The case of Basque at the *ixa* research group. In *SALTMIL workshop at LREC 2004: First steps in language documentation for minority languages*, pages 9–18, Lisbon. ELRA.

Borin, Lars. 2006. Supporting lesser-known languages: The promise of language technology. In Anju Saxena and Lars Borin, editors, *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*. Mouton de Gruyter, Berlin, pages 317–337.

Cunliffe, Daniel and Susan C. Herring. 2005. Introduction to minority languages, multimedia and the web. *New Review of Hypermedia and Multimedia*, 11(2):131–137.

Johannessen, Janne Bondi, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdóttir, Anders Nøklestad, Dimitrios Kokkinakis, Paul Meurer, Eckhard Bick, and Dorte Haltrup. 2005. Named entity recognition for the mainland Scandinavian languages. *Literary and Linguistic Computing*, 20(1):91–102.

Kann, Viggo and Magnus Rosell. 2005. Free construction of a free Swedish dictionary of synonyms. In *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, pages 105–110, Joensuu. University of Joensuu. Electronic resource: ⟨http://phon.joensuu.fi/lingjoy/01/kannrosell05F.pdf⟩.

Krauss, Michael. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Lewis, M. Paul, editor. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, sixteenth edition. Online version: ⟨http://www.ethnologue.com/⟩.

Maxwell, Michael and Anne David. 2008. Joint grammar development by linguists and computer scientists. In *Proceedings of the IJCNLP-08 workshop on NLP for less privileged languages*, pages 27–34, Hyderabad. Asian Federation of Natural Language Processing.

Maxwell, Mike and Baden Hughes. 2006. Frontiers in linguistic annotation for lower-density languages. In *Proceedings of the workshop on frontiers in linguistically annotated corpora 2006*, pages 29–37, Sydney. ACL.

Maynard, Diana, Valentin Tablan, Kalina Bontcheva, and Hamish Cunningham. 2003. Rapid customization of an information extraction system for surprise languages. *ACM Transactions on Asian language processing*, 2(3):295–300.

Nichols, David M., Ian H. Witten, Te Taka Keegan, David Bainbridge, and Michael Dewsnip. 2005. Digital libraries and minority languages. *New Review of Hypermedia and Multimedia*, 11(2):139–155.

Oard, Douglas W. 2003. The surprise language exercises. *ACM Transactions on Asian language processing*, 2(2):79–84.

Ostler, Nicholas. 2008. Is it globalization that endangers languages? In *UNESCO/UNU Conference 27–28 August 2008: Globalization and languages: Building our rich heritage*. UNU/UNESCO. ⟨http://www.unu.edu/globalization/2008/files/UNU-UNESCO_Ostler.pdf⟩.

Sarasola, Kepa. 2000. Strategic priorities for the development of language technology in minority languages. In *LREC 2000 workshop proceedings. Developing language resources for minority languages: Reusability and strategic priorities*, pages 106–109, Athens. ELRA.

Saxena, Anju. 2006. Introduction. In Anju Saxena and Lars Borin, editors, *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*. Mouton de Gruyter, Berlin, pages 1–28.

Smeaton, Alan F. 1997. Information retrieval: Still butting heads with natural language processing? In M. T. Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*. Springer, Berlin, pages 115–138.

Steinberger, Ralf and Bruno Pouliquen. 2007. Cross-lingual named entity recognition. *Lingvisticæ Investigationes*, 30(1):135–162.

Strassel, Stephanie, Mike Maxwell, and Christopher Cieri. 2003. Linguistic resource creation for research and technology development: A recent experiment. *ACM Transactions on Asian language processing*, 2(2):101–117.

Streiter, Oliver, Kevin P. Scannell, and Mathias Stuflesser. 2006. Implementing NLP projects for noncentral languages: Instructions for funding bodies, strategies for developers. *Machine Translation*, 20(4):267–289.

Trosterud, Trond. 2004. Porting morphological analysis and disambiguation to new languages. In *SALTMIL Workshop at LREC 2004: First Steps in Language Documentation for Minority Languages*, pages 90–92, Lisbon. ELRA.

Trosterud, Trond. 2006. Grammatically based language technology for minority languages. In Anju Saxena and Lars Borin, editors, *Lesser-known languages of South Asia: Status and policies, case studies and applications of information technology*. Mouton de Gruyter, Berlin, pages 293–315.

Wilks, Yorick. 2008. The semantic web as the apotheosis of annotation, but what are its semantics? *IEEE Intelligent Systems*, 23(3):41–49.

Wilks, Yorick and Christopher Brewster. 2009. Natural language processing as a foundation of the semantic web. *Foundations and Trends in Web Science*, 1(3–4):199–327.

Winter, Werner. 1993. Some conditions for the survival of small languages. In Ernst Håkon Jahr, editor, *Language conflict and language planning*. Mouton de Gruyter, Berlin, pages 299–314.

A LDC language density criteria

For the LDC low-density language survey, languages with at least one million native speakers were chosen, about 300 languages (excluding a handful of *a priori* high-density languages), covering about 80% of the world's population. Further, a set of criteria was defined, consisting of necessary prerequisites for creating language resources, as well as some core language resources, as reproduced below. The resulting survey, reporting a "yes", "no" or "no data" on each criterion for each language, is no longer available on the LDC website, but may still be published at some point (Strassel, Maxwell, and Cieri, 2003).

- Language written
- Words separated in writing
- Simple orthography
- Sentence punctuation
- Dictionary
- Newspaper
- Bible
- Standard digital encoding
- 100 kW news text
- 10 kW translation dictionary
- 100 kW parallel text
- Simple morphology
- Morphological analyzer