XXV Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN'09)

http://ixa2.si.ehu.es/saltmil/

SALTMIL 2009 WORKSHOP

September 7th, Miramar Palace
Donostia - San Sebastián

Editors
Iñaki Alegria
Mikel L. Forcada
Kepa Sarasola

Eraikuntza Hutsa, J. Oteiza

Information Retrieval and Information Extraction for Less Resourced Languages

IE-IR-LRL

Proceedings

+ Index

SEPLN-SALTMIL 2009 workshop

Information Retrieval and Information Extraction for Less Resourced Languages

IE-IR-LRL

Donostia, September 7th, 2009

Proceedings

Editors:
Iñaki Alegria
Mikel L. Forcada
Kepa Sarasola

SALTMIL - University of the Basque Country

Acknowledgements

We gratefully acknowledge the financial support from the University of the Basque Country and the Ministery of Education and Science as well as the collaboration of the organizers of the SEPLN 2009 conference.

Published by the University of the Basque Country and SALTMIL

Cover design: Xabier Artola and Maite Oronoz
(based on the sculpture *Eraikuntza Hutsa* by Jorge Oteiza)

Printed by Fotocopias Zorroaga

ISBN: 978-84-692-4940-6

SEPLN-SALTMIL 2009 workshop

Information Retrieval and Information Extraction for Less Resourced Languages

IE-IR-LRL

Donostia, September 7th, 2009

Program Committee

Iñaki Alegria, University of the Basque Country
Atelach Alemu Argaw, Stockholm University, Sweden
Xabier Arregi, University of the Basque Country
Jordi Atserias, Barcelona Media (yahoo! research Barcelona)
Shannon Bischoff, Universidad de Puerto Rico, Puerto Rico
Arantza Casillas, University of the Basque Country
Mikel L. Forcada, Universitat d'Alacant
Xavier Gomez Guinovart, University of Vigo
Lori Levin, Carnegie-Mellon University, USA
Climent Nadeu, Universitat Politècnica de Catalunya
Jon Patrick, University of Sydney, Australia
Juan Antonio Pérez-Ortiz, Universitat d'Alacant
Bojan Petek, University of Ljubljana, Slovenia
Kepa Sarasola, University of the Basque Country
Oliver Streiter, National University of Kaohsiung, Taiwan
Vasudeva Varma, IIIT, Hyderabad, India
Briony Williams, Bangor University, Wales, UK

Introduction

The phenomenal growth of the Internet has led to a situation where, by some estimates, more than one billion words of text is currently available. This is far more text than any given person can possibly process. Hence there is a need for automatic tools to access and process this mass of textual information. Emerging techniques of this kind include information retrieval (IR), information extraction (IE), and question answering (QA).

However, there is a growing concern among researchers about the situation of languages other than English. Although not all Internet text is in English, it is clear that non-English languages do not have the same degree of representation on the Internet. Simply counting the number of articles in Wikipedia, English is the only language with more than 20 percent of the available articles. There then follows a group of 17 languages with between one and ten percent of the articles. The remaining 245 languages each have less than one percent of the articles. Even these low-profile languages are relatively privileged, as the total number of languages in the world is estimated to be 6800.

Clearly there is a danger that the gap between high-profile and low-profile languages on the Internet will continue to increase, unless tools are developed for the low-profile languages to access textual information. Hence there is a pressing need to develop basic language technology software for less-resourced languages as well. In particular, the priority is to adapt the scope of recently-developed IE, IR and QA systems so that they can be used also for these languages. In doing so, several questions will naturally arise, such as:

- What problems emerge when faced with languages having different linguistic features from the major languages?
- Which techniques should be promoted in order to get the maximum yield from sparse training data?
- What standards will enable researchers to share tools and techniques across several different languages?
- Which tools are easily re-usable across several unrelated languages?

The contributors to the workshop were asked in the call for papers to address these issues in real-world applications, and indeed, some of the issues, particularly the first two, have been directly addressed in four of the papers selected: Ussishkin, Francom and Woudstra address the issues involved in the creation of corpora and lexica for Maltese and Hebrew, Yimam and Libsie describe a question-answering system for Amharic (a semitic language, just as Maltese and Hebrew, spoken in Ethiopia), and Fernandez, Alegria and Ezeiza deal with the translation of named entities, a basic task in cross- language information retrieval, and Alegria and coworkers show how to build a question-answering system from existing resources for a language.

But, as would be expected in view of our focus on lesser-resourced languages, we have also received papers addressing other aspects of lesser- resourced-language technology. The four papers accepted describe minority- language projects currently under way: Pereira-Varela and colleagues describe Babelium, a multimedia framework to learn minority languages; Chan, Jones and East describe a system to automate the writing of school reports in Welsh and English, Prys describes a special interest group for minority- language speech and language technologies, Humphreys describes a project to automatically subtitle TV programmes in Welsh and Moulin, Laluque and Ó Néill describe a “shell” to manipulate dictionaries.

This volume starts with a contribution by our invited speaker, Lars Borin, professor of linguistic computing at the University of Gothenburg in Sweden, “Linguistic diversity in the information society”. Prof. Borin, after giving some background about the current situation of the languages of the world and reviewing the concept of density of a language, discusses the main issues encountered when trying to develop written-language technologies for lower- density languages; in particular information extraction.

Invited Talk

Regular Papers

Projects

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 re-

source 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

Creating a Web-based Lexical Corpus and Information-extraction Tools for the Semitic Language Maltese*

La creación de un corpus léxico basado en Internet y herramientas para la extracción de información léxica para la lengua semítica maltés

Jerid Francom
Wake Forest University
323 Greene Hall, PO Box 7566
Winston-Salem, NC 27109 (USA)
francojc@wfu.edu

Dainon Woudstra, Adam Ussishkin
University of Arizona
Douglass Building, Room 200E
Tucson, AZ 85721 (USA)
{dainon, ussishki}@u.arizona.edu

Abstract: In this paper, we document the creation of a web-based lexical corpus and a set of lexical tools for the less-resourced Semitic language Maltese. The benefits and shortcomings of using the web as a source of textual information are discussed in addition to the practical steps taken to develop and evaluate the resulting resources. We believe this preliminary work sets the groundwork for further development of Semitic language resources, as well as for less-resourced languages in general, and contributes to a growing interest to apply corpus data in theoretical analysis.
Keywords: Semitic languages, Maltese, Web-based corpus, Lexical corpus, minority languages, theoretical linguistics

1 Introduction

For many languages other than English, the quality and quantity of available resources is quite limited. To address this gap, many researchers have recently focused their efforts on documenting and creating database resources for these less-studied languages (McEnery, Xiao, and Tono, 2006). In addition to the clear role that electronic text resources play in computational applications, corpora are increasingly playing a larger role in the testing and development of linguistic theories, and a wide range of languages is crucial to the success and applicability of these theories.

In what follows, we document the creation of a Maltese lexical corpus developed in order to further psycholinguistic research on the mental organization of Semitic lexicons. The primary goal of this project is to create a sizable lexical corpus and a set of lexical calculation tools capable of producing a filtered set of lexical items for psycholinguistic experimentation. The sources for this effort include text extracted from the web and collaborative efforts from other scholars working in the area. We discuss both theoretical and practical issues in creating corpora in general and this corpus more specifically, elaborate the steps taken to bring this project to fruition and report the statistical results of our efforts. The research reported in this paper contributes to corpus linguistics as well to other areas of linguistics, including formal approaches to language investigation and psycholinguistics.

2 Web as corpus

A corpus can be thought of as a collection of texts. Traditionally, collections of texts have been amassed from prose found in print (Francis, 1975; Johansson, 1982; Sinclair, 1987), though the web has become an increasingly popular source for corpus creation due to several factors. The web contains a large amount of text already in electronic form, obviating tedious and time-consuming processes for converting print to electronic media. The web also provides access to a wide range of languages, language varieties, and genres that may be difficult to acquire in print.

The web as a data source for linguistic corpus creation is not without its theoretical and practical pitfalls. One consequence of sampling is that the data may not be as representative of the body from which it is

* We gratefully acknowledge funding from the United States National Science Foundation (BCS-0715500) to Adam Ussishkin.

extracted as assumed both in terms of balance and sparseness. Thus, the corpus linguist must make pragmatic decisions based on the ultimate purpose of the corpus, and evaluate the degree to which the data sufficiently fills this function.

Another concern is that not all languages are equally represented on the web. Evidence from various web experiments show that a handful of languages dominate 90% of the web, most notably English with over 70% (Xu, 2000; Kilgarriff and Grefenstette, 2003). This exacerbates the sparseness problem for languages with less content on the web.

The web also presents a level of variability not typical for print sources. Languages where non-standard web characters are lexically contrastive raises the possibility of conflation and misrepresentation in token counts. Finally, text found on the web is inherently more variable than text in print. Rigorous publishing standards for print are not always adhered to on the web, resulting in a larger number of typographical errors (Ringlstetter, Schulz, and Mihov, 2006).

In sum, there are a number of practical advantages that motivate data extraction from the web along with concerns that must be addressed in web-based corpora construction in order to ensure its theoretical integrity.

3 Development of the corpus

In what follows, we describe the development of a web-based lexical corpus for Maltese. This corpus represents two stages: 1) bottom-up construction of Maltese data including seed selection, web extraction, data filtering and data indexing and 2) a collaborative effort with a Maltese computational linguist providing pre-filtered text prepared for tokenizing and data indexing.

3.1 Seed selection and web extraction

The first step in creating a web-based corpus is to select the sources that will serve as the seeds for web extraction. A preselection list of URLs was obtained through a Google search. The most prevalent sources were newspapers and blogs. As discussed in the previous section, there are a number of criteria that need to be negotiated in developing a corpus from the web. From the standpoint of this project and general considerations of size, non-target language contamination, proofreading standards and corpus balance, a decision was made to pursue the online newspaper sources as our primary extraction seeds. [1]

Another aspect critical to the validity of a Maltese lexical corpus is character encoding. Of the potential sources, not all of them fully encoded all Maltese characters, including Maltese-specific characters (ċ, ġ, ħ, ż). Given these considerations, this project produced three seed candidates for extraction: Illum (`http://www.illum.com.mt/`), L-Orizzont (`http://www.l-orizzont.com/`) and Malta Right Now (`http://www.maltarightnow.com/`).

Extraction from the web often employs one of three main approaches. 1) Web-based searches through popular search engines, 2) more advanced search-engine based extraction via API[2] interfaces and 3) independent web crawling.

Both web-based searches and API searches through search engines have inherent drawbacks. First, search-engine based extraction is limited in terms of the number of queries one can run and the type of queries that can be performed. This highlights the 'brittle' aspect of depending on commercial parties to provide academically relevant services. Maybe more importantly, search engines do not freely disclose the sources of their databases and indexing lists. Therefore there is no way to determine the relevance of the sample to the population of interest.

Given these shortcomings, the current project selected to perform a web-crawl on the selected seed URLs. This approach avoids the natural restrictions placed on search-engine based extraction as all aspects of the process are independently maintained including extraction, documentation and interactive searching. After investigating a number of web-crawling tools including Heritrix (Mohr et al., 2004), WIRE (Castillo and Baeza-Yates, 2005) and Nutch (Khare et al., 2004), we opted to employ the open-source UNIX utility Wget. This decision reflected a desire to develop feature-rich strategies that can be easily obtained, configured

[1] See (Ghani, Jones, and Mladenic, 2005) for methodologies for extracting minority languages more generally from the web.

[2] Application Program Interface: Google provides a set of tools for building software applications, which interface with search engine queries.

and deployed by other scholars without having to compromise adherence to web protocol standards and best practices such as respecting robots.txt[3] and easing remote host server load.

Our particular extraction efforts with Wget included a number of syntax flags illustrated in figure 1.

Figure 1: WGET Syntax

```
wget -r -w3 -U LabBot=host.address.edu/
-A htm, html, asp, php, cfm, shtml
http://www.sitename.com.mt/ -o log
--output-document=outputfile.htm
```

Through this implementation we recursively crawled the sites identified previously (-r). The server was hit on an interval of once every three seconds (-w3) and each time the crawler reported the name of our bot and our web address. Our efforts and intentions were documented here with expressing our full compliance to stop and destroy data on request, an effort to respect best practices. The crawler was specified to only extract pages that conformed to a pre-screening of page extensions found to contain viable prose in text readable form (-A). Finally, all activities were recorded in a log file and the output appended to a standard web-content file (output-document=outputfile.htm) stored offline for processing.

The results of this crawl produced two key file types, output files and log files. Output files contained raw source code from those pages found inside the root directory of each of the target seeds. The log files contained connection information, URL name, file size downloaded and connection status. In the case of two of the three site seeds used in this corpus construction, the URL also contained the publishing date of the article or piece downloaded. This conveniently provided date range information for archived articles.

3.2 Data filtering, indexing and results

An attempt to remove non-target textual information including links, titles to articles and other redundant site text was conducted. This 'boilerplate' information contributes to 'noisy' or 'dirty' data that skews and obscures relevant corpus frequency data. A 'wrapper' approach to collecting relevant data can be used to extract text which occurs between certain open and close HTML tags (i.e., $<$ *body* $>, <$ *div* $>, <$ *span* $>$, etc.). Some of these tags may include specific CLASS or ID values, which may be useful for isolating sections of the HTML code for extraction. JavaScript and HTML comments, which occur sporadically between open and close tags, required additional filtering.

Several strategies to combat this issue were considered and tested. Machine learning techniques, such as CFR++,[4] used to recognize and tag webpage content, may be used to filter data. The most effective for this project, given reduced number of sites to filter, was to create the above mentioned 'wrapper' effectively identifying unique tags surrounding relevant text. The source code under scrutiny, for both news sources, showed consistent CSS tagging for ID and CLASS descriptors. This strategy provided robust exclusion of irrelevant data, but unfortunately did not exclude all irrelevant non-target text. Regular expressions were used to minimize these types of text as a final processing adjustment.

The strategy of shingling (Gibson, Wellner, and Lubar, 2008), which computes the resemblance between two documents. Shingling would effectively reduce the chances of processing near duplicate or identical texts. The extracted text does contain words from non-target languages which was not filtered out of the final data. The application of text categorization tool (TextCat (Cavnar and Trenkle, 1994), for example) would potentially reduce the amount of foreign text. Our original implementation does not include these technique/tools, but it may provide useful in future improvements of this data.

The tokenization process included the data obtained from the web-crawling procedures described in addition to a sizable set of data from Dr. Albert Gatt.[5] To tokenize both the webpage content and Gatt corpora, words were split according to morphologi-

[3] For more information about web ethics and web crawling see (Eichmann, 1995)

[4] `http://crfpp.sourceforget.net/`

[5] This supplementary data included text from Kulħadd, Leħen is-Sewwa, Il-Mument and In-Nazzjon online newspapers. Of note, Kulħadd, Leħen is-Sewwa and In-Nazzjon data were extracted from non-overlapping date ranges to the web-crawl described here.

cal boundary characters such as the apostrophe and any non-Maltese alphabet characters such as hyphens, slashes, among a few others. This converted any complex words into single token strings. All special characters and numbers were removed from the corpora. Multiple white space characters were minimized to a single space and punctuation was removed. The resulting long string data was split into an array at each remaining white space character. The resulting array was reprocessed into a hash table with corresponding counts for each token.

The database structure was designed according to a simple set of criteria. First, a token column is needed to textually represent each unique token, second a total count column and, finally one column of the token counts for each corpus processed. The database was designed to keep data collected from different corpora separate. For instance, if the user wanted to query only frequencies which occur in In-Nazzjon, its respective count information must be retained separately.

The two sources of data results in 3,323,325 total tokens, of which 53,396 are unique. Web-crawling produced 58.9%[6] of the total database and 40.2% from the corpora provided by Gatt. The largest percent of unique words was found in In-Nazzjon and the smallest percent from Leħen is-Sewwa (In-Nazzjon = 79.1%, Malta Right Now = 31.3%, Kulħadd = 15.3%, L-Orizzont = 9.3%, Leħen is-Sewwa = 8.6%). The following corpus counts are illustrated by (total count—unique count): Malta Right Now (1,927,598—8,165), In-Nazzjon (1,240,923—42,240), Kulħadd (69,908—8,165), L-Orizzont (60,982—4,944), Leħen is-Sewwa (24,914—4,577).

4 Corpus interface

In this section we describe the creation of a web interface to provide access to the lexical corpus.[7]. One goal was to provide international, cross-platform and requirement-free access to this collection. Thus, the interface was implemented using PHP,[8] an open-source technology with robust compatibility with data-driven websites that does not require any special software on the user end.[9] Another goal was to construct an intuitive graphic interface that provided encoding neutral interactive queries and facilitated seamless comparisons between various lexical calculations. In what follows we describe in detail these aspects of the user interface.

4.1 Basic tools

At the most basic level the interface includes detailed documentation concerning the sources, citations, counts and date ranges of the collection. This information can be found outside the registered-users area of the site in order to provide a good-faith effort to document our efforts.

Once inside the registered-users area the user is presented with a language selector and corresponding virtual keyboard (Figure 2). The design of this interface and the database underlying the web portal is designed to be extensible. In this way, any future plans to incorporate other languages can proceed without major modification to the interface.[10] On selecting a language, the relevant database token counts are presented. These counts are dynamically updated as the database is refreshed with new content.

The interface is also composed of a general interactive search field and a set of lexical calculation tools. The search field supports full use of POSIX regular expressions and is supported by a graphical web keyboard layout. The keyboard layout serves as an encoding-neutral input source that will allow users to query the Maltese and other language corpora with appropriate graphemes that may not happen to be installed on the local machine. In addition, the keyboard includes a number of characters employed in regular expression syntax.

The search fields maintain previous query strings, and thus provide access to quick comparisons between the three calculators: a lexical frequency calculator, a lexical uniqueness point calculator, and a neighborhood density calculator. These tools provide the user with

[6]The Illum data is not included in the database to date.

[7]This site can be found at `http://dingo.sbs.arizona.edu/~psycol/resources/` Registration is required.

[8]PHP is server-side HTML embedded scripting language for Hypertext Preprocessing.

[9]One exception is that users must have cookies enabled on the client machine in order to interact fully with the database.

[10]Currently access to a Hebrew corpus and Khalkha Mongolian corpus is available. Hebrew: 60,052,261 tokens; Mongolian: 259,264 tokens.

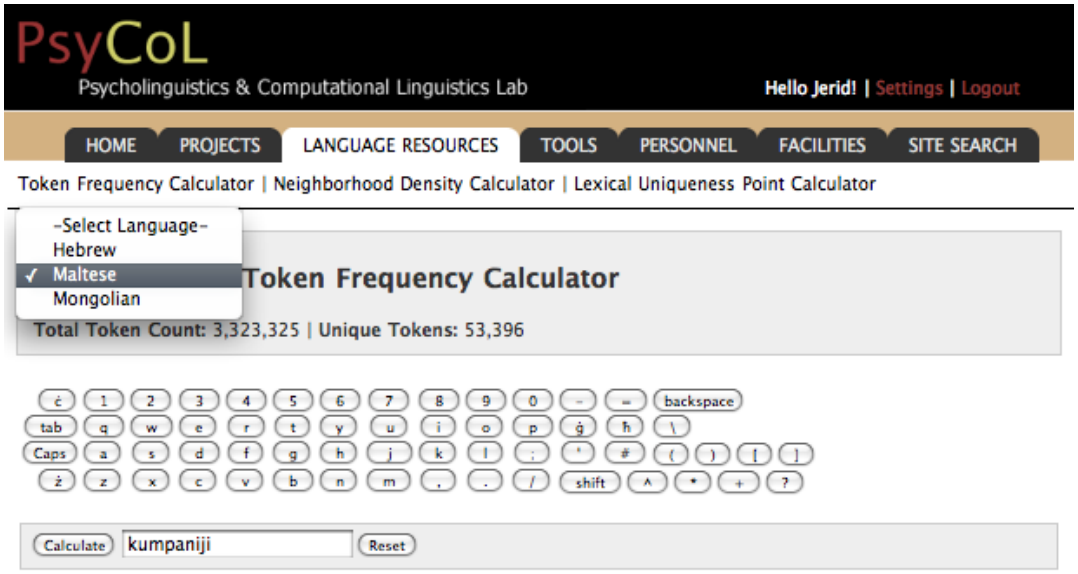


Figure 2: Basic web interface with language selector, query window and virtual keyboard

the opportunity to retrieve information useful for lexical statistics, item selection for experiments, and other uses.

4.2 Lexical frequency

Using POSIX regular expressions the lexical frequency calculator returns the total token count and number of unique tokens. The number of queried token counts divided by the total number of token counts in the corpus $\frac{c}{N}$ is known as lexical frequency. The returned query displays all tokens that match the regular expression with their respective total count and individual corpus counts. The return values for each token includes its counts per million and natural log frequency (Figure 3).

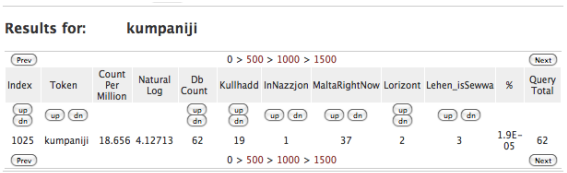


Figure 3: Token frequency sample output.

4.3 Lexical uniqueness point

The point at which a set of graphemes is no longer a subset of some other set of graphemes is known as a lexical uniqueness point. The uniqueness point can be indexed from left to right, or right to left. Marslen-Wilson's Cohort model (1978) suggests lexical processing makes optimal use of the lexicon during real-time use. Subsequent studies (Wurm, 2007) have shown reaction time effects corresponding to the Cohort model. Both calculations are included in the current web interface. This lexical tool queries the database for the desired string, which may be a word of the language but need not be. This string is then compared to each and every entry that contains it. If the string is not unique, the number and list of overlapped words is returned. If the string is unique, the point at which the input no longer overlaps is highlighted and two indices are provided (Figure 4). Not all unique strings are words, and this fact is also reported.

Figure 4: Lexical uniqueness sample output.

4.4 Neighborhood density

Research has shown that lexical access is sensitive to the number of lexical neighbors a given target has (Goldinger, Luce, and Pisoni, 1989; Cluff and Luce, 1990; Luce and Pisoni, 1998). The neighborhood density tool was designed to provide users with the ability to retrieve the number of neighbors of a given query string. The corpus density calculations are computed per query. The initial method of preprocessing this information was

rejected due to repopulation of data. As more data is added to the site, the neighborhood density would require reprocessing. To make this calculation on-the-fly, a language specific alphabet generates all combinations of token neighbors based on individual characters: insert a letter, remove a letter and replace a letter (i.e., a Hamming distance of 1). Each resulting possible neighbor is added to the query string. The resulting output contains the database counts for each neighboring token and the corpus density (weight) of these counts (Figure 5).

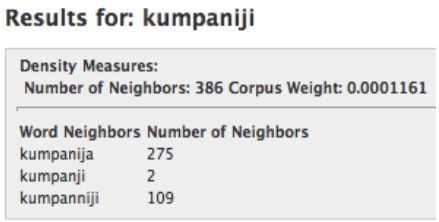


Figure 5: Neighborhood density sample output.

5 *Corpus evaluation*

The Maltese corpus described here shows both strengths and limitations as a language resource. First, the size of this collection represents quite a large sampling of the language, though "large" is a relative term. The size of electronic corpus data for English is quite humbling (i.e., Web 1T 5-gram Version 1, 1 trillion word tokens; North American News Text Corpus, 350 million word tokens; English Corpus Concordance, 18 million). To our knowledge, our Maltese corpus constitutes the largest lexical corpus in existence for Maltese.[11]

Another strength found in these resources is their general accessibility. Users can perform a number of calculations and queries without concern for special client-side software, and there is no need for local data storage. Finally, the corpora and lexical tools developed here were constructed with extensibility in mind. The underlying database structure is abstract enough that any language conforming to the encoding and database design can be accessed by the lexical tools with minor modification.

[11] Work reported by Kevin Scannell `http://borel.slu.edu/crubadan/` constitutes another large Maltese data set at 518,275 words.

There are potential limitations to our work that highlight crucial issues in corpus linguistics. First, all corpus creation must deal with the inadvertent inclusion of 'dirty' or 'noisy' data. These data can arise from various sources including author error in the case of misspelled or mistyped words from the original source and faulty filtering strategies attributable to the corpus designers in the form of web programming artifacts HTML, JavaScript, CSS, etc. We are actively engaged in improving our filtering process in order to address this issue.

Another limitation inherent in corpus development concerns the representativeness of word frequencies. The text extracted from the web is a pseudo-random sample of the target language, which provides a certain level of assurance that the corpus will contain a 'natural' balance of the language and the frequency of the words therein. It is important to bear in mind, however, that this limitation extends beyond the work presented here and must be assumed more generally as a part of all corpus development and analysis.

There are several limitations of concern that hold more specifically for our current enterprise and which constitute continuing areas of work. First, our efforts to design a maximally extensible resource are limited when data is collected via the web due to the idiosyncratic nature of web programming and coding practices. The wrapper approach adopted to filtering the raw web data retrieved cannot be performed automatically. Given the limited number of seeds (2) in our first web-crawl a machine-learning approach was not a requirement. As our project grows, however, we hope to explore more automatic strategies for teasing apart target data from extraneous web noise including wrapper (Prasad and Paepcke, 2008) and machine-learning techniques (Baroni and Ueyama, 2006; Spousta, Marek, and Pecina, 2008).

A second shortcoming is more specific to our primary goal of retrieving a filtered set of lexical items for psycholinguistic experimentation. Currently queries cannot be processed in batch. The ability to obtain lexical calculations for a set of items in one process will facilitate operations with the system. Plans to enable more robust queries and batch output is earmarked as a next step in the development of this set of language re-

source tools.

6 Conclusion

We have documented the creation of corpora and corresponding lexical tools for the Semitic language Maltese. These efforts coincide with a growing interest in developing corpus resources for less-resourced languages. In this project we highlighted the benefits and shortcomings of using the web as a source of textual information and pointed to emerging methods that may facilitate data extraction in future work. We believe this preliminary work sets the groundwork for further development and contributes to a growing interest to apply corpus data in theoretical analysis.

References

[Baroni and Ueyama2006] Baroni, M. and M. Ueyama. 2006. Building general-and special-purpose corpora by Web crawling. In *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, pages 31–40.

[Castillo and Baeza-Yates2005] Castillo, Carlos and Ricardo Baeza-Yates. 2005. Wire: an open-source web information retrieval environment. In *Workshop on Open Source Web Information Retrieval (OSWIR)*.

[Cavnar and Trenkle1994] Cavnar, W.B. and J.M. Trenkle. 1994. N-gram-based text categorization. *Ann Arbor MI*, 48113:4001.

[Cluff and Luce1990] Cluff, M.S. and P.A. Luce. 1990. Similarity neighborhoods of spoken two syllable words: Retroactive effects on multiple activation. *The Journal of the Acoustical Society of America*, 87:S125.

[Eichmann1995] Eichmann, D. 1995. Ethical web agents. *Computer Networks and ISDN Systems*, 28(1-2):127–136.

[Francis1975] Francis, WN. 1975. Problems of Assembling, Describing, and Computerizing Corpora. *Research Techniques and Prospects. Papers in Southwest English*, (1).

[Ghani, Jones, and Mladenic2005] Ghani, R., R. Jones, and D. Mladenic. 2005. Building Minority Language Corpora by Learning to Generate Web Search Queries. *Knowledge and Information Systems*, 7(1):56–83.

[Gibson, Wellner, and Lubar2008] Gibson, J., B. Wellner, and S. Lubar. 2008. Identification of Duplicate News Stories in Web Pages. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*.

[Goldinger, Luce, and Pisoni1989] Goldinger, S.D., P.A. Luce, and D.B. Pisoni. 1989. Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language*, 28(5):501–518.

[Johansson1982] Johansson, S., editor. 1982. *Computer corpora in English language research*. Bergen: NAVF.

[Khare et al.2004] Khare, R., D. Cutting, K. Sitaker, and A. Rifkin. 2004. Nutch: A flexible and scalable open-source web search engine. *Oregon State University*.

[Kilgarriff and Grefenstette2003] Kilgarriff, A. and G. Grefenstette. 2003. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 29(3).

[Luce and Pisoni1998] Luce, P.A. and D.B. Pisoni. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing*, 19(1):1.

[Marslen-Wilson and Welsh1978] Marslen-Wilson, W.D. and A. Welsh. 1978. Processing Interactions and Lexical Access during Word Recognition in Continuous Speech. *Cognitive Psychology*, 10(1):29–63.

[McEnery, Xiao, and Tono2006] McEnery, T., R. Xiao, and Y. Tono. 2006. *Corpus-based language studies: an advanced resource book*. Routledge.

[Mohr et al.2004] Mohr, G., M. Kimpton, M. Stack, and I. Ranitovic. 2004. Introduction to heritrix, an archival quality web crawler. In *4th International Web Archiving Workshop (IWAW04), Bath, UK*.

[Prasad and Paepcke2008] Prasad, J. and A. Paepcke. 2008. Coreex: content extraction from online news articles. In

Proceeding of the 17th ACM conference on Information and knowledge management. ACM New York, NY, USA.

[Ringlstetter, Schulz, and Mihov2006] Ringlstetter, C., K.U. Schulz, and S. Mihov. 2006. Orthographic errors in web pages: Toward cleaner web corpora. *Computational Linguistics*, 32(3):295–340.

[Sinclair1987] Sinclair, J.M. 1987. *Looking up: an account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins ELT.

[Spousta, Marek, and Pecina2008] Spousta, M., M. Marek, and P. Pecina. 2008. Victor: the Web-Page Cleaning Tool. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 12–17.

[Wurm2007] Wurm, L.H. 2007. Semantic processing in auditory lexical decision: Ear-of-presentation and sex differences. *Cognition & Emotion*, 99999(1):1–26.

[Xu2000] Xu, Jack. 2000. Multilingual search on the World Wide Web. In *Proceedings of the Hawaii International Conference on System Sciences HICSS*, volume 33.

TETEYEQ: Amharic Question Answering For Factoid Questions

"TETEYEQ: sistema de respuesta a preguntas factoides en lengua amárica"

Seid Muhie Yimam
Haramaya University, Ethiopia
seidyimam@haramaya.edu,
seidymam@gmail.com

Mulugeta Libsie
Addis Ababa University, Ethiopia
mlibsie@aau.edu,
mlibsie@aau.edu

ABSTRACT: The number of Amharic documents on the Web is increasing as many newspaper publishers started their services electronically. People were relying on IR systems to satisfy their information needs but it has been criticized for lack of delivering “readymade” information to the user, so that the Question Answering systems emerge as best solution to get the required information to the user with the help of information extraction techniques. The language specific issues in Amharic are extensively studied and hence, document normalization was found very crucial for the performance of our Question Answering system. The performance on normalized documents is found to be higher than on un-normalized ones. A distinct technique was used to determine the question types, possible question focuses, and expected answer types as well as to generate proper Information Retrieval query, based on our language specific issue investigations. An approach in document retrieval focuses on retrieving three types of documents (Sentence, paragraph, and file). An algorithm has been developed for sentence/paragraph re-ranking and answer selection. The named-entity-(gazetteer) and pattern-based answer pinpointing algorithms developed help locating possible answer particles in a document. The rule based question classification module classifies about 89% of the question correctly. The document retrieval component shows greater coverage of relevant document retrieval (97%) while the sentence based retrieval has the least (93%) which contributes to the better recall of our system. The gazetteer-based answer selection using a paragraph answer selection technique answers 72% of the questions correctly which can be considered as promising. The file based answer selection technique exhibits better recall (91%) which indicates that most relevant documents which are thought to have the correct answer are returned.

KEYWORDS: Amharic Question Answering, Answer Selection Techniques, Sentence/paragraph Re-ranking, Question Answering Evaluation

1. INTRODUCTION

Amharic documents on the web increases gradually as many newspaper agencies provide their service electronically. The traditional Information retrieval techniques were considered insufficient in retrieving precise information to the user. While information retrieval is effective by itself, users these days demand a better tool. First, they want to reduce the time and effort involved in formulating effective queries for search engines (users are required to formulate queries that should maximize document matching, and the search engine processes the query as submitted), and secondly they want their results to be real answers - not the list of relevant links. Automatic question answering has become an interesting research area and has resulted in a substantial improvement in its performance [1]. The aim of question answering (QA) is to retrieve exact

information from a large collection of documents, such as the Web. The main initiative behind QA system development is that users in general prefer to have a single answer or a couple of answers for their questions rather than having a number of documents to be read as it happens with the output of search engines [2]. Having a number of documents such as the World Wide Web or a local collection, a QA system should be able to retrieve answers to questions formulated in natural language. QA systems have already been developed in different languages such as Chinese [3, 4, 5], English [6, 7] and so on. This research is about Amharic Question Answering (AQA) System (**ተጠየቅ**), which is the first of its kind. Our QA system has been given a name ***ተጠየቅ*** *(Be questioned),* a historical verbalism in Ethiopia where two people appear before a judge used to ask a question for the defendant which are of kind ironic**.** Amharic is written with a version of the Ge'ez script known as **ፊደል** (Fidel). The Amharic language has its specific way of grammatical construction, character (fidel) representation and statement formation [8, 9, 10] where question answering system depends on both for question processing and answer selection techniques.

The question construction and answering techniques in Amharic language are different from English and other languages. In English, questions will be developed, for example, using "***wh***" words such as "who is the prime minister of Ethiopia?" and so on. But this same question will have different structure in Amharic such as a difference in character and word formation as well as grammatical arrangement and type of question particles (terms used to ask questions) used. For example, the above question will be translated as (**የኢትዮጵያ ጠቅላይ ሚኒስትር ማን ይባላሉ? -** *ye-ethiopia Teqlay minister man yibalal*). This question needs a special consideration to exactly return the correct answer, which is very different from English and other languages question answering techniques. There is no QA system developed for Amharic so far. In this study, we will investigate the problem and limitations of an Amharic search engine, the effect of developing QA system, analyze the strengths and weaknesses of QA with search engine and try to develop an Amharic question answering system.

2. THE AMHARIC LANGUAGE

Amharic is a Semitic language spoken in many parts of Ethiopia. It is the official working language of the Federal Democratic Republic of Ethiopia and thus has official status nationwide. It is also the official or working language of several of the states/regions within the federal system, including Amhara and the multi-ethnic Southern Nations, Nationalities and Peoples region. Outside Ethiopia, Amharic is the language of millions of emigrants (notably in Egypt, Israel and Sweden), and is spoken in Eritrea [11]. It is written using a writing system called fidel or abugida, adapted from the one used for the now-extinct Ge'ez language.

Ethiopic characters (fidels) have more than 380 Unicode representations (U+1200-U+137F) [12]. In every language, questions are constructed with the help of question particles (interrogative words) and question marks (?) which is placed at the end of the question. Table 1 shows some of the Amharic question particles.

Question word	Transliteration	Description
ማን	*man*	Who related questions
ለማን	*leman*	to whom …
ማነው	*manew*	Who is …..
የት	*yet*	Where …
ስንት	*Sint*	How many ….
ለምን	*Lemin*	Why
…	…	…

Table 1: Amharic Question Particles

3. DESIGN OF AQA

Every question answering system will have basic components of Question Analysis, Document retrieval and Answer Extraction [13, 14]. Our QA system has mainly five components, document

pre-processing, question processing, document retrieval, sentence/paragraph re-ranking, and answer selection modules.

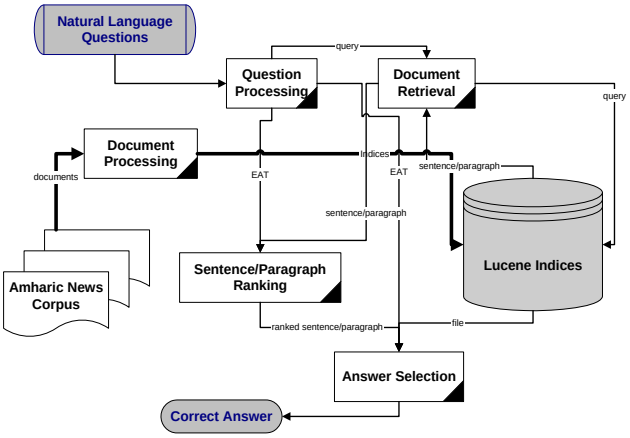


Figure 1: Architecture of the system

4. IMPLEMENTATION

In the **document preprocessing** module, documents will be normalized to show similar standards for document retrieval and answer selection processing. Amharic is too specific in having different character representation with the same reading and writing style. For example, the character ሀ (*ha*) has nearly five equivalent characters possess the same reading style, and people use it interchangeably with it, that are ሃ, ሐ, ሓ, ኀ, and ኃ and all occurrences these characters should be replaced with ሀ (*ha*). The research shows that document processing improves the performance of our system nearly by 12 percent (see Section 5). Besides character normalization, we also did number normalization. Numbers in Amharic represented in Arabic, Ethiopic, and alphabetic ways. The number normalization help as to detect all possible numeric answer particles (expected answers) in the document which otherwise left un-matched. To delimit documents in to sentence and paragraph, we have used different techniques. First sentences are detected with the Amharic full stop (።) if the document uses this punctuation mark. If the document uses group of Amharic word spaces (፡) or group of colons, we replace it with Amharic full stop. If the document is prepared with none of the punctuation marks mentioned, we use a frequently sentence finishing words such as (ነው-*newu*, ታውቋል-*tawuquwal*, ተባለ-*tebale*, ገልጿል-*geltsuwal*, etc.). Similarly paragraphs are detected by the normal paragraph separator (new line followed by a blank line) or an average number of sentences that can make up a paragraph. Once the document is normalized, sentences and paragraphs are delimited; then, our final task is create sentence, paragraph and file indexes using Lucene.

The **question processing** module accepts the user's question and performs tasks such as question type determination, question focus (important terms about the question) identification, and expected answer type determination. The question type will be determined based on the question particles and the question focuses. Since most of the question particles in Amharic are multipurpose, the question focus plays the greater role in determining the question type. We have developed a question typology that will be used to determine the expected answer type. The question processing module also generate the proper IR query that will be submitted to the document retrieval component of AQA.

The **document retrieval** component retrieves relevant documents so that the sentence/paragraph re-ranking module will process on it. For document retrieval module, different techniques were used. SpanNearQuery and RegexQuery, the Lucene contribution packages, have been used to maximize retrieval of documents with possible answer particles present. The RegexQuery was specially used to retrieve documents specifically for date and numeric related documents. The SpanNearQuery helps to filter out relevant documents by considering how far the query terms are present in the document. In addition to these techniques, we have also regulated the number of query terms presence in the document to be considered relevant to maximize relevant document retrieval. If the number of query terms is less than 3, the document is required to contain

all of the query terms to be considered relevant. If the query term varies from 4 to 6, at least 3/4 of the query terms should be present in the document and if the number of query terms is greater than 7, the document is required to consist at least half the query terms to be considered relevant. The rules designed indicate that as more number of query terms is present in a document, it is considered as better answer bearing document. Hence, the document retrieval component retrieves the sentence/paragraph and presents these documents to the sentence/paragraph re-ranking module and it also retrieve the total file and present the document directly to the answer selection module.

The **sentence/paragraph re-ranking** module first detects a possible answer particle in the retuned document. We have used two techniques to pinpoint a candidate answer in a document. The first one is Named Entity based (using a gazetteer for place names and person names, and regular expressions for numeric and date question types). The second one is pattern based answer pinpointing where a generic pattern is determined especially for person names. Once answer particles are identified, the best answer is determined based on query term-answer particle distance computation, if multiple answers are detected. The candidate answer in a document which seems very near to the query terms will be considered possible best answer. Once all possible candidate answers are identified from all documents, then another computation is done based on the number of query terms present in the document. When re-ranking, the document which shows more number of the query terms will be ranked atop, while the one with least number of query terms receive the least ranking weight.

The **answer selection** module selects the best top 5 answers from the previously ranked documents. Beside the already determined rank, the answer selection module also checks for possible repetition of answer particles from the candidate answer pool. If a given answer particle is repeated, the rank of the two will be summed to give a newer rank. Answer particles with the maximum rank value will be selected as an exact answer. The answer selection module considers two answers as equivalent if one of the other is the short form of it. For example **ጠቅላይ ሚኒስትር አቶ መለስ ዜናዊ** (Prime Minister Ato Meles Zenawi), **አቶ መለስ ዜናዊ**(Ato Meles Zenawi), **ጠቅላይ ሚኒስትር መለስ** (Prime Minister Meles), and **አቶ መለስ** (Ato Meles) are all considered equivalent.

5. EXPERIMENT

Java Programming language, the Lucene API, and a number of other third-party Java libraries such as *Fileutils* are used in developing our prototype. Figure 2 shows the user interface of our prototype.

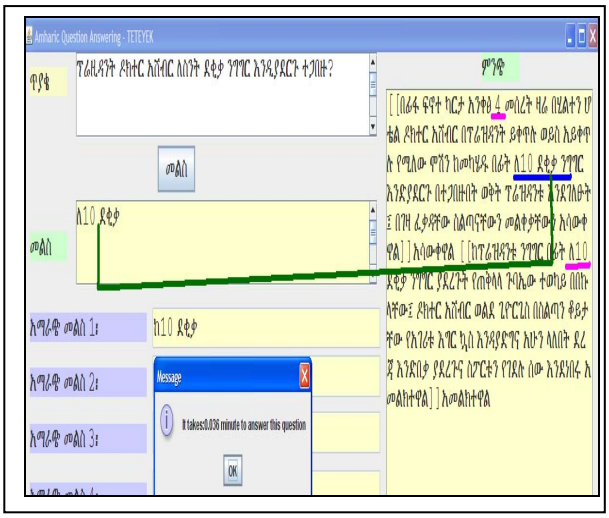


Figure 2: User Interface for AQA

Nearly 12000 question sets have been collected from the Web, Ethiopian Television games and from questionnaire respondents. A total of 15600 Amharic news articles (42 MB) corpus has been collected and normalized. Out of 12000 total questions, nearly 500 factoid questions are selected for the experiment. Hence, the experiment is conducted on the designed sample question and answer sets. The evaluation criterion we have used were correct answer accuracy. Hence the accuracy of our system is evaluated for recall, precision, percentage, and mean reciprocal rank (MRR). The evaluation formula for these criteria is as follows:

Recall: It is calculated as the total number of correct answers over the total of correct and missed answers.

$$\text{Recall} = \frac{correct}{correct + missed\ answers} X100\%$$

Precision: It is calculated as the percentage of correct answers over the total of correct answers, wrong answers, and No answers.

$$\text{Precision} = \frac{correct}{correct + wrong + No\ \ answers} X100\%$$

percentage: It is calculated as the total number of correct answers over all responses, wrong answers over all responses, and No answers over all responses

$$\text{Percentage} = \frac{correct}{total\ answers}, \frac{wrong}{total\ answers}, \frac{No\ answers}{total\ answers}$$

Mean reciprocal rank (MRR): It is also computed to evaluated average rank of answers; where rank is from top one to top five.

$$\text{MRR} = \frac{\sum_{i}^{n} \frac{1}{Ri}}{n}$$

Where Ri is the rank of a given answer which ranges from 1 to 5, and n is the total number of answers (correct +wrong + No answer).

The effect of document normalization is shown in table 2.

Document	Before normalization		After normalization	
	Precision	Recall	Precision	Recall
Sentence	53.3%	60.3%	66.6%	82.4%
Paragraph	55.4%	63.1%	63.7%	80.6%
File	51.4%	63.9%	55.3%	75.6%

Table 2: Effects of Document normalization

Table 2 shows that document normalization has a Performance gain of 7% for precision and 12% for recall. The Question processing module correctly classifies 89% of the questions using the rule based classification, while 62% are correctly classified by the IR based question classification techniques where question sets and answer sets are indexed so that an unseen question will be matched with the help of document similarity computations later.

Index type	Correct answer particles present	Wrong answer particles present
Sentence	465 (93 %)	35 (7%)
Paragraph	477 (95.4 %)	23 (4.6 %)
File	486 (97.2 %)	14 (2.8 %)

Table 3: document retrieval performance

Our document retrieval component also shows an excellent performance as shown in table 3.

Table 4 shows our Named-entity-based answer selection performance for person and numeric question types.

The pattern based answer selection outperforms the named entity based answer selection techniques as the named entity based answer selection technique fails to address all possible answer particles.

6. CONCLUSION

This research work attempted to identify the basic language specific issues in question answering. The first task we have tackled is normalizing the document so that a standard document will be indexed and matching relevant documents during searching will be maximized. We have also identified proper question particles as well as question focuses that will help in classifying the

Document	Number of correct answers	Number of wrong answers	Number of No Answers	Missed Answers	Precision	recall	MRR
Sentence	60 (56.6 %)	30(28.3 %)	16 (15.1%)	11	56.6%	84.5%	49.3%
Paragraph	72 (67.9%)	20 (18.9%)	14 (13.2%)	8	67.9%	90.0%	57.5%
file	60 (56.6%)	34 (32.1 %)	12 (13.3%)	6	56.6%	90.9%	43.8%

Table 4: Gazetteer based correct answer performance

question. Gazetteer based and pattern based answer selection algorithms have been developed to maximize correct answer selection.

Our algorithm first identifies all possible answer particles in a document. Once the answer particles are identified, the distance of every question particles toward the question terms will be calculated. The one with the minimum distance from the query terms will be considered the best candidate answer of that document. Once candidate answers are selected from every document, candidate answers which have been repeated more than once (i.e. appeared in more than one document) will be given higher rank. Candidate answers with maximum number of query terms matched in a document will be given higher priority in case a similar rank is given for two or more candidate answers. The evaluation of our system, being the first Amharic QA system, shows very well performance. The rule based question classification module classifies about 89% of the question correctly. The document retrieval component shows greater coverage of relevant document retrieval (97%) while the sentence based retrieval has the least (93%) which contributes to the better recall of our system. The gazetteer based answer selection using a paragraph answer selection technique answers 72% of the questions correctly which can be considered as promising. The file based answer selection technique exhibits better recall (91%) which indicates that most relevant documents which are thought to have the correct answer are returned. The pattern based answer selection technique has better accuracy for person names using paragraph based answer selection technique while the sentence based answer selection technique has outperformed the performance in numeric and date question types. In general, our algorithms and tools have shown good performance compared with highly resourced language QA systems such as English.

7. CONTRIBUTION OF THE WORK

The main contributions of this thesis work are summarized as follows:

- ✓ The study has adopted the efforts made towards English QA systems techniques to Amharic.
- ✓ The study has paved the way to identify language dependent components specific to Amharic question answering.

- ✓ The study identified key components of Amharic QA systems which can be considered a framework for factoid questions.
- ✓ The study showed the strategy, algorithms, and techniques in developing Amharic QA system.
- ✓ The study showed how questions in Amharic can be classified hierarchically (coarse and fine grained based), what are the specific question focuses for different questions, and the function of question particles to determine question type and expected answer types.
- ✓ This study also showed how information extraction can be accomplished in Amharic based on the standard off-the-shelf information retrieval techniques available.
- ✓ The study identified basic challenges in developing Amharic QA systems and the possible strategies to solve those challenges.

8. FUTURE WORK

Question answering is a very complex task, which consumes more time, and needs a number of different NLP tools. Hence, there are a number of rooms for improvement and modification for Amharic question answering. Below are some of the recommendations we propose for future work.

- Developing automatic named entity recognizer: The gazetteer we have used has limitations such as usage of a single named entity for multiple entities (such as person and place). Developing an automatic named entity recognizer will help the QA system to automatically detect the expected answer.
- Incorporating a parser and part of speech tagger: The NER will detect named entities in a document. A sentence parser will further help the QA system to know the structure of the question and the expected answer sentence. Besides, there is no POS tagger available publicly to integrate with our QA system. Integrating POS tagger will help the answer processing component of the QA system so that wrong answer particles, such as considering a verb as proper noun, will be eliminated.
- Developing Amharic WordNet: Word synonym, hyponym, antonym, metonym, meronym and so on help to match wider number of relevant documents. By using Amharic synonyms and the like, we believe that Amharic WordNet is very beneficial.
- Enhancing the Amharic stemmer: The stemmer that we have used brought some drawbacks both for document retrieval and answer processing algorithms. It will be better to develop a state-of-the-art stemmer which we believe will bring a significant change to the performance of QA systems.
- Incorporating Machine learning and statistical Question classifications: the rule based and IR based question classifications have some limitations. The rule based approach does not include all possible patterns of questions and the IR approach also does not help as the number of questions and question types indexed are very small. The machine learning and statistical approaches show better performance for other QA systems such as

English [60] and we hope it will also help for Amharic QA systems as well.

- Integrating with other search engines: for this research work, documents have been collected manually with the help of third party tools such as **DownThemAll** of Firefox and **WinHTTrack website copier**[1]. It will be better to incorporate a crawler component which will interact with the main search engines (Google, Yahoo, etc.) and Amharic Websites for collecting relevant documents.
- Extending to other question types: This research work shows that, even with minimal NLP tools, it would be possible to handle other question types such as list, define, and so on. Extending this work to other question types will be beneficial for wider applications where only a piece of information is not sought.
- Incorporating Amharic spell checker: most of the wrong answers and wrong documents returned are due to spelling errors. Incorporating spell checker will enhance the performance of our system.

Implementing for specific applications: The QA system can be easily implemented to satisfy the needs of some organizations for specific projects. It can be developed for customer service support such as e-commerce and e-governance.

REFERENCE

[1] Hu, H. Jiang, P. Ren, F. Kuroiwa, S. 2005. Web-based Question Answering System for Restricted Domain Based of Integrating Method Using Semantic Information Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference.

[2] Anne-Laure Ligozat, Brigitte Grau, Anne Vilnat, Isabelle Robba, Arnaud Grappy, 2007. Towards an automatic validation of answers in Question Answering, LIMSI-CNRS 91403 Orsay CEDEX France.

[3] Dongfeng Cai Yanju Dong Dexin Lv Guiping Zhang Xuelei Miao, 2004. A web based Chinese Question Answering System, Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference.

[4] Shouning Qu, Bing Zhang , Xinsheng Yu , Qin Wang, 2008. The Development and Application of Chinese Intelligent Question Answering System Based on J2EE Technology, Proceedings of the 1st international conference on Forensic applications and techniques in telecommunications, information, and multimedia and workshop.

[5] Zheng-Tao Yu Yan-Xia Qiu Jin-Hui Deng Lu Han Cun-Li Mao Xiang-Yan Meng , 2007, Research on Chinese FAQ questions Answering System in Restricted Domain, Machine Learning and Cybernetics, 2007 International Conference.

[6] Jignashu Parikh, M. Narasimha Murty, 2002. Adapting Question Answering Techniques to the Web, Proceedings of the Language Engineering Conference (LEC'02).

[7] Sameer S. Pradhan, Valerie Krugler, Wayne Ward, Dan Jurafsky and James H. Martin, Using Semantic Representations in Question Answering, Center for Spoken Language Research University of Colorado Boulder, CO 80309-0594, USA.

[8] http://en.wikipedia.org/ wiki/Amharic_language, last accessed on October 1, 2008

[1] HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility, http://www.httrack.com/

[9] **ጌታሁን አማረ፣ 1989 የአማርኛ ሰዋሰው በቀላል አቀራረብ** - *Getahun Amare, 1989 Ye-Amarigna sewasew beqelal aqerareb.*

[10] **ባየ ይማም፤1987 የአማርኛ ሰዋሰው፤ ት.መ.ማ.ማ.ድ**. – *Baye Yimam, Ye-AmariGna sewasew, T.M.M.M.D.*

[11] http://www.lonweb.org/link-amharic.htm, last accessed on March 30, 2009.

[12]http://jrgraphix.net/research/unicode _blocks.php?block=31, last accessed on March 31, 2009.

[13] Matthew W. Bilotti, Boris Katz, and Jimmy Lin, 2004, What Works Better for Question Answering: Stemming or Morphological Query Expansion?, Massachusetts Institute of TechnologyCambridge, Massachusetts, USA.

[14] Cheng-Wei Lee, Cheng-Wei Shih, Min-Yuh Day, Tzong-Han Tsai, Tian-Jian Jiang, Chia-Wei Wu, Cheng-Lung Sung, Yu-Ren Chen, Shih-Hung Wu, Wen-Lian Hsu, 2005, ASQA: Academia Sinica Question Answering System for NTCIR-5 CLQA, Proceedings of NTCIR-5 Workshop Meeting, December 6-9, 2005, Tokyo, Japan

Using Wikipedia for Named-Entity Translation

Izaskun Fernandez
Tekniker-IK4
Eibar, Basque Country
ifernandez@tekniker.es

Iñaki Alegria
IXA Group, EHU
Donostia, Basque Country
i.alegria@ehu.es

Nerea Ezeiza
IXA Group, EHU
Donostia, Basque Country
n.ezeiza@ehu.es

Abstract: In this paper we present a system for translating named-entities from Basque to English using Wikipedia's knowledge. We can exploit interlingual links from Wikipedia (WIL) to get named-entity translation, but entities without interlingual links can be translated using the Wikipedia as a corpus, suggesting new interlingual links. In this second case the interlingual links can be used as a test corpus in order to evaluate the translation process. We just need Wikipedia articles in both languages (specially in the target language) and a bilingual dictionary to apply this methodology to other language pairs.
Keywords: Named-Entity Translation, exploiting Wikipedia

1 Introduction

Person, location and organization names, are the main types of named entities (NEs), and they are expressions commonly used in all kinds of written texts. Recently, these expressions have become indispensable units for many applications in the area of information extraction, as well as for many searching engines. Named-entity translation task also has an increasing interest in the NLP community, since this kind of systems might help in the improvement of multilingual systems, such as machine translation or question answering systems. The proper processing of named-entities might not only improve numerical results in machine translation but also comprehensiveness of translations. Most systems dealing with NE translation are based on parallel corpora, which are aligned to extract the necessary information about different kinds of phrases, including NEs. However, and as it is widely known, obtaining parallel corpora is not an easy task, and it is even harder when one of the languages in the language pair is a minority language, as it is the case of Basque.

Our main goal is to build a multilingual NE database that would be used in multilingual or cross-lingual systems in general. Since getting the information for that multilingual NE database was a complex task, we decided to work in the field of NEs translation, designing a system for translating those expressions between different language pairs.

Wikipedia[1] is a free on-line multilingual encyclopedia written collaboratively by volunteers, where anyone can add and change articles. Each article in Wikipedia is uniquely identified by its title. Normally, the title is the most common name for the entity described in the article. Those forms that refer to an entity but are not the common forms, are represented in the Wikipedia through redirect and disambiguation pages.

Since Wikipedia is a multilingual resource, we can find entries in the Wikipedias for different languages, representing the same entity in each corresponding language. For instance, the Basque entry *Euskal Herria* and the English *Basque Country* represent the same entity in different languages. Wikipedia uses interlingual links (WIL)[2] in order to relate those forms in different languages. So

[1] http://en.wikipedia.org

[2] Links for each Wikipedia entry that connect them to the corresponding entries in the Wikipedias for other languages.

an exhaustive translation proccess may be avoided if we exploit WILs. For those entities without WILs, we propose a translation system based on the contents of Wikipedia in two different languages, following similar steps described in (Alegria *et al.*, 2008) for translation based on comparable corpora.

The paper is structured as follows. Section 2 presents the related works. Section 3 presents how to exploit Wikipedia for named-entity translation task. In section 4 we describe the development of the NE translation system using a limited amount of linguistic knowledge. In section 5, we present the results of the experiments, and finally, section 6 presents some conclusions and future work.

2 Related Works

Considerable research effort has been recently focused on machine translation systems (MT). Even though, most of the MT systems will translate the Spanish form *escuela de derecho de Harvard* into *school of the right of Harvard* instead of *Harvard Law School* which is the correct English form (Reeder, 2001). So, besides being a good way to obtain multilingual NE information, NE translation can be also considered a helpful task for MT improvement.

Concerning the resources, despite the difficulty to get bilingual parallel corpora for many languages, most NE translation systems work with parallel datasets. Those bilingual corpora are aligned not only at the paragraph level but also at the sentence level. For example, Moore's work (Moore, 2003) uses bilingual parallel English–French aligned corpora, and he obtains a French form for each English entity applying different statistical techniques.

Although comparable corpora have been less studied, there are some known systems designed to work with them as well, such as the system that translates entity names from Arabic to English (Al-Onaizan and Knight, 2002a) (Al-Onaizan and Knight, 2002b), and the Chinese–English translation tool presented in ACL 2003 (Chen *et al.*, 2003).

The main goal of both systems is to obtain the equivalent English form, taking Chinese and Arabic respectively as source language. Two kinds of translations can be distinguished in both systems: direct/simple translations and transliterations[3]. However, each tool uses a different technique. Frequency-based methods are used in Chinese–English translations, while in the Arabic–English language pair, a more complex combination of techniques is applied.

Similar techniques are applied in (Sproat *et al.*, 2006) and (Tao *et al.*, 2006), which transliterate English–Chinese NEs using comparable corpora. The former combines a supervised phonetic transliteration technique and a phonetic frequency correlation approach, while the latter combines those techniques, but applying the phonetic approach in an unsupervised way, where the distance is determined by means of combining the substitution, insertion and deletion of characters.

Not only approaches to named-entity transliteration have been presented in this area. The system presented in (Poliquen *et al.*, 2005) integrated at the news analysis system NewsExplorer[4], tries to extract person names from multilingual news collections to match name variants referring to the same person, and to infer relationships between people based on the co-occurrence information in related news.

WIL links are used to try to enrich the German–English pair (Sorg and Cimiano, 2008). They show that roughly 50% of the articles in German are linked to their corresponding English version and only 14% from English to German. They present a classification-based approach based on text-based features and/or graph-based features for that enrichment. The experiments show that their approach has a recall of 70% with a precision of 94%.

Multilingual named-entity recognition based on Wikipedia is faced on (Richman and Schone, 2008), showing that English language data can be used to bootstrap the NER process in other languages. For multilingual categorization they use links among languages when possible, and categories with their English equivalents in the remaining cases.

Concerning Basque, in our previous work we have found two different approaches to translate Basque NEs into Spanish (Alegria *et al.*, 2006). The first one was a language de-

[3]Transliteration is the proccess of replacing words in the source language with their approximate phonetic or spelling equivalents in the target language.

[4]`http://press.jrc.it/NewsExplorer/entities/en/1.html`

pendent tool for translating NEs from Basque to Spanish using comparable corpora. That system used linguistic information for both transliteration and entity element rearrangement. This system was tested using a set of the most common entities, and it obtained an f-score of 78.7% in named-entity translation task.

However, as the development of a language dependent system for each language pair was very expensive, we tried a relatively language semi-independent tool following a similar strategy, and using comparable corpora and bilingual dictionaries. This tool was tested first in the Basque–Spanish language pair and it shows that the performance was quite close to the language dependent tool, obtaining an f-score of 77.5%.

To confirm that the methodology was general enough, we tried using the translation methodology for the Spanish–English language pair in (Alegria *et al.*, 2008) and we obtained almost 65% f-score, which is a considerable lower performance. After observing the errors in detail, it was detected that due to the bad quality of the comparable corpora the 13% of the English entity forms where badly defined in the target corpus, so by correcting them the system would get results that are as good as the ones got for Basque–Spanish language pair. So the methodology based on comparable corpora seemed to be a good choice for developing systems to translate NE for different language pairs.

Thus we have obtained the language semi-independent approach to develop the Basque–English NE translation tool but using Wikipedia as a corpus. We will describe in more detail the system architecture in the following section.

3 Exploiting Wikipedia

Many articles of Wikipedia can be found in different languages, and that is why this encyclopedia can be considered an interesting resource to get named-entity translations between different language pairs, specially if the target language is English.

In this work, we exploit Wikipedia contents in many different ways. As we have mentioned before, the encyclopedia has interconnected entries from different languages by means of interlingual links (WIL), which means that both entry forms represent the same entity in their corresponding language. So this is the most effective and cheapest way to get named-entity translations from Wikipedia. Unfortunately, some named-entity pairs lack an interlingual link, which means that other techniques must be used in order to translate them.

Most of the translation systems use a target language lexicon, and as we will see in the next section, we also use an English lexicon in our system. Since the English Wikipedia is very rich and a resource that is continuously growing, we considered it an interesting source for our target lexicon generation. But as we are dealing with NEs, we are only interested in words appearing in this kind of expressions and not just in any kind of words. Because of the wide coverage of the target Wikipedia we assume that most of the source words would have their corresponding translation in the target lexicon if we use this resource.

Yahoo! has a semantically annotated Wikipedia[5] version presented in (Atserias *et al.*, 2008), where NER task has been applied. We took this version and extracted all the tagged NEs. We constructed the target lexicon by means of including words in those NEs and excluding grammatical words such as prepositions, articles, etc. using a stoplist[6].

Combining the lexicon with some techniques that we will see in the following section, we constructed a system that proposes some translation candidates for each given Basque entity. In order to make sure if they are suitable proposals, we can use Wikipedia again, this time for searching if proposals have an entry in the on-line encyclopedia. So, the system only gives revised named-entity translation proposals.

4 System Description

The system proposed for named-entity translation task in this work uses three main modules: 1) a searching module with different resources for searching, 2) an entity elements translation module using a transliteration grammar combined with a bilingual dictionary for those words that cannot be translated only by applying transliteration but still need some translation and 3) an el-

[5] `http://barcelona.research.yahoo.net/dokuwiki`

[6] `http://www.lc.leidenuniv.nl/awcourse/oracle/text.920/a96518/astopsup.htm`

ement rearranging module for the construction of the whole entity from components, which will treat the possible differences in syntactic structures.

As Figure 1 shows these three modules are applied following four main steps when a Basque NE is given for translation:

- Step 0: Searching for WIL between the Basque entity and an English Wikipedia entry (*Searching Module*)
- Step 1: Searching for a translation for the entire Basque NE as a multiword lemma in the bilingual dictionary (*Searching Module*)
- Step 2: Searching the Basque entity in the English Wikipedia (*Searching Module*)
- Step 3: Translation/transliteration of entity elements themselves, finally constructing the entire translation proposal using the individual translations and searching these entire proposals in the English Wikipedia.

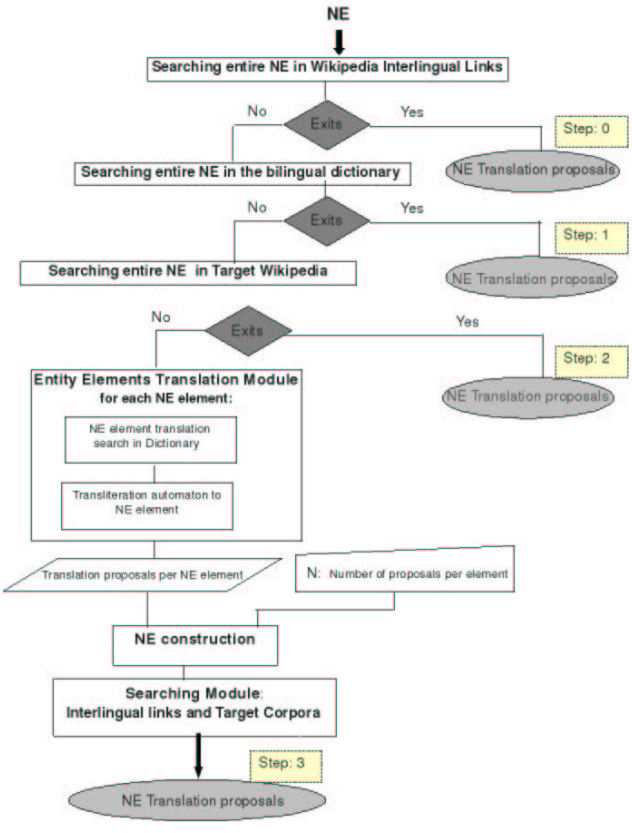


Figure 1: System Architecture

In the following subsections we will present each module in detail.

4.1 Searching Module

This module contains three main functionalities which are used in different steps of the system architecture: searching in Wikipedia's interlingual links, in the target Wikipedia, and in the bilingual dictionary.

As the previous step (0) the system exploits the interlingual links of Wikipedia and if a link exists between the input Basque entity and an English Wikipedia entry, the system suggests this English form as translation for the input expression. If no link is found, step (1) is applied.

In the step (1) the whole NE is searched in the bilingual dictionary. If the form is found the translation is obtained. This step is applied as a baseline in the experiments. For example the translation *Euskal Herria–Basque Country* is resolved using the dictionary. If no translation is found, step (2) is applied.

In the step (2) the system verifies if the input Basque form exists with the same form in the English Wikipedia. If it does, then the system will propose the entity in the target language as translation proposal.

When no translation proposal is obtained in the previous steps, and the system has to translate each element and construct then the entire translation posible forms, the last functionality of this module is applied in order to reduce the amount of suggested proposals and to improve the quality of them: in this step the system checks for every translation proposal if the proposed entry exists in the English Wikipedia, and only forms with entries will be given as possible translations.

As an additional functionality the system exploits the redirection links of Wikipedia. As it has been explained previously, these links connect the different forms in Wikipedia to refer to an entity in the same language. This way, the system returns not only those entity translations found or trusted to exist in the Wikipedia, but also all their connected forms.

The system uses the MediaWiki API [7] to exploit both the interlingual and the redirection links. For example the translation (*Alpeak*, *Alps*) can be obtained exploiting interlingual links and the result can be enriched with the pair *Alps–The Alps* using the redirection links.

[7] http://www.mediawiki.org/wiki/API

4.2 Entity element translation module

When no entity translation proposal is obtained from step (0), step (1) and step (2) the system applies the word-by-word translation process. The bilingual dictionary and the finite-state transducer, combined with the English lexicon, will be used in order to obtain translation proposals for each entity element.

As it is explained in (Alegria *et al.*, 2008), edit distance (Kukich, 1992) based on a finite-state grammar and a lexicon of the target language are neccesary for constructing trasliteration rules. Since each rule can be applied n times for each word, the set of all translated words that we obtain after applying rules independently and combining them, is too large. In order to reduce the size of the set of proposals, the system combines the grammar with the lexicon of the target language obtained from the Wikipedia, and it restricts the transformation rules to at most two applications per word, avoiding the generation of words with more than two transformations, as it is shown in the top of Figure 2.

With this transliteration automaton, the system will be able to translate *Txina* into *China*.

However, there are some translations that cannot be obtained applying only transliteration/edition rules. The system uses a source-target bilingual dictionary, converted into an transducer for this aim. The module strategy is shown in the bottom part of Figure 2 and is applied in the following order:

- get translation looking up the bilingual dictionary.
- suggest an identical word if it is in the target lexicon.
- propose words in the target lexicon with distance 1 from the source word.
- suggest words in the target lexicon with distance 2 from the source word.

So this module is able to translate not only the transliterated words such as *Kuba–Cuba*, but also, words that cannot be translated using transformation knowledge and need information from a bilingual dictionary, such as *Erakunde–Organization*.

For both transliteration and bilingual dictionary based automata, the system uses

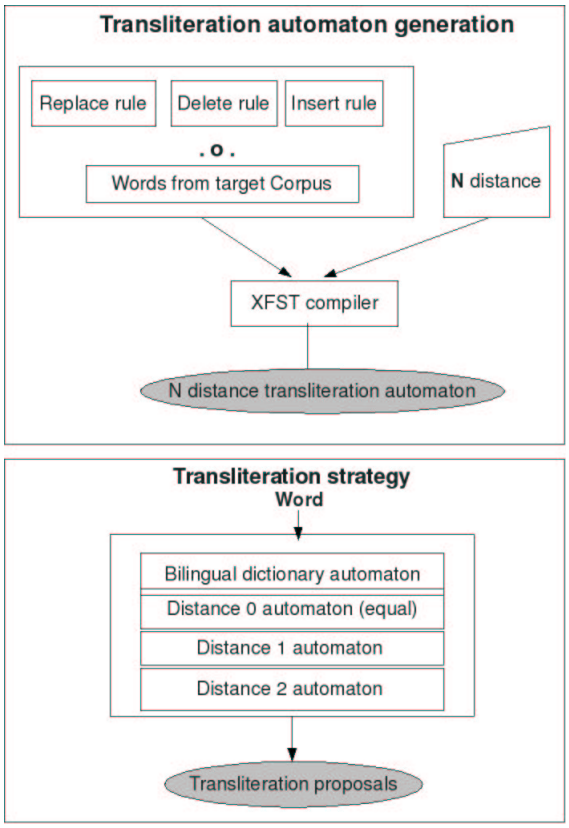


Figure 2: Transliteration automaton and strategy

the lemmatized form of the entity elements, applying Eustagger, the Basque lemmatizer/tagger (Ezeiza *et al.*, 1998) developed by IXA group .

Both kinds of automata are combined for translating entities like *Ipar Katalunia* into *North Catalonia*, using the dictionary for converting *Ipar* into *north*, *northern*, *north wind* and transliteration for transforming *Katalunia* into *Catalonia*, *Catalunya* and *Katatonia*. After looking up in the Wikipedia *North Catalonia* and *Northern Catalonia* forms are suggested.

4.3 Entire Entity Construction

Once each element is translated, the entire entity construction must be performed. For this work we cannot ignore the different syntactic patterns between languages, and this makes necessary to include some treatment for element rearrangement. This module is applied before searching for translation candidates in the target Wikipedia. As mentioned in (Alegria *et al.*, 2008), this module combines each proposed element with the rest, considering that each proposal can ap-

pear in any position within the entity.

Although in some cases prepositions and articles are needed to obtain the correct target form, the translation candidates for the whole entity will not contain any element apart from the translated words of the original entity. So, we will take into account the lack of these elements in the following step.

For reducing the amount of translation proposals, only the N most suitable translations for each word will be considered for the entire construction. For instance, when the system tries to translate *Itsaso Gorria* , the system gets candidate *Sea* for *Itsaso* and *Red* for *Gorria*. And this module generates the following entire candidates, considering that both elements can appear in any position: *Sea Red* maintaining the original position of words, and *Red Sea* inverting the positions. In this case, the correct form is the one obtained changing word order.

5 Experiments

We had no evaluation corpus for the system, so we considered convenient to generate an evaluation corpus in a semi-automatic way. We used two resources for this task: Wikipedia and the CLEF evaluation set. For the Wikipedia set we exploited the interlingual links in order to obtain the gold standard for testing; so, in this case the (0) step will be not applied.

For both evaluation sets we have used the same three measures:

- $Precision = \frac{correctly\_translated\_NEs}{Translated\_NEs}$
- $Recall = \frac{correctly\_translated\_NEs}{All\_NEs}$
- $f - score = \frac{2*Precision*Recall}{Precision+Recall}$

5.1 Evaluation with Wikipedia based corpus

For the construction of the first evaluation corpus, we have used a Basque article collection borrowed from *Euskaldunon Egunkaria*[8], which is a newspaper entirely written in Basque closed since February 20th 2003, and the WILs. The Basque corpus has 40,648 articles with 9,655,559 words and 142,464 NEs tagged in the Hermes project[9] (news databases: cross-lingual information retrieval and semantic extraction).

We selected the most frequent NEs from the Basque collection and searched the WILs in Basque–English direction to find linked English forms, we got a collection of 575 entity pairs interlinked in the Basque–English Wikipedia languages. Since interlingual links are used for the corpus generation, we will not use them for suggesting translations (Step 0 in Figure 1 is not carried out).

Steps	Total	Correct
Step (1)	17	11
Step (2)	391	375
Step (3)	59	48
No-Translation	108	0

Table 1: Translations distribution

Table 1 shows the number of translated entities in each step of the system, together with the amount of well translated entities. In the first row, we can see the number of entities that have been found in the dictionary. The second row shows how many Basque entities have been found in the English Wikipedia, thus they have no need to be translated element by element. In the third row we can see the entities that have been translated using the language semi-independent system[10]. Finally, we can see that around 19% of times the system did not obtain a translation.

	Pr.	R.	fs
Baseline	59.82%	59.82%	59.82%
Our system	93.36%	75.82%	83.68%

Table 2: Results for Wikipedia-based test set

In Table 2 we present the evaluation, and in the first row a baseline is shown. The baseline is calculated considering correct translations when Basque and English forms are identical.

The results are very encouraging, since we have obtained 83.68% f-score.

Analysing the errors in the development corpora we observed that sometimes WILs do not link the same entity form. For instance, if *Dorre Bikiak* is searched in the Basque Wikipedia and the interlingual link is used to obtain the English translation, the same way it has been used to build the test corpus, the Basque form found is *World Trade*

[8] http://www.unibertsitatea.net/blogak/ixa/egunkaria-hizkuntza-teknologiako-baliabideen-sortzailea

[9] http://nlp.uned.es/hermes/

[10] Translating each entity, constructing the English proposals with elements' translation and finding the best proposals with the searching module

Center, instead of *Twin Towers* which must be the interlingual linked form.

With the proposed translation system this kind of new links could be good suggestion to be added to the Wikipedia.

5.2 Evaluation with CLEF based corpus

Since Wikipedia has been used for constructing the named-entity translation system presented in this work, it can be considered that the evaluation corpus is biased in favor of the system. So, we considered interesting to evaluate the system using another NE set, independent from this encyclopodedia.

For that purpose, we used the ResPubliQA CLEF-2009[11] test set. This test set has 500 questions translated into Bulgarian, Basque, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish. We used Basque and English versions to construct the new evaluation set for the NE translation system.

Exploiting the questions set, we obtained 72 Basque–English NEs pairs, where 9 of them has no entry in the target Wikipedia. Since our system only proposes english translations trusted in the English Wikipedia, even if it gets correct English forms for that 9 entities, it would never propose them because they are not in the English Wikipedia. So, we can say that for this test set, our system's topline recall is 87.5%.

We have tested the system in two different ways for this evaluation corpus: When the system does not find out any of its proposals in the target Wikipedia,

- no translation is returned (silence-mode).
- the original Basque form is returned as translation proposal (talkative-mode).

	Pr.	**R.**	**fs**
Baseline	23.61%	23.61%	23.61%
Silence-mode	92.68%	52.77%	67.25%
Talkative-mode	55.5%	55.5%	55.5%

Table 3: Results for CLEF test set

Since CLEF test set does not belong to the Wikipedia, for this evaluation, we exploit the WILs between Basque and English Wikipedias(step0 in the system architecture), in order to evalaute the complete system architecture. Exploiting those links, we obtain translations proposals for 26 of the 72 entity colection, and all but one agree with the test set English forms.

As you can see in Table 3, this time the results are not as encouranging as the previous ones, but we want to highlight the improvement that the system gets respect to the baseline. So it would be pretty good to evaluate the system with a bigger and more extensive corpus.

6 System Improvement

Analysing the errors occurred with both evaluation sets, we detect that our biligual dictionary was not very suitable in many cases for translating words appearing in NEs. For example, if we try to translate *Nazio Batuak* into *United Nations* and we use our bilingual dictionary for it. First we will lemmatize the basque elements, then look up them into the dicitonary and finally we will get *Nation* and *Union* respectively. With this forms we will never get the correct English entire entity form.

But if we were able to enrich the dictionary with *Nazio–Nations* and *Batuak–United* pairs, the system will be able to obtain *United Nations* as translation candidate.

So, we decided to carry out an automatic dictionary lexical enrichment expoiting a small Basque–English WIL set, and then check if that enrichment improves our system performance, evaluating the system with CLEF test corpus.

We take as Basque–English WIL input set the wrong translated entities from the Wikipedia Based test corpus, concretly 84 entity pairs. For each entity pair, we try to match each Basque entity element with their corresponding English entity element mantaining the source Basque form, or translating with the existing bilingual dictionary. When every element in the Basque entity is parsed, if only one basque element has not been matched and in the target entity there is only one element too that has not been assigned, we will consider those elements as translations, and we will enrich the bilingual dictionary with them. This metholodogy will be applied iteratively, enriching the dicitonary in each step and using it in the following one, until no new elements' translation proposals are obtained.

[11] `http://celct.isti.cnr.it/ResPubliQA/`

For instance supose that in the input WILs set we have *Europako Parlamentua–European Parliament*. Carrying out the previously explained matching, we will not get any matching for *Europako* because it does not exist in bilingual dictionary, and it does not match identically with any of the English elements. Even though, from the bilingual dictionary we get that *Parlamentua* matches with *Parliament*. So *Europako* is the unique element in the Basque form without matching, and *European* in the English form. Applying the previous assumption, we will consider *European* a possible translation form for *Europako*, and we will entrich the dictionary with it.

	Pr.	**R.**	**Fs**
Silence-mode	93%	55.5%	69.51%
Talkative-mode	58.33%	58.33%	58.33%

Table 4: Results for CLEF test set with lexical enrichment

As you can see in Table 4, observing a very small set of entity pairs, we obtain a sligth improvement in the system. So it could be interesting to consider the entire Basque–English pairs linked with WILs to get a better lexical enrichment.

7 Conclusions and Further Work

We have presented an approach to translate NEs using the Wikipedia encyclopedia as main resource. It has been shown that exploiting Wikipedia might benefit in two directions: on the one hand it may help in building a good quality named-entity translation system; on the other hand, new interlingual links for Wikipedia might be suggested.

The evaluation gives us promising results but a deeper evaluation and error analysis is needed, for studying solutions for entities with different number of elements in each language. It would be also interesting to test this technique in other languages.

As further work we want to disambiguate NE written in minority languages such as Basque. Since the resources for that kind of languages are very limited, we are intending to use the translation system proposed in this paper for exploiting the information of the languages with much more resources like English, and the Wikipedia's disambiguation links.

Acknowledgment

This was partially supported by the Spanish Ministry of Education and Science (FIT-340000-2007-157 carried out at Tekniker and TIN2006-15307-C03-01)

References

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. *Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages*. COLING-ACL'98. Pgs.380–384 Vol 1. Montreal(Canada). August 10-14,1998.

Alegria I., Ezeiza N., Fernandez I. 2006. *Named Entities Translation Based on Comparable Corpora*. Proceedings of Multi-Word-Expressions in a Multilingual Context Workshop in EACL 2006. W06–2401.

Alegria I., Ezeiza N., Fernandez I. 2008. *Translating Named Entities using Comparable Corpora*. Proceedings of Building and Using Comparable Corpora Workshop in LREC 2008.

Al-Onaizan Y., Knight K. 2002. *Translating Named Entities Using Monolingual and Bilingual Resources*. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics 2002. Pgs. 400–408.

Al-Onaizan Y., Knight K. 2002. *Machine Transliteration of Names in Arabic Text*. Proceedings of the ACL-02 workshop on Computational approaches to semitic languages. Pgs. 1–13.

Atserias J., Zaragoza H., Ciaramita M., Attardi G. 2008. *Semantically Annotated Snapshot of the English Wikipedia*. Proceedings of LREC 2008. L08–1165.

Beesley K.R., Karttunen L. 2003. *Finite State Morphology: Xerox Tools and Techniques.* CSLI Publications

Chen H., Yang C., Lin Y. 2003. *Learning Formulation and Transformation Rules for Multilingual Named Entities*. Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition. Vol. 15 Pgs. 1–8.

Kukich K., 1992. *Techniques for automatically correcting word in text. ACM Computing Surveys* Vol. 24 No. 4 377-439

Moore R. C., 2003. *Learning Translations of Named-Entity Phrases from Parallel Corpora*. Proceedings of EACL 2003. Vol. 1 Pgs. 259–266.

Poliquen B., Steinberger R., Ignat C., Temnikova I., Widiger A., Zaghouani W., Žižka J. 2005. *Multilingual person name recognition and transliteration*. CORELA - COgnition, REpresentation, LAnguage, Poitiers, France, CERLICO. ISSN 1638-5748, 2005, vol. 3/3, no. 2, pp. 115-123.

Reeder F. 2001. *The Naming of Things and the Confusion of Tongues*. MT Evaluation: Who Did What To Whom Workshop on MT Summit VIII. Pgs. 55–59.

Richman A. E., Schone P. 2008. *Mining Wiki Resources for Multilingual Named Entity Recognition* Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, 2008. Pgs. 1–9.

Sorg P., Cimiano T. 2008. *Enriching the Crosslingual Link Structure of Wikipedia - A Classification Based Approach*. Proceedings of the AAII 2008, Workshop on Wikipedia and Artificial Intelligence, 2008.

Sproat R., Tao T., Zhai C. 2006. *Named Entity Translation with Comparable Corpora*. Proceedings of the 21st International Conference on Computational Linguistic and 44th Annual Meeting of the ACL 2006. Pgs. 73–80.

Tao T., Yoon S., Fister A., Sproat R., Zhai C. 2006. *Unsupervised Named Entity Translation Using Temporal and Phonetic Correlation*. Proceedings of the 2006 EMNLP. Pgs. 250–257.

Ihardetsi: A Question Answering system for Basque built on reused linguistic processors

Iñaki Alegria, Olatz Ansa, Xabier Arregi, Arantxa Otegi, Ander Soraluze
IXA Group. University of the Basque Country
xabier.arregi@ehu.es

Abstract: This paper presents *Ihardetsi*, a question answering system oriented to Basque. We describe the main architecture of the system, paying special attention to the use of linguistic resources and tools. The system has been built reusing such tools, on the basis that general linguistic processors can be adapted to satisfy the requirements of the question answering task. This methodology can be suitable for other projects, specially when lesser resourced languages are involved. Along with the description of the system, we outline its performance presenting some experiments and the obtained results.

Keywords: QA, Question Answering, Basque

1 Introduction

Question answering systems tackle the task of finding a precise and concrete answer for a natural language question on a document collection.

This task involves the use and adaptation of IR (Information Retrieval) and NLP (Natural Language Processing) resources, techniques and tools.

The current version of *Ihardetsi*,[1] a Basque question answering system, takes Basque questions as input and the corpora on which the answers are searched are written in Basque too.

The system incorporates tools and resources developed previously in the IXA group, like a lemmatizer/tagger, *Morfeus* (Ezeiza et al., 1998), and a recognizer and classifier of named entities (NERC) for Basque, *Eihera* (Alegria et al., 2004). Additionally the Basque Wordnet (Agirre et al., 2006) has been used in order to improve the results of the system.

The remainder of the paper is organized as follows. Section 2 is devoted to introduce the general architecture of the system. In section 3 we describe the main modules of the QA system. Then, in section 4 it is explained how the tool has been adapted to a new domain and to a multilingual environment. Finally, evaluation issues are discussed and some conclusions and suggestions for future research are pointed.

2 General Architecture

The principles of versatility and adaptability have guided the development of the system. It is based on web services, integrated by the SOAP (Simple Object Access Protocol)[2] communication protocol. As we have already remarked, some tools previously developed in the IXA group are reused as autonomous web services, and the QA system becomes a client that calls these services when it needs them. This distributed model allows to parametrize the linguistic tools, and to adjust the behaviour of the system during the development and testing phases.

The communication between the web services is done using XML documents. This model has been adopted by some other systems (Tomás et al. 2005, Hiyakumoto 2004).

The global features of each run are described in a XML configuration file. The set of features is divided into two categories:

1. General requirements. It includes specifications such as the corpus to be used, the processing model of the corpus,

[1] The name of the system, *Ihardetsi,* comes from a Basque word, generally used in the North dialect, which means "to answer".

[2] *www.w3.org/TR/soap/*

the location of the list of questions to be answered, and the description of the type of questions.

2. Descriptors of the QA process itself. This subset of features represents the characteristics of the answering process. Mainly, it determines which modules act during the answering process, describes them and specifies the parameters of each module. In that way, the process is controlled by means of the configuration file, and different processing options, techniques, and resources can be easily activated/deactivated and adapted. These descriptors constitute the documentation support of the system.

As it is common in the question answering systems *Ihardetsi* is founded on three main modules: the question analysis module, the passage retrieval module and the answer extraction module.

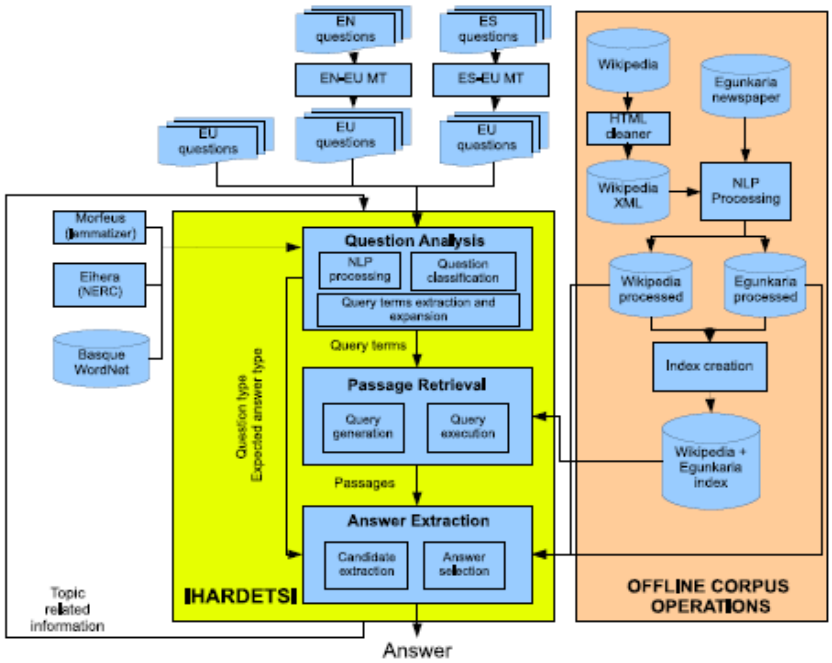


Figure 1: General architecture of the system.

3 Main modules of the system

The main modules in Fig. 1 are explained in the next subsections. Some of the off-line corpus operations are explained beside the Passage Retrieval section.

3.1 Question Analysis

The main goal of this module is to analyse the question and to generate the information needed for the next tasks. On the one hand, a set of search terms are extracted for the passage retrieval module (see section 3.2); on the other, the question type (*factoid*, *list* or *definition* mainly) and the expected answer type, along with some lexical information is passed to the answer extraction module.

The question analyser performs the following steps:

1. *Linguistic processing of the question*: The question analysis reuses a set of general purpose tools like the lemmatizer/tagger named *Morfeus*, and the NERC processor named *Eihera*.
2. *Question classification*: For identifying the main features of a question we attend to the question type, the question focus and the expected answer type. This process is carried out by means of a set of pattern rules that have been defined according to the Basque questions' structure.

   The taxonomy of answer types in our system distinguishes the following classes: *person*, *organisation*, *description*, *location*, *quantity*, *time*, *entity* and *other*. The assignment of a class to the analysed question is performed using the interrogative word, some heuristics of syntactic nature and the type of the question focus. The question focus is mapped with the semantic file characteristic of the Basque WordNet, and in this way we can determine more precisely the expected answer type.
3. *Extraction and expansion of query terms*: All nouns, verbs, adjectives and abbreviations of the question constitute the set of search terms. They are lemmatized and arranged in descending order by their *Inverse Document Frequency* (IDF)[3] value in the corpora.

[3] *Inverse Document Frequency*, a factor in the TF-IDF principle (http://en.wikipedia.org/wiki/Tf-idf)

Optionally, the search terms can be expanded using synonymy, hyponymy and hypernymy information. This expansion is carried out just in one level, without applying any word-sense disambiguation. The synonyms, hyponyms and/or hypernyms are selected by reusing a service that consults the *Basque WordNet* lexical-semantic database. This resource is the Basque version of *EuroWordNet*, and is integrated in the *Multilingual Central Repository* (MCR), which is a multilingual lexical database developed in the *Meaning* project (Atserias, 2004).

3.2 Passage retrieval

Basically an information retrieval task is performed, but in this case the retrieved units are passages and not entire documents.

This module receives as input the selected query terms and produces a set of queries that are passed to a search engine. We have tested different search engines, like *Swish-e*[4], *Jirs*[5] and *Indri* (Strohman et al., 2005). As long as none of them is equipped to deal with Basque, we need to process the corpus before indexing it, as will be explained in the following paragraph.

Since Basque is an agglutinative language, a given lemma takes many different word forms, depending on the case (genitive, locative, etc.) or the number (singular, plural, indefinite) for nouns and adjectives, and the person (me, he, etc.) and the tense (present, past, etc.) for verbs. For example, the lemma *lan* ("work") forms the inflections *lana* ("the work"), *lanak* ("works" or "the works"), *lanari* ("to the work"), *lanei* ("to the works"), *lanaren* ("of the work"), *lanen* ("of the works"), etc. This means that looking only for the given exact word, or the word plus an "s" for the plural, is not enough for Basque. And the use of wildcards, which some search engines allow, is not an adequate solution, as it can be returned occurrences of not only inflections of the lemma, but also derivatives, unrelated words, etc. For example, looking for *lan** would also return all the forms of the words *lanabes* ("tool"), *lanbide* ("job"), *lanbro* ("fog"), and many more. So, a stemmer or a lemmatizer/tagger is almost indispensable for this kind of languages. *Snowball*[6] can be a good option for stemming for some languages since it is open-source. We made an attempt for Basque, but it was not successful.

In our case, the entire document collection was lemmatized/tagged with part-of-speech and named entities. In this task we reused the lemmatizer/tagger named *Morfeus*, that returns only one lemma and one part-of-speech for each lexical unit. The NERC processor, *Eihera*, captures entities such as *person*, *organization* and *location*. The numerical and temporal expressions are captured by the lemmatizer/ tagger. Up to date, no semantic pre-processing has been performed.

3.3 Answer extraction

In this module two tasks are performed in sequence: the candidate extraction and the answer selection. Basically, the candidate extraction consists of extracting all the candidate answers from the retrieved passages, and the answer selection consists of choosing the best answers among the considered as candidates.

Candidate extraction: Firstly, all candidate answers are detected from each retrieved passage and a set of windows are defined around them. The selected window for each candidate answer is the smallest one which has all the query terms.

In order to extract the candidate answers the system addresses each question type in a different manner, as follows:

- Question type is *factoid*:[7] the answer selection depends on the named entities.
- Question type is *definition*: a set of rules have been defined to extract definitions from retrieved text passages.
- Question type is *list*: we followed a heuristic looking for lists of candidate answers in the same passage.

Answer selection: In order to select the best answers from the set of candidates, the same answers that appear in different passages must

[4] http://swish-e.org/

[5] http://sourceforge.net/projects/jirs/

[6] http://snowball.tartarus.org/

[7] As far as we know, there is no formal definition of *factoid question*. Intuitively, a *factoid* is asking about a simple *fact* or relationship, and the answer is easily expressed, usually by means of a named entity.

be combined. We try to map as identical those answers that refer to the same entity. For instance, "Miguel Indurain" and "Indurain" are different strings, but the system must detect that both refer to the same person. This point is relevant when assigning weights to the candidate answers. The formula used to compute the final score of each answer is as follows:

$$S(C) = \frac{\sum_{i=1}^{p} w_i}{N}$$

where *S(C)* is the score of the candidate *C*, *p* is the number of answers identical to *C*, w_i is the confidence score of the *i*-th identical answer, and *N* is the number of all candidate answers.

4 Applications

In this section we present several scenarios where *Ihardetsi* has been used and tested. The nature and the conditions of these scenarios are quite heterogeneous. The evaluation of some of them is reported in the next section.

The first version of the tool was carried out using a news corpus (*Euskaldunon Egunkaria*), which was previously processed in a project on IR.[8] *Euskaldunon Egunkaria* was the only Basque language newspaper in the world for years, but it was closed in 2003. We use a corpora from 2000, 2001 and 2002 years, with about 7 million words.

In 2008 we were involved in the CLEF-QA 2008 evaluation,[9] using the previous corpus and the Basque Wikipedia as evidence for the questions. The architecture of the system remained the same, and only a small change was introduced when the Wikipedia was preprocessed: the headword of an entry was inserted at the beginning of every paragraph of the entry, with the aim of considering all the paragraphs when answering questions about such entry. We participated in the cross-lingual QA evaluation too, answering questions in Spanish and English based on the Basque repository. We just translated the questions by reusing the *Matxin* technology (Alegria et al., 2007) with a small adaptation in order to improve the translation of the questions. Taking into account the specific structure of some questions, we performed an automatic post-edition (based on regular expressions) process for repairing some translations.

After, during the *Anhitz* project we had to adapt *Ihardetsi* to new features: cross-language, in the same way as in CLEF, new domain, science and technology, and multimodal process, question were processed by a speech recognizer.

The aim of the *AnHitz* project (Arrieta et al., 2008), whose participants are research groups with very different backgrounds, is to carry out research into language, speech and visual technologies for Basque. Several resources, tools and applications were integrated into a prototype of a 3D virtual expert on science and technology. It includes our QA system.

5 Evaluation

We have combined quantitative evaluations (mainly those that are framed in the CLEF campaigns), with more qualitative ones (specifically we tackle them in the *Anhitz* project).

5.1 CLEF 2008

This section describes the results we obtained in our first participation in the CLEF 2008 campaign, specifically in the Basque to Basque monolingual QA task (Ansa et al., 2008).

The exercise of the main QA task consisted of topic-related questions, i.e. clusters of questions which were related to the same topic and contained co-references between one question and the others. Moreover, besides the usual news collections provided by ELRA/ELDA, articles from Wikipedia were considered as an answer source. Some questions could have answers only in one collection, i.e. either only in the news corpus or in Wikipedia.

A Basque corpus and a set of Basque questions were offered for the first time in this edition.

The methodology we employed targeted precision at the expense of recall, therefore we always choose NIL answers for those questions

[8] HERMES project: News databases. cross-lingual information retrieval and semantic extraction (TIC-2000-0335-C03-03), founded by the Spanish Government.

[9] www.clef-campaign.org/2008/working_notes/

we could not reliably locate a candidate answer in the retrieved passage. Table 1 illustrates the results achieved by *Ihardetsi* in the monolingual run.

It is clear that the best results were achieved for factoid questions. This is due to the fact that we focused on this type of questions in the development of the system. A set of 145 factoid questions was processed and, taking into account the first three answers, we obtained the following results: 50 questions had a correct or inexact[10] answer in the proposed three answers, 22 had a wrong NIL answer[11] and 73 had a wrong answer. Analysing these 73 questions we detected that for 17 the correct passage was detected but the system did not extract the correct answer.

	R	W	I	ACC
OVERALL	26	163	11	13.0%
FACTOID	23	113	9	15.9%
DEFINITION	3	36	0	7.7%
LIST	0	14	2	0.0%

(R: Right, W: Wrong, I: Inexact, ACC: Accuracy)

Table 1: Results for the first answer obtained in the monolingual run (Ansa et al., 2008).

There are not correct answers for LIST questions because at the time of sending the runs we had not yet implemented the heuristics for answering such questions.

The system answered NIL for 57 questions but only 4 of them were correct. Analysing the reasons for this we can group them in 5 groups:

- The expected answer type detection failed: 6 questions.
- No passage was retrieved: 14 questions
- The passage had the answer but the system could not extract the answer: 13 question
- Retrieved passage had not the answer: 16 questions
- Some other reasons: 4 questions

It is remarkable that no other system took part in the Basque target task, so the obtained results could not be directly compared with another Basque system. Nonetheless, it is interesting to contrast our results with some other languages. For that purpose, we choose QA@CLEF 2007 (Giampiccolo et al., 2008) results as a reference because that was the first time that topic-related questions and the Wikipedia corpus were included. Although our results are far from the best ones, with overall accuracy of 54%, we realized that almost 40% of all the runs got worse results than those of our system.

5.2 The *Anhitz* Project

The *Ihardetsi* system, included as the QA functionality in the demo prototype developed in *AnHitz,* has been evaluated in order to measure its performance and weigh the impression of potential users about it. A group of 50 users formulated 3 questions and 3 cross-lingual searches each, making 300 tests in total. During the interaction of the testers with the system, some objective observations were noted down, such as the number of failures and successes of the QA system.

As it can be observed in Table 2, *Ihardetsi* answered correctly 30.61% of the times, and in another 15.30% the correct answer was among the first five possible answers given. 54.08% of the times the system did not give a correct solution or did not answer at all. We could not evaluate whether the correct answer was in the corpus or not.

Correct answer	**%**
In the 1st place	30,61
In the 2nd place	8,16
In the 3rd place	1,02
In the 4th place	3,06
In the 5th place	3,06
The right answer was not among the possible answers	36,73
The system did not answer at all	17,35

Table 2: Results for qualitative evaluation in *Anhitz*-QA

[10] An answer is incorrect if it contains less or more information than that required by the query.

[11] A NIL answer can be correct if really the question has no answer in the corpus.

5.3 Addressing cross-linguality

Although the main aim of *Ihardetsi* is to deal with Basque questions and Basque documents, we have carried out some cross-lingual experiments, such as the Spanish-Basque and English-Basque bilingual tasks at QA@CLEF 2008, and the Basque-English task in the RespubliQA exercise at QA@CLEF 2009.

5.3.1 Bilingual tasks at QA@CLEF

Three cross-lingual runs, two for Spanish–Basque and one for English–Basque, have been performed. The aim of the second run for Spanish–Basque was to test if the semantic expansion of the question could compensate the lost of precision in the translation process. The results of the three runs are quite poor. The loss of precision respect to the monolingual system is more than 50% (Ansa et al., 2008).

The main conclusions we want to remark are:

- Very similar results are obtained for the basic Spanish–Basque and for the English–Basque runs (in both there are 11 right answers, 7 right answers in 2nd or 3rd place and 7 inexact in the first place). Due to the better quality of the Spanish–Basque translator we hoped better results for this run.
- Although the results are similar in average, the right results do not correspond always to the same questions. Only five of the eleven right answers are common.
- The semantic expansion in the second run for Spanish-Basque does not achieve better results. A slightly smaller precision is observed, because some right answers are lost. In compensation to this, new right or inexact answers appear, but not in the first place. In view of these figures, one might think that at least a higher number of "passages" are recovered, but it is not true, because the number of recovered "passages" remains at same level (about 40 of 200).

5.3.2 ResPubliQA at QA@CLEF 2009

ResPubliQA[12] has been presented as a new task at CLEF 2009, and it consists of retrieving a passage string (small snippet of text) containing the answer to a question in natural language. *JRC-Acquis*[13] is the reference corpus to generate the questions and search for answers. In this corpus aligned documents are available in Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish.

Although no Basque documents are available, organizers have arranged a Basque–English task, so that, given a pool of Basque questions, systems must retrieve passages from the English document collection. This is a new cross-lingual experience for us, given that Basque is the source language.

Our system analyses the Basque questions, translates and disambiguates the query terms and searches for the passages that are relevant to the query. The techniques that have been used in the translation/disambiguation of the query terms are described in (Saralegi et al., 2008), and the approach for passage retrieval is based on the ideas of (Otegi et al., 2008).

We submitted two runs, but we have not received yet the evaluations.

6 Conclusions

This article shows a general architecture for Question-Answering, where general linguistic processors are reused and integrated. The underlying idea is that QA systems can be built, even for languages with fewer resources, reusing existing linguistic tools.

We present different evaluations of the system, which have been carried out in different scenarios.

In the QA@CLEF 2008 campaign we compared the performance of our system with other monolingual and cross-lingual systems on a heterogeneous document collection (news articles and Wikipedia). Although the results might look not so good, our general conclusion is positive considering that it was our first participation and taking into account the particularities of the Basque language. Moreover, we have been able to identify some of the strengths and weaknesses of each module of the system.

In the context of the *Anhitz* project, our QA system has been integrated and tested in a prototype that has received ample media

[12] http://celct.isti.cnr.it/ResPubliQA/

[13] JRC-Acquis is the total body of European Union (EU) law applicable in the EU Member States.

coverage and has been welcomed by Basque society. The system has been evaluated by 50 users who have completed a total of 300 tests, showing good performance and acceptance. We consider that this kind of qualitative evaluations are quite interesting.

The ResPubliQA experience at the QA@CLEF 2009 campaign has showed us that we can deal with cross-lingual tasks even if the target language is not Basque.

All these experiences constitute the background of future improvements. It would be useful to apply other techniques, such as the syntactic pattern matching and the anaphora resolution. As these tools are developed for Basque, *Ihardetsi* will use them.

Acknowledgements

This research was supported in part by the Spanish Ministry of Education and Science (Know TIN2006-15049-C03-01) and the Basque Government (AnHITZ 2006IE06-185).

References

Agirre E., Aldezabal I., Etxeberria J., Iruskieta M., Izagirre E., Mendizabal K., Pociello E.. Improving the Basque WordNet by corpus annotation. Proceedings of Third International WordNet Conference. pp. 287-290. Jeju Island (Korea). 2006.

Alegria I., Arregi O., Balza I., Ezeiza N., Fernandez I., and Urizar R.. Development of a Named Entity Recognizer for an Agglutinative Language. In IJCNLP, 2004.

Alegria I., Diaz de Ilarraza A., Labaka G., Lersundi M., Mayor A., and Sarasola K.. Transfer-based MT from Spanish into Basque: reusability, standardization and open source. LNCS. Springer. Vol. 4394/2009. pp. 374-384. 2007.

Ansa O., Arregi X., Otegi A., Soraluze A.. Ihardetsi question answering system at QA@CLEF 2008. Working Notes of the Cross-Lingual Evaluation Forum, Aarhus, Denmark. 2008.

Arrieta K., Diaz de Ilarraza A., Hernáez I., Iturraspe U., Leturia I., Navas E., Sarasola K.. AnHitz, development and integration of language, speech and visual technologies for Basque. Second International Symposium on Universal Communication Osaka. pp.338-343. 2008.

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., and Vossen P. The MEANING Multilingual Central Repository. In Proc. of the 2nd Global WordNet Conference, pp. 23-30. 2004.

Bilotti M. Query Expansion Techniques for Question Answering. Master's thesis, Massachusetts Institute of technology, 2004.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., and Urizar R.. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In COLING-ACL, pp.380–384, 1998.

Giampiccolo G., Herrera J., Peñas A., Ayache C., Forascu C., Jijkoun V., Osenova P., Rocha P., Sacaleanu B., and Sutcliffe R. Overview of the CLEF 2007 Multilingual Question Answering Track. LNCS. Springer. Volume 5152/2008. pp. 200-236. 2008.

Hiyakumoto L. S. Planning in the JAVELIN QA System. In CMU-CS-04-132, 2004.

Otegi A., Agirre E., Rigau G.. IXA at CLEF 2008 Robust-WSD Task: using Word Sense Disambiguation for (Cross Lingual) Information Retrieval. Working Notes of the Cross-Lingual Evaluation Forum, Aarhus. 2008.

Saralegi X., San Vicente I., Gurrutxaga A.. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain. 6th International Conference on Language Resources and Evaluations (LREC) - Building and using Comparable Corpora workshop. Marrakech. 2008.

Strohman H. T., Metzler D. and Croft W.B. Indri: A language model-based search engine for complex queries. Proceedings of the International Conference on Intelligence Analysis. Poster. 2005.

Tomás D., Vicedo J.L., Saiz M., and Izquierdo R.. Building an XML framework for Question Answering. Working Notes of the Cross-Lingual Evaluation Forum. Alacant. 2005.

Babelium Project. Promoting the Use and Learning of Minority Languages

Juan Antonio Pereira
Dept. of Computer Languages and Systems
University of the Basque Country
Pº Manuel de Lardizabal, 1, 20018 San Sebastian
juanan.pereira@ehu.es

Silvia Sanz
Dept. of Computer Languages and Systems
University of the Basque Country
Pº Manuel de Lardizabal, 1, 20018 San Sebastian
silvia.sanz@ehu.es

Julián Gutiérrez
Dept. of Computer Languages and Systems
University of the Basque Country
Pº Manuel de Lardizabal, 1, 20018 San Sebastian
julian.gutierrez@ehu.es

Abstract: Babelium Project is a new collaborative e-learning system for practicing speaking second languages that helps people who want to learn minority languages. The multimedia method this system offers tries to break barriers like timetable and economic problems and also difficulties for finding people for practicing speaking. So, there is no matter where the language is spoken or how many people master it. Now language students can learn new languages from their home with free software, with their own PC, without additional software in a multimedia environment.
Keywords: language learning, speaking practice, minority languages, collaborative environment, free software, multimedia

1 Introduction

The need and interest for learning new languages is not a new issue. There are multiple reasons for that including travelling, career development or just interest in knowing new cultures. Millions of people around the world are studying other languages and this number is increasing continuously.

Generally, although there are still some problems we will comment later, it is quite easy to study popular or extended languages as English, French, German, Italian, Spanish and others. But, when it comes to minority languages, it is more difficult because of the lack of resources.

The number of people and researchers working in the field of minority language is increasing continuously. Thanks to them, there is possible to find some resources that allow people to learn these languages. Nevertheless, there is still a lot of problems to practice speaking due to the fact that these languages are spoken by a limited number of users and also because they use to be concentrated in very concrete geographical areas.

E-learning technologies can be a solution for bringing people closer. The existing e-learning systems for language learning usually offer resources about grammar, vocabulary, structures for improving writing and so on. But generally, when people are studying languages, their final aim is to communicate with other people. So, theoretical studying must be complemented with practice. Speaking is the key to communication and it is recognized as critical skill, both by teachers and by learners (Cunningham, 1999). Besides, speech has its own skills, structures, and conventions different from written language (Burns & Joyce, 1997; Carter & McCarthy, 1995; Cohen, 1996). Therefore, speaking practice is very necessary, but existing systems only offer to establish a meeting with a teacher or a native person for

carrying out these sessions. This aspect breaks the flexibility of e-learning systems about studying anywhere at anytime. In addition, in the case of minority languages, the number of people available for speaking comes down drastically.

This paper describes a new e-learning system to help people to practice speaking with flexibility (anywhere at any time), using their existing installed software, in a collaborative learning environment and talking about their favourite themes or the issues they need for developing their job or hobbies. Section 2 puts in context the current solutions for foreign language speaking practice and comments briefly our proposal. Section 3 explains in detail the functionality of the Babelium Project, showing the foundations of the project and the current development status. Finally, some conclusions and future work are given.

2 Context Situation and Proposal

Nowadays there are two main possibilities to practice language (speaking), both in language schools and with e-learning systems: (i) offline conversations, that is face to face discussions; and (ii) online conversations, like videoconferences, audio chats or similar.

Offline interpersonal conversations consist mainly in a meeting of two or more people, the one that wants to learn a language and the one that have that language as native (or advanced knowledge about it). Although it is possible to establish this type of meetings, either with an academy teacher or with a non-professional native person, there are still some problems: as economic difficulties for paying for that service (more serious if we think in an ideal situation of lifelong learning), timetable problems due to job, family or motivation aspects and difficulties like going to the country/zone where the language is spoken or the number of people available for practicing minority languages.

Online interpersonal conversations imply the use of a computer and Internet to hold a virtual meeting, either with video and audio (videoconference) or just voice. There are two possible methods: (1) conversation between students (for example a Basque student learning Irish talking to an Irish student learning Basque), in which case there must be an agreement between both students to practice in both languages, making corrections each other; and (2) student-teacher conversation, where students contract a professional teacher for practicing speaking with him. Once again, the second method presents similar problems as in the offline conversations. An additional limitation in the first case is that there is no script or methodology to establish a conversation.

So, the objective is to practice speaking in real-life situations but without some limitations that come from engaging in face to face conversations. The idea would be to practice speaking with someone with no time problems. What about using a computer as speaker? Is it possible? Nowadays, voice recognition and artificial intelligence are research areas that are not developed enough to have a real human-computer conversation. So, we propose an intermediate solution between this approach and face to face one: an e-learning system to practice speaking in real-life situations, but using a computer as speaker, using videos from different sources with the aim of having human-computer 'conversations'.

At present there are several technical, social and economic conditions that makes possible the development and exploitation of a free software system for live long language learning in a collaborative environment with interactive multimedia. The next sections explain in detail these conditions and the modules and functionality of the system.

3 Babelium Project

The main and simple idea is that minority languages are also mother tongues of some people. So, these people master those languages. Even more, each one has different knowledge due to they live in a particular place, works in a specific area and has their own hobbies. So, the amalgam that makes up one person language (mother tongue) knowledge skills is different for other person with the same mother tongue.

The other important concept for the Babelium Project development is the collaborative environment idea. We think there is a great potential from people all around the world to contribute and collaborate with their own language knowledge (mother tongue) for improving their knowledge in other languages. An essential idea is that collaborative learning should be carried out in the web with free software so as to allowe every developer to improve it collaboratively too. We want the

Babelium Project not as a commercial enterprise product but as a system for a user community that manages and improves it continuously.

Bearing in mind all of this key aspects, the Babelium Project tries to offer minority language students the possibility to improve their speaking abilities in a comfortable way, without moving or setting up a meeting with any native person and with no need to be afraid of pronunciation and understanding problems. Even more, this system offers users the possibility of selecting their favourite topics and/or real-life situations they need to practice or improve.

The project is based in the use of multimedia videos with real-life conversations between two or more people. Users/students can watch videos in native language with subtitles (in various languages). The objective is to take the role of one of these people in the video and practice speaking when their turn comes. Users can record their conversations and publish them in order to be evaluated. The evaluation is carried out collaboratively by other users whose native language is the one spoken in the video. These referees must assess the performance of the student. Also it is possible to send feedback to the student adding commentaries about the assessment. This is the basic loop for learning to speak. Users collaborate not only in the evaluation phase but also uploading new videos and creating the subtitles for those videos, which in the case of non-native users is another way of learning that implies listening and writing.

3.1 How the Babelium Project Works

In order to explain more precisely all the functionality and roles of the system next paragraphs describes, step by step, how the Babelium Project works and how users collaborate for improving speaking in languages learning.

1. Using the Compiler Video component, users upload interesting videos with short real-life conversations between two or more people.
2. Native users (or those that master the language spoken in the video) add metadata and subtitles to the video using the Label component. Metadata consist on: difficulty level, keywords, roles, and so on. Subtitles are added for each role (people who talk in the conversation) and for each moment (from second X to second Y) that role talks. This work is published for helping the rest of the community to learn.
3. A user that wants to practice speaking a second language looks for videos using some keywords he is interested on. The system returns a list with those videos that fulfill his query. This step is carried on by the Video and Recording components.
4. This user can watch the selected video including metadata and subtitles created for it by other users. When various subtitles are available for the same conversation moment, the interfaces shows the most voted one. All this process is inside the Recording component.
5. The next step is to select a role (person) of the video conversation. This is the role that is going to be played by the user who wants to practice speaking.
6. Video starts reproducing. When the selected role turn comes (it is marked with a red arrow in the video timeline), audio channel stops and video (image) channel continues reproducing. This is the point where user must start talking as if he were that person. The system will record his speaking. This process is carried out every time that role takes part in the conversation. If user wants, he has the support of subtitles (like in a karaoke).
7. At this point, the user can watch the original video again, watch both the original video and his recording, watch only the recording video, record the same video again (with the same role or another one) or publish it for evaluation phase. The last 3 steps are carried on by the Recording component.
8. The evaluation phase is developed collaboratively by a number of referees, who are system users that are native or master in the language spoken in the video. The evaluation task consist on reviewing a speaking exercise (watching both the original and the recorder videos) and score it as good, regular or bad. Also, and may be more important, it is possible to send feedback to the speaker with comments about pronunciation, vocabulary or whatever helps a referee considers for speaker to improve or just for congratulating

him for the performance. This process is developed using the Evaluation component.

This is a collaborative environment where people work to help other users and also benefit from the work of these users. This project is being developed basing on the power of collaborative environments but also in the strength of motivation. Mainly this last idea is the reason for creating a credit system that we call Karma. The operation is simple: users who collaborate (in any way) with the system receive Karma and users that consume any resource of the system pay with Karma. This is an ordinary manner of supporting collaborative environments avoiding user abuses, both in virtual communities but also in real-life communities (time banking). In the Babelium Project, users receive Karma by uploading videos, labelling them, editing subtitles and evaluating other users, and consume Karma when they want to be evaluated. The Credit component manages all this process.

3.2 Why the Babelium Project can Work

Some years ago this project could be impossible to develop, but nowadays there is a mixture of conditions, including technical, social and economic possibilities that makes this moment perfect to develop it. The next paragraphs describe the key factors that support the project:

- Videos with real conversations. There are several web pages that store and distribute videos, and the very well-known and most important one is YouTube (http://www.youtube.com). Millions of users (non-professional) download and upload thousands of videos per day. Even, other similar web pages have emerged with the same objective and relatively great success: Blip.TV (http://blip.tv/), Metacafe (http://www.metacafe.com/), Hulu (http://www.hulu.com/), imeem (http://www.imeem.com/), revver (http://revver.com/) or archive.org (with permission licenses) (http://www.archive.org/index.php). There is also another huge category of videos useful for our purpose: advertisement videos and movie trailers.
- Open Licenses. Over the last years, there have being an increase in the amount of free and open digital resources licenses. From these ones, Creative Commons licenses (by sa) are the most important for this project due to it permissiveness in reproduction of videos without any royalties.
- Subtitles. Deaf people have fought hard to preserve their rights. And their efforts have being rewarded because a big amount of multimedia resources (like movies, TV programs or advertisements) include subtitles today.
- Technological issues. Some years ago it was impossible for an ordinary person to seat at home and to maintain a videoconference with webcam and microphone or simply watch a video through internet due to technical constraints: no broadband connections. But today the technological revolution makes these situations possible: broadband is available to a growing number of users and also users are able to have a complete multimedia experience through internet using only a navigator and free software.
- Collaborative environments. There have been various studies and researches that emphasize the goodness of the *social constructivism* (Vygotsky, 1978) learning method. Nowadays, with the expansion of Internet, collaborative philosophy is growing within web users around the world. There are different types of collaborative communities (like Facebook (http://www.facebook.com/), Wikipedia (http://www.wikipedia.org/) or more concretely sites for language learning like Palabea (http://www.palabea.net/), Livemocha (http://www.livemocha.com/)or yappr (http://es.yappr.com/)) that grow day by day.

3.3 Current Development State

At present, the Babelium Project is a prototype that let users to practice Basque, English and Spanish speaking. All the components that take part in the architecture explained above are being developed, except the Credit and Statistics components. Both are under development due to we are still thinking about Karma management and help statistics for users and also for administrations purpose.

The system is being developed using an architecture based on some known free software tools. Adobe Flex SDK for the presentation layer. It deals also with all the multimedia resources (mainly video displaying and recording, audio synchronization, and user provided data management). Using Flex SDK we can develop easily a Flash based multiplatform compatible video powered web site. Other major component is Red5 (an open-source Flash media server) that we use to store

the videos that users generate. When one user records his video-exercise, that video is saved in Red5. In the same way, when a referee-user wants to watch a video-exercise to assess its accuracy, the video is streamed through Red5. Finally, an Apache Web Server is responsible for storing and serving static HTML content and executing dynamic PHP scripts (used for access and manage MySQL database connections)

4 Conclusions and Future Work

As commented in the introduction section, speaking is a fundamental skill when learning a foreign language. It is the key for communication, but everybody has experienced a lack of fluency when trying to speak in a foreign language, moreover when these speaking moments take place once in a while. In the case of minority languages, the opportunities for practicing speaking can be reduced drastically, so the lack of fluency must be a very serious problem. There are several reasons for that, including economic problems, lack of time and specifically difficulties for finding people to practice. More over, there is another remarkable problem: people are afraid of speaking in a language in which they are not fluent in. This is a serious fact in the sense that if you are afraid of speaking you do not practice and if you do not practice you do not improve your abilities; so, the fear loop starts again.

The key concept is to mix two important aspects that are currently happening: a very important trend toward the use of social networks and a common necessity/wish to speak other languages. Collaborative environments like Wikipedia, in the general culture area, and also language learning systems like Livemocha or Palabea with more than 200.000 users have demonstrate that collaborative communities for exchanging knowledge works properly and that people are willing to enter and work into these environments. Adding the fact that each person has a deep knowledge about his mother tongue, there is a perfect combination to exchange language knowledge/language learning among people around the world.

The Babelium Project tries to cover the gap of minority language speaking practice taking advantage of the current technical and social development moment and offering speaking practice without difficulties for finding people to practice, time, money or fear limitations. It offers a new learning method based on well-known web environment, but using innovative video treatment techniques that improve the existing solutions for practicing speaking. The great amount of videos (with or without subtitles) uploaded in Internet (like in YouTube, Metacafe, arvhives.org, etc.), all the resources with Creative Commons licences and the simplicity for common people to access to broadband Internet, makes this project to be a reality. Moreover, the Babelium Project is a free software tool for people to practice foreign language speaking in a collaborative web environment, obtaining help from other people and helping those people to learn their mother tongue.

Among other advantages commented before, we consider video a very good media resource for practicing speaking due to the fact that speech is not always unpredictable. Patterns that tend to recur in certain discourse situations (e.g. asking for the time, paying something in a shop, requesting help to go to a specific address), can be identified and charted (Burns & Joyce, 1997). Speaking requires that learners not only know how to produce specific points of language such as grammar, pronunciation, or vocabulary *(linguistic competence)*, but also that they understand when, why, and in what ways to produce language *(sociolinguistic competence)* (Cunningham, 1999). Furthermore, videos of real-life situations allow learners to see facial expressions and body language at the same time as they hear the stress, intonation, and rhythm of the language (Bello, 1999).

Although the Babelium Project is still a prototype for Basque speaking practice we expect to be working on next July, we are working inside various projects to expand Babelium with new languages and functionalities. Another step ahead is to automate some process that currently must be done by users but their execution does not involve any learning; for example, recognition of conversations inside a video to upload it to the system, recognition of different roles in a conversation in order to separate them for subtitles edition help or voice recognition to support the subtitles creation. All of these are complex task of machine learning that require a deep research so we expect to work on in a later future.

Bibliografía

Bello, T. 1999. New avenues to choosing and using videos. *TESOL Matters,* 9(4): 1-20.

Burns, A., & H. Joyce. 1997. Focus on speaking. Sydney: National Center for English Language Teaching and Research, Macquarie University.

Carter, R., & M. McCarthy. 1995. Grammar and spoken language. *Applied Linguistics,* 16 (2): 141-158.

Cohen, A. 1996. Developing the ability to perform speech acts. *Studies in Second Language Acquisition,* 18 (2): 253-267.

Cunningham, M. 1999. Improving Adult English Language Learners' Speaking Skills. Washington, DC: National Center for ESL Literacy Education.

Vygotsky, L.. 1978. Mind in Society. London: Harvard University Press.

A web-based system for multilingual school reports

David Chan
Canolfan Bedwyr
Bangor University
Safle'r Normal
Bangor, Gwynedd
LL57 2PX UK
d.chan@bangor.ac.uk

Dewi Jones
Canolfan Bedwyr
Bangor University
Safle'r Normal
Bangor, Gwynedd
LL57 2PX UK
d.b.jones@bangor.ac.uk

Oggy East
Semantise Ltd
Unit 4, Coed y Parc Ind. Est.
Bethesda, Gwynedd
LL57 4YY UK
oggy@semantise.com

Abstract: Schools in Wales have a statutory duty to produce bilingual Welsh/English school reports. Since computerization of the reporting process, this has mainly been achieved with English-only reporting software, using a labour-intensive and error-prone process of disparate add-ons and ad-hoc post-processing to work round the underlying monolingual data model. In the scope of this project a reporting tool, osisADRODD, with a fully multilingual data model was designed from scratch to support the various bilingual scenarios operating in the schools of Wales. A multilingual co-ordinated comment bank system was designed with support for gender-dependent pronoun substitution and agreement in any language. A general library for language-specific text proofing within an HTML form was implemented and incorporated into the system, and a grammar checker and spell checker were included for Welsh and English respectively.

Keywords: School report, Multilingual data model, comment bank, text proofing, html

1 School Reporting in Wales

1.1 Welsh Language Act

The Welsh Language Act (United Kingdom 1993) requires all state-run schools in Wales to treat Welsh and English "on a basis of equality [...] so far as is both appropriate in the circumstances and reasonably practicable".

1.2 Language use in schools

Both Welsh and English are used in every school in Wales. The distribution of both the medium of instruction and the linguistic proficiency of the teachers will vary depending on the school.

Irrespective of these factors, the Welsh Language Act requires school reports to be provided in Welsh and/or English according to parental preference.

1.3 Monolingual systems in use

Since the introduction of Student Information Systems (SISs) into education in Wales, schools have generally been using reporting systems designed for a monolingual environment. These typically have a user interface and data model which cannot accept and store parallel bilingual text.

In many cases, the monolingual SISs have been customised with extensions which accept text in an alternate language and store it in a duplicate record or a separate database table. Because the core system is unaware of the extension there is usually a risk of corrupting data or destroying the interlanguage correspondence, which is managed by painstaking care and manual verification.

1.3.1 Comment Banks

Comment banks (ready-made, pre-entered sentences of general utility) reduce the effort required to write reports. However, the monolingual SISs do not generally support the maintenence of parallel bilingual comment banks. This is in some cases addressed by institutions manually creating an alternate language version of the comment database with parallel comments stored using the same internal database record ID as in the main language record. Naturally, this

correspondence is fragile and manual care is required when editing either data set.

Furthermore, comment bank systems generally have a language-specific pronoun substitution mechanism; for example, a comment such as "^ does ~ homework" may be rendered as "Adam does his homework" or "Eve does her homework" depending on the student's name and gender. Major international software vendors have generally declined to support Welsh-language pronouns due to the small size of the market, notwithstanding the simplicity of such textual subsitutions. It therefore seems there is little prospect of support for more complex variations, such as the requirement in Welsh to vary the word for "homework" to agree with the gender of the student.

In some cases institutions have worked round these limitations by paraphrasing comment bank entries to avoid the use of pronouns or gender agreement altogether; the disadvantage is a final report containing unnatural, forced sentences which sound excessively formal and are harder to understand.

1.3.2 Post processing

Due to the limitations of the authoring systems, most institutions perform some amount of post-processing, whereby the system outputs word processor documents which are then modified manually. This compensates for the missing flexibility in the main system, but at the expense of losing the convenience, data integrity and uniformity which is the advantage of a reporting system.

1.4 Lack of market power

As indicated above, multinational vendors have proven unwilling to customise their software to support the market in Wales due to its comparatively small size. There are, nevertheless, nearly 2000 schools in Wales (Welsh Assembly Government 2007) and therefore it might be expected that ample funding exists to pay for the required customisations, if implemented for example by small to medium-sized enterprises, which may have lower overheads and greater flexibility than the international software vendors which dominate the market for SIS software.

The feasibility of such a solution depends on an institution's chosen SIS being sufficiently modular or open that a third party extension can interface with it.

2 Linguistic use cases for a reporting system

In order to be useful to schools with varying linguistic profiles, a reporting system must support a variety of usage scenarios. Teachers may author in Welsh and English simultaneously, or they may write content in a primary language which is later translated by themselves or by clerical staff. Indeed, a teacher's proficiency in one of the languages may be non-existent or insufficient to author confidently. Content may be composed from scratch, selected from comment banks or a combination of the two processes. Schools may have differing workflows for creation, editing and approval of content. Staff should be able to depend on convenient access to language proofing tools at any stage in the workflow.

3 osisADRODD: A Multilingual Reporting System

In order to maximise compatibility with existing systems and minimise data exchange issues, the data flow from the school's SIS to the reporting system is one way; that is, the reporting system acts as a consumer of data only. Existing monolingual reporting functionality in the SIS is disregarded, due to insufficient support for a multilingual data model.

The system's user interface is implemented in HTML and Javascript using the Google Web Toolkit (Google 2009), with a SQL database backend. This minimises dependence on a school's particular choice of SIS.

3.1 Comment banks

The data model stores parallel comment bank entries as separate records sharing a database ID, with the comment's language code forming part of the record key. This enables multiple language storage (not just two languages).

Comment bank entries contain metalinguistic strings to enable the specification of name- and gender-dependent text. For example, an entry such as "%studentName% thinks hard about %(his|her)% work" specifies two sets of sentences (male and female). This is more powerful than the substitution mechanism described above, since any gender-

dependent variation can be specified, not just a few language-specific cases hardcoded into the software.

3.2 The content authoring process

The content authoring process follows three stages, as follows.

Firstly, the teacher constructs a sequence of sentences from a comment bank; these are displayed in the teacher's preferred language but the equivalent text can also be constructed in any other language which the comment bank supports. The teacher's selections are saved.

Secondly, the sentences are concatenated into paragraphs of free text, one for each language version, which the teacher can then modify. The modifications are highlighted for clarity. Note that the teacher may have no proficiency in one or more of the languages; in this case, the paragraph versions in those languages are not displayed to the teacher, in accordance with stored preferences.

The modified free text is then saved in each language, including any versions which were not displayed.

Thirdly, the free text with highlighted modifications is recalled, and displayed in each language, in order that an editor or translator can proofread the translation correspondence and make any neccessary changes.

3.3 Language proofing subsystem

Where editable free text fields appear, the user can invoke language proofing tools to check grammar and spelling. A common application programming interface has been defined for passing text in a specified language via an HTTP POST request to a separate proofing server, which then returns a list of spelling and/ or grammar errors with requisite details and suggestions. We have implemented a proofing server conforming to this API which wraps Bangor University's Cysill grammar checker (Prys *et al.* 2002) to proof Welsh language text, in addition to the standard Hunspell library (Németh 2009) for spellchecking English and other languages.

The entire proofing subsystem is independent of the other components of osisADRODD in general, and could in principle be used to provide language proofing tools for virtually any HTML forms. An exciting possibility is to invoke the proofing subsystem via a Javascript bookmarklet (a script which can be run on the content of another webpage at the click of a button). A user could thereby proof text entered into a third party content provider's web form without any action from the content provider to integrate the tools. This may provide an alternative to persuading providers of online services to spend time and resources providing proofing tools for lesser-used languages.

3.4 Security

School reports are inherently confidential personal data and require protection as such. In the original security model, osisADRODD is served from the same security context as the school's SIS, typically running from servers physically located within a school, firewalled and accessible only from a staff intranet. The same authentication credentials used in the SIS are used to restrict users' data access by relevance.

However this model is not generally appropriate for primary schools (for students aged 4-11 years), as these seldom have the budget or expertise to manage application servers internally. Further development is now in progress to provide a hosted solution for such schools; here the system will be accessed across a wide area network, or even directly via the public Internet, and secured in the usual way by a combination of an encrypted HTTP session and digital certificates.

References

United Kingdom. *Welsh Language Act 1993 (c. 38)*, London.

Schools Management Division, Welsh Assembly Government (2007). *Defining schools according to Welsh medium provision*, Cardiff.

Prys D., Hicks B., Jones D. B., Morgan M. (2002), *Cysgliad*, University of Wales, Bangor. http://hdl.handle.net/10242/13190

Németh L. (2009). *Hunspell: open-source spell checking, stemming, morphological analysis and generation*, Hungary, http://hunspell.sourceforge.net .

Google Inc (2009). *Google Web Toolkit*. http://code.google.com/webtoolkit .

The SALT Cymru Feasibility Report and the resulting Special Interest Group

Gruffudd Prys
Bangor University
Bangor, Gwynedd, Wales
g.prys@bangor.ac.uk

Abstract: This paper examines the findings of a 2008 project examining the feasibility of establishing a Special Interest Group (SIG) focusing on Speech and Language Technology (SALT) research in Wales. It will also report on the current progress of the SALT Cymru Special Interest Group, which was established as a direct result of the feasibility study's recommendations.
Keywords: SALT, Wales, Welsh, Minority Language, Speech and Language Technology

SALT Cymru Feasibility Project

Background

The SALT Cymru Feasibility Project was the result of a successful application from Bangor University to the Welsh Assembly Government's EU-funded Knowledge Exchange Programme (KEF). KEF's aim was to stimulate Wales's economy by promoting knowledge transfer between universities and businesses. The study would explore the feasibility of establishing a special interest group that would bring together representatives of academia and industry in Wales who share an interest in speech and language technology. The final report would form the basis of an application for Welsh Assembly administrated European funding for the proposed special interest group.

The project was undertaken by Bangor University's Language Technologies Unit (LTU), which specializes in the development of multilingual language technology, in particular those that serve the needs of the Welsh language as a modern, living language.

Despite the unit's experience with developing Welsh speech and language technology, the SALT Cymru feasibility project and the proposed special interest group were not conceived as a language-specific project nor as a minority language project. Rather, the project's main aim, aligned with the Welsh Assembly's priorities, was to stimulate the Welsh economy by encouraging the participation of Welsh industry in the growing worldwide multilingual SALT market, whatever the language. However, the statutory requirement for Welsh-language provision from public bodies in Wales was also regarded by the LTU as an economic opportunity for business in Wales.

Due to the multidisciplinary nature of SALT, the project team featured a variety of specialists. Delyth Prys, a linguist specializing in terminology, served as the project director and Rhys James Jones, a speech technology specialist, managed the project together with Gruffudd Prys, a linguist and web developer who served as the north Wales Co-ordinator. Additional expertise was provided by a team of software engineers including Dewi Bryn Jones, Ambrose Choy and David Chan, whose experiences include a number of natural language processing technologies.

Defining SALT

A report examining issues to do with SALT naturally requires that SALT be precisely defined. However, as SALT is considered in academia to be a multidisciplinary area, its precise definition is not universally defined. A survey of the definitions of SALT used by various research institutions (see the *SALT Feasibility Report*, *Appendix A*) failed to identify a common description. As the SALT Cymru Feasibility Report (the project's ultimate output) required the creation of a standard definition for SALT for the sake of clarity and consistency, it was decided to follow the definition found in the *Survey of the State of the Art in Human Language Technology* edited by Varile and Zampolli (1997) and use the following:

For the purposes of the project, SALT (speech and language technology) is defined as the inclusion of human language in software for processing text, speech and knowledge. It includes, but is not limited to the following fields: speech technology; written language input; language analysis, understanding and translation; automatic document processing; machine translation; multimodality; electronic language resources and SALT evaluation.

The categories found in Survey of the State of the Art were also emulated, although for ease of a definition the thirteen categories were reduced in number to the following eight:

1. Speech technology (speech recognition, speaker recognition, text to speech techniques, speech coding and enhancement, multilingual speech processing)
2. Written language input (optical character recognition, handwriting recognition)
3. Language analysis, understanding and generation (grammar, semantics, parsing, discourse and dialogue)
4. Document processing (text and term extraction, interpretation, summarization)
5. Machine translation (including computer aided translation, multilingual information retrieval)
6. Multimodality (gesture and facial movement recognition, visualisation of text data)
7. Language resources (written and spoken corpora, lexica, terminology)
8. Evaluation (of all of the above)

Project Overview

In order to ascertain the feasibility of establishing a Special Interest Group for SALT in Wales, the SALT Cymru Feasibility Project proposed to investigate the:

- Worldwide economic impact of SALT
- International Best Practice and State of the Art in SALT
- Academic SALT research base in Wales
- Enterprise research base in Wales
- SALT needs of industry in Wales

These are discussed in more detail below.

The economic importance of the SALT Sector worldwide is often talked about but the establishment of a funded SALT SIG required proof of the global economic significance of SALT. Several presentations and papers at the LangTech conference attended by SALT Cymru representatives provided figures demonstrating the significance of the SALT sector to the global economy. These included:

- The worldwide market in speech technology deployments alone is estimated to be worth $3.2 billion by 2010 (*Understanding the Market Movements in Network Speech*)
- In 2005 the translation market, particularly in multilingual websites and software localization, generated $8.8 billion in worldwide revenues. (*Venture Capital and Language Technologies*)
- 4 of the top 20 international companies site their HQs in London, the North of England and Ireland employing over 2300 people. (*Venture Capital and Language Technologies*)

This provided a strong economic argument for ensuring that academia and industry in Wales are involved in SALT research and development.

Although it was clear that SALT formed a significant sector internationally, the focus of the SALT Cymru feasibility report project was its potential for the Welsh economy. As a project to be funded through the Welsh Assembly Government's Academic Expertise for Business (A4B) programme the special interest group's purpose was to facilitate the transfer of expertise from Welsh Universities to businesses in Wales for their economic benefit.

However, before the special interest group could be funded, the feasibility study had to establish that SALT research was to be found

within academia in Wales, and that the provision of funds for establishing a special interest group would be of benefit to both industry and academia in Wales.

As a result, the project was required to locate and quantify the existing SALT research to be found within academia and industry in Wales. Therefore, an audit of the SALT research in Wales was undertaken by members of the SALT Cymru team.

The relevant departments in all Higher Education institutions in Wales were identified and contacted to establish what activities, if any, they were undertaking in the field of SALT. It was discovered that the SALT development knowledge base in Wales was relatively small. In Welsh higher education institutions, it was estimated that the equivalent of fewer than ten full-time academics worked directly on SALT. Of these, only about half were permanently contracted to do so.

However, despite this small size a wide variety of SALT were found to be under development. They encompassed both speech and language technology, and seven of the eight research areas defined by the project as part of SALT were found being researched within Wales, namely:

- Speech technology (speaker recognition in Swansea, text-to-speech techniques and a speech recognition project in Bangor)
- Language analysis, understanding and generation (spelling and grammar checking, and a library of language tools in Bangor)
- Document processing (text and term extraction, in Bangor)
- Machine translation (a small pilot project in Bangor)
- Multimodality (SALT-associated research in Swansea)
- Language resources (developed in Bangor, Swansea and Lampeter)
- Evaluation (standardization work in Bangor and by Geolang, undertaken with ISO)

It was also noted that there was very little duplication of subject areas within Welsh SALT research. Only rarely were multiple companies or academic institutions found to be active in the same subfield of SALT, with each group seeming to have found their own individual niche. This finding was a positive one for the envisaged special interest group, as a lack of conflicts of interest within the group would likely lead to a more receptive environment for sharing ideas, and working together in a collaborative environment. Based on the growth of SALT internationally, it was felt that this knowledge base could thrive with appropriate funding and the establishment of greater ties with industry, and would be well placed in a multilingual marketplace due to the bilingual nature of Wales.

In addition to the pure SALT research and development that was identified, a significant amount of SALT-associated development work was found in Wales, both in academia and within industry. A number of companies were found to be exploiting SALT development together with other technologies, especially in the field of web design and in the multimedia and creative industries. Enabling these industries to benefit from the research underway at Welsh universities was seen as a priority for the intended Special Interest Group. This prompted an audit of the present state of play of industry in Wales and an assessment of its needs.

In examining the needs and current state of industry in Wales, special attention was paid to Welsh SMEs and their markets/potential markets.

These SMEs fell into three categories:

(a) developers of software products and services e.g. software developers, web developers

(b) developers of language-based assistive technology products e.g. disability aids

(c) users and potential users of SALT-enhanced applications e.g. translation companies

Outside academia, many of those questioned about SALT were unfamiliar with the SALT label or grouping, even though they were daily users of SALT technologies. A number of SALT Cymru survey respondents reported that they did not use SALT despite mentioning in conversation that they used technologies such as spellcheckers and OCR relatively frequently.

One of the challenges when encouraging knowledge transfer between academia and business is therefore that of overcoming the language barrier formed by the specialized use of terminology in academia, which is unfamiliar to businesses and end users.

In addition to an audit by the project's researchers, information was also collected by means of a bilingual online survey. Participation was by invitation, and responses were used to identify potential focus group participants. A total of 48 complete forms were received, which was an encouraging number considering its narrow scope. Respondents included users, developers and potential developers of SALT, and the sample included a wide range of interest in various SALT subfields.

Survey Results

The survey results provided a valuable insight into the SALT developed in Wales, with most types of SALT being represented, and no specific SALT type dominating research. Accessibility issues and the linguistic needs of a bilingual country were highlighted, as were the improvement of interfaces between users and objects or information, and of the identification of what constitutes relevant information to the user, and the meaningful categorization of information. Pure SALT development work in Wales appeared to be relatively fragmentary, geographically dispersed, and was in general accomplished by individuals or small teams. Associated SALT development within Wales appears to have a somewhat stronger research base.

Of those who responded:

- 40% stated that they were SALT developers
- 31% stated that they were prospective SALT developers, who might be interested in developing such technologies in the future
- 29% stated that they were users or prospective users of SALT: they did not develop SALT and had no intention to do so in the future, but were interested in the technology from a user perspective

A wide variety of SALT areas were mentioned by respondents. However these were the areas which appeared to be of consistent interest to developers, potential developers and users.

- Text-proofing tools
- Speech-enabled technologies (whether speech synthesis as a module, or speech-enabled communication aids as finished products)
- Speech recognition
- Intelligent web searching
- Keyword and trend spotting from text
- Machine translation

It would therefore appear that current SALT development in Wales is aimed at:

- the accessibility needs of disabled users
- the linguistic needs of a bilingual country
- the improvement of interfaces between users and objects or information
- the improvement of the identification of what constitutes relevant information to the user
- the meaningful categorization of information

Interviews, investigations **and findings**

In addition to the research undertaken by the SALT Cymru project's own researchers, the advice and opinions of experts working in the field of SALT were sought through the medium of face-to-face interviews. A sample of

individuals actively involved in SALT R&D from across Wales was selected, and a list of key players was compiled, together with an overview of their interest(s) in SALT. Interviews were carried out with individuals in internationally regarded laboratories researching the state of the art in the discipline. These interviews reflect the wide-ranging nature of SALT. Key findings included the importance of multimodality to future developments, and the need to nurture a knowledge base from the earliest opportunity, i.e. in schools in addition to universities. A summary of these interviews can be found in the SALT Cymru Feasibility Report (see http://www.saltcymru.org).

To further investigate the state of the art, the annual LangTech international conference was attended. The conference provided a broad overview of a wide range of SALT technologies in research and commercial contexts, and addressed future directions in international SALT development. A focus group with professional translators, held in Wales provided further opportunities to examine the needs of an important group of SALT users and their specific needs and priorities.

The project discovered a significant appetite for speech and language technologies (SALT) within Wales, both by end users and amongst SMEs that are currently developing such technologies or which have an interest in doing so in future. Investment, a building up of the knowledge base, and an increase in awareness activities were seen to be required in order that this sector of the economy may capitalize on Wales's privileged position as a bilingual nation and grow to fulfil its full potential in exploiting worldwide markets.

An important part of the feasibility report was the evaluation of relevant open source software and standards. Particular attention was paid to the potential of resources to be enhanced for use in Wales in a pre-competitive research stage, including their suitability for further development by industry in a non state-aid environment.

Five key components of SALT were examined, based on the main areas of interest highlighted in the survey results. These areas were:

- Speech synthesis
- Speech recognition
- Intelligent web searching; keyword and trend spotting from text
- Optical character recognition

Open source examples of software resources corresponding to the above were investigated, and a series of papers were published as appendices to the SALT Cymru Feasibility Report.

The software investigated were as follows:

- Moses (machine translation)
- Festival (speech synthesis)
- Sphinx (speech recognition)
- UIMA (intelligent web searching; keyword and trend spotting from text)
- Tesseract (optical character recognition)

It is important to note that a sensitive approach to licensing requirements is needed to cater for the complex and differing needs of private sector companies. Flexibility and choice of licences for individual circumstances is welcome by developers working in a commercial environment. In selecting the above resources the project found that the more permissive BSD or MIT-style licensed software were generally preferred by developers. As a project aimed to promote commercial SALT development, it was important that the licensing terms for the software should be permissive enough to be attractive to as wide a range of commercial enterprises as possible.

The LTU sees facilitating commercial developers to create commercial minority language end products as being a significant step in normalizing such languages in relation to their more resourced counterparts. Currently, the LTU is involved with commercial developers working on a number of minority language products and the catalyst for many of these collaborations has been the availability of the software with a permissive open source license.

Report recommendations and terms of reference

In light of the research undertaken during the course of the project, the SALT Cymru Feasibility Report recommended that:

1. A SALT Cymru specialist interest group (SIG) be established.
2. The SIG shall have an international watching brief for SALT.
3. The SIG shall work to strengthen the research base in Wales**.**
4. The SIG shall draw up a prioritized programme to develop a toolkit of basic language resources**.**
5. The SIG shall maintain and further develop the SALT Cymru website to include a resource portal.
6. The SIG shall guide a Welsh SALT education and training programme.
7. The SIG shall address the training and communication needs of its own members.
8. The SIG shall address evaluation and quality control issues for SALT developers.
9. The SIG shall seek adequate funding to enable it to fulfil its terms of reference.

The Establishment of the SALT Cymru Special Interest Group

The above recommendations are now being put into action, as, following the report, funding was secured from the Welsh Assembly Government to finance the envisaged Special Interest Group.

Coordinated by Bangor University's Language Technologies Unit, the SIG commenced in February of 2009 and was officially launched by the Deputy First Minister of Wales on the 20th of March. Funding will run for two years, by which time it is hoped the SIG will be sufficiently mature to be sustainable in the long term.

The international watching brief has been maintained through visits to UNESCO's Infoterm offices, Austria, and the NAACLT (North American Association of Celtic Language Teachers) conference and Welsh Office in New York. A surprising amount of international interest in Welsh SALT has been noted since the SIG's inception, and interest from amongst the traditionally TV based media sector in Wales is growing as the boundaries between old and new media become blurred.

A foundation for strengthening the research base has been established by the setting up of a SALT Cymru steering group comprising of representatives from many Welsh Universities. Work is continuing on the development of the SALT Cymru website into a SALT resource portal, which will hopefully serve as a catalyst for SALT research and development in Wales. The website is to be found at www.saltcymru.com.

A series of quarterly seminars, networking events and newsletters is addressing the communication and knowledge transfer needs of the members of the SALT Cymru special interest group, and an annual conference is currently in the planning stages.

The SALT Cymru Special Interest group also offers SALT consultation to business which is paid for by the Welsh Assembly Government. These consultation sessions are tailored to the needs of the individual enterprises, and have benefitted from the research into open source resources undertaken as part of the SALT Cymru Feasibility Report.

So far, 6 enterprises (from a target of 48) have been assisted by the SALT Cymru SIG project team, and the project is currently on track.

These include:

- An enterprise specialising in education wishing to target its products at the multilingual market
- An enterprise specializing in subtitling and translation wishing to develop machine translation capabilities
- A translation company wishing to improve proofing tools for software translation
- A software developer specialising in development for mobile devices with an interest in SALT
- A software developer with an interest in proofing tools and translation memory systems
- A SALT developer interested in the SALT capabilities of mobile devices

Interestingly from a minority language perspective, the majority of these enterprises wished to apply SALT to Welsh-language business opportunities, with a number using the language as a step towards multilingualism in their products that would facilitate entry into international markets in the future.

Some of the consultations have resulted in Knowledge Transfer Partnerships (KTPs) between Bangor University and Welsh Enterprises. These are partnerships between an enterprise and a university who team together to accomplish a project that will be of benefit to the company, but would not have been feasible without the KTP. The enterprise receives 60% funding towards the employment costs of an 'associate', a suitable graduate identified and supervised throughout the course of the project by the university whilst working on location at the enterprise. The university also receives funding from the government for its part in the KTP, which contributes to the ongoing survival of the language Technologies.

Conclusion

The SALT Cymru Feasibility Report successfully led to the securing of funding to establish a SALT special interest group for Wales by highlighting the increasing economic significance of SALT globally, and identifying the potential of the existing knowledge base in Wales whilst drawing attention to its precarious future. The feasibility study also highlighted the fact that SALT is not a label that is recognised within businesses that do not have strong links to academia. This has led to refocusing discussions between SALT team member and enterprises on individual relevant technologies rather than on SALT as a whole, as enterprise are understandably more interested in technologies useful to themselves than on a complete overview of SALT as discipline.

This paper has not focused on a specifically minority-language project as neither the feasibility report nor the special interest group project are specifically minority language projects. However it is relevant to a minority language discussion as it illustrates that when enterprises are able make a business case for minority language services and products, economic stimulus projects such as these may validly address the issues facing a minority language such as Welsh. This is seen by the LTU as a step towards normalising the Welsh language by funding aspects of its development through funding streams not specifically set aside for minority language or 'cultural' purposes.

In fact, most of the companies so far assisted that have been interested in SALT technologies for economic reasons have been interested in its application for the Welsh language. This interest has primarily been driven by the demands of the public sector where it is legislated that, 'where practicable', Welsh and English should be treated on the basis of equality. By increasingly making Welsh-language support 'practicable' within the information and communication systems of today, SALT Cymru is playing an important role in the continuing relevance of Welsh as a modern language.

For more information about SALT Cymru, including the full report and appendices, see www.saltcymru.org.

Bibliography

Varile, Giovanni Battista, and Antonio Zampolli. 1997. *Survey of the state of the art in human language technology*. Linguistica computazionale, v. 12-13. Cambridge [England]: Cambridge University Press.

Jones, Rhys, and Prys, Gruffudd et al. 2008. *Project Closure Report for SALT Cymru.* Bangor University. On-line version available at: http://saltcymru.org/english/adroddiadsalt.html (accessed 24.08.09)

Automated English subtitling of Welsh TV Programmes

Llio Humphreys
Testun Cyf.
Norfolk House
57-59 Charles Street
Cardiff
Wales, UK
CF10 2GD
llio@testun.co.uk

Abstract: This paper discusses a government-funded Knowledge Transfer Partnership (KTP) project between Testun and Canolfan Bedwyr. The aim of the project is to develop an automated system for automating English subtitles for Welsh TV programmes, integrating a Welsh speech recognizer and a Welsh-English translation module within a subtitling environment. The paper discusses the motivation and mechanism, challenges, project plan and scope of this ambitious project as well as references to relevant prior research.
Keywords: subtitling, speech recognition, machine translation

1 Introduction

1.1 Motivation

Testun is a translation, subtitling and teletext company based in Cardiff, Wales. Testun's subtitling department edits prepared subtitles, and provides live and late subtitles, for programmes on S4C's three Welsh TV channels.

S4C has a target to subtitle 100% of programmes by the end of 2009, where copyright permits, so that more non-Welsh speakers and hard-of-hearing people can enjoy more Welsh programmes. The number of programmes with live sections is also increasing. There is an increase in demand for Testun's work. The problem is that translation, subtitling and teletext is labour-intensive.

Testun envisages that an automated translation and subtitling software solution would help Testun provide a more efficient subtitling service.

1.2 Mechanism

The Knowledge Transfer Partnership (KTP) Scheme helps UK businesses increase their competitiveness and productivity by accessing knowledge, technology and skills within research institutes. This project's research institute, Canolfan Bedwyr's Language Technology Unit, develops language resources for the Welsh language, the Celtic languages, and for multilingual situations in general. The Unit is responsible for standardising terminology, dictionaries, a Welsh language spellchecker and grammarchecker, computer-based learning and speech technology.

1.3 Challenges

The project involves challenges in intellectual property, security and software integration. Most of Testun's subtitling work takes place in S4C's offices using S4C equipment chosen and purchased by S4C. The automated subtitling software must be compatible with existing systems. Comprehensive consideration of these issues is beyond the scope of this paper.

The critical functional challenge is making best use of advanced, but brittle technologies, to produce subtitles comparable in quality to human-generated subtitles.

Speech recognition is made difficult by:-

- differences in voice quality, between men and women, young and old
- differences in accents and pronunciation
- mumbling
- interference: music, coughing etc.
- unknown words

Challenges for translation are:-

- good use of bilingual corpora - accurately aligning words or phrases where two languages express things

differently, using different parts of speech
- processing grammatically incorrect spoken text
- segmenting sentences with disfluencies
- dealing with unknown words
- producing subtitles that conform to industry-standard regulations

The challenge of providing NLP (Natural Language Processing) solutions involving less resourced languages has necessitated this project. The lack of available NLP resources does not mean that lower standards will be acceptable to S4C or its audience, who are used to high-quality human-generated subtitles. S4C's contracts for live and late subtitling (where pre-recorded programmes are received for subtitling shortly before transmission) expressly require a high standard of translation, and a spelling and grammar accuracy rate of 95%.

One solution is to view the problem of producing subtitles as much as information retrieval as language generation. For each sentence spoken, a search is conducted for the most similar Welsh sentence from a domain-specific parallel corpus, from which the English translation provides the subtitle output. Another solution is to limit the language domain, since both speech recognition and translation software perform best on sublanguages. Uncertainty from the speech technology side can also be reduced by loading a script for the TV programme. The speech module then aligns the spoken text to the script, to ensure appropriate subtitle timings.

2 Literature Review

2.1 Multilingual Subtitling Systems

MUltilingual Subtitling of multimediA content (MUSA)

MUSA (Piperidis, Demiros, and Prokopidis, 2004) combines speech recognition, advanced text analysis, and machine translation to help generate multilingual subtitles in English, French and Greek. Subtitling processes include tokenisation and subtitle text splitting, calculation of cue-in/cue-out timecodes, handcrafted rules, and shallow parsing. Subtitles must to some extent summarise the speech, as people can listen to more than they can read in the same time. MUSA's sentence compression module (Daelemans & Höthker, 2004) uses shallow-parsing information and handcrafted deletion rules to:-

- remove disfluencies such as repetitions introduced by hesitation
- replace part of the input sentence by a shorter paraphrase

eTITLE

eTITLE (Melero, Oliver and Badia, 2006) is a web-based multilingual subtitling service for the English, Spanish, Czech and Catalan. The system is configured as a distributed environment, so that different translation memories and machine translation modules can be located on different computers. A script and video file is inserted before any processing takes place. Like MUSA, eTITLE's translation module integrates translation memories with machine translation, because their shortcomings and advantages are complementary:

> *Since the output [of translation memories] is originally generated by humans, the typical errors and noisy output that MT systems sometimes produce are avoided. However, they are less robust than MT. (p2., Melero, Oliver & Badia, 2006).*

2.2 Speech Recognition

2.2.1 HMM-based Speech Recognition

Myers and Whitson's (1995) implementation of Rabiner's statistical speech recognizer (1989) is described as a set of programs that:

1. converts audio/wave files to sequences of multi-dimensional feature vectors. eg. DFT (discrete Fourier transform), PLP (Perceptual Linear Prediction), etc.
2. quantizes feature vectors into sequences of symbols eg. VQ (Vector Quantization)
3. trains a model for each recognition object (ie. word, phoneme) from the sequences of symbols. e.g. HMM)
4. constrains models using grammar information.

Phonemes are the standard unit of sound used, being the smallest unit of speech that can differentiate one word from another. English phonemes include 'ue' sound in 'moon', 'blue', 'grew' and 'tune', and 'f' in 'field' and 'photo'.

2.2.2 Grapheme-based Speech Recognition

A grapheme-based system (Killer, Stüker, and Schultz, 2003) can be used for languages with strong grapheme-phoneme relations. The advantage for a less-resourced language with lack of trained phoneticians (Williams, 2008) is that audio files and transcripts are sufficient data to train the system. Tests reveal the relative strength of phoneme-grapheme relations in different languages:

	Word Error Rate (WER)		
Language	**English**	**German**	**Spanish**
Phoneme	12.7%	17.7%	24.5%
Grapheme	19.1%	17.0%	26.8%

Table 1: Phoneme-based vs Grapheme-based recognition (p.3142, Killer, Stüker, and Schultz, 2003)

Canolfan Bedwyr's research on developing a text-to-speech synthesiser confirmed that "[i]n contrast to English, Welsh orthography is a reliable guide to pronunciation." (p.2, Williams, 1995). In addition, "Welsh, unlike English, does not exhibit stress-related vowel reduction or vowel lengthening." (p.2, Williams, 1995).

2.2.3 Speech Recognition and Disfluencies

Other research of relevance is the varying performance of speech recognition depending on the context of spoken text. Duchateau, Laureys, and Wambacq (p.1, 2004) found that:

> *the recognition accuracy of freely spoken language is quite poor when compared to that of dictated speech: while the word error rate(WER) for large vocabulary speaker-independent dictation is about 5%, the WER for spontaneous speech recognition ranges from 15% for broadcast news.. to 40% for meeting and telephone conversation transcription.*

2.2.4 Welsh Speech Recognition

There is no Welsh speech recognizer currently available. However, a significant output of Canolfan Bedwyr's (Welsh and Irish Speech Processing Resources) WISPR Project (Williams, Jones and Uemlianin, 2006) was a large grapheme-to-phoneme dictionary, suitable for use in a speech recognizer. The KTP adviser for this project is a speech recognition expert who was involved in Canolfan Bedwyr's WISPR project. We aim to develop a Welsh HMM-based speech recognizer that can be integrated into the automated subtitling system.

2.3 Translation

2.3.1 Rule Based Machine Translation (RBMT)

Apertium

Apertium is a sophisticated rule-based machine translation system that uses a sequence of finite-state automata:-

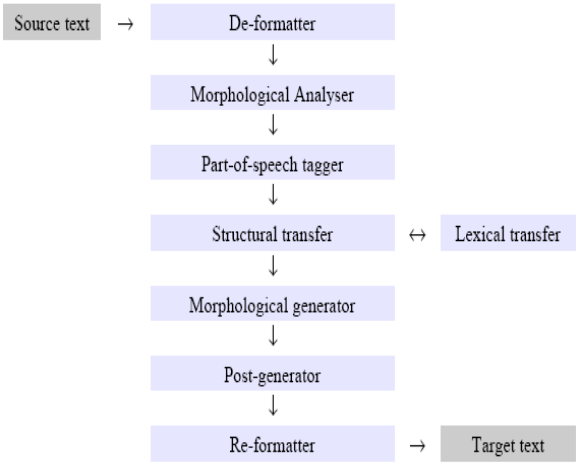


Figure 1: The eight modules of the Apertium system (p.4, Corbí-Bellot et al., 2005)

The deformatter extracts text from formatted documents and outputs a lemma, lexical category and morphological inflectional information. The HMM part-of-speech tagger is trained on representative source language text. The lexical transfer module outputs dictionary entries for target language words and multi-word units. The structural transfer module identifies chunks or phrases requiring additional grammatical processing to account for structural differences. The morphological generator inflects target language surface forms. The post-generator performs contractions and inserts apostrophes. The re-formatter provides formatting as per the original file.

2.3.2 Statistical Machine Translation (SMT)

Moses

Moses (Koehn et al., 2007) is one of the leading statistical-machine translation software available. The package includes corpus preparation (tokenisisation, converting to lowercase), statistical data (n-grams, statistically-derived word classes), machine translation steps (word and phrase alignment, translation), tuning and evaluation. A key attraction is that that the system is language independent. In practice, the success of SMT depends on the language pair and domain.

2.3.3 Translation Memories and Example Based Machine Translation (EBMT)

Armstrong et al. EBMT

Armstrong et al. (2007) researched using translation memories and Example Based Machine Translation (EBMT) for subtitling films. Notwithstanding user feedback that post-edited EBMT generated German subtitles of film dialogue displayed good colloquial phrases, had the correct tone and register and "felt like German", users preferred the output of rule-based Babelfish MT output over raw EBMT. The translation difficulties encountered may stem largely from developing a system for a broad subject-matter and spoken dialogue.

Leplus, Langlais and Lapalme

Leplus, Langlais and Lapalme (2004) developed a limited-domain translation memory and EBMT system for weather reports. They created a bilingual corpus from Environment Canada forecast reports in French and English. With the full memory (about 300,000 source sentences), 87% of sentences were found verbatim in the memory, and 89.5% always exhibited the same translation, possibly because these translations were generated by the rule-based METEO machine translation system.

Déjà Vu

Testun has some understanding of Atril's Déjà Vu, as it is the translation memory used by their Translation Department. Déjà Vu uses three types of resource:-

- translation memory of source and translated sentences,
- terminology database, and
- lexicon of translated word or multi-word terms from common words in files to be translated

With this information, Déjà Vu can:-

- search for source sentences that have been translated before
- search for similar sentences and highlight differences
- assemble a translation

This last function appears uses EBMT. If the source text contains the sentence:

Prometheus, the heavy equipment and engine manufacturer

the following French sentence from the translation memory would partly correspond:

Prometheus, the heavy equipment and engine producer

Prometheus, le producteur de matériel lourd et de moteurs

If the terminology database contains "producer" and "manufacturer", Déjà Vu can assemble a translation by replacing "*producteur*" with "*fabricant*".

The key to successful implementation would therefore be supplying the system with an extensive terminology database and lexicón

2.3.4 Welsh-English Machine Translation

English-Welsh machine translation systems have been developed – for example Phillips (2001) implemented a statistical system, Jones and Eisele (2006) tested the Pharaoh (Koehn, 2004) system (precursor to Moses), and Tyers and Donnelly (2009) adapted the rule-based Apertium system for Welsh-English translation. Both Apertium (Tyers and Donnelly, 2009) and Moses (Koehn et al., 2007) can produce excellent gist translations:-

Welsh original sentence	Mae Heddlu'r De yn ymchwilio i farwolaeth dyn 41 oed o Abertawe.
Apertium translation	South Wales Police is investigating death man 41 years old Swansea.
Moses translation	the south wales police investigation into the death of a man 41 years of age of abertawe.

Table 2: Apertium and Moses output for Welsh-English translation (Donnelly, 2009)

The output does not satisfy S4C's quality standards. However, Welsh-English machine translation is expected to improve – there is a project under the fifth Google Summer of Code to combine Moses and Apertium.

We would be most interested in using machine translation where output meets S4C's quality criteria and is acceptable to end-users.

2.4 Useful Techniques from Information Retrieval (IR)

IR techniques could play an important part in this project, both in the preparation of suitable data, and selection of appropriate sentences to be used as subtitles.

2.4.1 Term Extraction

2.4.1.1 Word Clustering

Word clustering techniques can be useful to form equivalence classes that can be used as interchangeable variables during language generation (Brown, 1999). For example, the sentence

> *John Miller flew to Frankfurt on December 3rd*

can yield the following pattern:

> *<firstname> <lastname> flew to <city> on <month> <ordinal>*

Words that have a similar context of surrounding words are deemed to belong to the same class. Brown's (1999) system looks at previous and succeeding words up to three positions apart. Equivalence is measured by cosine similarity. The standard formula for cosine similarity is:-

$$sim(s,y)=\cos(\vec{x},\vec{y})=\frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}=\frac{1\cdot1+1\cdot0+0\cdot1}{\sqrt{(1^2+1^2+0^2)*(1^2+0^2+1^2)}}$$

As a simplified example, if word **x** has as its context the preceding word **a** and following word **b**, and word **y** has context words **a** and **c**, a table can be formed as follows:-

	a	**b**	**c**
x	1	1	0
y	1	0	1

Table 3: Equivalence by cosine similarity

Brown adapted the cosine similarity formula to add increasing weight to increasingly adjacent words. An aligned word pair is made on the basis of dictionary entries, and must meet a certain threshold of similarity to join an existing bilingual cluster.

2.4.1.2 Multi-Word Units

Statistical measures can also be useful for extracting multi-word units. Based on the observation that words within multi-word units have high n-gram probability and low entropy, while there is considerable variance in words surrounding the multi-word units, Shimohata et al. (1997) devised an entropy-based formula for extracting multi-word terms:

$$H(str)=\sum_{i=1}^{n}-p(w_i)\log p(w_i)$$

To avoid uninteresting n-grams such as 'label for the', Merkel and Andersson (2000) compiled four stop-word lists:

- non-starters cannot begin a multi-word unit
- non-enders cannot end a multi-word unit
- prohibited words cannot be part of a multi-word unit
- ignored words are statistically insignificant and are ignored for entropy calculations

This 'linguistic-lite' approach is flexible and portable, which benefits less-resourced languages. The methodology may be adapted to use part-of-speech categories rather than word lists e.g. non-enders could be prepositions, articles, conjunctions and verbs.

2.4.2 Measuring Sentence Similarity

Finding a Welsh sentence in a parallel corpus that is similar to an input spoken sentence is an IR task. Jin and Barrière (2005) evaluated four formulae that measure sentence similarity by treating sentences as pure strings, stating that:

> *"The advantage of such an approach compared to more linguistically motivated approaches is that the system can quickly retrieve similar sentences from a large size corpus (over one million sentences), work well with ill-structured sentences, and work across different human languages." (p.1, Jin and Barrière, 2005).*

For an automated subtitling system, sentences need to be retrieved quickly to generate live subtitles. Also, syntactic analysis may not work with input sentences that, even when scripted,

are formulated for spoken rather than written well-formedness. The measures are also language-neutral which render them suitable for use with a less-resourced language.

In the formulae below, Q represents a query or input sentence, and S a sentence from the corpus. Three of the formulae evaluated are IR similarity-ranking functions:

- the Dice Coefficient (Hersh, 2003)

$$Dice(Q,S) = \frac{(2 * N_{common})}{(N_Q + N_S)}$$

- Vector Space Model (Cosine) (Salton and McGill, 1983), and

$$COSINE(Q,S) = \frac{\sum_{k=1}^{t}(w_{qk} \cdot w_{sk})}{\sqrt{\sum_{k=1}^{t}(w_{qk})^2 \cdot \sum_{k=1}^{t}(w_{sk})^2}}$$

- Lin's Information Theory Similarity (Aslam and Frost, 2003)

$$IT - Sim(Q,S) = \frac{2\sum_{w \in Q \cap S} \log \pi(w)}{\sum_{w \in Q} \log \pi(w) + \sum_{w \in S} \log \pi(w)}$$

The fourth formula evaluated was BLEU (Papineni et al., 2002), used to compare the output of machine translation with reference translations:

$$\text{Log BLEU} = \min(1 - r/c, 0) + \sum_{n=1}^{N} w_n \log p_n$$

Each formula was tested on English, French and Chinese text, and four sentences were retrieved for each input sentence. The output of the retrieved sentences were evaluated by human testers:

Grade	Explanation
4	The sentence exactly matches the input
3	The sentence provides enough information about the whole input
2	The sentence provides information about some part of the input
1	The sentence provides no information about the input

Table 4: Accuracy grades (p.6., Jin and Barrière, 2005)

The scores across languages were:

Algorithm	Average similarity score received (Highest score is 4)
Dice	2.73
Cosine	2.75
Lin	2.73
BLEU	2.64

Table 5: Average similarity score for different algorithms across languages (p.6, Jin and Barrière, 2005)

Ranking of the four sentences retrieved by each formula were compared with human ranking of the same sentences:

Algorithm	Percentage of agreement with human rating
Dice	100%
Cosine	93%
Lin	67%
BLEU	80%

Table 6: Average similarity score for different algorithms across languages (p.6, Jin and Barrière, 2005).

3 Project Plan

A suitable starting point for an automated Welsh-English subtitling system would be weather bulletins. S4C currently broadcasts weather bulletins six times a day during weekdays with at least three bulletins on Saturday and four bulletins on Sunday. The bulletins are not currently subtitled, but they will be by the end of the year to meet S4C's subtitling targets. The contract for subtitling weather bulletins will specify S4C's quality requirements, but not the implementation. For Testun, employing subtitling staff for short bulletins at otherwise non-busy times would be expensive. An automated subtitling system could help the company offer a competitive subtitling service for this programme.

Here is a summary of the project plan:

- prepare script and audio files for the script-audio alignment module
- build a suitable bilingual parallel corpus
- extract or compile domain-specific terminologies
- integrate translation memories and summarizer

- integrate a Welsh speech recognizer
- testing, debugging, refinement and rollout

The client and developer's view of the system are represented in figures 2 and 3:

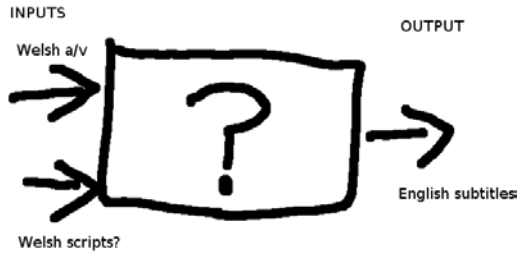


Figure 2: Client view of Testun's automated subtitling system

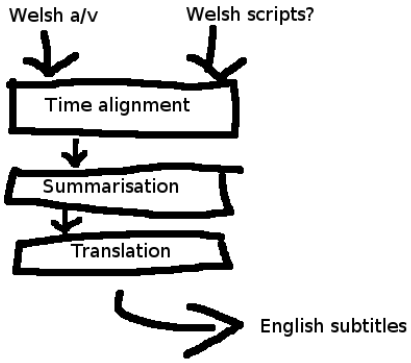


Figure 3: Developer view of Testun's automated subtitling system

3.1 Preparation of Script and Audio Files

Testun uses Sysmedia's WinCAPS subtitling system for its subtitling work for S4C. Sysmedia has also developed the SpeechFollower script-audio module. This module can be used for any language. Audio files to train the system should contain a range of text spoken by six different speakers, including men and women, and people with various regional accents. This speaker-independent system can take as input the speech of actual presenters rather than re-speakers. SpeechFollower handles timecodes to ensure that subtitles are generated in good time and at a readable pace.

3.2 Build a Bilingual Corpus

Testun not only provides subtitling, but also Sbectel teletext services for S4C – textual information pages available for view on a TV. Teletext weather bulletins are available twice a day in English and Welsh, and are exact translations. Testun has three months of archive teletext bulletins available for a parallel corpus. Teletext pages are concise, and the short sentence structures are highly suitable for use as subtitles.

3.3 Extraction of Domain-Specific Terminologies

Preliminary analysis of bulletins indicate that weather, temporal and location terms would be useful. We will experiment with the clustering and term extraction techniques outlined in section 2.4.1 to extract interchangeable words and multi-word units such as 'sunny intervals' and 'over the course of the week'. Following alignment and term categorisation, new sentences can be generated to expand the parallel corpus.

3.4 Integration of Translation and Summarizer Components

A comprehensive literature review and testing of machine translation and translation memory systems will be conducted. The component must be capable of seamless integration with WinCAPS subtitling software. We will also look into the feasibility of matching scripted sentences with Welsh sentences from the teletext corpus, using a sentence similarity metric (see section 2.4.2). Text summarisation may involve selecting sentences for which the closest match is found and/or which score highly for domain-specific terms. The equivalent English sentences can be output as a subtitle with confidence in their suitability and grammaticality. However, the results will need to be evaluated carefully to ensure coherent sequence and no significant loss in information. Alternatively, an English text summarizer may be used.

3.5 Integration of a Welsh Speech Recognizer

Some presenters do not prepare scripts and others are inclined to ad-lib than follow their

scripts faithfully, rendering the script-speech alignment module inoperable. It will therefore be necessary to integrate a Welsh speech recognizer. We may use a freely available speech recognition toolkit to develop our own time-alignment software, such as the open-source speech recognition toolkit Sphinx and/or the free-of-charge HTK (Hidden Markov Model Toolkit). Practical experience of working with the SpeechFollower module will indicate the procedures necessary to achieve time-alignment.

3.6 Testing, debugging, refinement and rollout

While Testun's Company Director and the Head of Subtitling Unit will be involved in all stages of requirements, design and evaluation of the system, a thorough evaluation of overall performance will be undertaken at this stage. Criteria for evaluation will include use of resources and speed. Extensive evaluation of overall performance will also be undertaken. Standard measures such as WER (Tillmann et al. 1997) and BLEU (Papineni et al., 2002) will be used to measure speech recognition performance and translation performance respectively. Human evaluation will be provided by Testun's subtitling staff.

4 Expected Results

We hope to develop usable resources as soon as possible. Some examples are:

- Shortcuts for subtitlers. Shortcuts are abbreviated terms that a subtitler can use within WinCAPS to aid fast typing. We will extract Welsh phrases used frequently in the forecasts and create English shortcut dictionaries for WinCAPS.
- Audio and data files for training the speech-audio alignment module
- A bilingual parallel corpus of weather forecasts
- Single and multi-word domain-specific, interchangeable, weather, temporal and location terms.
- A sentence similarity matcher and/or summarizer
- A Welsh speech recognizer

5 Conclusions

This paper has outlined the scope and challenges of a two-year KTP project for the development of automated English-Welsh subtitling software. Prior research indicate ways of mitigating the challenges, so that limited but practical systems can be developed. There is an incentive to get a good working system in use in the very near future, while providing opportunities for incremental development. The project will be deemed successful if careful and appropriate use is made of the automated subtitling system, helping the Welsh language TV industry increase services for social inclusion and accessibility.

Bibliography

Apertium <http://www.apertium.org/> Viewed 15 July 2007

Armstrong, S., Way, A., Caffrey, C., Flanagan, M., Kenny, D., O'Hagan, M., 2007. Leading by Example: Automatic Translation of Subtitles via EBMT. In *Perspectives, 14:3,163 — 184*

Aslam, J. Frost, M. 2003 An Information theoretic Measure for Document Similarity. In *Proceedings of the 26the Annual International ACM SINGIR Conference on Research and Development in Information Retrieval (ACM Press) 449-450*

Atril Déjà Vu X <http://www.atril.com/> Viewed 15 July 2007

Brown, R. D. 1999. Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22-32.

Canals R., Esteve, A., Garrido, A., Guardiola, M. I., Iturraspe-Bellver, A., Montserrat, S., Pérez-Antón, P., Ortiz, S., Pastor, H., and Forcada, M. 2001. InterNOSTRUM: a Spanish-Catalan Machine Translation System. In *Machine Translation Review, 11:21-25.*

Cancelo, P., 2000, [Reseña a] Herramientas Mágicas / Word Magic Tools de Word Magic Software (Word Magic Tools 2000

Deluxe 2.1.). In *Revista de Lexicografía, VI (1999-2000)*, pages 235-238

Corbí-Bellot, A.M., Forcada, M. L., Ortiz-Rojas, S., Pérez-Ortiz, J. A., Ramírez-Sánchez, G., Sánchez-Martínez, F., Alegria, I., Mayor A., Sarasola, K. 2005. An open-source shallow-transfer machine translation engine for the romance languages of Spain. In *Proceedings of the European Associtation for Machine Translation, 10th Annual Conference* (Budapest, Hungary, 30-31.05.2005), pages 79—86

Daelemans, W. & Höthker. A .2004. Automatic Sentence Compression in the MUSA project. In *Languages & The Media*, Berlin

Duchateau, J., Laureys, T., Wambacq, P. 2004. Adding Robustness to Language Models for Spontaneous Speech Recognition. In *the Proceedings of Robust2004, COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*

Eurfa <www.eurfa.org.ok> Viewed 15 July 2009

Hersh, W. 2003. Information Retrieval: A Health & Biomedical Perspective (Second Edition, Springer-Verlag), chap. 8.

Hidden Markov Model Toolkit (HTK) <http://htk.eng.cam.ac.uk/> Viewed 15 July 2009

Huggins-Daines, D. The CMU Sphinx Group Open Source Speech Recognition Engines. <http://cmusphinx.sourceforge.net/html/cmusphinx.php> Viewed 15 July 2009

Jin, Z. Barrière, C. 2005. Exploring Sentence Variations with Bilingual Corpora. In *Corpus Linguistics 2005 Conference*, Birmingham, United Kingdom

Jones, D. and Eisele, A. 2006. Phrase-based statistical machine translation between English and Welsh. Strategies for developing machine translation for minority languages in *5th SALTMIL workshop on Minority Languages, LREC-2006,* Genoa

Killer, M., Stüker, S., and Schultz, T. 2003. Grapheme Based Speech Recognition. In *Proceedings of the Eurospeech.* Geneva, Switzerland.

Koehn, P. 2004. Pharaoh: a beam search decoder for statistical machine translation. In *6th Conference of the Association for Machine Translation in the Americas, Lecture Notes in Computer Science. AMTA,* Springer.

Koehn, P., Hoang, H., Birch, A., Callison Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A. and Herbst, E. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL 2007, demonstration session.*

Leplus, T., Langlais, P. and Lapalme, G., 2004. Weather Report Translation using a Translation Memory. In *AMTA 2004: 154-163*

Merkel, M., Andersson, M. 2000. Knowledge-lite extraction of multi-word units with language filters and entropy thresholds. In *Proceedings of 2000 conference user-oriented content-based text and image handling* (pages 737-746), Paris, France.

Melero, M. Oliver. A and Badia, T. 2006. Automatic Multilingual Subtitling in the eTITLE Project. in *Proceedings of ASLIB Translating and the Computer 28*

Microton Intelligent Software: <http://www.eurotran.cz/> Viewed 15 July 2007

Myers, R. And Whitson, J. 1995. Hidden Markov Model for automatic speech recognition: <http://read.pudn.com/downloads71/sourcecode/graph/254695/hmm-1.03/README.hmm__.htm> Viewed 15 July 2007

Papineni, K. Roukos, S. Ward, T. Zhu, W. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation, in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, 311-318*

Piperidis, S., Demiros, I.,Prokopidis. P. 2004. Multimodal multilingual information processing for automatic subtitle generation: Resources, Methods and System Architecture (MUSA). In *Languages & The Media,* Berlin

Piperidis, S. Demiros, I. Prokopidis, P. 2006. Infrastructure for a multilingual subtitle generation system. in *Linguistics in the*

Twenty First Century, pages 369-378. Cambridge Scholars Press

Planas, E. SIMILIS. Second-Generation Translation Memory Software. in *ASLIB CONFERENCE, Translating and the Computer 27 Conference Programme*. <http://www.aslib.co.uk/conferences/ programme27.html> Viewed 15 July 2007

Phillips, J. D. 2001. The bible as a basis for machine translation. In *Proceedings of PACLing 2001*.

Prys, D., Williams, B., Hicks, B., Jones, D., Chasaide, A.N, Gobl, C., Berndsen, J., Cummins, F., Chiosáin, M. N., McKenna, J., Scaife, R., Dhonnchadha, E. U. , 2004. WISPR: Speech Processing Resources for Welsh and Irish. In *Pre-Conference Workshop on First Steps for Language Documentation of Minority Languages, 4th Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal.

Rabiner, L. R. 1989. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In *Proceedings of the IEEE*.

Salton, G. McGill, M. 1983. Introduction to Modern Information Retrieval, McGraw-Hill Book Company.

Systran < http://www.systran.co.uk/> Viewed 15 July 2009

Sysmedia <http://www.sysmedia.com/> Viewed 15 July 2009

Tr-AID Translation Memory <http://www.ilsp.gr/traid1_eng.html> Viewed 15 July 2009

Tillmann, C., S. Vogel, H. Ney, A. Zubiaga & H. Sawaf: 1997, Accelerated DP based search for statistical translation. In *Proceedings of 5th EUROSPEECH*, pp. 2667–2670.

Tyers, F. M. and Donnelly, K. 2009. *apertium-cy - a collaboratively-developed free RBMT system for Welsh to English*. In The Prague Bulletin of Mathematical Linguistics No. 91, pp. 57–66

Donnelly, K. 2008. Cyfieithu awtomatig a'r Google Summer of Code. <http://ilazki.thinkgeek.co.uk/~donnek/extras/pr2009_05_12.php> Viewed 15 July 2009

Williams, B., 1995. Text-to-speech synthesis for Welsh and Welsh English. In *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 95)*, Madrid, Spain.

Williams, B., Jones, R. J. and Uemlianin, I. 2006. Tools and resources for speech synthesis arising from a Welsh TTS project. In *Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, 24-26.

Williams, B and Jones, R. 2008. Acquiring Pronunciation Data for a Placenames Lexicon in a Less-Resourced Language. In *Proceedings of the 6th LREC (International Conference on Language Resources and Evaluation).*

Dictionary System Shell

Diccionario Sistema

Florie Moulin,
Polytech' Montpellier / Ollscoil Luimnigh,
Montpellier / Luimneach,
France / Éire
floriemoulin@gmail.com

Laura Laluque
Polytech' Montpellier / Ollscoil Luimnigh,
Montpellier / Luimneach,
France / Éire
laura_laluque11@hotmail.fr

Gearóid Ó Néill
Ollscoil Luimnigh,
Luimneach
Éire
Gearoid.oneill@ul.ie

Resumen

En los últimos años, se ha trabajado sobre los sistemas de diccionario en la Universidad de Limerick. Actualmente se está trabajando para proporcionar un sistema de diccionario que puede ser utilizado por no-especialista para generar diccionarios de "menos recursos" idiomas.

El sistema propone distintos niveles de funcionalidad, desde el acceso a través de microfichas hasta un completo acceso de texto, pasando por el acceso a los imágenes de un diccionario gracias a los números de página. El acceso al sistema se realiza desde la web.

Palabras clave

Diccionario, sistema de generación, automatización, meta-sistema.

Summary

Over the past few years, there has been work on dictionary systems here at the University of Limerick. Currently work is in progress to provide a dictionary system which can be used by non-computer specialist to generate dictionaries for "less-resourced" languages.

The system provides for different levels of functionality, from microfiche type access through page number access to images of a book dictionary through to fairly comprehensive text access, available on the web.

Keywords

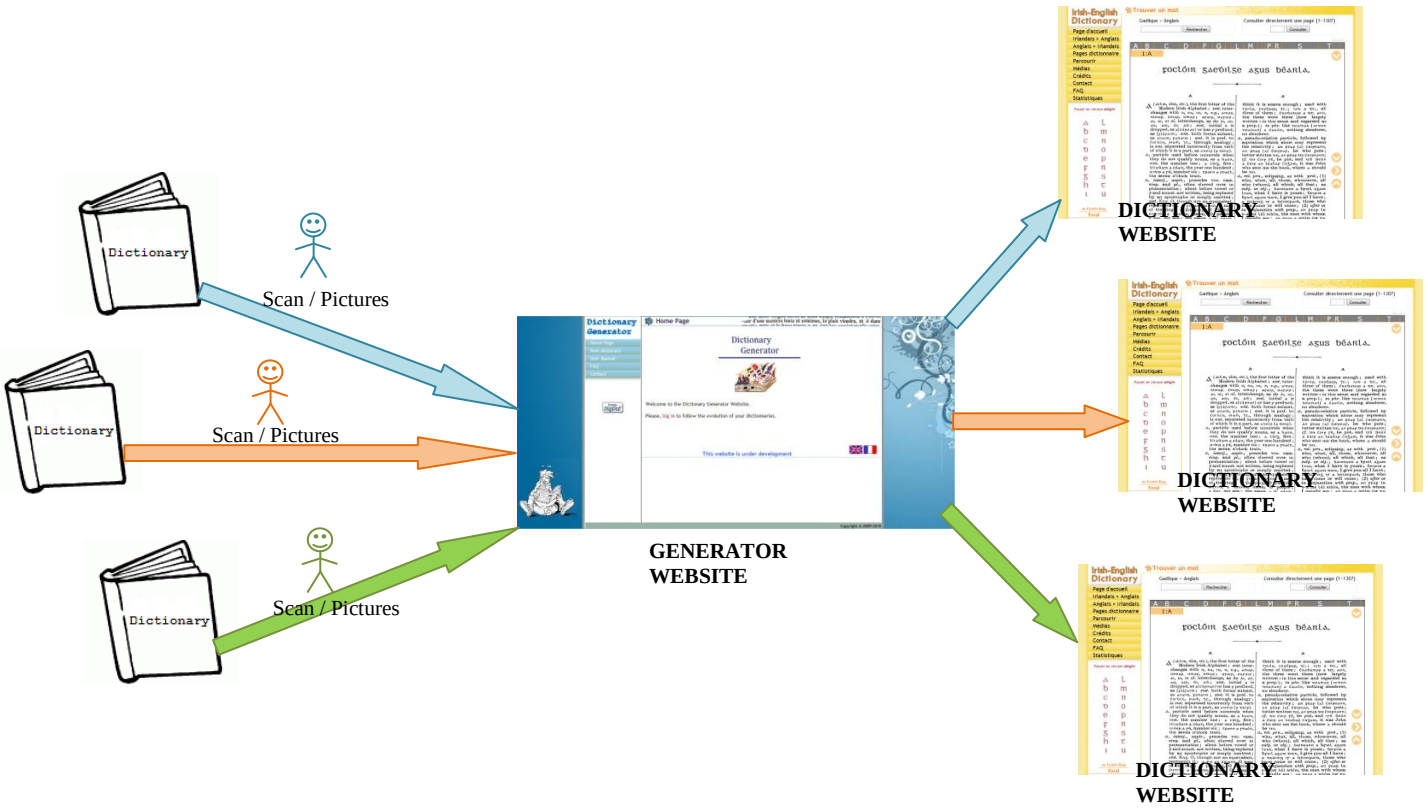
Dictionary, system generation, automation, meta-system

Background

In the early 90s a monolingual "small Irish dictionary" was published by An Gúm. McElligott et al. computerised the dictionary adding the inflected forms of the words.
It was the first mono-lingual Celtic language dictionary to be available on the world wide web. The Dinneen Irish-English dictionary, published in 1927, has been made available on the world wide web (Both are available on suitably web enabled mobile phones, for example, Samsung's Tocco).

The Dinneen dictionary is an Irish-English dictionary. It uses an Irish script for the Irish and a Latin script for the English. Irish and English words appear in the body of an entry.

The system is being adapted with a view to making a "shell" available such that a dictionary system can be set up requiring little computing expertise and be a monolingual or bilingual dictionary.



Dictionary Structure

For our purposes, a dictionary can be regarded as having the structure
(<D>|(<N,P>)|(<N,H,P>)|(<K,N,H,M>)|(<K,A,N,H,M>),

where <D> is an image of the whole dictionary;

where <N,P> is an ordered pair, for example, the page number and the page image;

where <N,H,P> is a triple, for example, N is the page number, H is the first headword on the page and P is the page image;

where (<K,N,H,M>) | (<K, A,N,H,M>) are ordered set of entries, described below.

K is a system generated key. N is the page number where the word occurs. H is a 'headword' structure and M is an 'information' structure.

K is generated to allow for (generous) increases in entries. Although once established, dictionaries tend to change rather slowly, the need for the generous allowance of keys is that it is expected, at least initially, there could be copious errors in an OCR phase. This, from our experience with the Dinneen dictionary, leads to entries being subsequently split or joined.

H is a minimalist 'headword' structure, namely,

((((W|S)(B1*))+)B2*)*

where each W, S is separated from another by B1 ‘a basic common’ separator, for example, a space and B2 an optional headword structure separator.

W is a string of the language (word) which can stand by itself. S is a substring of the language, possibly with a non-letter symbol of the language attached but which – ‘ordinarly’ – cannot stand by itself.

The headword can be null.

M purports to give information about the headword (if there is any). M itself has a structure, namely,

(I|F|E)*

where I is ‘metadata’, F is ‘definittion’ and E ‘examples’. Each piece of data within a structure may use separators. The information may be in any (identifiable) order. Grammar information, for example, could be included in I.

The meaning may be explicitly null.

<A>

The information can be ‘augmented’, with extra tables (see below), including multi-media data. This includes the facility to search by images, for example, there is a picture of a dog and on clicking on the dog, the entry for a dog appears.

Information Retrieval

From a dictionary can be generated a spell-checker. The spell-checker would then facilitate both the user in giving him or her more confidence and by improving the chances of getting results from a query.

The generation of the dictionary could be done in conjunction with a web browser, both to find putative headwords and examples of use and verification (probabilistic) of grammar.

Translation

One of the authors has an interest in automatic translation (see Ó Néill) but in this talk we are mainly concerned with the dictionary structure and one particular practical application, namely the facilitation of the computerisation of dictionaries.

Page Images

The “lesser-resourced” languages vary greatly in the state of the language in relation to published material.

<D>

For those languages with printed dictionaries, one way to make the dictionary available is to scan or photograph the pages of the dictionary. The dictionary would be simply accessible through the system, rather like the microfiche systems of old.

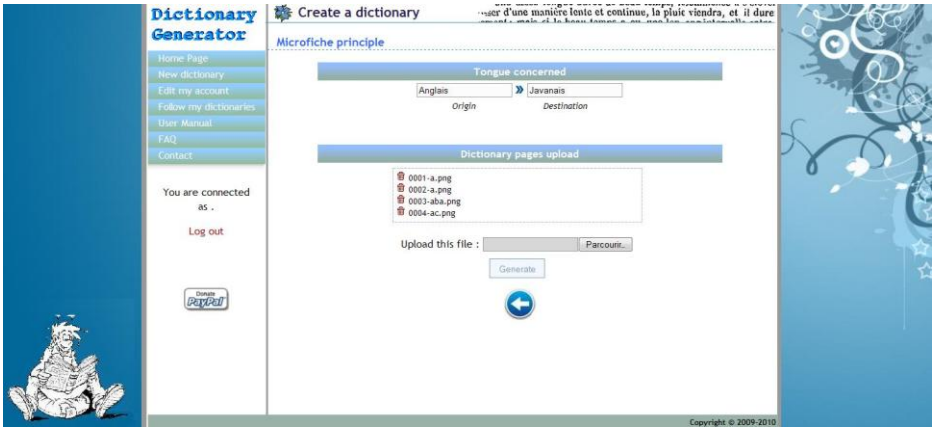


Figure 1 : User upload of dictionary pages

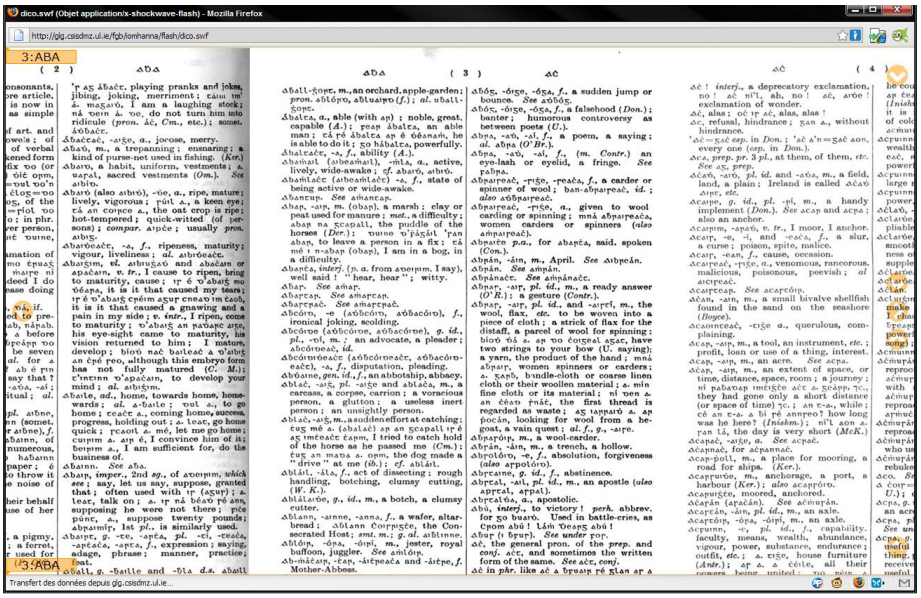


Figure 2 : Example of microfiche system (Dineen Dictionary)

<N,P>

The next stage up is to provide page numbers for the scanned pages. As well as providing the "microfiche" style of access, it also allows for access by page.

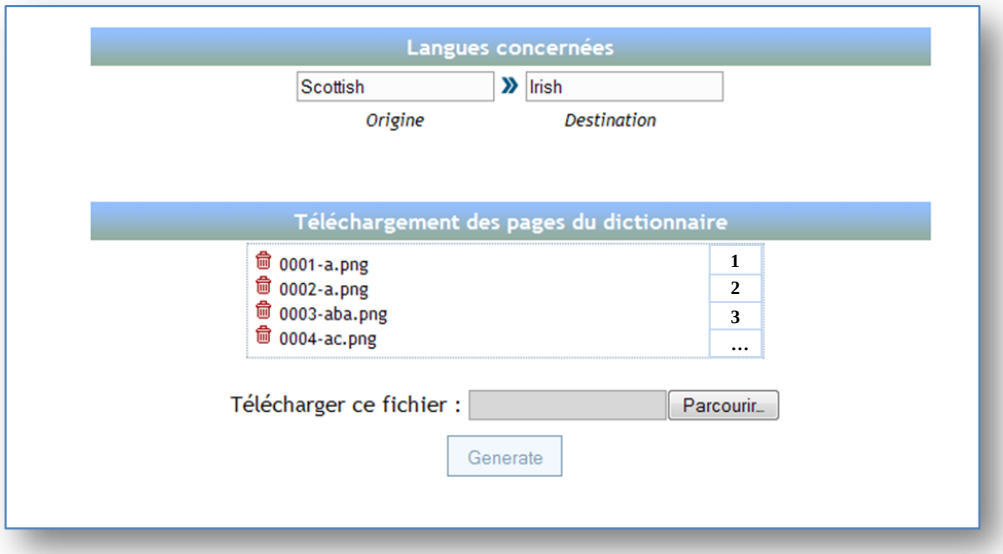


Figure 3 : User upload of pages and page numbers association

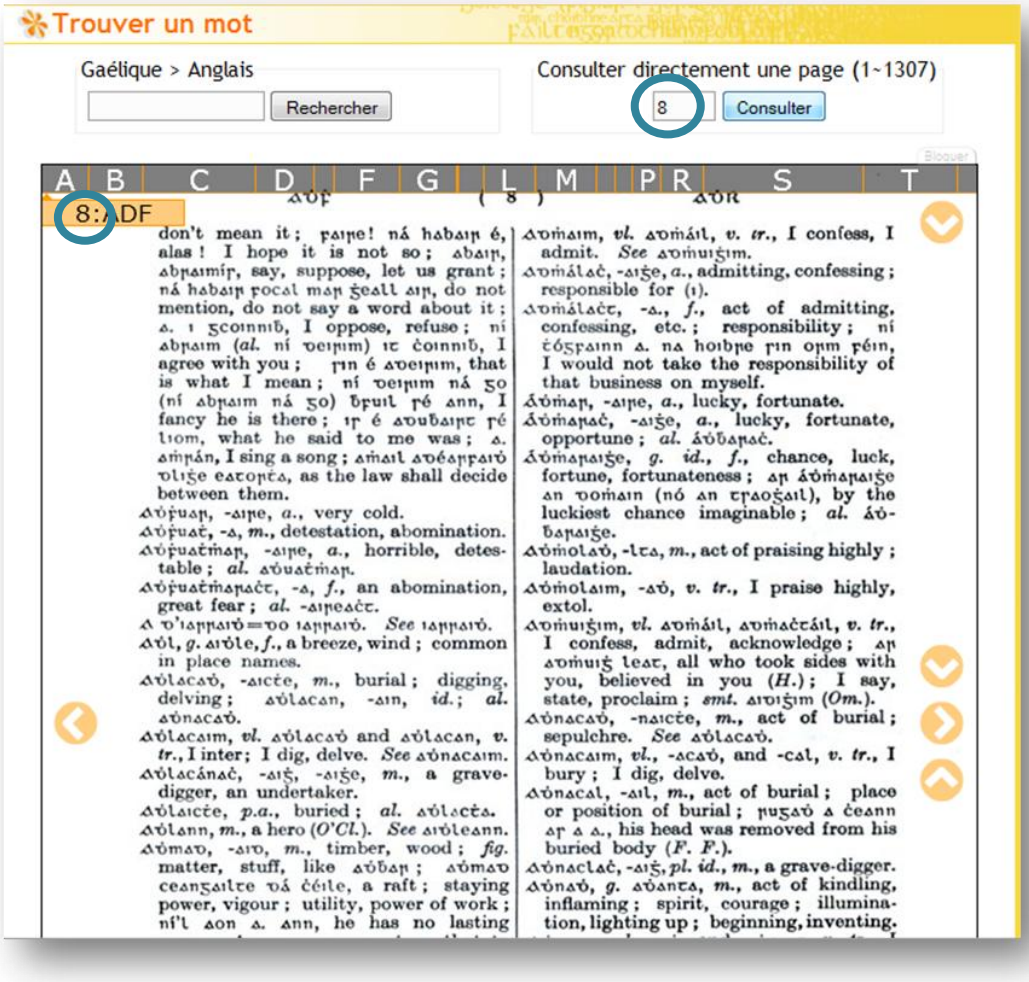


Figure 4 : Example of page access (Dineen Dictionary)

<N,H,P>

The third level of access for the scanned image is by word. This is achieved by specifying the first headword on each page, along with the page number. It requires a lexicographical (alphabetic type) ordering. When presented with a word, it finds the page within which the word would occur. It is then up to the reader to search the page visually to locate the word or determine that the word is not in the dictionary.

The above levels of dictionary usage is relatively easily and cheaply achieved – it does not require optical character recognition. It can be achieved by someone with no expertise in computing. It does require the user to be able to scan or photograph the pages and to enter them into a computer.

The user of the dictionary can search by substrings, so exact spelling is not a requirement.

Figure 5 : User upload of pages, page numbers and headwords association

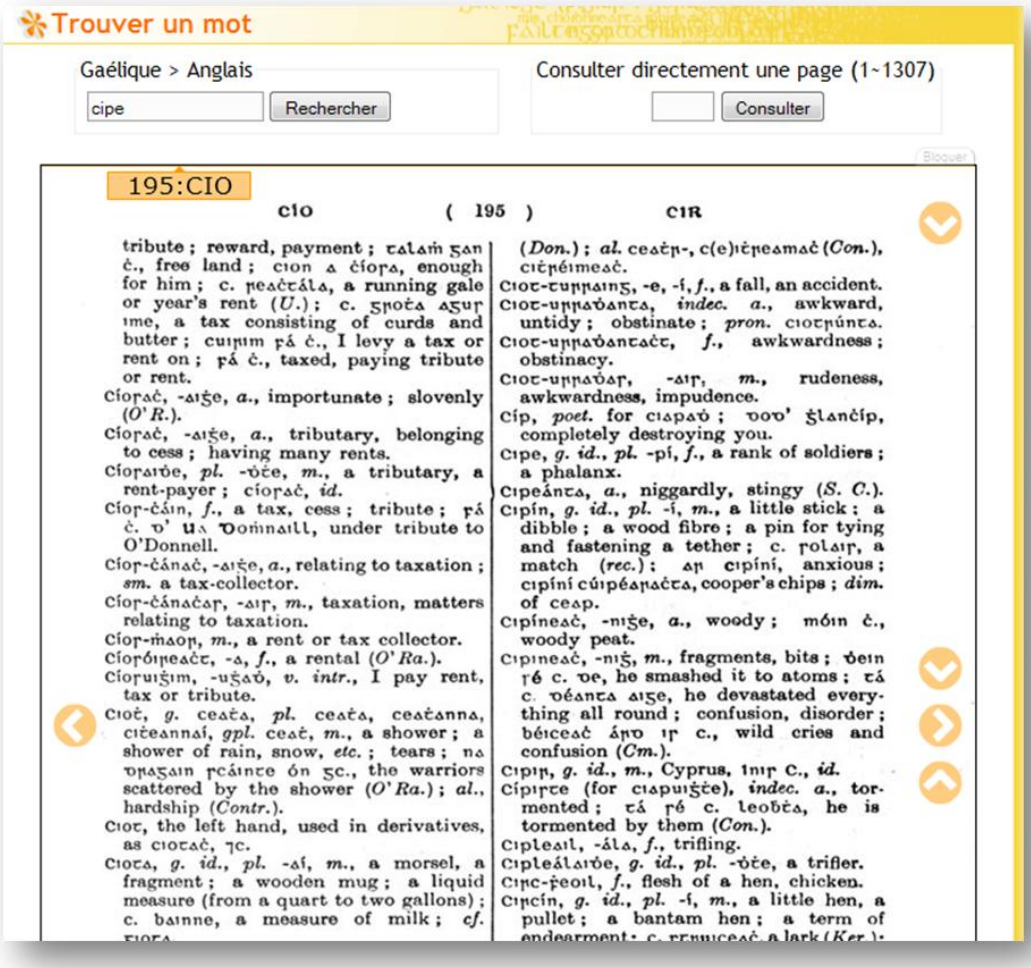


Figure 6 : Example of word access (Dineen Dictionary)

Text Entries

<K,N,H,M>

The next processing level up – perhaps ironically - is searching by text. To search by text requires either the dictionary to be typed in, collected electronically or to use optical character recognition (OCR). OCR works fairly well, especially on "well-known" alphabets and, of course, "well-resourced" languages. Although we are concerned here with "less-resourced" languages, the dictionary might well be a bilingual dictionary into or from a "well-resourced" language. If the lesser-resourced language has a distinctive script, wich is not available directly in the OCR software, then there is a likelihood of increased errors.

The system can take text entries, ranging from untagged to extensively tagged. It can also be related to the scanned images, allowing for results from comprehensive text searches to be in the form of the original scanned image.

This latter feature pre-empts, to some extent, disputes about an entry - relative to the source - being correct or not. It can also facilitate the use of a cherished script.

The system also allows for "reverse" look-up. The entry can be scanned for words or substrings and the context, as well as the headword, shown.

The user is prompted for the type of information to be presented. At the level of the scanned pages, the file name of the images. The default is the text used to name the images, a suitable extension and then the system will expect the pages to be in lexicographic order. The user is free to specify otherwise.

Figure 7 : Example of OCR use (Dinneen Dictionary)

Meta-tables

The arrangement of information for storing in a database is often done by arranging data into tables. The system allows for the user to specify a particular feature of a language to be incorporated into the actual dictionary, explicitly. This facility is also used in the Irish-English dictionary to add other facilities. It is currently being used to add an option to search by phonetic representation.

As mentioned earlier, the dictionary may be searched using pictures.

An Aid

A tool which will be supplied along with dictionary is a tool for correcting spelling, using dynamic programming and n-grams. To be used effectively, this tool would require some knowledge of the language or languages in the dictionary.

A quasi-static system, with possible errors cannot be used for automatic self-correction but some degree of error correction can be achieved. If the user has access to (correct) word lists, then quite a lot can be done by way of auto-correction.

In the Dinneen case, many errors were corrected for the English, using a source external to the Dinneen dictionary. The spelling corrector used a dynamic programming approach (see, for example, Levenshtein_distance).

Tagging

The text, if supplied untagged, will be treated as consisting of a headword followed by the body of the entry. Since reverse look-up is possible, no information is lost but information retrieval might be more difficult and less meaningful.

At the other extreme, the text may be fully tagged.

Once created, the system can be edited.

Rules for Distinguishing Languages

If the dictionary is a bilingual dictionary, with languages mixed in the body of the entry, then rules can be provided for the languages or alternatively ‘carefully’ selected samples can be provided. The rules would be of the type single, double or triple letter combinations that occur in one language but not in the other language.

Some System Details

A system prototype was developed in Prolog and then converted to MySQL because variants of SQL are more commonly taught than Prolog.

Test

Our first test will be to set up a system for a monolingual Scottish Gaelic dictionary. (It is hoped to have the first tests completed by the middle of July.)

Free

Most of the functionality described is currently available for the Dinneen Irish-English dictionary and work is in progress to make the system available as a “shell”, from which one can build a particular web-based language dictionary. It will be free to non-profit organisations.

In the first instance the ‘new’ dictionaries, will be made available, free of charge, on a Department of Computer Science server, in the University of Limerick.

Future Work

It would be interesting to develop a facility to allow for entries to be searched for by sound, graphically or by gesture.

Acknowledgements

The following have made major contributiions to the development of the system:
Etienne Guilloy, Kevin Monmousseau, Laurent Legraverand, Quentin Stratement (Ollscoil Luimnigh/IUT, Montpellier), Clément Lambilliotte, Ugo Mathieu, Christophe Clerque (Ollsoil Luimnigh/Polytech Montpellier), Raphael Plaut, (Ollscoill Luimnigh/Insa Toulouse) John Sturdy(Ollscoil Luimnigh), André LeMeur(Université de Rennes).

References

An Gúm. 1991. An Foclóir Beag, Baile Átha Cliath: An Roinn Oideachais, Dublin, Ireland.
Available at the Dinnen site or http://www.csis.ul.ie/focloir.
Dinneen, P..1927. Foclóir Gaedhilge agus Béarla. *The Irish Texts Society*, Dublin.
Available at *http://glg.csisdmz.ul.ie/fgb/iomhanna/index.php*
McElligott, A., and Ó Néill, G. 1993. A Dictionary For CALL, *EuroCall 93,* Hull:CTI Centre for Modern Languages, England.
http://en.wikipedia.org/wiki/Michael_Wideniu. 2009. MySQL. June.
http://en.wikipedia.org/wiki/ 2009. Prolog system was developed in 1972 by Alain Colmerauer and Phillipe Roussel. June.
http://en.wikipedia.org/wiki/Levenshtein_distance. 2009. June.
Ó Neill, G. 2002. Preliminary Investigation into Irish<>Manx Transcription. *NAACLT Conference*, St.Francis Xavier University, Canada.
Ó Néill, G. 2003. An Investigation into CALG (computer assisted lexicon generation).*CALICO* 2003 Conference, University of Ottawa, Ottawa, Canada.
Ó Néill, G. 2005. Comhaisnéis agus Foclóirín Gaeilge<>Manainnis. *Teangeolaíocht na Gaeilge*, DIAS, Dublin, Ireland.
Sutcliffe, R.F. E., McElligott, A. and Ó Néill, G. 1993. Irish-English Lexical Translation Using Distributed Semantic Representations. *Proceedings of the AAAI-93 Spring Symposium on Artificial Intelligence and Cognitive Science.* Queen's University, Belfast, Ireland.