elsnews 10.1

The Newsletter of the European Network in Human Language Technologies Spring 2001

Special Issue on Minority Languages

As announced in ELSNews *9.4 (Winter 2000), this issue is dedicated to work associated with minority languages. But what are these? Below, Geoffrey Sampson discusses the problems of trying to give a definition to this elusive and rather ambiguous term.*

What is a Minority Language?

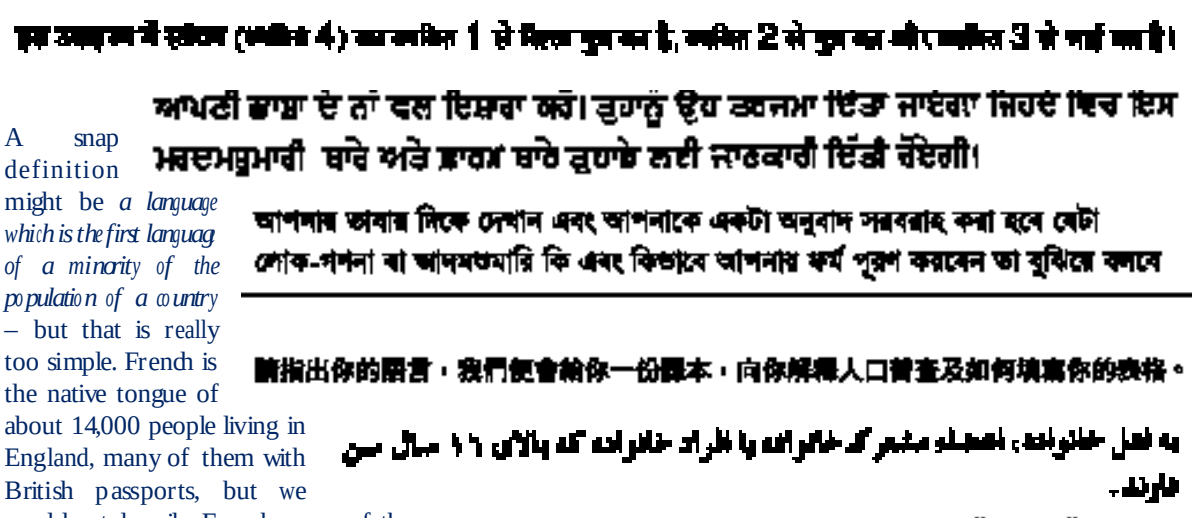
Geoffrey Sampson, *University of Sussex*

Some time after the Editorial Team decided to have an *ELSNews* Special Issue on Minority Languages, the Editor found herself asking me *"What IS a minority language?"* – and I realised that the phrase is less straightforward than it sounds.

A snap definition might be *a language which is the first language of a minority of the population of a country* – but that is really too simple. French is the native tongue of about 14,000 people living in England, many of them with British passports, but we would not describe French as one of the British minority languages, because French-speakers are thoroughly mixed up with the rest of the British population (often they are married to them). *Minority language*, surely, implies at least a language community with a certain cohesiveness.

Beyond that, though, there are quite a lot of different situations to which the term can apply. The first case most of us think of, probably, is languages like Basque in Spain and France, or Welsh in Britain – languages spoken by indigenous groups of the population of a country, but which are not the most widely-spoken language of that country. This definition is relative to particular countries, of course: Hungarian is a minority language in Rumania, but not in Hungary.

A definition that is almost but not quite equivalent would contrast minority languages not with the *most widely-spoken* but with the *official* language of a country. This would be significant particularly for Irish, because for historical reasons this is the official language of the Irish Republic, though the number of people who actually use it in their everyday lives is very small – far fewer than in the case of



Some of the officially recognised Minority Languages in Britain today, both non-indiginous ...

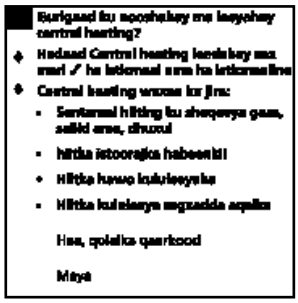
... and indiginous

Spring 2001

elsnet

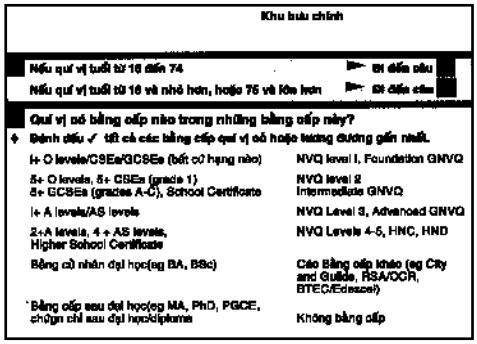
ISSN 1350-990X

Welsh on the other side of the Irish Sea, for instance, although until very recently Welsh had no *official* status at all in Britain.



It's difficult to imagine the concept 'central heating' *being part of life in Somalia ...*

But then, if official status were a central criterion, we might be forced to describe English too as a minority language, because in Britain, unlike some other European countries, linguistic matters have traditionally not been seen as a proper field for State action. (The UK has never had an Academy charged with overseeing language standards, and English dictionaries have been purely private-enterprise affairs.) Surely official status is a side-issue; for all purposes that matter, Irish is a minority language in Ireland, and English is certainly not a minority language in Britain.



... although one might have expected a term for 'school certificate' *in Vietnamese*

A different kind of case is languages spoken by non-indigenous groups who arrived in the relevant region – Europe, for *ELSNews* – within recent generations: for instance, Arabic in France, South Asian languages such as Gujarati or Bengali in Britain. These differ from the first group of minority languages not only in the sense that their historical roots in the region are shallower, but because on a world scale they may be majority languages: there are a lot more Bengali speakers in the world than there are Italian speakers, though in a European context Bengali is a minority language and Italian is not. On the other hand, even within Europe these languages are often *larger* than the indigenous minority languages. There are more Gujarati speakers in Britain than speakers of Gaelic, though Gaelic is spoken nowhere but in Scotland and Ireland.

Even the principal language of a State is sometimes called a *minority language*, if the State is one with a small population. In a European context one sometimes hears languages such as Lithuanian or Slovak referred to as minority languages, though they are the standard languages of Lithuania and Slovakia. But it seems to me that this usage tends in practice to be coloured also by considerations of level of economic development. I have never heard Danish called a minority language, though there are actually slightly fewer speakers of Danish than of Slovak.

All in all, the concept is certainly more tangled than it seemed at first blush. It is not one on which *ELSNews* wishes to impose a specific definition. This Special Issue includes material on languages which are *minority languages* in different senses. Some minority language issues relate only to one class of languages, others are common to all of them; we hope that all are interesting.

ELSNews is published at the School of Cognitive and Computing Sciences, University of Sussex. It is printed by the University of Sussex Print Unit.

Editor: Jenny Norris.

Editorial Team: Geoffrey Sampson, Steven Krauwer and Brigitte Burger.

ISSN 1350-990X

FOR INFORMATION Contributions to *ELSNews*, and corrections, should be sent to: jennyn@cogs.susx.ac.uk

Tel: +44 1273 678659
Fax: +44 1273 671320

Material for the next issue is due: 15 July 2001

FOR INFORMATION

Geoffrey Sampson is Professor of Natural Language Computing at the University of Sussex. He is a member of the executive board of ELSNET and is on the editorial team of *ELSNews*.

Email: geoffs@cogssusx.ac.uk

URL: http://www.cogs.susx.ac.uk/users/geoffs/

Illustrations

We are grateful to the Office for National Statistics in the UK for multilingual copies of their Census forms, from which the extracts in this article appear.

(continued from page 3)

This 'Concede' version of the resources has recently been released. Version 2 of MULTEXT-East resources contains: the revised and expanded MULTEXT- and EAGLES-based morphosyntactic specifications, both in print form and as (over 5000) TEI feature structures; the morphosyntactic lexica, totalling at least 15,000 lemmas per language; and the corrected and TEI encoded *1984* annotated corpus, with about 100,000 words per language. The corpus includes 2-way and 7-way sentence alignments in CES (Corpus Encoding Standard) format.

In the same spirit as version 1, the second release is also being made available to the research community free of charge. The resources will be incorporated in the TRACTOR archive and also mounted on the MULTEXT-East web site, from where interested parties will be able to download them after completing a web-based licensing agreement for non-commercial use. Commercial exploitation is more complex, not least because the resource owners span seven countries. However, we hope to reach an agreement with ELRA, which was set up especially to make such dissemination possible

MULTEXT-East Resources Revisited

Tomaž Erjavec, *Institute Jožef Stefan, Ljubljana, Slovenia*

The MULTEXT-East project (Multilingual Text Tools and Corpora for Eastern and Central European Languages) developed from the EU MULTEXT project and was financed under the INCO-Copernicus programme. The project ran from 1995 to 1997 and developed language resources for six languages: Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene, as well as for English, as the 'hub' language of the project. The main results of the project were morpholexical resources and an annotated multilingual corpus for the seven languages. The centrepiece of the corpus is the Orwell novel *1984* in the English original and translations; the novel is sentence-aligned and its words annotated for context-disambiguated lemmas and morphosyntactic descriptions.

This makes the corpus a unique dataset for studying word-class syntactic tagging, bi-lingual lexicon extraction, and other issues relevant to language engineering applications for a number of Eastern and Central European Languages. With free word-order and rich inflection or agglutination, these languages present significantly different linguistic problems than do those of Western Europe.

One of the objectives of MULTEXT-East was to make its resources freely available for research purposes. In the scope of the TELRI concerted action (Trans European Language Resources Infrastructure), the results of MULTEXT-East were released in 1998 on CD-ROM, and have recently been made available via the TRACTOR (TELRI Research Archive of Computational Tools and Resources) web site. In the years since the CD-ROM was released, the MULTEXT-East resources have served as models for reference corpora, and have been applied to new languages. They have been used in a number of experiments, e.g., in evaluating part-of-speech tagger performance, developing new taggers and lemmatisers, automatic extraction of bi- and multi-lingual lexicons, and studies on multilingual sense disambiguation.

For most of the languages in question, the original MULTEXT-East annotation work was a pioneering effort, so it was hardly surprising that during use a number of errors and inconsistencies were discovered in the data and specifications. These errors were subsequently corrected, but because the work was done at different sites and in different ways, the corpus encodings had begun to lose consistency.

The EU project Concede (Consortium for Central European Dictionary Encoding), which ran from 1998 to 2000 and comprised most of the MULTEXT-East partners, offered the opportunity to return the versions to a common footing. Although Concede was primarily devoted to machine-readable dictionaries and lexical databases, one of its work packages did consider the integration of the dictionary data with the MULTEXT-East corpus. In the scope of this work package, the corrected *1984* corpus was normalised and the primary data re-encoded according to the TEI (Text Encoding Initiative) guidelines and, largely, XML recommendations.

(continued on page 2)

The Concede team posing above the Danube. Photo taken during a project meeting held in April 1999.

FOR INFORMATION

Tomaž Erjavec is Senior Research Fellow in the Dept. of Intelligent Systems at the Institute Jožef Stefan in Ljubljana, Slovenia. He is also currently president of SDJT, the Slovenian Language Technologies Society

Email: tomaz.erjavec@ijs.si
Web: http://nl.ijs.si/et/

MULTEXT-East: http://nl.ijs.si/ME/
Concede: http://www.itri.brighton.ac.uk/projects/concede/
TELRI and **TRACTOR**: http://www.telri.de/
TEI: http://www.hcu.ox.ac.uk/TEI/
CES: http://www.cs.vassar.edu/CES/

Feature

Developing Language Technology for a Minority Language: Progress and Strategy

E. Agirre, I. Aldezabal, I. Alegria, X. Arregi, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, K. Sarasola, A. Soroa, *University of the Basque Country*

The development of language technology for minority languages differs in several aspects from its development for widely used languages. The high capacity and computational power of present computers, combined with the scarcity of human and linguistic resources, motivates the design of new and different strategies. This proposal presents the conclusions resulting from twelve years of experience with the automatic processing of Basque.

Human Language Technologies will make an indispensable contribution to the success of the information society, but most of the working applications are available only in English. For those working with minority languages, a great effort is needed to face this challenge.

The IXA Group was created in 1988 with the aim of promoting the modernisation of the Basque language by means of developing basic language resources for it. Today, the IXA group is composed of seventeen computer scientists and ten linguists from the University of the Basque Country, and as a result of their work four applications are available for common use: a spell-checker; a lemmatisation-based web-crawler; a lemmatisation-based on-line bilingual dictionary; and a generator of weather reports.

Some features of Basque have to be appreciated in order to evaluate the utility of our strategy for other minority languages. There are 700,000 Basque speakers, and these comprise about 25% of the total population of the Basque Country – but they are not evenly distributed. There are six dialects, but since 1968 the Academy of Language has been involved in a standardisation process. At present, morphology, which is very rich, is completely standardised, but the lexical standardisation process is still in progress.

From our twelve years' experience we present here an open proposal for making progress in Human Language Technology. The steps here proposed do not correspond exactly with those observed in the history of the processing of English, because the high capacity and computational power of present computers facilitates different approaches to the problems.

Language foundations and research are essential to the creation of any tool or application; but in the same way, tools and applications will be very helpful in the research and improvement of language foundations. Therefore, these three levels (language foundations, tools, and applications) need to be developed incrementally, in a parallel and coordinated way, in order to get the best benefit from them. We propose five phases as a general strategy to follow in the processing of a language.

First Phase: Laying Foundations

- corpus I – a collection of raw text with no tagging marks
- lexical database I – this could be simply a list of lemmas and affixes
- machine-readable dictionaries
- morphological description
- speech corpus I
- description of phonemes

Second Phase: Basic Tools

- statistical tools for treatment of the corpus
- morphological analyser/generator
- lemmatiser/tagger
- speech processing at word level
- corpus II – word forms are tagged with their part of speech and lemma
- lexical database II – lexical support for constructing general applications, including part-of-speech, and morphological information

Third Phase: Tools of Medium Complexity

- environment for tool integration: for example, following the guidelines defined by TEI using XML
- spell-checker and -corrector (in morphologically simple languages a word list may be sufficient)
- web-crawler – traditional search engine that integrates lemmatisation and language identification
- surface syntax
- structured versions of dictionaries
- bilingual dictionary integrated with a common text-processor to be consulted on-line. When a user selects a word form in the text, its equivalents in the other language are shown (considering all the possible combinations of lemma and part-of-speech for that word form)
- lexical database III – version II is enriched with multiword lexical units

Fourth Phase: Advanced tools

- corpus III – syntactically tagged text
- grammar- and style-checkers
- integration of dictionaries in text editors
- lexical-semantic knowledge base – creation of a taxonomy of concepts (such as WordNet)
- word-sense disambiguation
- speech processing at sentence level
- language learning systems

Fifth Phase: Multilinguality and General Applications

- corpus IV – semantically tagged text after word senses have been disambiguated
- information retrieval and extraction
- translation aids. Integrated use of multiple on-line dictionaries; translation of noun phrases and simple sentences
- dialogue systems
- knowledge base of multilingual lexico-semantic relations and its applications

At IXA we are now working on the fourth phase described above. The foundations, tools, and applications developed in the previous three phases are all of great importance in facing new problems and applications. The spell-checker and the lemmatiser are particularly active tools in the ongoing standardisation of Basque.

FOR INFORMATION

The authors are all members of the IXA Group for Natural Language Processing in the Dept. of Computer Languages and Systems at the University of the Basque Country

Email: jipsagak@si.ehu.es
Web: http://ixa.si.ehu.es

Announcement

9th ELSNET European Summer School on Language and Speech Communication

This year, the ELSNET European Summer School on Language and Speech Communication will be held from **16-27 July**, and has the topic of **Text and Speech Corpora**. It is organised by the Institute of Formal and Applied Linguistics and the Center for Computational Linguistics at **Charles University, Prague**. The School is aimed at advanced undergraduate students, PhD students, postdocs, and academic and industrial researchers and developers with an interest in the following courses:

- Annotation graphs in theory and practice (Steven Bird, UPENN)
- Text encoding initiative (Lou Burnard, Oxford)
- Validation of speech databases (Henk van den Heuvel & Eric Sanders, Nijmegen)
- Dialogue corpora (MATE) (Amy Isard, Edinburgh & Ole Bernsen, Odense)
- Speech resources & industrial applications (Jan Odijk, Lemout & Hauspie)
- Speech tools for database processing (Uli Türk, Munich)
- Multimodal resources, including speech, etc. (Chalapathy Neti, IBM)
- Annotation at the grammatical level (Geoffrey Sampson, Sussex)
- Prosodic annotation; IVIE extensions to ToBI (Esther Grabe, Oxford)
- Linguistic annotation of a large corpus: from morphology to syntax (Jan Hajic, Prague)

For details:
Web http://ufal.ms.mff.cuni.cz/~ess2001/
Tel. +420 - 2 - 2191 4278 **Fax** +420 - 2 - 2191 4309
Email ess2001@ufal.ms.mff.cuni.cz

Feature

What is this Technology ever Going to Do for Minority Languages?

Nicholas Ostler, *Linguacubun Ltd, Bath, UK*

Nicholas Ostler

The term *minority language* is part of sociolinguistics, rather than language typology or computing science. There is no formal property which defines a minority language: rather, it is any language that happens to be spoken by a smaller group, in the context of a larger group such as a nation. (Indeed if any languages are to be considered the odd ones out here, it is the majority languages. Since the top tenth of a percent of all the languages there are in the world (say five – Chinese, English, Spanish, Bangla and Hindi), accounts for one third of the world's 5.6 billion people, and the top one percent of the world's languages (say 65) for three quarters (figures from SIL's *Ethnologue*, 1999), it is statistically and scientifically reasonable to equate the set of human languages simply with the set of minority languages.)

But very few of even the highly populous languages have been the subject of language technology. The vast majority of languages in which serious work has been done have been national, or at least official, languages of major European or East Asian powers, probably no more than a dozen or two.

There are so many reasons why languages have failed to figure in this charmed circle. They could have been official languages that do not happen to be national languages. (Two such major languages in China, Shanghainese and Cantonese, and three in India, Bengali, Telugu, and Tamil, in fact figure in the world's Top 20 by population.) They could have been dialects of major languages, more or less different linguistically, but somehow identified as part of the same community. They could have been unofficial vernaculars, which are often comparable in size and arguably importance with official regional languages. (Javanese is the largest language in Indonesia, and other examples are Uighur in China, Occitan in France, and even Belorussian in Belarus.) All these language statuses (national, dialectal, regional, or mere vernacular) vary in size from hundreds of millions to a few thousand. There is even less parity, or a natural ranking in terms of size, among languages than there is among nations. Endangered languages, likewise, can figure in any of these statuses, differing only in having poor prospects of survival.

In political reality, what these *minority* languages do have in common is an absence of large-scale government support; this disadvantage is compounded in many, perhaps most, cases, by a declining speaker population. And in economic reality, their populations are all too small, or too poor, to be interesting markets for outside investors at the present time.

It is interesting, though difficult, to speculate on what the potential effects of the spread of language technology will be on these vast numbers of languages. This Western cultural artefact, a product of linguistic analysis and computer technology, is different from previous cultural imports with linguistic implications (such as the global spice trade, missionary religions, or colonial impositions) in that it concerns a set of new methods and modalities in which language can be used, not a new set of values to talk about. In Marshall McLuhan's terms, it is a new medium, not a new message. Everything will depend on how widely compatible this new digital medium turns out to be.

What the technology does, or is capable of doing, is very various. But in essence it offers readier penetration to the content of what is said or written, by methods that may not require the users themselves to understand, or even be aware of, the words and language at source. This is the language processing side. It also implies and requires, and hence tends, over time, to create, a much greater compatibility among the digital representations used for different languages. (This is not something that comes about immediately, as the current undignified and disorderly scramble for non-ASCII URLs in Asian languages is showing.[1])

Different languages will take their place as simple media of access to the Internet. This enables "*nation to speak unto nation*", much as radio and TV have been doing over the past 50 years and more. But it also enables far more two-way communication than the mass media ever did: village will be able to speak unto village, and person to person, not only all across the world but all across a single region, and without the mediation, helpful or intrusive, of large-scale institutions like governments or multinational companies.

This combination of greater penetration and wider range will mean that, even as outsiders find it easier to penetrate the social and market barriers which have kept foreigners out (these often being the same things which, in the past, have held small language communities together), the insiders who speak the minority languages should increasingly find that the world is their oyster, and available to them on something like their own terms.

If they use this new freedom mostly as a substitute for contact with those nearby, the result may be to weaken intercourse in their own languages. The dynamics here would be much the same as the forces that weaken neighbourhood shops and markets when consumers get

access to larger supermarkets and department stores. But if they maintain their local links, and go on to use their freedom to keep in touch with others far off who used to be part of the local unit, or who are close by but out of ready contact, the result may be to provide new channels for use of their language, and so strengthen it. The Welsh and the Maoris are not only keeping in touch with fellow-speakers in diaspora across the world; they are even providing each other with ideas and inspiration about how to use this new freedom.

But the initial mismatch between language technology and the internet on the one hand, and smaller groups (from nation-states to villages) with their languages and traditions on the other, can be significant. Quasi-universal contact of individual with individual all across the world is not an easy development for communities which have been used to keeping a low profile, and running their own shows without much outside interest or interference. There is a great scope for misunderstanding, and a great need for caution as the new links are established.

There are plenty of examples of these early difficulties in the language world.

Transparent Language, a US software company with some concern for smaller languages and a considerable body of expertise and technology in computer-aided language learning, offered to develop a language tuition CD for any language community that could provide the services of a language expert and $100,000.[2] Although the intent behind this was benign, the approach caused too stark a confrontation between the two worlds. It raised questions which were painful because their answers were as yet unknown, or perhaps indecent. What use could computer-aided language learning have in supporting transmission of a language when children were not picking it up naturally? Was it ethical that a language's chance of survival be somehow weighed in the balance against a company's judgement of a fair return on its investment?

There is a general temptation for those with technical expertise (and hence nowadays a programmed solution) to offer it to solve what they see as an immediate linguistic problem. Hence in Mexico, since 1989, CELIAC in Oaxaca[3] has offered a facility for individuals to come for training in IT and the basic principles of writing systems, and start to create written materials in their languages without reference to any pre-existing standards. Whether the resulting plethora of spelling systems will converge to a standard, and whether this matters anyway, remains a political problem, i.e, one that can only be solved, at some level, in the community. The more recent proposal of a common alphabet for the Mayan languages from the *Academia de las Lenguas Mayas de Guatemala* may be seen as a properly-elaborated (hence slow and painful) community response to a similar problem.[4]

These are cases where the purveyors of language technology appear to be disinterested, almost naïve in meddling in problems with a community dimension. But in other cases, there are grounds for seeing the introduction of language technology as rather more self-interested. A recent example of this is the NICE project, funded by the US DARPA in collaboration with the Organisation of American States, to provide rapid development of machine translation systems for such minority languages as Mapudungun (with 400,000 speakers) in Chile, Inupiaq (with 3,500 speakers) in Alaska, Siona (with 300 speakers) in Colombia.[5] Although the case is made that this will provide a new resource for indigenous communities, and so give access to a wide range of materials through the indigenous languages, it is surprising that one of the very few languages chosen is that of the tiny community that happens to live in the Putumayo area of Colombia, where the USA is assisting military fumigation of illegal coca crops. Probably, this sort of development needs to be seen as part of the current US imperative (served also by the Expedition project[6]) to provide language decoding (i.e., machine translation) at short notice for any area where they may identify a threat or security problem.[7]

The fact that others are looking to their own interests, or may misunderstand the complexity of a small community, does not deny potential for benefit to the minority languages. But a bridge needs to be built quite consciously, between the view of a language as dissected digitally, and the view of it taken by its speaker community, a community which, when the language is small, is likely to play a much more important role. This can be thought of as an aspect of literacy as it presents itself in the modern world – and it should be remembered that two thirds of the world's languages are still without any written literature.

References

[1] See *Far Eastern Economic Review*, 22 February '01.
[2] http://carmen.murdoch.edu.au/lists/endangered-languages-l/ell.arcs/ell-arcs-1999/endangered-languages-l.9904
[3] http://zapotec.agron.iastate.edu/celiac.html
[4] http://iisd1.iisd.ca/50comm/commdb/desc/d37.htm
[5] www.cs.cmu.edu/~sfarce/NICE/NICE_index.html
[6] crl.NMSU.Edu/expedition/
[7] As discussed, e.g., in the *New York Times* of 16 April '01, www.nytimes.com/2001/04/16/world/16LANG.html?ex=988411880&ei=1&en=8724869165f28729

FOR INFORMATION

Nicholas Ostler is Director of Linguacubun Ltd, a language technology consultancy, and also President of the Foundation for Endangered Languages (FEL), a registered charity. FEL is holding a conference in Agadir, Morocco, on 21-24 September 2001 on Endangered Languages and the Media.

Email: nostler@chibcha.demon.co.uk
Linguacubun **URL:** http://www.chibcha.demon.co.uk
FEL **URL:** http:// www.ogmios.org

SIGdial Page

Language Technology as aid to preserving linguistic diversity

Jens Allwood, *Göteborg University*

Jens Allwood

There are between 4000 and 8000 languages in the world. The reason a more exact figure cannot be given is that linguistic factors are not sufficient to define a language. Rather, political factors are intrinsic in the concept of language. Today a dialect (language), tomorrow a language (dialect). As an example, consider the situation in Yugoslavia. Twenty years ago there was an attempt to make Serbo-Croatian the national language. Today linguists are helping to make Serbian, Croatian and Bosnjak into distinct *languages*.

Leaving difficulties of definition aside, it might not be unreasonable to claim that there are, say, 6000 languages in the world. Many of these are disappearing or are threatened by extinction, often because they are only spoken by a small, ageing population. Some languages today are used by only one or two speakers who are all over 60 years of age.

If we want to preserve the linguistic diversity of our planet, the key factor is the usability of the languages. There have to be opportunities and desires as well as needs to use the languages. The desires and needs are created by, for example, loyalty, tradition, group membership, and by the possibility of sharing thoughts with others. However, a language will only really be usable if it satisfies practical needs (such as work, food provision, career progression, etc.).

The opportunities for using a given language are created by, among other things, the technological support for communication in that language. This is where language technology and dialogue systems come in.

Language technology can play an important role both in preserving, and in supporting active use of, the world's languages. Even if we cannot retain active use of all languages, we should perhaps at least try to preserve a record of them for future generations. Today, we definitely have the technological resources to do this.

The basic need here would be to create corpora of written (even today there are languages that don't have writing systems) and spoken language. These corpora should preferably be in multimedial form, so that bodily communication and typical situational contexts would also be preserved. Since so much of language technology depends on written language, multimedial corpora should also, where possible, be accompanied by transcribed versions. Once fairly large-scale corpora have been created, they can then be subjected to many different kinds of linguistic analysis.

With a more ambitious goal than preservation, there are many more things that can be done using language technology. However, this does require the basic resources of large corpora of spoken and written language, out of which other tools can be developed.

Let me end by listing some examples of the types of language support that can be given:

- linguistic interfaces for operating systems and communication programs
- word processing systems (with dictionaries, support for spelling, hyphenation, and grammar)
- speech analysis and speech synthesis tools
- dialogue systems
- information retrieval via translation into larger languages
- multimodal communication systems

The list can be made much longer, but points to real and important tasks for all of us involved in language technology and dialogue research. Let us hope that more linguists interested in dialogue systems and language technology become aware of the field of language preservation as an interesting area of application.

FOR INFORMATION

Jens Allwood is Professor in the Department of Linguistics at Göteborg University, Head of SSKKII (the Center for Cognitive Science) at Göteborg, and President of the Immigrant Institute. He is also a member of SIGDial, the Special Interest Group on Discourse and Dialogue of the ACL.

Email: jens@ling.gu.se
Web: http://www.ling.gu.se/~jens/
SIGdial Website: http://www.sigdial.org/

Europe's Ignored Languages

Tony McEnery, *Lancaster University*

Feature

Corpus building in Europe has traditionally focussed on languages that are indigenous to European countries: English, French, Spanish, German, Italian, etc. It follows that most of the corpus-based human language technology research undertaken in Europe has also focussed on those languages. Consequently, speakers of such languages benefit from an extensive range of computational resources such as fonts, word-processors, spell-checkers, online dictionaries, thesauri, automatic translation utilities, and a host of other language processing products. However, in the UK and other European countries there are sizeable communities of speakers of non-indigenous minority languages (NIMLs). For example, in the UK, Bengali, Cantonese, Gujarati, Panjabi, and Urdu are spoken by a sizeable proportion of the population. The existence of these NIML speech communities means that the domestic translation market in the UK is currently focussed around NIMLs, with South Asian languages predominating.

Tony McEnery

There is no reason to believe that this state of affairs will change in the near future, especially as the continued migration of speakers from South Asia to the UK means that the demand for the translation of these languages will be sustained across time. In other European countries a similar state of affairs exists, though differing patterns of immigration mean that, to some degree, the NIMLs of importance vary from country to country. Arabic is a much more important NIML in France than the UK, for example. However, across Europe one factor remains consistent. The focus of human language technology research is almost exclusively on indigenous European languages. This emphasises a 'digital divide' which exists between Europe's indigenous and non-indigenous languages. Computational resources are scant for NIMLs (as shown by Somers, 1997, for example).

This situation is exacerbated by a number of factors. First, corpus resources to enable further research into the machine processing of NIMLs are not readily available. With regard to South Asian languages, for example, there is no substantial spoken corpus of any South Asian language yet available, though such data is being constructed on the EMILLE project. (EMILLE – Enabling Minority Language Engineering – is a three-year EPSRC project at Lancaster University and Sheffield University, designed to build a 63 million word electronic corpus of South Asian languages, especially those spoken in the UK.) Written corpus resources for South Asian languages, though now slowly becoming available, bear no comparison to what is available for Basque, let alone English.

Secondly, and as a consequence of the first point, research into the machine processing of such languages has to date been fitful at best. While some of the languages, notably Arabic and Chinese, fare better than others, some languages are only just beginning to be researched using corpus-based approaches.

Finally, the lack of a European focus on the study of NIMLs is a major problem, at least from the point of view of translators. Imagine that parallel corpora were developed in India covering a range of Indian languages (you will have to imagine that because such corpora have not yet been created). In the UK context, the development of terminology databases created using such parallel corpora will be of little use – much of the domestic translation of these languages focusses around concepts not necessarily shared between the UK and South Asia. Consequently parallel corpora need to be developed in the UK to meet the needs of the UK. Good examples of this can be found in social security leaflets, all of which are translated into nine different UK NIMLs from their English originals. Many of the terms in the leaflets, such as *winter fuel heating benefit* or *supplementary benefit* would not be found in a parallel corpus gathered in South Asia, as these terms are specific to the UK social security context.

While the list given here of factors bedevilling the development of NIML language processing research is of necessity brief and incomplete, it does at least show that NIML language processing faces problems which most European languages have long since solved.

It seems strange that, given a need for such research, and a relative lack of relevant research in many countries where these NIMLs are indigenous or majority languages, there is so little being done across Europe to support research into NIMLs. This becomes even stranger when one considers, for example, that common UK NIMLs are some of the world's largest languages: Bengali has 189 million speakers; Gujarati, 44 million; Hindi, 182 million; Panjabi, 56 million; and Urdu, 58

million (figures from *Ethnologue*). Major world languages in need of serious language processing research are spoken across Europe by Europeans. Action is needed, urgently in my view, to widen the scope of those languages that European language processing researchers see fit to study in order to meet the needs of the communities in which they are conducting their research. Currently, significant sectors of these communities are being sidelined and disadvantaged by the prevailing focus on indigenous languages.

There are moves afoot to end this uneven approach to research into languages spoken in Europe. In the UK, the EPSRC has adopted multilinguality as a priority in its Information Technology and Computer Science research programme. The EPSRC's focus on multilinguality explicitly covers both indigenous and non-indigenous UK languages. Such a move is to be warmly welcomed, and one would hope that other European research agencies would be far sighted enough to follow such a lead. However, before that happens it is important that those in the language processing communities remember Europe's forgotten languages and begin to tell funders that these are languages with which they want to work. Only then will the digital divide between NIMLs and indigenous languages in Europe begin to narrow.

References

Somers, H. 1997. Machine Translation and Minority Languages. Papers from the ASLIB Conference, 13-14 November 1997. In *Translating and the Computer*, vol. 19.

FOR INFORMATION

Tony McEnery is Head of Department and Reader in Multilingual Corpus Linguistics in the Department of Linguistics and Modern English Language, at Lancaster University.

Email: A.McEnery@lancs.ac.uk
Web: http://www.ling.lancs.ac.uk/staff/tony/tony.htm

For more information about the **EMILLE** project, visit http://www.emille.lancs.ac.uk

Letter

Letter to the Editor

To the Editor,

I don't want to get involved in the Wilks versus ACL debate [*see* ELSNews *issues 9.3 (Opinion) and 9.4 (Letters), available online via www.elsnet.org – Ed.*], but I would like to comment on the issue of anonymised reviewing, and ask readers if they share my experience.

I have been reviewing for ACL and indeed other conferences and journals in the field for many years, both before and after so-called anonymous reviewing was introduced. I can honestly say that of all the papers I have reviewed, the only ones where I have been unable to guess the authorship have been from newcomers to the field, in which case the newness (as evidenced by lack of references to standard works) has been just as effective a bias as the unknown author's name might have been.

Despite the request to anonymise their work, authors give the game away by any of the following ploys, deliberate or otherwise:

(a) Referring to their previous publications on the same on-going project
(b) Referring to the well-known name of their project
(c) Referring to unpublished works by themselves or their colleagues and students (papers awaiting publication, internal reports, PhD theses)
(d) Taking a well-known stance on some controversial issue
(e) Writing once again about work that they have reported elsewhere.

Notice that all these are perfectly legitimate things to do in a scholarly article.

I am not saying that efforts to review anonymously should be abandoned altogether. But the reality, in my experience, is that it simply does not work. As Editor of one of the journals in this field, I have taken the view that the author's identity is one of the factors, rightly or wrongly, which contributes to the acceptability of a paper.

For example, if someone well known says something outrageous, that might be more suitable for publication than if an unknown newcomer made the same statements. Which brings us back full circle to Yorick Wilks' article.

Yours sincerely,
Harold Somers

FOR INFORMATION

Harold Somers is Professor of Language Engineering at UMIST, Manchester, UK.

Email address for all correspondence:
Harold.Somers@umist.ac.uk

Opinion Column

Ypres' Keepers

John Nerbonne, *University of Groningen*

Interests can't remain pure: conservationists inspired by nature have to follow developments in synthetic fuel, sports fans into cutting-edge performance find out a lot about drugs, and technology freaks soon find themselves thinking about money. Filthy mammon, preferably in large amounts.

So it's natural, almost necessary, that we extend our interest from language and speech technology to the language and speech industry. We develop the technology to be useful, and in a free market that means someone ought to be willing to pay for it. You just need an office, production facilities, distribution channels, marketing strategies, legal representation, calculations of marginal return, etc. And so you buy a couple of suits, or you go into business with people that already have them. Or maybe they buy you.

For a while it looked as though the speech and language industry would have a flagship, a most prominent, eminently successful undertaking we could all look to for inspiration, i.e., financing. Lernout and Hauspie Speech Products (LHSP) was founded in Ypres in 1987 and grew steadily in its concentration of language and speech expertise. LHSP had product offerings in virtually all the language and speech application areas, especially speech dictation, text-to-speech synthesis, machine translation, and translation assistance. The colleagues at LHSP I've had contact with are technically serious. Things looked good in Ypres.

The rule of the 'new economy' is market share, and therefore growth, and LHSP grew. One lost track of all the companies they acquired. When LHSP took over Dictaphone and Dragon Speech Systems at the end of March 2000, its stock was traded at $65. Counting the two stock splits L&H had, its value had risen by a factor of 25 since 1995 (introduced at $11). It had become the fifth largest company in Belgium, and its two founders were minor folk heroes.

It unravelled fast. In August of last year the *Wall Street Journal* accused LHSP of listing sales in Korea too optimistically. LHSP promised a quick audit that kept everyone waiting – by the time it was done, the Securities and Exchange Commission had begun an investigation into the Korean connection, and also a second investigation into further allegations that LHSP had set up phony 'language development companies' (LDCs) in Singapore, listing them again as customers, but receiving no payment for the base systems the LDCs were to customise. Under pressure, LHSP admitted book-keeping irregularities over the previous 30 months. The long awaited audit by KPMG accused LHSP board members of fraud. The stock value fell regularly during the autumn until it was removed from trading in New York and Brussels, almost simultaneously. It was valued at about 1% of the high ($0.70) in unofficial trading in January. As well as LHSP, Flanders Language Valley – an investment fund that co-operated closely with LHSP – was accused of fraud. When a court appointed trustees for LHSP, it came to light that it was largely owned by L&H holding company, which effectively controlled the LHSP board. This made it resistant to reform, even recently. A picture emerges of murky business relations, deceptive practices, and massive conflicts of interest.

There are big losers in all of this, most conspicuously 1,000 (of 6,500) former LHSP employees who have been laid off; the former owners (stockholders) of Dragon and Dictaphone, who were paid in LHSP stock; and the approximately 15,000 small stockholders who are suing former board members. Who can blame the small shareholders? They were encouraged by the investments made by partners as serious as Dresdner Bank and Deutsche Bank, Artesia and Fortis, all of whom waited until court proceedings started before pulling out.

Language and speech technology? Lots of companies' stock has slid with LHSP, and this is translated rapidly into less money for R&D. My (few) business audiences ask about this, so people certainly associate us with the LHSP scandal. The interest is at the level of juicy gossip, so maybe there are not serious worries about the technology as a whole. But investors and agencies with subsidies are likely to be more cautious.

Jo Lernout has been the most colourful figure throughout. When we film it, we'll want Michael Gambon to play Lernout: ambiguous, uncomprehending, and blustery when the rest of the LHSP board had gone into hiding; out of his league in international finance, but belligerent even there. When the *Wall Street Journal* accused LHSP of inventing customers, Lernout responded with fantasies about a take-over plan among Wall Street's short sellers. And a Belgian judge pointedly told Lernout that he wanted financial disclosure, not speeches, whilst refusing LHSP's first request for protection under bankruptcy. Language technologists can take it on the chin.

FOR INFORMATION

John Nerbonne is Professor in the Department of Humanities & Computing at the University of Groningen in the Netherlands.

Email J.Nerbonne@let.rug.nl
Web: http://www.let.rug.nl/~nerbonne/

Feature

Human Language Technology Issues for Haitian Creole – a Minority Language

Marilyn Mason, *Mason Integrated Technologies, Boston, USA*

Marilyn Mason

This article builds upon a few surveys [1] and many other articles [2-6] that have been written about human language technologies (speech- and text-based systems and corpus collection projects) for a specific group of minority languages in the world today – Creole languages. In contrast to the world's international 'major' languages, we use the term minority language here to refer to vernacular languages; specifically, to low-density/sparse-data/less-prevalent languages which usually lack both electronic corpora and computational systems for automatic language processing.

One of the fundamental stumbling blocks in development efforts for natural language processing (NLP) and human language technologies (HLTs) for minority languages can be the assumption that the written form of such languages is naturally standardised at the same level as major international languages. We provide evidence in this article of the risk of such assumptions.

From our experience in developing different types of HLT systems for minority languages, we have noticed a number of extra-linguistic factors for vernacular languages which must be considered when attempting to provide automated processing techniques for languages in which linguistic variation permeates the entire lexicon, as in Haitian Creole. One of the key points comes from the fields of sociolinguistics and language planning, namely, the distinction made between the *standardisation of an orthography* and the *normalisation* of its use for those who wish to write in a given language. Standardisation (determining what forms should be used) is a decision-making process: normalisation (implementing what has been decided) is putting the decisions into practice. It has been noted that many vernacular languages are currently undergoing stages of standardisation [7, p. 6], but it is important to remember this distinction between the stages of standardisation and normalisation.

Although the thrust of the 'standardisation' of Haitian Creole in Haiti has taken place over several decades, it has, unfortunately, mainly focussed only on orthographic standardisation. In essence, the orthographic issues of standardisation were more or less resolved in the late 1970s and early 1980s, with the creation of the 'official' Institut Pédagogique National (IPN) orthography [8]. Yet the fact is that over many decades "in Haiti, there have often been two or more competing orthographies in the same territory" [9, p. 120]. Others have shown that Haitian Creole has had eleven known proposed spelling systems [10], as well as a dozen known hybrid spelling systems. Of all the orthographies that have been produced, the IPN orthography remains the official one, and is consequently the most widely accepted for Haitian Creole today. However, despite the existence of an official orthography, there is no guarantee that all texts follow it, nor that the Haitian Creole written language will naturally and automatically pass through the stage of wider-use normalisation whereby the lexicon standardises itself in written form. Standardisation of the lexicon, and not simply just of the orthography, is therefore crucial to the use of the written form of the language in all potential industrial user areas (authoring, publishing, translation, web site information, government administrative information, etc.), from which data can be used to develop human language processing tools.

One study [2] has provided detailed frequency counts on variation found for 27 Haitian Creole lexical items within texts collected from 13 independent sources. Listed below are a few examples of the initial study on variation in Haitian Creole spelling.

Frequency	Written form	Basic speech-to-text phonetic interpretation
(1) The word for *enemy*		
457	lènmi	{lEnmi}
2	lènnmi	{lEn:mi}
9	lenmi	{lɛ̃mi}
5	lennmi	{lɛ̃nmi}
9	ènmi	{Enmi}
6	enmi	{ɛ̃mi}
7	ennmi	{ɛ̃nmi}
(2) The word for *week*		
295	semèn	{semEn}
11	semènn	{semEn:}
20	semen	{semɛ̃}
28	semenn	{semɛ̃n}
2	senmenn	{sɛ̃mɛ̃n}
(3) The word for *government*		
10	gouvèman	{guvEmã}
8	gouvèmnan	{guvEmnã}
7	gouvènmam	{guvEnmam}
924	gouvènman	{guvEnmã}
5	gouvènnman	{guvEn:mã}
20	gouvenman	{guvɛ̃mã}

Announcement

Hundreds of additional examples [11] of the high level of variation in the Haitian Creole lexicon have been accounted for. This high amount of variation in spelling for the same lexical items has been shown to be both 'inter-textual' (i.e., between the many different editorial teams writing in Creole) and 'intra-textual' (i.e., within the same texts produced by the same editorial team).

Other researchers have noted similar lexical variation issues for other Creoles. Ken Decker [12, section 3.2] states that in "B[elize] C[reole] texts, I have often found the same word spelled different ways in the same text, or even the same sentence." Pierre-Louis Mangeard (personal e-mail communication, 15 October 1998), speaking of Reunion Creole, indicates that "la variation graphique atteint ici 100 % des unites lexicales" (our translation: every lexical item [of the language] has instances of graphemic variation).

HLT developers must be aware of both the inter-textual and intra-textual variation that can be found in written corpora of minority languages. In other words, the existence of a written corpus does not mean that the lexical forms of the data in it are inherently consistent, or even consistent with other corpora. All HLT development teams working on minority languages should consider such issues.

Lexical standardisation is one of the key issues for all HLT systems. For some of the international languages, such standardisation has been achieved over a period of many centuries. Normalisation is then reinforced with the recent help of integrated spell-checkers in Microsoft Word and other applications. The majority of the world's languages, being minority and vernacular languages, have not been able to benefit from such technology. Through the efforts of Mason Integrated Technologies, it is now possible to focus on lexical standardisation within existing and upcoming corpora for many of the world's Creole languages. By applying its orthography conversion and corpus cleaning technologies to the standardisation of corpora, it will then be possible to build more reliable HLT systems for treating the standardised information.

Techniques must be developed and implemented to provide for something as simple as lexical standardisation. If not, these minority languages will suffer greatly and will be unable to meet their users' information needs (authoring, translation, web site localisation, documentation, etc.). Nor will it be possible to develop the tools (for translation, desktop publishing, OCR, spell-checking, information retrieval, question-answering, speech recognition and synthesis, etc.) upon which the modern world is basing its current and future trends for information processing and communication.

References

[1] Mason, M. & Allen, J. 2000. The State of the Art of French Creole Language Resource Engineering. Minority Languages Workshop, LREC2000, Athens, 31 May-2 June 2000.

[2] Allen, J. and Hogan, C. 1998. Evaluating Haitian Creole orthographies from a non-literacy-based perspective. Society for Pidgin and Creole Linguistics conference, New York City, 9-10 Jan 1998.

[3] Mason, M. 2000. Authoring and Documentation Workflow Tools for Haitian Creole – a Minority Language. In *Technical Communicators' (TC) Forum* Magazine, vol. 1, 2000 (Jan-Mar 2000), pp. 8-9. (Online at http://www.tc-forum.org/topictr/tr17auth.htm)

[4] Mason, M. 1999. Orthographic Conversion and Lexical Standardization for Vernacular Languages. In *ELRA Newsletter*, Vol. 4, (4), Oct-Dec 1999. pp. 5-7.

[5] Allen, J. 1998. Lexical variation in Haitian Creole and orthographic issues for Machine Translation (MT) and Optical Character Recognition (OCR) applications. 1st Workshop on Embedded Machine Translation systems, AMTA98, Philadelphia, 28 Oct 1998.

[6] Mason, M. 2000. Issues from corpus analysis that have influenced the on-going development of various Haitian Creole text- and speech-based NLP systems and applications. Poster & demo at LREC2000, Athens, 31 May-2 June 2000.

[7] Tabouret-Keller, A. et al (eds). 1997. *Vernacular Literacy: A Re-evaluation*. Oxford: Clarendon Press.

[8] Bernard, J. 1980. Ki Jan Nou Ekri Kreyòl Ayisyen [Reprint of communiqué on Haitian Creole official orthography]. *Études Créoles* 3.1: 101-105.

[9] Baker, P. 1997. Developing Ways of Writing Vernaculars: Problems and Solutions in a Historical Perspective. In Tabouret-Keller et al (eds). pp. 93-141.

[10] Schiefflin, B. & Doucet, R.C. 1992. The 'Real' Haitian Creole: Metalinguistics and Orthographic Choice. In *Pragmatics* 2 (3), pp. 427-443.

[11] Eskenazi, M., Hogan, C., Allen, J, & Frederking, R. 1998. Issues in database design: recording and processing speech from new populations (poster session). In *Proceedings of LREC'98*, Granada. Vol. 2, pp. 1289-1293.

[12] Decker, K. 1996. Orthography Development for Belize Creole. In *1994 Mid-America Linguistics Conference Papers*, Vol. II, edited by Frances Ingemann. Lawrence, Kansas: The University of Kansas. pp. 351-362.

FOR INFORMATION

Marilyn Mason is President & Chief Operating Officer of Mason Integrated Technologies

Email marilinc@aol.com
Web: http://www.mit2usa.com

Comment

Fictional Family for Basque

Larry Trask, *University of Sussex*

One of the problems besetting minority languages is that people forsake their normal caution and believe the most sensational claims where they are concerned. **Larry Trask** *finds that recent statements made by leading British and French newspapers about Basque have tried his patience altogether too far.*

Basque, spoken at the western end of the Pyrenees, is a language with no relatives, but that fact doesn't stop fantasists from looking – and finding – what they want to find. Revolutionary new breakthroughs are announced every year.

The Spanish linguist Jorge Alonso has recently published a stream of books, claiming "decipherments" of four extinct ancient languages of the Mediterranean: Iberian, Tartessian, Etruscan, and Minoan. The Etruscan texts are partly readable, while the others are wholly unintelligible, and, for some of them, we don't even know how to pronounce the characters. But Sr Alonso is not dismayed: he has discovered that they are all Basque.

One of his books received glowing reviews from the *Times* and from *Le Monde*. Here's a typical example of what the reviewers – who were not linguists – drooled over. This is not an especially dubious example nastily picked out by me: it's an example trumpeted by Alonso himself as illustrating the success of his methods particularly well, and quoted delightedly by the *Times*.

Alonso tells us that an Etruscan word *dule*, of unknown meaning, is "found in graveyards", and is "virtually identical" to the Basque word *dulle* 'scythe', which is "commonly used as a synonym for 'death' " – just the sort of word we expect to find in a graveyard. Impressive, eh?

Maybe not. Let's look.

First, most of the surviving Etruscan texts are from tombs, and so finding a particular word on a tombstone is not interesting. In the context, this is rather like claiming that a word found in a dictionary must have something to do with dictionaries.

Second, Etruscan had no consonant /d/, and the Etruscan alphabet didn't use the letter D. So, that reported *dule* is impossible, and a search of the on-line Etruscan lexicon fails to reveal *dule* or anything even vaguely similar. The word doesn't exist: Alonso has made it up.

Third, the choice of Basque as a comparandum for a word of unknown meaning is arbitrary and unmotivated. Why not choose the Irish surname *Dooley*, which means 'black hero', or Greek *doule* 'female slave', or Turkish *dul* 'widow', or any of a thousand other things instead? Why not pick something colourful and dramatic? Why prefer a mundane Basque name for a farmyard tool? Why not select Welsh *dwl* – pronounced 'dool' – which means 'foolish, stupid', and which for some reason creeps into my mind at this point?

Fourth, the supposed Basque *dulle* doesn't exist either: Alonso has made this up, too. You were waiting for that one, weren't you?

Fifth, the word he is trying to cite is Basque *dallu* or *dalla* 'scythe'. Apparently Alonso can't even copy a word out of the dictionary correctly. This word is real. But, as it happens, no native Basque word ever begins with /d/, and this is a transparent mediaeval borrowing from Romance descendants of Latin *daculum* 'scythe' – compare, for example, Gascon *dalha* 'scythe'.

Sixth, it is not true that any Basque word for 'scythe' is used as a synonym for 'death': Alonso has made this up, too.

Seventh, the personification of death as a Grim Reaper wielding a scythe is an explicitly Christian image, and one not recorded before the Middle Ages. The Etruscans were not Christians, and neither, for that matter, were the Basques before the tenth century.

The diligent reviewers failed to spot any of these irritating details, and the *Times* went so far as to add an awe-struck leader admiring Sr Alonso's revolutionary new piece of truth, which it called "exciting" and "scientifically fascinating". Following Sr Alonso's lead, the leader went on to declare that the Basques are "obsessed with death". Well, the Basques are as football-mad as anyone else, and they are passionate about cycle-racing, rowing, and good food, but, as a Basque obsession, death is right up there with underwater shove-ha'penny.

Losing its tenuous grip on reality altogether, the *Times* leader goes on to express surprise that "the smiling, fun-loving feminist Etruscans" may be related to "the dark and misogynistic Pyreneans". As it happens, the Basques are fair-skinned and blue-eyed. And misogynistic? A Basque wife is the absolute mistress of the household, and she has an equal say with her husband in choosing the heir – and a daughter may be preferred to a son. These journalists should get out more.

It's a lot easier to make these scholarly breakthroughs when you're allowed to invent your own data. Real data can be so tiresomely disappointing.

FOR INFORMATION

Larry Trask is Professor of Linguistics in the School of Cognitive and Computing Sciences at the University of Sussex. He is a leading expert on the Basque language.

Email: larryt@cogs.sussex.ac.uk
Web: http://www.cogs.susx.ac.uk/users/larryt/

Future Events

May 30-June 4 *DIALOGUE 2001*: Moscow, Russia.
Email: info@dialog-21.ru URL: http://www.dialog-21.ru

June 3-4 *Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customisations* (in conjunction with NAACL2001): Pittsburgh, Pennsylvania, USA.
Email: wim@dcs.shef.ac.uk URL: http://www.seas.smu.edu/%7Emoldovan/mwnw/

June 3 or 4 *Workshop on Machine Translation Evaluation* (with NAACL2001): Pittsburgh, Pennsylvania, USA.
Email: freeder@mitre.org URL: http://www.isi.edu/natural-language/mt-eval-naacl.html

June 4 *Workshop on Adaptation in Dialogue Systems* (with NAACL2001): Pittsburgh, Pennsylvania, USA.
Email: timpaek@microsoft.com URL: www.cs.utah.edu/%7Ecindi/AdaptDial.html

June 2-7 *2nd Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL2001)*; *Language Technologies*: Pittsburgh, Pennsylvania, USA.
Email: naaclpgm@isi.edu URL: http://www.cscmu.edu/0.000000E+00ref/naacl2001.html

June 10-11 European Commission/Soros *3rd Annual Summit of the East-West Collaboration in the Development of Interactive Media*: Budapest, Hungary.
Email: vlevai@osi.hu URL: http://www.osi.hu/ep/im2001

June 14-16 *5th Workshop on the Semantics and Pragmatics of Dialogue (BI-DIALOG 2001)*: Bielefeld, Germany.
Email: bidialog@uni-bielefeld.de URL: http://www.uni-bielefeld.de/BIDIALOG/

June 25-27 *7th Bar-Ilan International Symposium on the Foundations of Artificial Intelligence (BISFAI-'01)*: Ramat Gan, Israel.
Email: ariel@cs.biu.ac.il URL: www.cs.biu.ac.il/%7Ebisfai

July 5-6 *2nd International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*: Toulouse, France. Email: phil@sharp.co.uk URL: http://www.sle.sharp.co.uk/senseval2

July 6-7 *Workshop on Human Language Technology and Knowledge Management* (with ACL 2001): Toulouse, France.
Email: pmmmac@mitre.org URL: http://www.elsnet.org/acl2001-hlt+km.html

July 6 *Workshop on Arabic Language Processing Status and Prospects* (with ACL/EACL 2001): Toulouse, France.
Email: steven.krauwer@elsnet.org URL: http://www.elsnet.org/acl2001-arabic.html

July 6 *Workshop on Evaluation for Language and Dialogue Systems* (with ACL/EACL 2001): Toulouse, France.
Email: pap@limsi.fr URL: http://www.limsi.fr/TLP/CLASS/eacl01.html

July 7 *Workshop on Sharing Tools and Resources for Research and Education* (with ACL/EACL 2001): Toulouse, France. Email: declerck@dfki.de URL: http://www.elsnet.org/acl2001-tools.html

July 7 *Workshop on Data-Driven Machine Translation* (with ACL/EACL 2001): Toulouse, France.
Email: deborahc@microsoft.com URL: http://www.cs.unca.edu/%7Ebruce/acl01/MT.html

July 6-11 *39th Annual Meeting of the Association for Computational Linguistics (ACL/EACL 2001)*: Toulouse, France.
Email: acl2001@dfki.de URL: http://www.irit.fr/ACTIVITES/EQ_ILPL/aclWeb/acl2001.html

July 16-27 *9th Annual ELSNET European Summer School on Language and Speech Communication: Text and Speech Corpora*: Prague, Czech Republic.
Email: ess2001@ufal.ms.mff.cuni.cz URL: http://ufal.ms.mff.cuni.cz/%7Eess2001/

July 16-20 *Training Workshop in Lexicography and Lexical Computing*: Brighton, UK.
Email: lexicom-request@itri.bton.ac.uk URL: http://www.itri.bton.ac.uk/lexicom

July 30-Aug 11 *5th Eurolan Summer School on Creation and Exploitation of Annotated Language Resources*: Iasi, Romania.
Email: eurolan@infoiasi.ro URL: http://www.clg.wlv.ac.uk/eurolan/

This is only a selection of events – see http://www.elsnet.org/cgi-bin/elsnet/events.pl for details of many more events, including additional workshops associated with both NAACL in Pittsburgh, and ACL/EACL in Toulouse.

ELSNET

Office

Steven Krauwer,
Co-ordinator
Brigitte Burger,
Assistant Co-ordinator
Monique Hanrath,
Secretary
Utrecht University (NL)

Task Groups

Training & Mobility
Gerrit Bloothooft,
Utrecht University (NL)
Koenraad de Smedt,
University of Bergen (NO)

Linguistic & Speech Resources
Antonio Zampolli,
Istituto di Linguistica Computazionale (I) and
Ulrich Heid, Stuttgart University (D)

Research
Niels Ole Bernsen, NIS
Odense University (DK)
and Joseph Mariani,
LIMSI-CNRS (F)

Executive Board

Steven Krauwer,
Utrecht University (NL)
Niels Ole Bernsen, NIS,
Odense University (DK)
Björn Granström,
Royal Institute of Technology (S)
Nikos Fakotakis,
University of Patras (EL)
Ulrich Heid,
Stuttgart University (D)
Joseph Mariani,
LIMSI/CNRS (F)
José M. Pardo,
Polytechnic University of Madrid (E)
Geoffrey Sampson,
University of Sussex (UK)
Antonio Zampolli,
University of Pisa (I)

ELSNET Participants

Academic Sites

A	Austrian Research Institute for Artificial Intelligence (ÖFAI)
A	Graz University of Technology
A	University of Vienna
A	Vienna University of Technology
B	Leuven University
B	University of Antwerp - UIA
BG	Academy of Sciences Institute of Mathematics
BY	Belorussian Academy of Sciences
CH	SUPSI University of Applied Sciences
CH	University of Geneva
CZ	Charles University
D	Christian-Albrechts University, Kiel
D	German Research Center for Artificial Intelligence (DFKI)
D	Institute of Applied Information Science (IAI)
D	Ruhr-Universität Bochum
D	Universität Erlangen Nürnberg - FORWISS
D	Universität Hamburg
D	Universität Stuttgart
D	Universität des Saarlandes
DK	Aalborg University
DK	Center for Sprogteknologi
DK	University of Southern Denmark
E	Polytechnic University of Catalonia
E	Universidad Nacional de Educación a Distancia (UNED)E
E	Polytechnic University of Madrid
E	Polytechnic University of Valencia
E	Universitat Autonoma de Barcelona
E	University of Granada
EL	Institute for Language & Speech Processing (ILSP), Athens
EL	NCSR 'Demokritos', Athens
EL	University of Patras
F	IRISA/ENSSAT, Lannion
F	Inst. National Polytechnique de Grenoble
F	Institute de Phonétique, CNRS
F	LIMSI/CNRS, Orsay
F	LORIA, Nancy
F	Université Paul Sabatier (Toulouse III)
GE	Tbilisi State University, Centre on Language, Logic and Speech
HU	Lóránd Eötvös University
HU	Technical University of Budapest
I	Consiglio Nazionale delle Ricerche
I	Consorzio Pisa Ricerche
I	Fondazione Ugo Bordoni
I	IRST, Trento
I	Università degli Studi di Pisa
IRL	Trinity College, University of Dublin
IRL	University College Dublin
LT	Institute of Mathematics & Informatics
NL	Eindhoven University of Technology
NL	Foundation for Speech Technology
NL	Leiden University
NL	TNO Human Factors Research Institute
NL	Tilburg University
NL	University of Amsterdam
NL	University of Groningen
NL	University of Nijmegen
NL	University of Twente
NL	Utrecht University
NO	Norwegian University of Science and Technology
NO	University of Bergen
P	University of Lisbon
P	INESC, Lisbon
P	New University of Lisbon
PL	Polish Academy of Sciences
RO	Romanian Academy
RU	Russian Academy of Sciences, Moscow
S	KTH (Royal Institute of Technology)
S	Linköping University
UA	IRTC UNESCO/IIP
UK	Leeds University
UK	SOAS, School of Oriental and African Studies
UK	UMIST, Manchester
UK	University College London
UK	University of Brighton
UK	University of Cambridge
UK	University of Dundee
UK	University of Edinburgh
UK	University of Essex
UK	University of Sheffield
UK	University of Sunderland
UK	University of Sussex
UK	University of Ulster
UK	University of York

Industrial Sites

B	Lernout & Hauspie Speech Products
D	ALPNET Technology Gmbh
D	DaimlerChrysler AG
D	Grundig Professional Electronics GmbH
D	IBM Deutschland
D	Langenscheidt KG
D	Novotech GmbH
D	Philips Research Laboratories
D	Sympalog Speech Technologies AG
D	Varetis Communications
D	aspect Ges. für Mensch-Maschine Kommunikation mbH
DK	Tele Danmark
E	Sema Group sae
E	Telefonica I & D
EL	KNOWLEGDE SA
F	Aerospatiale
F	LINGA s.a.r.l.
F	LexiQuest
F	Memodata
F	SCIPER
F	Systran SA
F	TGID
F	VECSYS
F	Xerox Research Centre Europe
FIN	Kielikone Oy
FIN	Nokia Research Center
HU	MorphoLogic Ltd
I	CSELT
I	OLIVETTI RICERCA SCpA
I	SOGEI
I	Tecnopolis CSATA Novus Ortus
LV	TILDE
NL	Cap Gemini Nederland BV
NL	Compuleer
NL	IP Globalnet Nederland BV
NL	KPN Research
NL	Knowledge Concepts BV
NL	Sopheon
RU	ANALIT Ltd
RU	Russicon Company
S	Sema Group Infodata
S	Telia Promotor
UK	20/20 Speech Ltd
UK	ALPNET UK Limited
UK	BICC Plc
UK	BT Adastral Park
UK	Cambridge Algorithmica Limited
UK	Canon Research Centre Europe Ltd
UK	Hewlett-Packard Laboratories
UK	Imagination Technologies plc
UK	Logica Cambridge Ltd
UK	SRI International
UK	Sharp Laboratories of Europe Ltd
UK	Vocalis Ltd

What is ELSNET?

ELSNET, the European Network of Excellence in Human Language Technologies, is funded by the European Commission's Human Language Technologies programme. Members are academic and public research institutes (81) and industrial companies (55) from all over Europe.

The long-term technological goal, which unites the members of ELSNET, is to build integrated multilingual natural language and speech systems with unrestricted coverage of both spoken and written language. However, the realistic prospect for commercial applications involves systems that are restricted in one way or another. Such systems are of crucial importance for Europe in that they allow implementation of, and access to, the emerging multilingual information infrastructure. These systems also contribute to the increase of European industry's competitiveness by giving better access to product and service markets across language barriers.

Building multilingual language and speech systems requires a massive joint effort by two pairs of communities: on the one hand, the natural language and speech communities, and on the other, academia and industry. Both pairs of communities are traditionally separated by wide gaps. It is ELSNET's objective to provide a platform which bridges both gaps, and to ensure that all parties are provided with optimal conditions for fruitful collaboration.

To achieve this, ELSNET has established an infrastructure for sharing knowledge, resources, problems, and solutions by offering (information) services and facilities, and by organising events which serve academia and industry in the language and speech communities.

Electronic Mailing List

elsnet-list is ELSNET's electronic mailing list. Email sent to elsnet-list@let.uu.nl is received by all member site contact persons, as well as other interested parties. This mailing list may be used to announce activities, post job openings, or discuss issues which are relevant to ELSNET. To request additions/deletions/changes of address in the mailing list, please send mail to elsnet@let.uu.nl

Subscriptions

To subscribe to *ELSNews* visit **http://www.elsnet.org** and follow the links to *ELSNews* and subscription.

FOR INFORMATION
ELSNET
Utrecht Institute of Linguistics OTS, Utrecht University,
Trans 10, 3512 JK, Utrecht, The Netherlands
Tel: +31 30 253 6039
Fax: +31 30 253 6000
Email: elsnet@elsnet.org
Web: http://www.elsnet.org