ISCA SALTMIL SIG:
Speech and Language Technology for Minority Languages

Climent Nadeu[1], Donncha Ó'Cróinín[2], Bojan Petek[3], Kepa Sarasola[4], and Briony Williams[5]

[1]TALP Research Centre
Univ. Politècnica de Catalunya
climent@talp.upc.es

[2]Linguistics Institute of Ireland
ITE
doc@ite.ie

[3]Interactive Systems Laboratory
University of Ljubljana
Bojan.Petek@Uni-Lj.si

[4]Faculty of Computer Science
University of the Basque Country
jipsagak@si.ehu.es

[5]c/o Centre for Speech Technology Research
University of Edinburgh, UK
briony@cstr.ed.ac.uk

Abstract

This paper presents International Speech Communication Association (ISCA) Special Interest Group (SIG, http://www.isca-speech.org/sig.html) on Speech And Language Technology for MInority Languages (SALTMIL). Overview of the group's mission, including its past and present activities are presented and discussed.

1. Introduction

The start of a new ISCA SIG on Speech And Language Technology for MInority Languages (SALTMIL) was announced in October 1999 [1,2]. Founding committee members were Dr Briony Willliams, formerly with CSTR, Edinburgh University, who became the SALTMIL Chair and Liaison representative, Professor Climent Nadeu, Universitat Politecnica de Catalunya, who took responsibilities of SIG Secretary and Dr Donncha Ó Cróinín, Linguistics Institute of Ireland who acted as the SIG Publicity Officer.

Initial activities of the SALTMIL SIG were aimed to establish an Email discussion list (now at [3]) and the web site (now at [4]).

2. SALTMIL Issues, Aims and Objectives

Local language could be defined as the human language non-visible in information system mediated natural interactivity of the information age. The local languages frequently cope with difficulties in appropriate amount of language resources in standardized electronic form, expertise and/or funding to overcome technological neglect in the future [5].

On the other hand, several recent research programs for global languages focused on the next generation of conversational interfaces. Their fundamental goal is to create speech enabled multimodal systems that scale gracefully across modalities. Such interfaces include speech, graphics, gesture and support complex conversational interaction principles comparable to the human-human interactivity. The systems typically integrate spoken language dialogue systems, including multimodal communication and web-based data handling tools. The long term goal of this research is therefore to transform the present computer systems to become transparent in communication task and to support similar communication patterns as experienced in common interpersonal communication [6, 7].

Given these premises the general aims of the SALTMIL SIG represent promotion of research, development and education in the area of Human Language Technology (HLT) for less prevalent (lesser-used, lesser-studied, minority, regional or local) languages. Main group activities include organization of workshops/conferences with the focus on issues pertaining to the less prevalent languages. Another important action points are dissemination of information between researchers in less prevalent and major language groups as well as provision of guidelines that should initiate and help to promote common projects between all researchers in the HLT area [5].

3. Past activities

Briony Williams proposed two notable workshops that were successfully organized so far. The first one was entitled the Language Resources for European Minority Languages, and was held in May 1998 in conjunction with the First International Language Resources and Evaluation Conference (LREC 1998) in Granada, Spain [8]. The next workshop was organized in May 2000, on the topic Developing Language Resources for Minority Languages: Re-usability and Strategic Priorities, just before the Second International Language Resources and Evaluation Conference (LREC 2000) in Athens, Greece [9].

The minority or lesser used languages of the world (e.g. Basque, Welsh, Breton) are under increasing pressure from the major languages. Some of them (e.g. Gaelic) are endangered, but others (e.g. Catalan) are in a stronger position. However, the situation with regard to language resources is fragmented and disorganized. Some minority languages have been adequately researched linguistically, but most have not, and the vast majority does not yet possess basic speech and language resources (such as text and speech corpora) which are sufficient to permit commercial development of products [10].

If this situation were to continue, the minority languages of the world would fall a long way behind the major languages, as regards the availability of commercial speech and language products. This in turn will accelerate the decline of those languages that are already struggling to survive, as speakers are forced to use the majority language for interaction with these products. To break this vicious circle, it is important to encourage the development of basic language resources [10].

The workshops were a step towards encouraging the development of such resources. The aim was to disseminate information on existing projects and possible future strategies, as well as to form personal contacts and share best practice.

This will make it easier for isolated researchers with little funding and no pre-existing resources to begin developing language resources that are maximally useful [10].

Both workshops succeeded as lively forums for exchange of information among researchers and brought unique opportunities that helped to shape the focus of forthcoming SALTMIL activities.

In November 2000, three members of the SALTMIL committee were invited as speakers to the II International Multimedia and Minority Languages Congress (http://www.gaia.es/multilinguae), which was held in Donostia-San Sebastián.

3.1. The LREC 1998 Workshop

The first workshop that also included the SALTMIL SIG proposal was organized on Wednesday May 27, 1998. About 35 people attended the workshop. Five oral papers and 12 poster papers were presented, all available online at [8].

Further details on the workshop are detailed at [8]. Dr Nicholas Ostler, President of Foundation for Endangered Languages, wrote an interesting online overview paper accessible at [9].

3.2. The LREC 2000 Workshop

One day workshop on the topic Developing language resources for minority languages: reusability and strategic priorities was held on the afternoon of May 30, 2000 in Athens, Greece. The aim of the workshop was to bring together researchers developing language resources for minority languages in order to build contacts and share experience [10].

About 39 people registered for the event. The workshop program included four oral papers and 14 poster presentations, detailed at [10], including the first meeting of the ISCA SALTMIL SIG, described in [11,12].

3.3. The Multilinguae Congress

GAIA, the Telecommunications Cluster Association of the Basque Country (http://www.gaia.es), supported by the European Commission and with collaboration from Eusko Ikaskuntza (http://suse00.su.ehu.es) and the IXA Group (http://ixa.si.ehu.es), organized the II International Multimedia and Minority Languages Congress (http://www.gaia.es/multilinguae), which was held on the 8th and 9th of November in the Miramar Palace in Donostia-San Sebastián.

The minority or lesser used by intuition that for a community such as the Basque Country (with a longstanding history, bilingual and bicultural wealth, with a desire for future projection in the Information Society, etc.), having a focal reference point for these experiences would be a highly important competitive and strategic advantage. Also, the consolidation of a line of studies in this field would be a reference point in Europe and internationally.

GAIA's general aim with regard to this Convention was to promote and boost the capacities of the new technologies, especially the Internet, in the development and preservation of minority languages, and to foment what is known as the Language Industry in the Basque Country.

Over 80 professionals from the world of technology, GAIA associates and collaborators, and specialists in minority language studies in Europe (three committee members of SALTMIL) came together at this congress, where European minority languages played a leading role. There were presentations on the following languages: Basque, Gaelic, Catalan, Mennonite (spoken in Germany, Canada and the United States), and Cimbro (spoken in Italy and considered to be the smallest minority language today in Europe, with less than 500 speakers), amongst others.

Among the large number of speakers who took part in this Congress were the Director-General of the Society for Basque Studies, José María Vélez de Mendizabal; Caroline Philips from Philips Consulting (a company which develops electronic commerce supports in Basque); and Alastair MacPhail (http://www.europa.eu.int/comm/education/callg.html), a European Commission official who provided information on financing for the development of future projects. Also participating in the convention were other specialists in minority language studies at an international level such as, Sara Scardoni (Italy), Peter Wiens (Germany), Jon Patrick (Australia) and three committee members of the SALTMIL: Donncha O'Croinin (Ireland), Bojan Petek (Slovenia) and Kepa Sarasola (Basque Country).

We therefore hope that the development of this kind of activities will encourage the advancement and use of the minority languages with the aid of the New Technologies.

4. Present activities

These are activities we are planning for in the middle term:

- 2001 Workshop in Ireland.
- 3rd SALTMIL Workshop simultaneously with the LREC in 2002.
- 2002 Summer School on Speech and Language Processing for Local Languages.
- Creation of a journal with two main features: merging speech and NLP areas, and with special emphasis on the less prevalent languages issues.

5. SALTMIL SIG Vision

By definition, a less prevalent language has a smaller resource base than the major languages. For some minority languages (those that are fighting for survival), there are not sufficient resources to support research and development in speech and language technology. This means that, in time, the language will fall even further behind the major languages and will be viewed as second-class and pre-technological, since all interaction with computers will need to take place in a major language. This is a critical factor in the survival of some languages. If minority languages die, then a valuable part of the world's cultural diversity will have been lost [13].

Therefore, the vision of the SALTMIL SIG is to provide support for research in speech and language technology for minority languages. It is probable that advances can be made by acting together that could not be made by any one language acting alone. In the case of certain minority languages, the state Government is not particularly sympathetic to their existence, and support must be found from outside or private agencies [13].

Therefore, the vision of the SALTMIL SIG is that sharing of information and the forming of a network of researchers is important to begin with. It is hoped that this networking will form the seed-bed out of which more substantial projects will grow. The adoption of common standards and procedures will

help to minimize costs and workload in research, and this can be promoted through the SALTMIL SIG network.

5.1. Portability Issues in Human Language Technology

Vision of the SALTMIL SIG is also to provide an expert forum that advises on emerging research issues in the HLT which could be of particular importance to the less prevalent languages.

One of the interesting recent research avenues the SALTMIL focuses on is research related to the portability issues in HLT. At the time of writing Lamel, Lefevre, Gauvain and Adda have published an insightful HLT'2001 conference paper on the portability issues for speech recognition technologies [14]. Since automatic speech recognition plays an important role in every system that enables natural interactivity, mastering the issues of portability appears to be of relevance to any language.

Aim is to develop generic core speech recognition technology that reduces otherwise huge manual efforts when porting the system to a new task or a new language. The authors [14] accessed the genericity of wide domain models by evaluating recognition performance under several tasks, investigated novel techniques for lightly supervised acoustic model training, and explored methods of generic model adaptation to a specific task that provide a higher degree of genericity [14].

Among many novel results they showed that unsupervised adaptation can reduce the performance difference between the task independent and task dependent acoustic models. They also showed that supervised adaptation of generic acoustic models can yield a better performance than that achieved with the task-specific models [14].

Another important result is that the recognition results using acoustic models trained on large amounts of automatically annotated data are comparable to results obtained with acoustic models trained on large quantities of manually labeled data [14].

In [15], Padrell reports another example of portability in the sense of partially avoiding the lack of a sufficiently large database by using similar resources available in another language. In that work, a subset of the acoustic models of a speech recognition system for the Catalan language are trained using both the Catalan and the Spanish SpeechDat databases. An improvement of the recognition performance was observed by properly choosing the subset of phonetic units that are trained using the databases from both languages.These results in the context of less prevalent languages bring new perspectives and possible bright future when addressing the very serious problem of the lack of resources and portability of technology towards these languages.

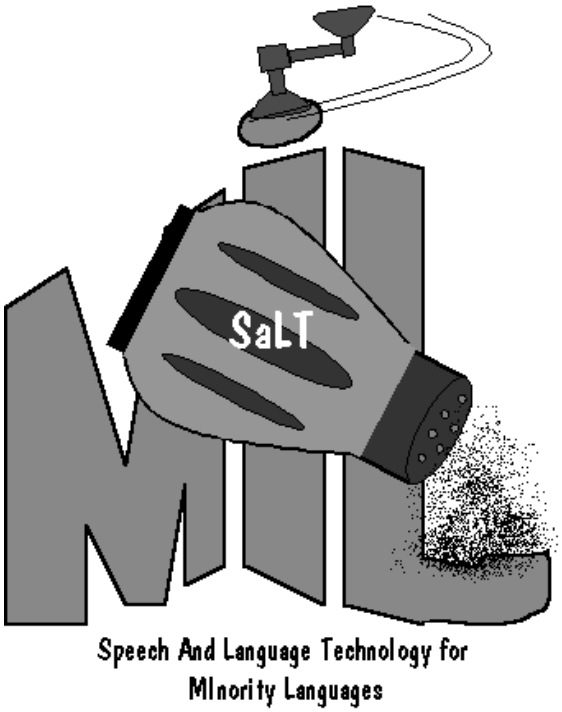
6. Discussion

In September 1999 the European Speech Communication Association internationalized to become truly global in the field of speech science and technology, and changed its name to ISCA. Therefore, widening the scope of the SALTMIL SIG by promoting links to the global efforts appears to be one of the natural steps forward. Specifically, links to other professional organizations worldwide in the field could increase the expertise and provide the positive impact on the technological maturity and status for less prevalent languages.

7. References

URL list was accessed in March 2001:

[1] http://www.isca-speech.org/escapads/iscapad18
[2] http://www.emich.edu/~linguist/issues/10/10-1542.html
[3] http://groups.yahoo.com/group/saltmil/
[4] http://isl.ntftex.uni-lj.si/SALTMIL/
[5] http://isl.ntftex.uni-lj.si/SALTMIL/aims.html
[6] http://www.hltcentral.org/projects/
[7] http://nespole.itc.it/
[8] http://isl.ntftex.uni-lj.si/SALTMIL/lrec98.html
[9] http://isl.ntftex.uni-lj.si/SALTMIL/review.html
[10] http://isl.ntftex.uni-lj.si/SALTMIL/lrec00.html
[11] http://isl.ntftex.uni-lj.si/SALTMIL/athmins00r.html
[12] Petek, B., "Developing Language Resources for Minority Languages: Reusability and Strategic Priorities", *ELRA Newsletter*, 20-21, July-September 2000. (ISSN 1026-8200)
[13] http://isl.ntftex.uni-lj.si/SALTMIL/sigprop.html
[14] Lamel, L., Lefevre, F., Gauvain, JL., Adda, G. "Portability Issues for Speech Recognition Technologies". Proc. HLT 2001, March 2001.
[15] Padrell, J., Robust speech recognition for a dialog system, PhD Dissertation, UPC, 2001.



SALTMIL ISCA SIG Speech And Language Technology for Minority Languages

Aims

The ISCA (International Speech Communication Association) Special Interest Group on Speech and Language Technology for Minority Languages has the overall aim of promoting research and development in the field of speech and language technology for lesser-used languages.

Activities

- The SALTMIL web site (http://isl.ntftex.uni-lj.si/SALTMIL). The site contains the following: aims, activities, history, member information (literature references) and links to other similar resources.

- An associated email list at http://www.egroups.com/group/saltmil. This list is currently used for members to give presentations on their work in progress, and so it is a great way to keep up with current work before it appears in published form.

  You can join the SALTMIL email list either over the Web (at the link above) or via email (send a blank email to saltmil-subscribe@egroups.com).

- In May 1998, a workshop was held in Granada, Spain, on the theme of "Language Resources for European Minority Languages". The full range of workshop papers can be downloaded from the SALTMIL web site.

- In May 2000, a second workshop was held in Athens, Greece, on "**Developing language resources for minority languages: re-useability and strategic priorities**" (http://isl.ntftex.uni-lj.si/SALTMIL/lrec00.html). 39 participants from 19 countries were registered. 15 posters and four invited talks were presented. The workshop proceedings are available from ELRA, ITE (Linguistic Institute of Ireland), and soon also from our web site.