Support

LREC 2002

Co-operating Organisation

ISCA SALTMIL SIG: “Speech and Language Technology for Minority Languages”

The Workshop Programme

Saturday, 1st June 2002

14:30	Registration, and preparation of posters	

Oral Session: Portability Issues in Human Language Technologies

14:50	Workshop Welcome and Introduction	Bojan Petek, University of Ljubljana, Slovenia
14:55	Multilingual Time Maps: Portable Phonotactic Models for Speech Technology	Julie Carson-Berndsen, University College Dublin, Ireland
15:20	Units for Automatic Language Independent Speech Processing	Jan Černocký, Brno University of Technology, Czech Republic
15:45	Some Issues in Speech Recognizer Portability	Lori Lamel, LIMSI-CNRS, France
16:10	Seven Dimensions of Portability for Language Documentation and Description	Steven Bird* and Gary Simons**, * Linguistic Data Consortium, University of Pennsylvania, USA ** SIL International, USA
16:35	Break	

Oral Session: HLT and the Coverage of Languages

17:00	Challenges and Opportunities in Portability of Human Language Technologies	Bojan Petek, University of Ljubljana, Slovenia
17:25	The Atlantis Observatory: Resources Available on the Internet to Serve Speakers and Learners of Minority Languages	Salvador Climent*, Miquel Strubell*, Marta Torres*, and Glyn Williams** *Universitat Oberta de Catalunya, Spain **Foundation for European Research, Wales, Great Britain
17:50	Towards the Definition of a Basic Toolkit for HLT	Kepa Sarasola, University of the Basque Country, Spain

Poster Session

18:15	Poster Session (see next page)	
20:00	End	

(continued next page)

(continued)

Poster Session

Workshop Organisers

Julie Carson-Berndsen, University College Dublin, Ireland
Steven Greenberg, International Computer Science Institute, USA
Bojan Petek, University of Ljubljana, Slovenia
Kepa Sarasola, University of the Basque Country, Spain

Workshop Programme Committee

Julie Carson-Berndsen, University College Dublin, Ireland
Steven Greenberg, International Computer Science Institute, USA
Bojan Petek, University of Ljubljana, Slovenia
Kepa Sarasola, University of the Basque Country, Spain

Table of Contents

Author Index

Multilingual Time Maps: Portable Phonotactic Models for Speech Technology

Julie Carson-Berndsen

Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
Julie.Berndsen@ucd.ie

Abstract

This paper addresses the notion of portability of human language technologies with respect to a computational model of phonology known as the *Time Map model*, focusing specifically on generic techniques for acquiring, representing and evaluating specific phonological information used by the model in multilingual speech technology applications. *Multilingual time maps* are multilevel finite state transducers which define various types of information with respect to their phonotactic context. A development environment for *multilingual time maps* is presented and an illustration of how such a multilevel finite state transducer can be constructed for a new language is given.

1. Introduction[*]

The extent to which human language technologies can be adapted for use in other application domains has become obvious in recent years with speech interfaces to a wide variety of information systems becoming more and more commonplace. However, the extent to which such technologies can be adapted to other languages, in particular minority languages, remains to be seen. Furthermore, little emphasis has been placed on developing generic technologies which can be applied not only to "new" languages but which can be employed in other task domains. While it is often considered practical to have a speech recognition system for a language if one is about to embark on developing a speech synthesis system for that language - the recognition can support data annotation - the fact that the linguistic knowledge which is being represented for the one task domain could be relevant for the other is largely ignored. In order to address the issue of portability adequately, linguistic representations must be integrated more explicitly into human language technologies. While a hidden Markov model can be trained as a speech recognizer perhaps for any language given a large data set, there is no potential for exploiting the commonalities of human languages or for using the same knowledge in the synthesis task domain.

This paper addresses the notion of portability with respect to a computational model of phonology known as the *Time Map model* focusing specifically on generic techniques for acquiring, representing and evaluating different types of phonological information used by the model in multilingual speech technology applications. Ubiquitous language technology concerns the development of language technologies for different purposes on different platforms so that they can be made available to everybody at all times rather than to a select group for specific purposes. Much of the further development of the *Time Map model* is aimed towards providing fine-grained representations for speech recognition and synthesis and developing computational models which will contribute to achieving this long-term goal. The techniques presented in this paper make way for the extension of current speech technology to languages which have received little attention thus far by modeling linguistic information at various levels of granularity.

In the context of this paper, portability refers to extending the functionality of a system to cater for another language. It does not cover issues such as re-applying the technology to new content domains (e.g. adapting an English spoken language interface for a football results information system to an English information system for accommodation in London). The model described below is not restricted to a specific application domain and therefore the main concern is adapting the system to another language. This involves parameterization of the system so that the language-specific components can be substituted in a "plug and play" fashion.

In the next section, the *Time Map model* is sketched briefly with particular attention to the language-specific knowledge components. Section 3 discusses how the model has been parameterized to allow extension to other languages by defining the notion of *multilingual time maps* specifying information at different levels of granularity and a development environment, *PhonoDeSK*, for acquiring and evaluating such *time maps* is presented. Section 4 describes an example illustrating the role which can be played by *PhonoDeSK* in the context of portability of human language technologies and section 5 concludes with some comments on future work.

2. Time Map Model

The *Time Map model* was proposed as a computational linguistic model for speech recognition by Carson-Berndsen (1998, 2000) and has been tested within a speech recognition architecture for German. More recently, the model has been extended to English and has been provided with an interface which allows users to define and evaluate phonotactic descriptions for other languages and sublanguages (Carson-Berndsen & Walsh, 2000). In extending the model to cater for

[*] This research was part funded by the HEA under the MMRP programme.

English, particular emphasis was placed on parameterizing the model so that knowledge components for other languages could readily be substituted.

The original motivation for the design of the *Time Map model* was to address specific problems in the area of speech recognition below the level of the word. In particular, the problem of out-of-vocabulary items, also termed the “new word” problem, is addressed explicitly in the model. This is done by including *complete* phonotactic descriptions of a language which describe not only those forms specified in some corpus lexicon, but also all potential forms which adhere to the phonotactic constraints imposed by the language. Another specific problem addressed by this approach is the modelling of coarticulation phenomena. This is done by assuming a non-segmental approach to the description and interpretation of speech utterances which avoids having to segment an utterance into non-overlapping units at any level of representation.

The *Time Map model* has two main language-specific components: the *phonotactic automaton* and the *time map lexicon*. These components each assume a particular representation of speech utterances in terms of a multilinear representation of features similar to an autosegmental score.

2.1. Multilinear Representations

Speech utterances are defined in the model in terms of a multilinear representation of tiers of features which are associated with signal time. The notion of tiers of features is not new in the area of phonology (cf. for example Goldsmith, 1990). However, recently there has been a significant upsurge in phonetic feature extraction and classification, and automatic transcription using the type of features proposed in our model (e.g. Chang, Greenberg & Wester (2001), Ali et al., (1999)). An example multilinear event representation using the Chang, Greenberg & Wester (2001) features is depicted in figure 1.

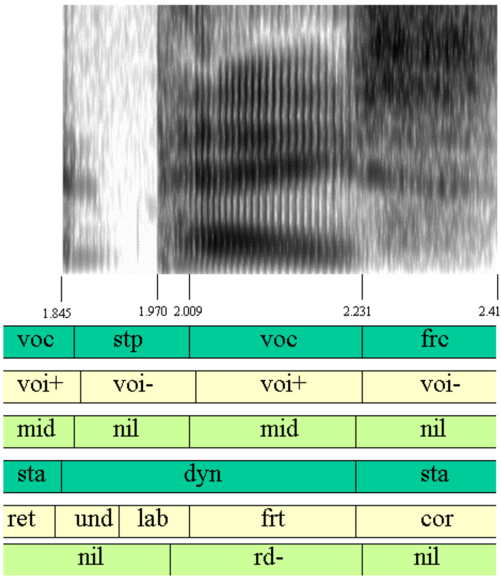


Figure 1: Multilinear representation of the word *pace*

As can be seen from figure 1, each feature in a multilinear event representation is associated with a specific tier (on the vertical axis) and with a specific time interval in terms of milliseconds (on the horizontal axis). The features do not all start and end simultaneously. An overlap of properties (coarticulation) exists in any time interval; for example, the feature *rd-* begins before the *voc* feature indicating that the lips have been spread during the plosive (*stp*) anticipating the following nonround vowel.

A multilinear event representation of a speech utterance is in fact highly constrained. It is not the case, that any combination of features can occur in any order. The allowable combinations of features are dictated partly by the phonological structure of the language, as defined by the phonotactics, and partly by predictable phonetic variation, which often results from limitations associated with human speech production.

2.2. Phonotactic Automata

The primary knowledge component of the *Time Map model* is a complete set of phonotactic constraints for a language which is represented in terms of a finite state automaton. A *phonotactic automaton* describes all permissible sound combinations of a language within the domain of a syllable. It can be phoneme-based (just specifying phonemes), feature-based (generalizing over phonemes) or event-based (specifying constraints on temporal relations between the features). A subsection of a phonotactic automaton depicting CC- clusters in English syllable onsets can be seen in figure 2.

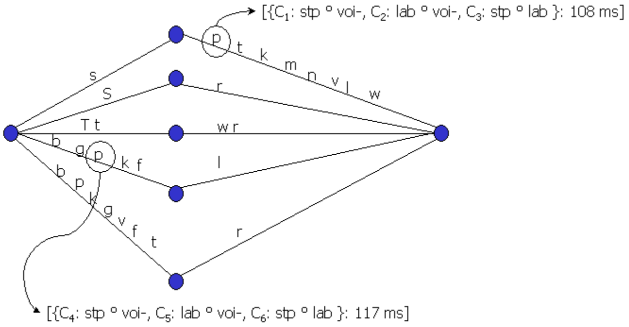

[{C_1: stp ° voi-, C_2: lab ° voi-, C_3: stp ° lab }: 108 ms]
s
S
p t k m n v l w
r
T t
w r
b g p k f
l
b p k g v f t
r
[{C_4: stp ° voi-, C_5: lab ° voi-, C_6: stp ° lab }: 117 ms]

Figure 2: Subsection of a phonotactic automaton

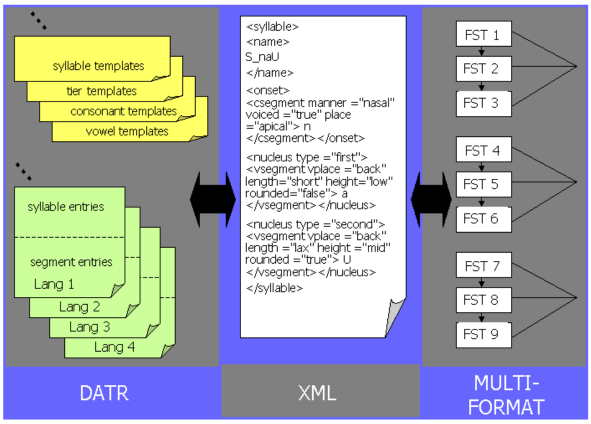
The arcs in an event-based phonotactic automaton define a set of constraints on overlap relations which hold between features in a particular phonotactic context (i.e. the structural position within the syllable domain).[1] In the phonotactic automaton of figure 2, the constraint C_1*: stp ° voi-*, for example, states that the feature *stp* (a plosive) on the manner tier should overlap the feature *voi-* (voiceless) on the phonation tier. The millisecond values refer to the average durations for the sounds in this particular phonotactic context which have been calculated from a large corpus.

[1] The monadic symbols written on the arcs in figure 2 are purely mnemonic for the feature overlap constraints they represent; the ° symbol represents the overlap relation.

2.3. Time Map Lexicon

The *time map lexicon* defines fully specified multilinear event representation of each syllable in the corpus (or each lexicalised syllable in the language) together with their phonemic and orthographic forms. The *time map lexicon* is used online by the model to distinguish between actual and potential syllables and used offline for evaluation purposes with respect to a particular corpus.

The *time map lexicon* is compiled from a generic lexicon model (Carson-Berndsen, 1999). Generic lexical information is represented in DATR, a simple language designed specifically for lexical knowledge representation that allows the definition of nonmonotonic inheritance networks with path/value equations (cf. Evans & Gazdar, 1996). Varying degrees of granularity (syllables, tiers in a multilinear representation, consonants, vowels etc.) are specified as templates in DATR. For each language (see figure 3) specific word, syllable and segment inventories are defined which contain information such as frequency and average duration. Specific entries inherit regularities and sub-regularities from the templates while exceptions are specified in the entries themselves. Either individual or cascades of finite state transducers are then applied to generate individual lexicons for speech applications in an application specific format (cf. Cahill, Carson-Berndsen & Gazdar, 2000).



```
<syllable>
<name>
S_naU
</name>
<onset>
<csegment manner ="nasal"
voiced ="true" place
="apical"> n
</csegment></onset>
<nucleus type ="first">
<vsegment vplace ="back"
length="short" height="low"
rounded="false"> a
</vsegment></nucleus>
<nucleus type ="second">
<vsegment vplace ="back"
length ="lax" height ="mid"
rounded ="true"> U
</vsegment></nucleus>
</syllable>
```

Figure 3: Generic lexicon architecture

2.4. Speech Recognition with the Time Map Model

In the context of speech recognition, input to the model is a multilinear representation of a speech utterance in terms of absolute time events, i.e. features with start and end points which are extracted from the speech signal. Phonological parsing in the *Time Map model* is guided by the *phonotactic automaton* which provides top-down constraints on the interpretation of the multilinear representation, specifying which overlap and precedence relations are expected by the phonotactics. If the constraints are satisfied, the parser moves on to the next state in the automaton. Each time a final state of the automaton is reached, a well-formed syllable has been found. This well-formed syllable may be underspecified, however, since some of the constraints in the phonotactic automaton may have been relaxed. It is then compared with a fully specified multilinear representation in the *time map lexicon* which allows the system to distinguish between actual (lexicalised) and potential syllables. The architecture of the model in the context of speech recognition is depicted in figure 4.

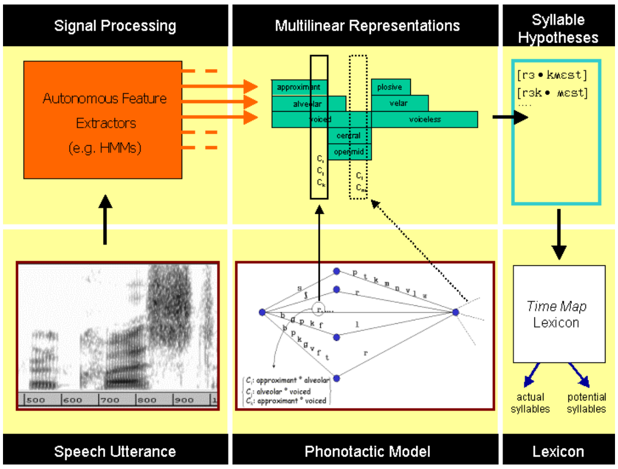


Figure 4: *Time Map* speech recognition architecture

Speech synthesis based on the *Time Map model* is also currently under investigation (see Bohan et al., 2001). This involves generating multilinear representations from a lexical representation of an utterance using a cascade of finite state transducers mapping from phonemes to allophones to event representations. The aim of this research is to investigate the application of the *Time Map model* in the synthesis domain and to a language with a significantly different phonology, namely Irish. The methodology used to port the model to Irish is discussed in section 4 below.

3. Multilingual Time Maps

The *Time Map model* has been parameterized to allow the language-specific components to be substituted by components describing other languages. There are two issues involved in this process. The first issue concerns how to represent the language-specific information in a uniform way so that it can be used immediately by the model and also be made available for use with other technologies. The second issue concerns the acquisition of the language-specific components. In what follows, both of these issues are discussed in turn with respect to the *phonotactic automaton* and the *time map lexicon*. The language-specific configuration of the model is defined by a *multilingual time map*. The *time map* defines mappings between different types of information and constraints on overlap relations between features. It is termed *multilingual* because on the one hand, it provides a framework for developing the language-specific knowledge components for the *Time Map model* either by using knowledge of a related language already available to the system to predict the relevant structures of a “new” language or by learning these directly. On the other hand, it has a uniform structure which allows for cross-language comparisons and the generation of *time maps* which cover a number of languages.

3.1. Representation

A *multilingual time map* comprises language-specific information at various levels of granularity represented as a multilevel finite state transducer. The advantage of this representation is that it is declarative, bidirectional and efficient to process. The multilevel finite state transducer can be viewed as an extension of the phonotactic automaton to include (at least) the following levels:

1. Graphemes
2. Phonemes
3. Allophones
4. Features
5. Constraints on Overlap Relations
6. Average Duration
7. Frequency
8. Probability

Each arc specifies information on all of these levels (although some of this information may not be available in all cases but can be readily updated at any time). For example, figure 5 depicts a single arc of a *multilingual time map* for English.

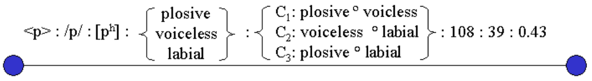


Figure 5: Arc in a *multilingual time map*

The first level is the grapheme level, the second is the phoneme level, the third is the allophone level (i.e. *p* is aspirated in this phonotactic context). The fourth level is the feature level specifying the features for this phoneme (which can be selected from a number of different possible feature sets). The fifth level specifies the constraints on the overlap relations between the features; the sixth level specifies the average duration of the [p] in this phonotactic context. The seventh level specifies the frequency of this sound in this phonotactic context and the eighth level specifies the probability of the arc.

A generic transducer interpreter[2] is used to extract the levels required for different purposes from the *multilingual time map*. To construct the phonotactic automaton for a speech recognition application of the *Time Map model*, for example, the generic transducer interpreter takes level 2 as input and outputs level 5 and level 6 from the transducer. Note that it is also possible to map between other levels in the transducer to obtain other types of information (e.g. input graphemes and output phonemes; input phonemes and output allophones etc.).

3.2. Acquisition

The real challenge for the portability of the *Time Map* model lies in acquisition of the *multilingual time maps* (i.e. not just to be able to use them to generate the *phonotactic automaton* and the *time map lexicon* for a particular language, but to be able to construct them efficiently). *PhonoDeSK* (see figure 6) is a suite of tools which has been designed specifically for acquiring and evaluating *multilingual time maps* (see Ashby, Carson-Berndsen & Joue, (2001 for an initial specification). These tools are used by *PhonoDeSK agents*[3] which collaborate with each other in order to define an optimal phonological description of the language.

PhonoDeSK foresees three strategies for structured data acquisition; user-driven, data-driven and data-driven with user prompting. That is to say, *multilingual time maps* can either be produced manually by a trained linguist or can be learned from a data set with or without user intervention. *PhonoDeSK* is web-based and can thus be accessed anywhere at any time. The user is also viewed as an agent in the context of *PhonoDeSK* – the *verification agent*.

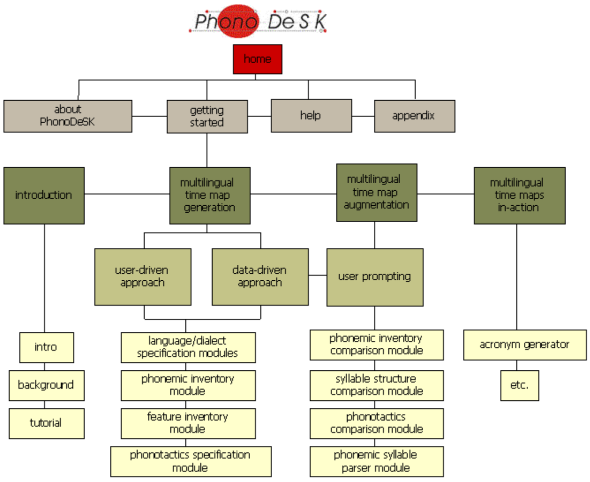

PhonoDeSK
home
about PhonoDeSK
getting started
help
appendix
introduction
multilingual time map generation
multilingual time map augmentation
multilingual time maps in-action
user-driven approach
data-driven approach
user prompting
intro
background
tutorial
language/dialect specification modules
phonemic inventory module
feature inventory module
phonotactics specification module
phonemic inventory comparison module
syllable structure comparison module
phonotactics comparison module
phonemic syllable parser module
acronym generator
etc.

Figure 6*: PhonoDeSK*

When constructing a *multilingual time map* for a "new" language, a number of inventories are created by an *inventory agent:*

1. Phoneme Inventory
2. Allophone Inventory
3. Feature Inventory
4. Syllable Inventory

In each case, any available resources may be used directly. For example, a phonemically labeled data set can be used to extract the phoneme inventory and, together with a *learning age*nt for phonotactic automata, PAL (Kelly, 2001), to predict the syllable inventory. Using an existing *multilingual time map* for a related language, predictions may be made which can be accepted or rejected by a native speaker (*verification agent*) of the language. The acceptances/rejections are then incorporated into the learning procedure. If no resources whatsoever are available for the "new" language then much more manual input is required by the user. The first pass *multilingual time map*

[2] This has been implemented by Robert Kelly, University College Dublin.

[3] The notion of agent will not be discussed further in this paper. Further details on the agent approach assumed here can be found at http://said.ucd.ie.

constructed using *PhonoDeSK* will be, in general, underspecified on some of the transducer levels.

This section has discussed *multilingual time maps*, how they are represented and how they can be acquired for new languages. The next section illustrates how a *multilingual time map* can be constructed using *PhonoDeSK agents,* taking Irish as an example.

4. An Example

In this section, an example of a *multilingual time map* is presented which has been constructed from an initial corpus of tri-syllabic Irish words. This is work in progress and therefore as stated above not all levels in the *time map* are fully specified at present. In *PhonoDeSK*, a *phonotactic agent* and a *lexicon agent* collaborate with each other and with other learning and generalization agents to construct a *multilingual time map* which can be used with the *Time Map model* for speech recognition and synthesis.

4.1. Observation

The corpus was recorded and labeled phonetically including syllable boundaries and a distinction between stressed and unstressed syllables was made. The phoneme inventory and the feature inventory were specified manually by an expert on Irish phonology (a human *verification agent* for Irish). The corpus was input to the *learning agent*. The first pass produced a deterministic phonotactic automaton which included average durations for each sound in each phonotactic context, frequency of each sound in each phonotactic context with respect to the corpus and a probability of each arc. The initial *multilingual time map* for Irish specifies levels 2 to 8.

Since this initial *multilingual time map* specifies all the forms in the corpus, it automatically contains all the forms which should be included in the *time map lexicon*. Here a distinction is being made between a corpus lexicon and a complete lexicon of the Irish language. The *lexicon agent* uses all the information which is available in the *multilingual time map* to define new syllable and segment entries together with syllable, tier, consonant and vowel templates in the DATR-based lexicon.

4.2. From Observation to Generalization

The initial *multilingual time map* is input to a *generalization agent* which extends and optimizes the *time map*. The *generalization agent* interacts with 4 other agents until an optimal description is reached.

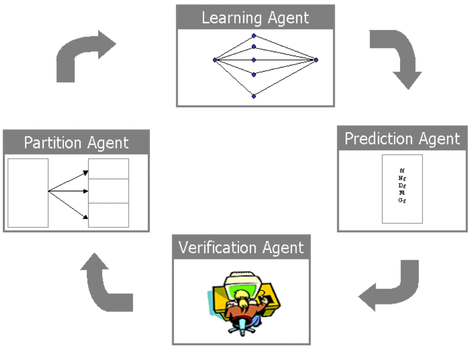


Figure 7: Generalisation cycle

For the purposes of generalization over substructures, the data is partitioned into initial consonant cluster (onset), vowel and final consonant cluster (rhyme) and further into CC- onsets; CCC- onsets etc. A phonotactic automaton for each of the partitions is learned separately by the *learning agent*. An example for the resulting structure for stressed syllable CC- onsets in Irish is depicted in figure 8, whereby only level 2 (phonemes) is specified. Note that Irish has both palatalized (represented in this figure by uppercase) and plain consonants.

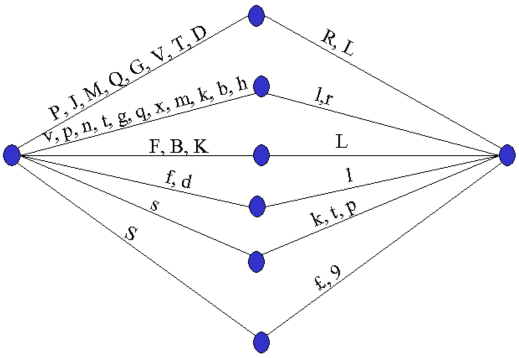


Figure 8: Learned CC- onset of Irish syllable

This onset automaton has been learned purely on the basis of the data set. It does not claim to cover all CC- onsets of Irish, only those represented in the data. In order to extend this to a complete onset description, (idiosyncratic) gaps must be identified in the representation which could also be permissible onsets of the language. There are two methods for identifying idiosyncratic gaps in the automaton. The first involves examining general distributional properties evident in the automaton. To the eye, one possible gap is obvious: there is a path representing the combinations [fl] and [dl] and [f] and [d] stand out as being the only plain consonants followed by [l] but not followed by [r] in the onset. The *prediction agent* identifies such gaps and the combinations [fr] and [dr] are presented to native speakers of the language (*verification agents*) to verify whether these are permissible combinations or not. The arcs in the *multilingual time map* are then generalized with respect to the feature level (level 4) in order to determine the commonalities between the phonemes in a particular phonotactic context. The *prediction agent* requests a *phonoclass agent* to group phonemes into natural classes (based on the intersection of the their features). Using the complete phoneme and feature inventories, the *prediction agent* presents other phonemes which are part of the natural class but are not found in that phonotactic position to the *verification agent*.

This is performed until all partitions of the data have been generalized or until the verification agent decides that the *multilingual time map* is optimal[4].

[4] In this context, optimal means deemed suitable for use in some application.

4.3. Alternative Routes

The description of the *multilingual time map* construction using *PhonoDeSK* has assumed thus far that phonemically labeled data is available as a starting point for learning. Clearly, this will not always be the case and there are two alternative routes which can be taken. Firstly, it may be possible to predict using the *multilingual time map* of a related language, what phonotactic combinations are permissible in the new language. These can be input directly to the *prediction agent* and the forms can be accepted or rejected by the *verification agent*. Secondly it is also possible for the user to specify the canonical form of the syllable and the phoneme inventory of the "new" language and this will be used by the *prediction agent* to elicit permissible combinations from the *verification agent*.

There are a number of additional tools available in *PhonoDeSK* which support the *verification agent* in constructing a new *multilingual time map* from an existing *time map*: for example, all possible forms represented by the *time map* can be generated; two descriptions may be compared directly with each other at the phoneme and syllable level (cf. Ashby, Carson-Berndsen & Joue, 2001); a single parse or all possible parses of a given phonemic representation can be generated and presented to the user for verification. The *verification agent* thus provides important information for preferences which are used in turn to update the probability of a particular parse.

The level which remains to be included in the *multilingual time map* after generalization of the phonotactics is complete is the grapheme level (level 1). This task is performed by the *lexicon agent*. The phonemic forms are presented to the *verification agent* to elicit a correct orthographic form for the lexicon, possibly using the *prediction agent* to suggest mappings based on the original corpus. Once the orthographic forms are available in the lexicon, the *phonotactic agent* requests the *learning agent* to learn the grapheme-phoneme mapping of the words in the corpus. This can later be used to estimate a grapheme-phoneme mapping for new forms.

5. Conclusion

This paper has presented the concept of a *multilingual time map* which has evolved out of a desire for portability of a computational phonological model for use in various human language technology tasks. The development environment, *PhonoDeSK*, has been designed specifically for acquiring, representing and applying phonological information at various levels of granularity. It combines finite state techniques with automatic and manual data acquisition through the use of agents which collaborate to instantiate the various levels of the *multilingual time map*. The *multilingual time maps* have been designed specifically for use with the *Time Map model* but they also represent an important step on the road to the realisation of ubiquitous language technology in general, by providing a framework which allows portability to new languages. However, the information represented in the *multilingual time maps* can be used directly by other technologies for structural fine tuning. Future work is concerned with extending *PhonoDeSK* agents and with applying the technology to other languages.

6. References

Ali, A.M..A.; J. Van der Spiegel; P. Mueller; G. Haentjaens & J. Berman (1999): An Acoustic-Phonetic Feature-Based System for Automatic Phoneme Recognition in Continuous Speech. In: *IEEE International Symposium on Circuits and Systems (ISCAS-99)*, III-118 - III-121, 1999.

Ashby, S.; J. Carson-Berndsen, & G. Joue (2001): A testbed for the development of multilingual phonotactic descriptions. In: *Proceedings of Eurospeech 2001*, Aalborg.

Bohan, A.; E. Creedon, J. Carson-Berndsen & F. Cummins (2001): Application of a Computational Model of Phonology to Speech Synthesis, In: *Proceedings of AICS2001*, Maynooth, September 2001.

Cahill, L; J. Carson-Berndsen & G. Gazdar (2000), Phonology-based Lexical Knowledge Representation. In: F. van Eynde & D. Gibbon (eds.) *Lexicon Development for Speech and Language Processing*, Kluwer Academic Publishers, Dordrecht.

Carson-Berndsen, J. (1998): *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition*. Kluwer Academic Publishers, Dordrecht, 1998.

Carson-Berndsen, J. (2000): Finite State Models, Event Logics and Statistics in Speech Recognition, In: Gazdar, G.; K. Sparck Jones & R. Needham (eds.): *Computers, Language and Speech: Integrating formal theories and statistical data*. Philosophical Transactions of the Royal Society, Series A, 358(1770), 1255-1266.

Carson-Berndsen, J. (1999): A Generic Lexicon Tool for Word Model Definition in Multimodal Applications. *Proceedings of EUROSPEECH 99, 6th European Conference on Speech Communication and Technology*, Budapest, September 1999

Carson-Berndsen, J. and Walsh, M. (2000): Generic techniques for multilingual speech technology applications, *Proceedings of the 7th Conference on Automatic Natural Language Processing*, Lausanne, Switzerland, 61-70.

Chang, S.; S. Greenberg & M. Wester (2001): An Elitist Approach to Articulatory-Acoustic Feature Classification. In: *Proceedings of Eurospeech 2001*, Aalborg.

Evans, R & G. Gazdar (1996), DATR: A language for lexical knowledge representation. In: *Computational Linguistics* 22, 2, pp. 167-216.

Goldsmith, J. (1990): *Autosegmental and Metrical Phonology*. Basil Blackwell, Cambridge, MA.

Kelly, R. (2001): PAL: Phonotactic Automaton Learning. Technical Report, Department of Computer Science, University College Dublin.

Units for Automatic Language Independent Speech Processing

Jan Černocký

Brno University of Technology, Faculty of Information Technology
Božetěchova 2, 61266 Brno, Czech Republic
cernocky@fit.vutbr.cz

Abstract

Many current systems for automatic speech processing rely on sub-word units defined using phonetic knowledge. Our paper presents an alternative to this approach – determination of speech units using ALISP (Automatic Language Independent Speech Processing) techniques. Such units were experimentally tested in a very low bit rate phonetic vocoder, where mean bit rates of hundreds bps for unit encoding were achieved. Improvements of the proposed coder and some links to "classical" approaches of speech synthesis are discussed. Based on the results of comparison of an ALISP segmentation with a phonetic alignment, we comment on the potential use of automatically derived units in speech recognition, speaker verification and language identification.

1. Introduction

The International Phonetic Association (IPA) sets up as one of its objectives the definition of a symbolic representation of speech for any of the speakers of any language in the world: the International Phonetic Alphabet[1]. However, despite efforts devoted to this topic, some substantial problems persist in the adequacy of this alphabet for spoken speech.

Recent advances in ALISP (Automatic Language Independent Speech Processing) (Chollet et al., 1999) led us to the idea of defining such a set of units *automatically*, without an a-priori knowledge; to let it emerge uniquely from the speech data. For this purpose, a number of tools which proved their efficiency in automatic speech processing (coding, recognition, synthesis, language identification, speaker verification) have been developed: temporal decomposition (TD), non-supervised clustering, Hidden Markov Models (HMM) and others. Basic information about these tools with references are given in Section 2.

On contrary to IPA, where it is difficult to find an objective criterion, the set of units can be evaluated using *very low bit rate (VLBR) speech coding* at about 200 bps (Černocký et al., 1998). At these rates, a symbolic representation of the incoming speech is required. If the decoded speech is intelligible, one must admit that the symbolic representation is capable of capturing the significant acoustic-phonetic structure of the message. Moreover, the coding rate in bps and dictionary size give an idea of *efficiency* of the description while the quality of decoded speech is related to its *precision*. Section 3. gives an overview of our VLBR coding experiments in three languages and their results. It also contains a description of recent advances in VLBR coding using ALISP units.

However, the domain with the greatest need of optimized and automatically derivable units is the large vocabulary continuous speech recognition (LVCSR) based in current systems on phones or their derivatives (context-dependent phones, syllables). Section 4. presents a comparison of two alignments of data: phonetic and ALISP in terms of a confusion matrix. It also contains some reflections on encoding of target vocabulary using data-driven units.

Section 5. contains conclusions and some comments on the use of ALISP units in other domains (speaker verification and language identification).

2. ALISP tools

Classical speech processing suffers from the need of large phonetically labeled, or at least orthographically annotated corpora. This is making the current algorithms unpractical when used in a condition for which a database does not exist, or is very costly (rare language, environmental noise, channel, application domain). The main goal in ALISP processing is to find data-driven units with as little supervision as possible. We will see, that for coding, the process can be fully automated. Steps that should be taken to use such units in recognition are discussed in section 4. The expected result of unit creation is:

- a set of units (that can be compared to a set of phonemes).
- labeling of the training data using those units.
- models of units to detect them automatically in unseen data.

The tools used to find units are:

The *temporal decomposition (TD)* is a representative of algorithms able to detect quasi-stationary parts in the parametric representation of speech. This method, introduced by Atal (Atal, 1983) and refined by Bimbot (Bimbot, 1990), approximates the trajectories of parameters $x_i(n)$ by a sum of m *targets* a_{ik} weighted by *interpolation functions* (IF):

$$\hat{x}_i(n) = \sum_{k=1}^{m} a_{ik}\phi_k(n), \quad \text{for} \quad i = 1, \ldots, P, \qquad (1)$$

where P is the dimension of the parameter vectors. Equation 1 can be written:

$$\begin{array}{cccc} \hat{\mathbf{X}} & = & \mathbf{A} & \mathbf{\Phi} \\ (P \times N) & & (P \times m) & (m \times N), \end{array} \qquad (2)$$

[1] http://www.arts.gla.ac.uk/IPA/ipa.html

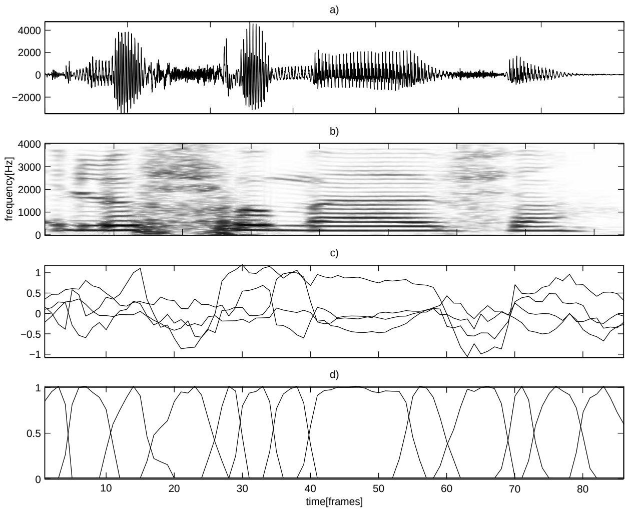


Figure 1: Illustration of temporal decomposition of French word "le chômage": a) signal. b) spectrogram. c) trajectories of first 4 LPCC parameters. d) TD interpolation functions.

where the lower line indicates matrix dimensions. The initial interpolation functions are found using local Singular Value Decomposition with adaptive windowing, followed by post-processing (smoothing, decorrelation and normalization). Target vectors are then computed by: $\mathbf{A} = \mathbf{X}\mathbf{\Phi}^{\#}$, where $\mathbf{\Phi}^{\#}$ denotes the pseudo-inverse of IFs matrix. IFs and targets are locally refined in iterations minimizing the distance of $\mathbf{X}$ and $\hat{\mathbf{X}}$. Intersections of interpolation functions permit to define speech segments. An example of TD can be seen in Fig. 1. Critically speaking, any of automatic segmentation procedures, based for example on spectral variation function (SVF), could be used. We chose TD because the algorithm and software were readily available in the lab.

Unsupervised clustering assigns segments to classes. Vector quantization (VQ) is used for automatic determination of classes: class centroids are minimizing the overall distortion on the training set. The VQ codebook $\mathbf{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_L\}$ is trained by K-means algorithm with binary splitting. Training is performed using vectors positioned in gravity centers of TD interpolation functions, while the *quantization* takes into account entire segments using cumulated distances between all vectors of a segment and a code-vector. TD with VQ produce a phone-like segmentation of speech.

Hidden Markov models (HMM) can be used to model the units. HMM parameters are *initialized* using context-free and context-dependent Baum-Welch training (Young et al., 1996) with TD+VQ transcriptions, and *refined* in successive steps of corpus segmentation (using HMMs) and model parameters re-estimation. The speech represented by observation vector string can then be aligned with models by standard likelihood maximization. At this point, we obtain the three desired outputs of the unit determination algorithm: units, their models and training data alignments. The units can be used in further processing.

3. Very low bit rate coding

The VLBR coding using ALISP units has been the first verification of our approach. It turned out however, that after some modifications to improve the output speech quality (discussed later in this section), it can have potential applications.

The coder performs the *recognition* of input unseen speech into ALISP units, that we call *coding units*. For the *synthesis* in the decoder, however, another type of units called *synthesis units* can be defined – units can be for example designed in such a way that the synthesis units spans a speech segment between two spectrally stable parts, so that the concatenation becomes easier. Finally, the decoder must dispose of a certain number of *representatives* of each synthesis unit. The coder must send the index of best-matching representative (DTW-distance was used as distortion measure) and information on the prosody: timing and pitch and energy contours.

The *decoder* receives the information on coding units and derives the information on synthesis units, then it retrieves the representative from its memory. The synthesis modifies the prosody of the representative and produces output speech.

3.1. Basic coding tests

This approach was first tested in speaker-dependent experiments on American English (Černocký, 1998), French (Černocký et al., 1998) and Czech (Černocký et al., 1999). The speech parameterization was done by a set of LPC-cepstral coefficients on 20 ms frames with 10 ms frame-shift. Temporal decomposition was set to produce 15–17 targets per second in average (corresponding to average phoneme rate). The VQ codebook had 64 code-vectors that were trained using the original vectors (not TD-targets) located in gravity centers of TD interpolation functions. After initial labeling using the TD+VQ tandem, first "generation" of HMMs (3 emitting states, no state-skip, single-Gaussian) was trained. The training corpus was aligned with those models, and 5 iterations of retraining-alignment were run.

In the coding, synthesis units corresponded to the coding ones, and for each, 8 longest representatives were searched in the training data. The number of bits per unit was therefore $\log_2 64$ (unit) + $\log_2 8$ (representative) = 9. This led to the average bit rate for unit encoding of 100–200 bps. The prosody was not coded in those experiments, and the physical synthesis in the decoder was done by a rudimentary LPC synthesizer.

Intelligible speech was obtained for the three languages – low speech quality was attributed mainly to rudimentary LPC synthesis rather than the units themselves. Those experiments justified our approach – they proved that a "phonetic-like" speech coder can be trained without ever seeing any transcriptions of the speech data.

3.2. Harmonic Noise Model synthesis

In basic structure of the coder, LPC synthesis has been used to produce the output speech. It was found to be highly responsible for the low quality of the resulting speech (that can be proved by a copy LPC analysis-synthesis). Therefore, the Harmonic-Noise Model (HNM) which brings

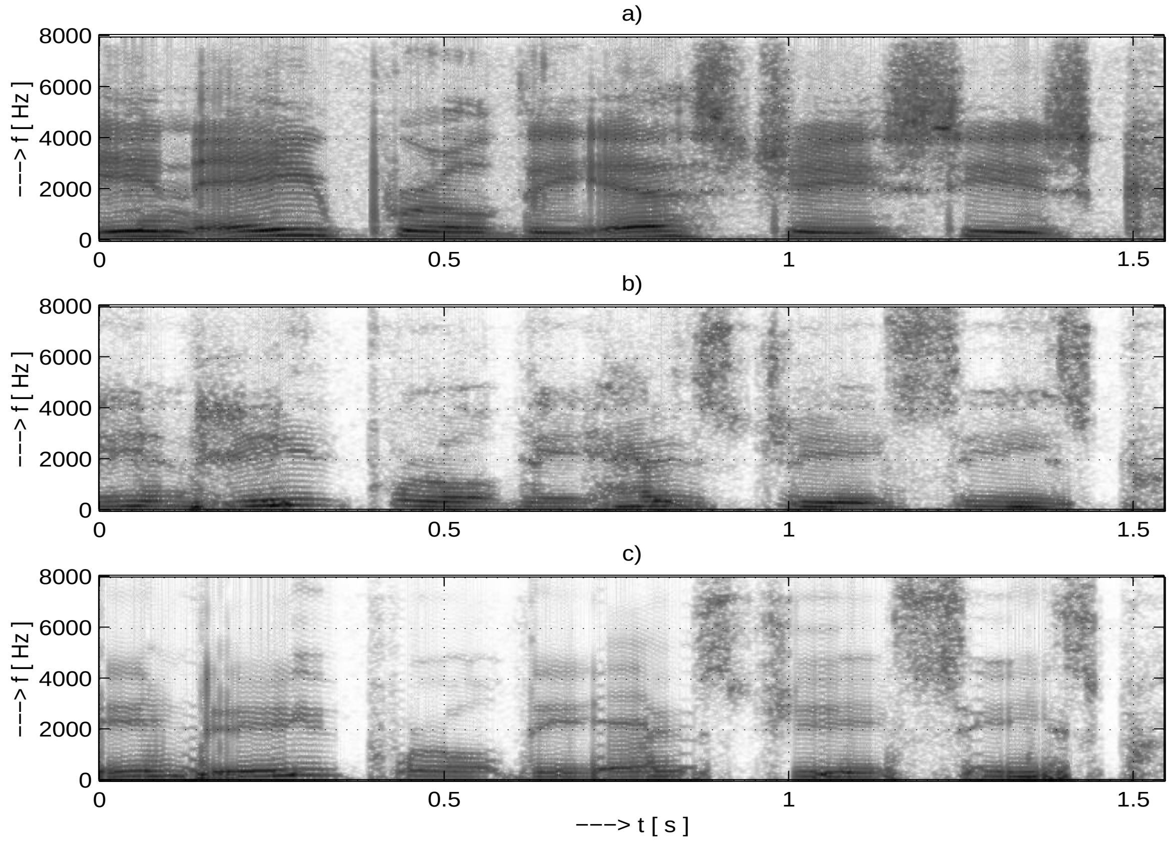


Figure 2: Spectrograms. a) original speech signal. b) coded speech synthesized by HNM. c) coded speech synthesized by LPC.

much higher quality of the synthesized speech, is applied. The principle of HNM is in detail described in (Oudot, 1996; Stylianou, 1996). The HNM is built on following representation of signal $x(n)$:

$$x(n) = \underbrace{\sum_{k=1}^{P} \alpha_k \cos(2\pi f_k n - \phi_k)}_{\text{Harmonics}} + \underbrace{b(n)}_{\text{Noise}} , \tag{3}$$

where P is the number of harmonics, α_k are the amplitudes, f_k the multiples of pitch and ϕ_k phases of harmonic part. $b(n)$ expresses components of noise.

Eq. 3. describes both parts of HNM. The first part "Harmonics" decomposes the speech signal into a sum of sinusoids. In fact, a combination of harmonically related and non-harmonically related sinusoids can also be used. "Noise" in Eq. 3 represents non-harmonic part of speech signal. The parameters for the noise and harmonic part are estimated separately. The fundamental frequency estimation is isolated from the estimation of amplitudes and phases and the interdependence of the parameters in neighboring frames is alleviated through the hypothesis of the quasi-stationary signal. Thus, the first step of the analysis process consists of estimating the fundamental frequency for the voiced frames. In our work, a classical method based on normalized cross correlation function (NCCF) (Talkin, 1995) has been applied.

The estimation of amplitudes and phases of the harmonics is done using the method of least mean squares (Charbit and Paulsson, 2000). The noise of *voiced frames* is obtained by subtracting the previously computed harmonics from the input signal. Its spectrum is modeled by LPC auto-recursive filter of 12^{th} order. In *unvoiced frames*, only parameters of the noise model are estimated. Auto-recursive filter of 12^{th} order is used, as above. In synthesis, the source signal is represented by white noise filtered by the estimated LPC filter.

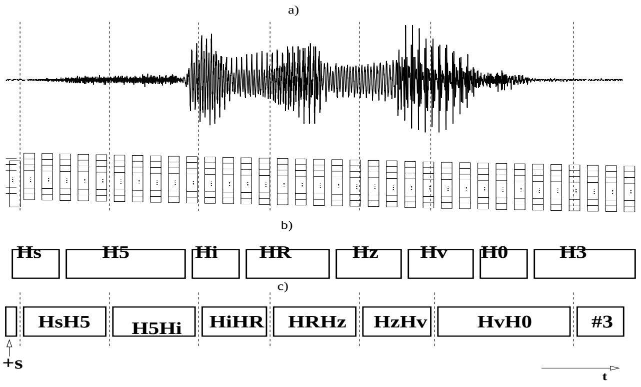


Figure 3: Example of re-segmentation according to middle frames of original units. Minimal length of new units is 4 frames: a) speech signal with its splitting into the frames. b) original segmentation recognized by HMMs. c) new re-segmentation.

Spectral envelope is needed to perform pitch modification in the synthesized speech. The log of spectral envelope is computed from the estimated amplitudes of the harmonics using real-cepstrum coefficients (Charbit and Paulsson, 2000).

The results (Motlíček et al., 2001) have demonstrated, that the replacement LPC synthesis by HNM version is highly responsible for great improvement of quality of resulting speech, as can be seen in Fig. 2, where spectrograms from the same part of speech signal are compared[2].

3.3. Synthesis units

The units initialized by the temporal decomposition are inherently unstable at their boundaries (remember, that the center of TD-units tends to be stable). Such units are therefore not very suitable for synthesis as they do not have good concatenation properties. We have therefore tested two approaches to make *synthesis units* units closer to diphone-based or corpus-based speech synthesis.

First, *selection of longer synthesis units* based on the original coding ones was tested (Motlíček et al., 2001). This approach is illustrated in Fig. 3. These long units can be constructed by aggregation of short ALISP coding units with re-segmentation in spectrally stable parts of the extremity units. The synthesizer is similar to a diphone one. The results were however not satisfactory, some concatenation noise was still audible and due to the limitation of the training corpus, some synthesis units were missing and difficult to replace.

Therefore, a different method called *short synthesis units with dynamic selection* was developed (Baudoin et al., 2002). Here, for each ALISP class, a large number of representatives is extracted from the training corpus. These synthesis representatives are determined in order to fulfill criteria of good representation of a given segment to be coded and criteria of good concatenation of successive segments.

For each coding unit H_j, we define sub-classes called H_iH_j containing all the speech segments of class H_j that were preceded by a segment belonging to the class H_i in the training corpus. It is possible to keep as synthesis representatives all the segments of the training corpus organized in classes and sub-classes as described above or to limit the size of each sub-class to some maximal value K.

During coding, if a segment is recognized as belonging to class H_j and is preceded by a segment in class H_i, the representative is searched in the subclass H_iH_j of class H_j. The selection of the best representative in the sub-class is done on the distance D_C of good representation of the segment. The D_C distance is based on a spectral comparison by DTW between the segment to code and the potential synthesis representatives. The distance D_C can also include a distance on prosody parameters.

We have verified that this approach provides superior speech quality than the "short" coding units or re-segmented longer ones.

3.4. Toward speaker independent ALISP coder

First results of speaker-independent (SI) coding on large French database BREF have been reported in (Baudoin et al., 2002). Coding units were trained on 33 male speakers from this corpus, and the corresponding representatives were selected from all available speakers in similar fashion to speaker-dependent coding presented in section 3.1. The resulting speech was intelligible, though with lower quality than the speaker-dependent counterpart. This confirmed the possibility to use the ALISP scheme also in SI environment.

Two problems are crucial for the SI operation: speaker clustering or speaker adaptation in the coder and voice modification in the decoder. For the first problem, the TSD article (Baudoin et al., 2002) presents speaker-independent coding with VQ-based speaker clustering. Here, the reference speakers are *pre-clustered*, in order to select the closest speaker or the closest subset of speakers for HMM refinements and/or adaptation of synthesis units. A VQ-based inter-speaker distance using the unsupervised hierarchical VQ algorithm was used (Furui, 1989). The basic assumption is that training speech material from the processed speaker is available during a short training phase for running the VQ adaptation process. The inter-speaker distance is defined as the cumulated distance between centroids of the non-aligned code-books, using the correspondence resulting from the aligned code-books obtained through the adaptation process. This distance is used in the off-line pre-training phase for clustering the reference speakers, and during the on-line training phase for selecting the closest cluster to the user. From the distance matrix, subclasses are extracted using a simplified split-based clustering method.

The proposed concept has been validated on the BREF corpus (phonetically balanced sentences), 16 LPCC coefficients and 64 classes were used. Illustration of the clustering process is given for the largest class, (left panel of Fig. 4), a typical class (middle panel) and an isolated speaker (right panel) in terms of relative distance to the other speakers. One could note the similar positioning of speakers belonging to the same cluster.

The obtained results in terms of speaker clustering using a small amount of data are encouraging. In our future works, we will study a speaker-independent VLBR structure derived from this concept, by adding HMM adaptation at the encoder, and voice conversion techniques at the decoder.

4. ALISP units in recognition

4.1. ALISP–phonetic correspondence

To investigate the potential usability of ALISP units in speech recognition, we performed several experiments on the comparisons of ALISP and phonetic alignments (Černocký et al., 2001).

Such alignments were available with the Boston University corpus of American English (a database that we used for the initial VLBR coding experiments). They were obtained at BU using a segmental HMM recognizer constrained by possible pronunciations of utterances (Ostendorf et al., 1995). The measure of correspondence was the relative overlap r of ALISP unit with a phoneme (see Fig. 5 for illustration). The results are summarized in *confusion*

[2] http://www.fee.vutbr.cz/∼motlicek/speech_hnm.html contains examples of speech after coding/decoding.

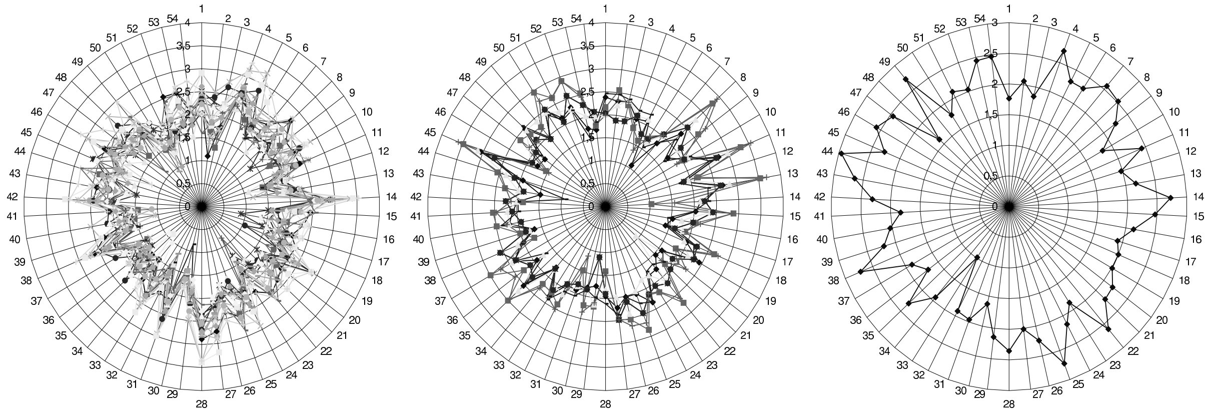


Figure 4: Left panel: Relative distance of speakers from the largest cluster. Middle panel: Relative distance of speakers from a typical cluster (indexes 6, 14, 21, 29, 31, 43). Right panel: Relative distance of speakers from an isolated speaker (index 33).

matrix $\mathbf{X}$ ($n_p \times n_a$), whose elements are defined:

$$x_{i,j} = \frac{\sum_{k=1}^{c(p_i)} r(p_{i_k}, a_j)}{c(p_i)}. \qquad (4)$$

n_p and n_a are respectively the sizes of phoneme and ALISP unit dictionaries, p_i is the i-th phoneme, a_j is the j-th ALISP unit, $c(p_i)$ is the count of p_i in the corpus and $r(p_{i_k}, a_j)$ is the relative overlapping of k-th occurrence of p_i with ALISP unit a_j. The columns of $\mathbf{X}$ are rearranged to let the matrix have a quasi-diagonal form. As for the phoneme set, on contrary to BU alignments, where stressed vowels are differentiated from unstressed ones, we used the original TIMIT set. The ALISP set had 64 units. The resulting matrix is given in Fig. 6.

This matrix shows, that the correspondence between ALISP units and phonemes is consistent, but not unique. We can for example see, that the ALISP unit `a` corresponds to closures, but also to the pause. The unit `$` has a strong correlation with `SH` but it is also linked to its voices counterpart `ZH` and to affricates `JH` et `CH`, which are acoustically very closed.

4.2. Using ALISP units in recognition

Although the above mentioned experiments showed a correlation of phonemes and ALISP units, an ALISP recognition system should probably not be based on direct phoneme–ALISP mapping. Stochastic mapping of *sequences* of phonemes to *sequences* of ALISP units would be one solution. This approach was studied in (Deligne, 1996): likelihood maximization is applied to joint segmentation of two streams of observations, where the first can be the phonetic labeling and the second the sequence of automatically derived units. When the testing data are processed, the method finds the segmentation together with optimal "transcription" into sequences of phonemes. The observations can be either symbolic (in this case, the method is "discrete") or vectorial (here, not only statistics of sequences, but also emission PDFs come into mind).

Another option is the *composition* of ALISP units into word and phone models, proposed in (Fukada et al., 1996). Here, the basic units are first derived in an unsupervised manner. Then, phonetic transcription is compared to ALISP segmentation and composite models are constructed for the *words* of the corpus. In case the data do not contain sufficient number of examples of a word, the method can "back-up" to *phoneme models* composed in similar manner as the word ones.

Third solution was proposed for triphone models, but it would generalize well also with ALISP units. This approach does not require phonetic transcriptions but a large database with word boundaries. ALISP labels are generated for this DB and the ALISP-pronunciation dictionary is created. It is however necessary to develop an expert system for the transcription of unseen words in terms of ALISP units.

5. Conclusions

The algorithm of unit search produces set of consistent units but is far from optimal. As for the feature extraction, we have for example not investigated the perceptually motivated features used by Hermansky and his group (Hermansky, 1997). The distance used in VQ could be replaced by the Kullback-Leibler one, that has shown superior performances in selection of units for synthesis (Stylianou and Syrdal, 2001). The training of unit models could be done completely without initialization of time boundaries (currently temporal decomposition) and of labels (VQ) by using an Ergodic Hidden Markov model (EHMM) for both tasks simultaneously. Finally, it is necessary to think about "shaping" the units for the target application.

The first part of the paper demonstrates that *speech coding*, at transmission rate lower than 400 bps, can be achieved using automatically derived units. The drawback of our proposal is the size of the memory required both in coder and decoder and the delay introduced by the maximal duration of the segments (several hundreds msec). There are many applications which could tolerate both a large memory (let say 200 Mbytes) and the delay. Among such applications are the multimedia mobile terminal of the future (including the electronic book), the secured mobile phone, the compression of conferences (including distance education), etc. More work is necessary on voice transformation so that only typical voices will be kept in memory.

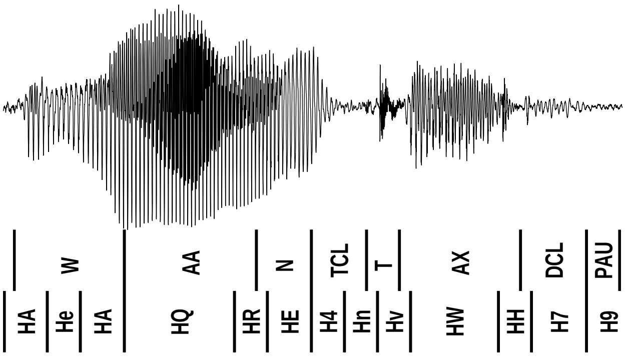


Figure 5: Illustration of comparison of ALISP and phonetic segmentations: word “wanted” from female speaker of Boston University corpus.

Characterization of a voice based on limited data and use of this characterization to transform another voice is an interesting research topic.

As for the *recognition*, we can conclude that building of ALISP-based recognizer will not be a straightforward task. The invested efforts should however be generously recompensed by the limitation of human efforts needed to create or modify such a recognizer. If we obtain an efficient scheme, and in the same time we succeed in limiting the human labor (annotations, pronunciation dictionaries, etc.), it will be a great step toward the *real automating* of speech processing, and it will also open the way to its easier implementation in different languages.

ALISP unit use should not be limited to coding or recognition. In (Petrovska-Delacrétaz et al., 2000), we have reported results of a speaker-verification system with pre-segmentation in ALISP units before actual scoring. The performance of our system on 1999 NIST data was not optimal, but we believe that pre-segmentation of speech into classes and determination of their speaker-characterization performances can aid the verification system. Results obtained from class-specific models can be then combined using appropriate weighting factors before taking the decision.

The last proposed application domain is the *language identification*. Most current language identification (LI) systems are based on the approach of extracting the phonotactic language specific information. The phonotactics is related to the modeling of the statistical dependencies inherent in the phonetic chains. Unfortunately, transcribed databases should be available to train the required phonetic recognizer, and the transcription step is a major bottleneck for the adaptation of systems to new languages or services (as it is for the other domains). We propose to replace the widely used phonetic-based recognizers by an ALISP-based recognizer, and to extract from the automatically segmented speech units the necessary information for solving the problem of language identification. The advantage of the proposed method is its portability to new languages, for which we do not have annotated databases.

6. Acknowledgments

Several colleagues have contributed to this article. Fréderic Bimbot (IRISA Rennes, France) contributed the `td95` package used for the temporal decomposition. Geneviève Baudoin (ESIEE Paris) supervised initial stages of the work and with Fadi El Chami (Université Libanaise Tripoli and ESIEE), she is the author of dynamic unit selection for the synthesis. Petr Motlíček (VUT Brno) worked on the creation of long synthesis units and HNM synthesis. François Capman (Thales Communications) is the author of VQ-based speaker clustering. Dijana Petrovska-Delacrétaz (University of Fribourg, Switzerland) did the speaker-independent experiments and is working on ALISP units deployment in speaker verification. Maurice Charbit and his students (ENST Paris) contributed to the HNM synthesis. Vladimír Šebesta and Richard Menšík (VUT Brno) helped with the visualization of confusion matrices. Finally, thanks to Gérard Chollet (ESNT Paris) for having been at the original ALISP idea and for many constructive comments and supervision.

This research has been supported by Grant Agency of Czech Republic under project No. 102/02/0124 and by the RNRT (Réseau National de Recherche en Télécommunications) project SYMPATEX.

7. References

B. S. Atal. 1983. Efficient coding of LPC parameters by temporal decomposition. In *Proc. IEEE ICASSP 83*, pages 81–84.

G. Baudoin, F. Capman, J. Černocký, F. El Chami, M. Charbit, and Gérard Chollet. 2002. Advances in very low bit rate speech coding using recognition and synthesis techniques. In *submitted to TSD 2002*, Brno, Czech Repbublic, September.

F. Bimbot. 1990. An evaluation of temporal decomposition. Technical report, Acoustic research department AT&T Bell Labs.

M. Charbit and N. Paulsson. 2000. Boîte à outils harmoniques plus bruit. Technical report, ENST Paris, France, November.

G. Chollet, J. Černocký, A. Constantinescu, S. Deligne, and F. Bimbot. 1999. Towards ALISP: a proposal for Automatic Language Independent Speech Processing. In K. Ponting, editor, *Computational models of speech pattern processing*, NATO ASI Series, pages 375–388. Springer Verlag.

S. Deligne. 1996. *Modèles de séquences de longueurs variables: Application au traitement du langage écrit et de la parole*. Ph.D. thesis, École nationale supérieure des télécommunications (ENST), Paris.

T. Fukada, M. Bacchiani, K. Paliwal, and Y. Sagisaka. 1996. Speech recognition based on acoustically derived segment units. In *Proc. ICSLP 96*, pages 1077–1080.

S. Furui. 1989. Unsupervised speaker adaptation method based on hierarchical spectral clustering. In *Proc. ICASSP'89*, pages 286–289.

H. Hermansky. 1997. Should recognizers have ears? In *Proc. Tutorial and Research Workshop on Robust speech recognition for unknown communication channels*, pages 1–10, Pont-a-Mousson, France, April. ESCA-NATO.

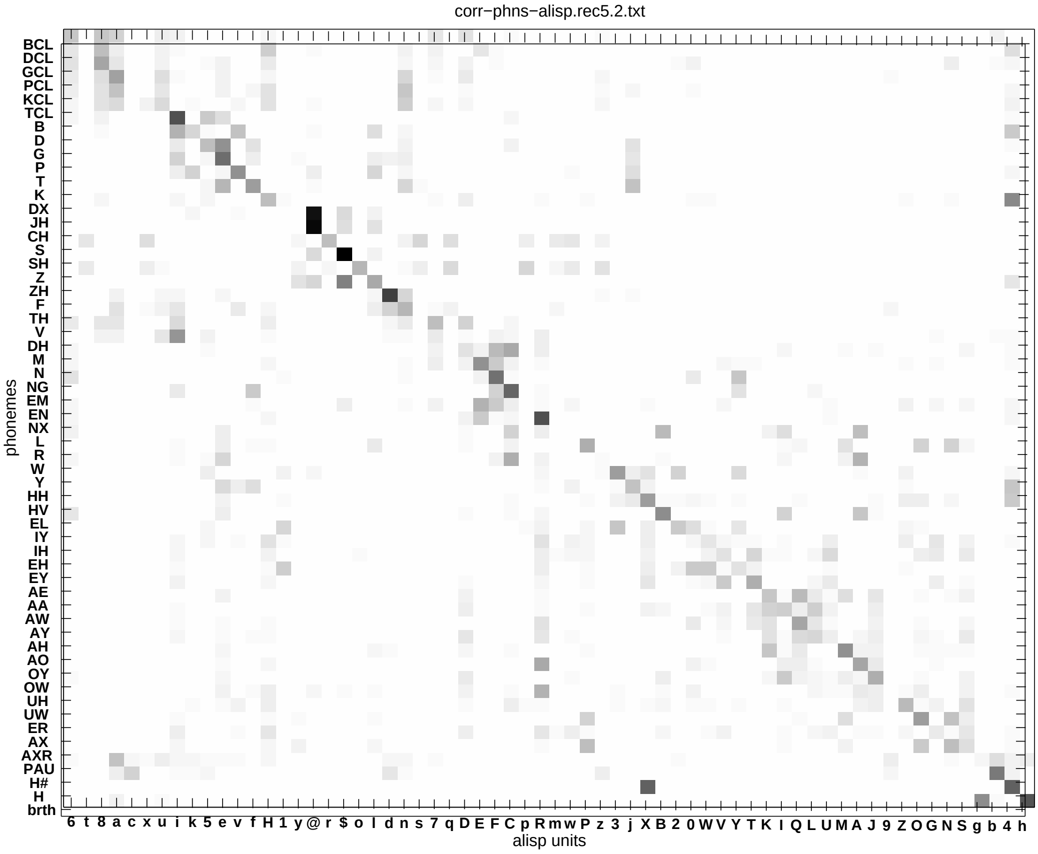


Figure 6: Correspondence of ALISP segmentation and phonetic alignment for speaker F2B in BU corpus. White color corresponds to zero correlation, black to maximum value $x_{i,j}$=0.806

P. Motlíček, J. Černocký, G. Baudoin, and G. Chollet. 2001. Minimization of transition noise and HNM synthesis in very low bit rate speech coding. In V. Matoušek, P. Mautner, P. Mouček, and K. Taušer, editors, *Proc. of 4th International Conference Text, Speech,Dialogue - TSD 2001*, number 2166 in Lecture notes in artificial intelligence, pages 305–312, Železná Ruda, Czech Republic, September. Springer Verlag.

M. Ostendorf, P.J. Price, and S. Shattuck-Hufnagel. 1995. The Boston University radio news corpus. Technical report, Boston University, February.

M. C. Oudot. 1996. *Etude du modèle "Sinusoides et bruit" pour le traitement des signaux de parole, estimation robuste de l'enveloppe spectrale*. Ph.D. thesis, École nationale supérieure des télécommunications (ENST), Paris.

D. Petrovska-Delacrétaz, J. Černocký, J. Hennebert, and G. Chollet. 2000. Segmental approaches for automatic speaker verification. *Digital Signal Processing*, 10:1–3, January/April/July. Special Issue: NIST 1999 Speaker Recognition Workshop.

Y. Stylianou and A.K. Syrdal. 2001. Perceptual and objective detection of discontinuities in concatenative speech synthesis. In *Proc. ICASSP'01*, volume 2, pages 837–840, Salt Lake City, Utah, USA.

I. Stylianou. 1996. *Modèles harmoniques plus bruit combinés avec des méthodes statistiques, pour la modification de la parole et du locuteur*. Ph.D. thesis, École nationale supérieure des télécommunications (ENST), Paris, January.

D. Talkin. 1995. A robust algorithm for pitch tracking (rapt). In W. B. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis*, New York. Elseviever.

J. Černocký, G. Baudoin, and G. Chollet. 1998. Segmental vocoder - going beyond the phonetic approach. In *Proc. IEEE ICASSP 98*, pages 605–608, Seattle, WA, May. http://www.fee.vutbr.cz/˜cernocky/Icassp98.html.

J. Černocký, I. Kopeček, G. Baudoin, and G. Chollet. 1999. Very low bit rate speech coding: comparison of data-driven units with syllable segments. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Proc. of Workshop on Text Speech and Dialogue (TSD'99)*, number 1692 in Lecture notes in computer science, pages 262–267, Mariánské Lázně, Czech Republic, September. Springer Verlag.

J. Černocký, G. Baudoin, D. Petrovska-Delacrétaz, and G. Chollet. 2001. Vers une analyse acoustico-phonétique de la parole indépendante de la langue, basée sur ALISP. *Revue Parole*, 2001(17,18,19):191–226.

J. Černocký. 1998. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*. Ph.D. thesis, Université Paris XI Orsay, December.

S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. 1996. *The HTK book*. Entropics Cambridge Research Lab., Cambridge, UK.

Some Issues in Speech Recognizer Portability

Lori Lamel

Spoken Language Processing Group,
LIMSI-CNRS, France
lamel@limsi.fr

Abstract

Speech recognition technology has greatly evolved over the last decade. However, one of the remaining challenges is reducing the development cost. Most recognition systems are tuned to a particular task and porting the system to a new task (or language) requires substantial investment of time and money, as well as human expertise. Todays state-of-the-art systems rely on the availability of large amounts of manually transcribed data for acoustic model training and large normalized text corpora for language model training. Obtaining such data is both time-consuming and expensive, requiring trained human annotators with substantial amounts of supervision. This paper addresses some of the main issues in porting a recognizer to another task or language, and highlights some some recent research activities aimed a reducing the porting cost and at developing generic core speech recognition technology.

1. Introduction

Speech recognition tasks can be categorized by several dimensions: the number of speakers known to the system, the vocabulary size, the speaking style, and the acoustic conditions. Concerning speakers, the most restrictive is when only one speaker can use the system and the speaker is required to enroll with the system in order to be recognized (speaker-dependent). The system may be able to recognize speech from several speakers, but still requires enrollment data (multiple speaker) or the system can recognize the speech from nominally any speaker without any training data (speaker-independent).

A decade ago the most common recognition tasks were either small vocabulary isolated word or phrases or speaker dependent dictation, whereas today speech recognizers are able to transcribe unrestricted continuous speech from broadcast data in multiple languages with acceptable performance. The increased capabilities of todays recognizers is in part due to the improved accuracy (and increased complexity) of the models, which are closely related to the availability of large spoken and text corpora for training, and the wide availability of faster and cheaper computational means which have enabled the development and implementation of better training and decoding algorithms. Despite the extent of progress over the recent years, recognition accuracy is still quite sensitive to the environmental conditions and speaking style: channel quality, speaker characteristics, and background noise have a large impact on the acoustic component of the speech recognizer, whereas the speaking style and discourse domain largely influence the linguistic component. In addition, most systems are both task and language dependent, and bringing up a system for a different task or language is costly and requires human expertise.

Only for small vocabulary, speaker-dependent isolated word or phrase speech recognizers, such as name dialing on mobile telephones, portability is not really an issue. With such devices, all of the names must be entered by the user according to the specific protocol - such systems typically use whole word patterns and do not care who the speaker or what the language is. For almost all more complex tasks, portability is a major concern. Some speech technology companies have been addressing the language localization problem for many years, and some research sites have also been investigating speech recognition in multiple languages (4; 13; 14; 21; 35; 37) as well as speech recognition using multi-lingual components (19; 33). Multi-lingual speech processing has been the subject of several special sessions at conferences and workshops (see for example, (1; 2; 3; 20)). The EC CORETEX project (http://coretex.itc.it) is investigating methods to improve basic speech recognition technology, including fast system development, as well as the development of systems with high genericity and adaptability. Fast system development refers to both language support, i.e., the capability of porting technology to different languages at a reasonable cost; and task portability, i.e. the capability to easily adapt a technology to a new task by exploiting limited amounts of domain-specific knowledge. Genericity and adaptability refer to the capacity of the technology to work properly on a wide range of tasks and to dynamically keep models up to date using contemporary data. The more robust the initial generic system is, the less there is a need for adaptation.

In the next section an overview of todays most widely used speech recognition technology is given. Following subsections address several approaches to reducing the cost of porting, such as improving model genericity, and reducing the need for annotated training data. An attempt is made to give an idea of the amount of data and effort required to port to a different language or task.

2. Speech Recognition Overview

Speech recognition is concerned with converting the speech waveform into a sequence of words. Today's most performant approaches are based on a statistical modelization of the speech signal (16; 31; 32; 38). The basic modeling techniques have been successfully applied to a number of languages and for a wide range of applications.

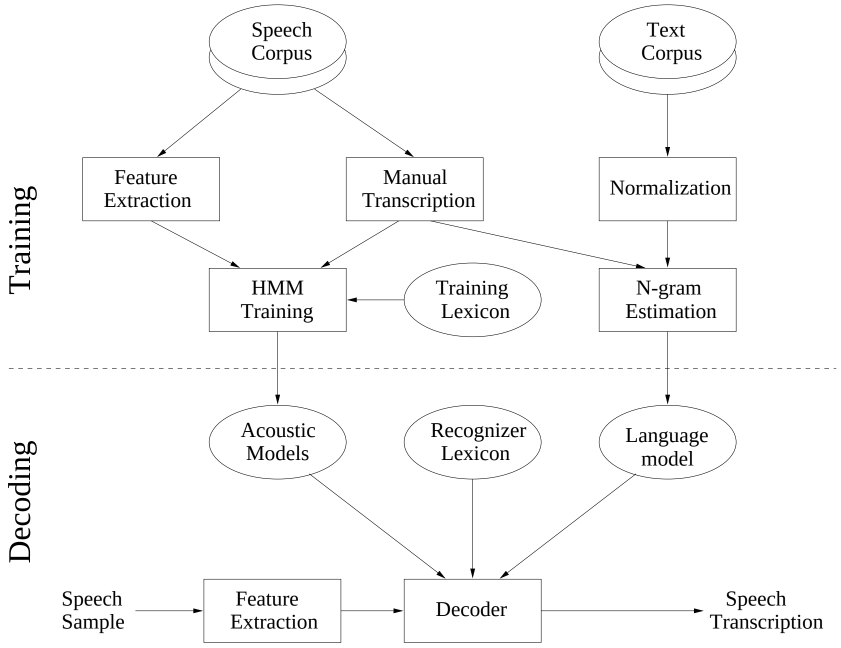


Figure 1: System diagram of a generic speech recognizer based using statistical models, including training and decoding processes.

The main components of a speech recognition system are shown in Figure 1. The elements shown are the main knowledge sources (speech and textual training materials and the pronunciation lexicon), the feature analysis (or parameterization), the acoustic and language models which are estimated in a training phase, and the decoder. The training and decoding algorithms are largely task and language independent, the main language dependencies are in the knowledge sources (the training corpora).

The first step of the acoustic feature analysis is digitization, in which the continuous speech signal is converted into discrete samples. Acoustic feature extraction is then carried out on a windowed portion of speech [1], with the goal of reducing model complexity while trying to maintain the linguistic information relevant for speech recognition. Most recognition systems use short-time cepstral features based either on a Fourier transform or a linear prediction model. Cepstral parameters are popular because they are a compact representation, and are less correlated than direct spectral components. Cepstral mean removal (subtraction of the mean from all input frames) is commonly used to reduce the dependency on the acoustic recording conditions, and delta parameters (obtained by taking the first and second differences of the parameters in successive frames) are often used to capture the dynamic nature of the speech signal. While the details of the feature analysis differs from system to system, most of the commonly used analyses can be expected to work reasonably well for most languages and tasks.

Most state-of-the-art systems make use of hidden Markov models (HMM) for acoustic modeling, which consists of modeling the probability density function of a sequence of acoustic feature vectors (32). These models are popular as they are performant and their parameters can be efficiently estimated using well established techniques. The Markov model is described by the number of states and the transitions probabilities between states. The most widely used acoustic units in continuous speech recognition systems are phone-based[2], and typically have a small number of left-to-right states in order to capture the spectral change across time. Since the number of states imposes a minimal time duration for the unit, some configurations allow certain states to be skipped. The probability of an observation (i.e. a speech vector) is assumed to be dependent only on the state, which is known as the 1st order Markov assumption.

Phone based models offer the advantage that recognition lexicons can be described using the elementary units of the given language, and thus benefit from many linguistic studies. It is of course possible to perform speech recognition without using a phonemic lexicon, either by use of “word models” (a commonly used approach for isolated word recognition) or a different mapping such as the fenones (7). Compared with larger units, small subword units reduce the number of parameters, and more importantly can be associated with back-off mechanisms to model rare or unseen, contexts, and facilitate porting to new vocabularies. Fenones offer the additional advantage of automatic training which is of interest for language porting, but lack the ability to include *a priori* linguistic models.

A given HMM can represent a phone without consideration of its neighbors (context-independent or mono-

[1] An inherent assumption is that due to physical constraints on the rate at which the articulators can move, the signal can be considered quasi-stationary for short periods (on the order of 10ms to 20ms).

[2] Phones usually correspond to phonemes, but may also correspond to allophones such as flaps or glottal stop.

phone model) or a phone in a particular context (context-dependent model). The context may or may not include the position of the phone within the word (word-position dependent), and word-internal and cross-word contexts may or may not be merged. Different approaches can be used to select the contextual units based on frequency or using clustering techniques, or decision trees, and different types of contexts have been investigated. The model states are often clustered so as to reduce the model size, resulting in what are referred to as "tied-state" models.

Acoustic model training consists of estimating the parameters of each HMM. For continuous density Gaussian mixture HMMs, this requires estimating the means and covariance matrices, the mixture weights and the transition probabilities. The most popular approaches make use of the Maximum Likelihood criterion, ensuring the best match between the model and the training data (assuming that the size of the training data is sufficient to provide robust estimates). Since the goal of training is to find the best model to account of the observed data, the performance of the recognizer is critically dependent upon the representativity of the training data. Speaker-independence is obtained by estimating the parameters of the acoustic models on large speech corpora containing data from a large speaker population. Since there are substantial differences in speech from male and female talkers arising from anatomical differences it is thus common practice to use separate models for male and female speech in order to improve recognition performance (requiring automatic gender identification).

2.1. Lexical and pronunciation modeling

The lexicon is the link between the acoustic-level representation and the word sequence output by the speech recognizer (34). Lexical design entails two main parts: definition and selection of the vocabulary items and representation of each pronunciation entry using the basic acoustic units of the recognizer. Recognition performance is obviously related to lexical coverage, and the accuracy of the acoustic models is linked to the consistency of the pronunciations associated with each lexical entry. Developing a consistent pronunciation lexicon requires substantial language specific knowledge from a native speaker of the language and usually entails manual modification even if grapheme-to-phoneme rules are reasonably good for the language of interest. The lexical units must be able to be automatically extracted from a text corpus or from speech transcriptions and for a given size lexicon should optimize the lexical coverage for the language and the application. Since on average, each out-of-vocabulary (OOV) word causes more than a single error (usually between 1.5 and 2 errors), it is important to judiciously select the recognition vocabulary. The recognition word list is to some extent dependent on the conventions used in the source text (punctuation markers, compound words, acronyms, case sensitivity, ...) and the specific language. The lexical units can be chosen to explicitly model observed pronunciation variants, for example, using compound words to represent word sequences subject to severe reductions such as "dunno" for "don't know". The vocabulary is usually comprised of a simple list of lexical items as observed in the text. Attempts have been made to use other units, for example, to use a list of root forms (stems) augmented by derivation, declension, composition rules. However, while more powerful in terms of language coverage, such representations are more difficult to integrate in present state-of-the-art recognizer technology.

These pronunciations may be taken from existing pronunciation dictionaries, created manually or generated by an automatic grapheme-phoneme conversion software. Alternate pronunciations are sometimes used to explicitly represent variants that cannot be easily modeled by the acoustic units, as is the case for homographs (words spelled the same, but pronounced differently) which reflect different parts of speech (verb or noun) such as *excuse, record, produce*. While pronunciation modeling is widely acknowledged to be a challenge to the research community, there is a lack of agreement as to what pronunciation variants should be modeled and how to do so. Adding a large number of pronunciation variants to a recognition lexicon without accounting for their frequency of occurrence can reduce the system performance. An automatic alignment system is able to serve as an analysis tool which can be used to quantify the occurrence of events in large speech corpora and to investigate their dependence on lexical frequency (5).

2.2. Language modeling

Language models (LMs) are used in speech recognition to estimate the probability of word sequences. Grammatical constraints can be described using a context-free grammars (for small to medium size vocabulary tasks these are usually manually elaborated) or can be modeled stochastically, as is common for LVCSR. The most popular statistical methods are *n*-gram models, which attempt to capture the syntactic and semantic constraints by estimating the frequencies of sequences of *n* words. The assumption is made that the probability of a given word string $(w_1, w_2, ..., w_k)$ can be approximated by $\prod_{i=1}^{k} \Pr(w_i|w_{i-n+1}, ..., w_{i-2}, w_{i-1})$, therefore reducing the word history to the preceding $n - 1$ words. A back-off mechanism is generally used to smooth the estimates of the probabilities of rare *n*-grams by relying on a lower order *n*-gram when there is insufficient training data, and to provide a means of modeling unobserved word sequences (17).

Given a large text corpus it may seem relatively straightforward to construct *n*-gram language models. Most of the steps are pretty standard and make use of tools that count word and word sequence occurrences. The main differences arise in the choice of the vocabulary and in the definition of words, such as the treatment of compound words or acronyms, and the choice of the back-off strategy. There is, however, a significant amount of effort needed to process the texts before they can be used.

One of the main motivations for text normalization is to reduce lexical variability so as to increase the coverage for a fixed vocabulary size. The normalization decisions are generally language-specific. Much of speech recognition research for American English has been supported by ARPA and has been based on text materials which were processed to remove upper/lower case distinction and compounds. Thus, for instance, no lexical distinction is made between *Gates, gates* or *Green, green*. However with increased interest in going beyond transcription to information extraction tasks (such as finding named entities or locating events in the audio signal) such distinctions are important. In our work at LIMSI for other languages (French, German, Portuguese) capitalization of proper names is distinctive with different lexical items for the French words *Pierre, pierre* or *Roman, roman*.

The main conditioning steps are text mark-up and conversion. Text mark-up consists of tagging the texts (article, paragraph and sentence markers) and garbage bracketing (which includes not only corrupted text materials, but all text material unsuitable for sentence-based language modeling, such as tables and lists). Numerical expressions are typically expanded to approximate the spoken form ($150 → one hundred and fifty dollars). Further semi-automatic processing is necessary to correct frequent errors inherent in the texts (such as obvious mispellings *milllion, officals*) or arising from processing with the distributed text processing tools. Some normalizations can be considered as "decompounding" rules in they modify the word boundaries and the total number of words. These concern the processing of ambiguous punctuation markers (such as hyphen and apostrophe), the processing of digit strings, and treatment of abbreviations and acronyms (ABCD → A. B. C. D.). Another example is the treatment of numbers in German, where decompounding can be used in order to increase lexical coverage. The date 1991 which in standard German is written as *neunzehnhunderteinundneunzig* can be represented by word sequence *neunzehn hundert ein und neunzig*. Generally speaking, the choice is a compromise between producing an output close to correct standard written form of the language and lexical coverage, with the final choice of normalization being largely application-driven.

In practice, the selection of words is done so as to minimize the system's OOV rate by including the most useful words. By useful we mean that the words are expected as an input to the recognizer, but also that the LM can be trained given the available text corpora. There is the sometimes conflicting need for sufficient amounts of text data to estimate LM parameters and assuring that the data is representative of the task. It is also common that different types of LM training material are available in differing quantities. One easy way to combine training material from different sources is to train a language model per source and to interpolate them, where the interpolation weights are estimated on some development data.

2.3. Decoding

The aim of the decoder is to determine the word sequence with the highest likelihood given the lexicon and the acoustic and language models. Since it is often prohibitive to exhaustively search for the best solution, techniques have been developed to reduce the computational load by limiting the search to a small part of the search space. The most commonly used approach for small and medium vocabulary sizes is the one-pass frame-synchronous Viterbi beam search which uses a dynamic programming algorithm. This basic strategy has been extended to deal with large vocabularies by adding features such as dynamic decoding, multipass search and N-best rescoring. Multi-pass decoding strategies progressively add knowledge sources in the decoding process and allows the complexity of the individual decoding passes to be reduced. Information between passes is usually transmitted via word graphs, although some systems use N-best hypotheses (a list of the most likely word sequences with their respectives scores). One important advantage of multi-pass is the possibility to adapt the models between decoding passes. Acoustic model adaptation can be used to compensate mismatches between the training and testing conditions, such as due to differences in acoustic environment, to microphones and transmission channels, or to particular speaker characteristics. Attempts at language model adaptation have been less successful. However, multi-pass approaches are not well suited to real-time applications since no hypothesis can be returned until the entire utterance has been processed.

3. Language porting

Porting a recognizer to another language necessitates modification of some of the system parameters, i.e. those incorporating language-dependent knowledge sources such as the phone set, the recognition lexicon (alternate word pronunciations), and phonological rules and the language model. Different languages have different sets of units and different coarticulation influences amomg adjacent phonemes. This influences the way of choosing context-dependent models and of tying distributions. Other considerations are the acoustic confusability of the words in the language (such as homophone, monophone, and compound word rates) and the word coverage of a given size recognition vocabulary.

One important aspect in developing a transcription system for a different language is obtaining the necessary resources for training the acoustic and language models, and a pronunciation lexicon. The Linguistic Data Consortium (LDC http://www.ldc.upenn.edu) and the European Language Resources Association (ELRA http://www.elda.fr) have greatly aided the creation and distribution of language resources. The number and diversity of language resources has grown substantially over recent years. However, most of the resources are only available for the most interesting languages from the commercial or military perspectives.

There are two predominant approaches taken to bootstrapping the acoustic models for another language. The first is to use acoustic models from an existing recognizer and a pronunciation dictionary to segment manually annotated training data for the target language. If recognizers for several languages are available, the seed models can be selected by taking the closest model in one of the available language-specific sets. An alternative approach is to use a set of global acoustic models, that cover a wide number of phonemes (33). This approach offers the advantage of being able to use the multilingual acoustic models to provide additional training data, which is particularly interesting when only very limited amounts of data ($<$ 10 hours) for the target language are available.

A general rule of thumb for the necessary resources for speaker independent, large vocabulary continuous speech recognizers is that the minimal data requirements are on the order of 10 hours transcribed audio data for training the acoustic models and several million words of texts (transcriptions of audio if available) for language modeling. Depending upon the application, these resources are more or less difficult to obtain. For example, unannotated data for broadcast news type tasks can be easily recorded via standard TV, satellite or cable and data of this type is becoming more easily accessible via the Internet. Related text materials are also available from a variety of on-line newspapers and new feeds. The manual effort required to transcribe broadcast news data is roughly 20-40 hours per hour of audio data, depending upon the desired precision (8).

Data for other applications can be much more difficult to obtain. In general, for spoken language dialog systems, training data needs to be obtained from users interacting with the system. Often times an initial corpus is recorded from a human-human service (should it exist) or using simulations (Wizard-of-OZ) or an initial prototype system. The different means offer different advantages. For example, WOz simulations help in making design decisions before the technology is implemented and allow alternative designs to be simulated quickly. However, the amount of data that can be collected with a WOz setup is limited by the need for a human wizard. Prototype systems offer the possibility of collection much larger corpora, albeit somewhat limited by the capacity of the current system. We have observed that the system's response generation has a large influence on the naturalness of the data collected with a prototype system.

Other application areas of growing interest are the transcription of conversational speech from telephone conversations and meetings, as well as voicemail. Several sources of multilingual corpora are available (for example, the CallHome and CallFriend corpora from LDC). This data is quite difficult to obtain and costly to annotate due to its very spontaneous nature (hesitations, interruptions, use of jargon). The manual effort involved is higher than that required for broadcast news transcription, and the transcriptions are less consistent and accurate.

The application-specific data is useful for accurate modeling at different levels (acoustic, lexical, syntactic and semantic). Acquiring sufficient amounts of text training data is more challenging than obtaining acoustic data. With 10k queries relatively robust acoustic models can be trained, but these queries contain only on the order of 100k words, which probably yield an incomplete coverage of the task (ie. they are not sufficient for word list development) and are insufficient for training *n*-gram language models.

At LIMSI broadcast news transcription systems have been developed for the American English, French, German, Mandarin, Spanish, Arabic and Portuguese languages. The Mandarin language was chosen because it is quite different from the other languages (tone and syllable-based), and Mandarin resources are available via the LDC as well as reference performance results from DARPA benchmark tests. To give an idea of the resources used in developing these systems, the training material are shown in Table 1. It can be seen that there is a wide disparity in the available language resources for a broadcast news transcription task: for American English, 200 hours of manually transcribed acoustic training were available from the LDC, compared with only about 20-50 hours for the other languages. Obtaining appropriate language model training data is even more difficult. While newspaper and newswire texts are becoming widely available in many languages, these texts are quite different than transcriptions of spoken language. Over 10k hours of commercial transcripts are available for American English (from PSMedia), and many TV stations provide closed captions. Such data are not available for most other languages, and in some countries it is illegal to sell transcripts. Not shown here, manually annotated broadcast news corpora are also available for the Italian (30 hours) and Czech (30 hours) languages via ELRA and LDC respectively, and some text sources can be found on the Internet.

Some of the system characteristics are shown in Table 2, along with indicative recognition performance rates for these languages. State-of-the-art systems can transcribe unrestricted American English broadcast news data with word error rates under 20%. Our transcription systems for French and German have comparable error rates for news broadcasts (6). The character error rate for Mandarin is also about 20% (10). Based on our experience, it appears that with appropriately trained models, recognizer performance is more dependent upon the type and source of data, than on the language. For example, documentaries are particularly challenging to transcribe, as the audio quality is often not very high, and there is a large proportion of voice over.

4. Reducing the porting cost

4.1. Improving Genericity

In the context of the EC CORETEX project, research is underway to improve the genercity of speech recognition technology, by improving the basic technolgoy and exploring rapid adaptation methods which start with the initial robust generic system and enhance performance on particular tasks. To this extent, cross task recognition experiments have been reported where models from one task are used as a starting point for other tasks (24; 9; 15; 26; 30; 11). In (26) broadcast news (BN) (28) acoustic and language models to decode the test data for three other tasks (TI-digits (27), ATIS (12) and WSJ (29)). For TI-digits and ATIS the word error rate increase was shown to be primarily due to a linguistic mismatch since using task-specific language models greatly reduces the error rate. For spontaneous WSJ dictation the BN models out-performed task-specific models trained on read speech data, which can be attributed to a better modelization of spontaneous speech effects (such as breath and filler words).

Methods to improve genericity of the models via multi-source training have been investigated. Multi-source training can be carried out in a variety of ways – by pooling data, by interpolating models or via single or multi-step model adaptation. The aim of multi-source training is to ob-

	Audio			*Text (words)*	
Language	*Radio-TV sources*	*Duration*	*Size*	*News*	*Com.Trans.*
English	ABC, CNN, CSPAN, NPR, PRI, VOA	200h	1.9M	790M	240M
French	Arte, TF1, A2, France-Info, France-Inter	50h	0.8M	300M	20M
German	Arte	20h	0.2M	260M	-
Mandarin	VOA, CCTV, KAZN	20h	0.7M(c)	200M(c)	-
Portuguese	9 sources	3.5h	∼35k	70M	-
Spanish	Televisa, Univision, VOA	30h	0.33M	295M	-
Arabic	tv: Aljazeera, Syria; radio: Orient, Elsharq, ...	50h	0.32M	200M	-

Table 1: Approximate sizes of the transcribed audio data and text corpora used for estimating acoustic and language models. For the text data, newspaper texts (News) and commercial transcriptions (Com.Trans.) are distinguished in terms of the millions of words (or characters for Mandarin). The American English, Spanish and Mandarin data are distributed by the LDC. The German data come from the EC OLIVE project and the French data partially from OLIVE and from the DGA. The Portuguese data are part of the 5h, 11 source Pilot corpus used in the EC ALERT project (data from 2 sources 24Horas and JornalTarde were reserved for the test set). The Arabic data were produced by the Vecsys company in collaboration with the DGA.

		Lexicon		*Language Model*		*Test*	
Language	*#phon.*	*size (words)*	*coverage*	*N-gram*	*ppx*	*Duration*	*%Werr*
English	48	65k	99.4%	11M fg, 14M tg, 7M bg	140	3.0h	20
French	37	65k	98.8%	10M fg, 13M tg, 14M bg	98	3.0h	23
German	51	65k	96.5%	10M fg, 14M tg, 8M bg	213	2.0h	25(n)-35(d)
Mandarin	39	40k+5k(c)	99.7%	19M fg, 11M tg, 3M bg	190	1.5h	20
Spanish	27	65k	94.3%	8M fg, 7M tg, 2M bg	159	1.0h	20
Portuguese	39	65k	94.0%	9M tg, 3M bg	154	1.5h	40
Arabic	40	60k	90.5%	11M tg, 6M bg	160	5.7h	20

Table 2: Some language characteristics. Specified for each language are: the number of phones used to represent lexical pronunciations, the approximate vocabulary size in words (characters for Mandarin) and lexical coverage (of the test data), the language model size and the perplexity, the test data duration (in hours) and the word/character error rates. For Arabic the vocabulary and language model are vowelized, however the word error rate does not include vowel or gemination errors. For German, separate word error rates are given for broadcast news (n) and documentaries (d).

tain generic models which are comparable in performance to the respective task-dependent models for all tasks under consideration. Compared to the results obtained with task-dependent acoustic models, both data pooling and sequential adaptation schemes led to better performance for ATIS and WSJ read, with slight degradations for BN and TI-digits (25).

In (9) cross-task porting experiments are reported for porting from an Italian broadcast news speech recognition system to two spoken dialogue domains. Supervised adaptation was shown to recover about 60% of the WER gap between the broadcast news acoustic models and the task-specific acoustic models. Language model adaptation using just 30 minutes of transcriptions was found to reduce the gap in perplexity between the broadcast news and task-dependent language models by 90%. It was also observed that the out-of-vocabulary rates for the task-specific language models are 3 to 5 times higher than the best adapted models, due to the relatively limited amount of task-specific data and the wide coverage of the broadcast news domain.

Techniques for large-scale discriminative training of the acoustic models of speech recognition systems using the maximum mutual information estimation (MMIE) criterion in place of conventional maximum likelihood estimation (MLE) have studied and it has been demonstrated that MMIE-based systems can lead to sizable reductions in word error rate on the transcription of conversational telephone speech (30). Experiments on discriminative training for cross-task genericity have made use of recognition systems trained on the low-noise North American Business News corpus of read newspaper texts and tested on television and radio Broadcast News data. These experiments showed that MMIE-trained models could indeed provide improved cross-task performance (11).

4.2. Reducing the need for annotated training data

With today's technology, the adaptation of a recognition system to a new task or new language requires the availability of sufficient amount of transcribed training data. When changing to new domains, usually no exact transcriptions of acoustic data are available, and the generation of such transcribed data is an expensive process in terms of manpower and time. On the other hand, there often exist incomplete information such as approximate transcriptions, summaries or at least key words, which can be used to provide supervision in what can be referred to as "informed speech

Amount of training data		*Language Model*
Raw	*Usable*	News.Com.Cap
10min	10min	53.1
1.5h	1h	33.3
50h	33h	20.7
104h	67h	19.1
200h	123h	18.0

Table 3: Supervised acoustic model training: Word error rate (%) on the 1999 evaluation test data for various conditions using one set of gender-independent acoustic models trained on subsets of the HUB4 training data with detailed manual transcriptions. The language model is trained on the available text sources, without any detailed transcriptions of the acoustic training data. The raw data reflects the size of the audio data before partitioning, and the usable data the amount of data used in training the acoustic models.

recognition". Depending on the level of completeness, this information can be used to develop confidence measures with adapted or trigger language models or by approximate alignments to automatic transcriptions. Another approach is to use existing recognizer components (developed for other tasks or languages) to automatically transcribe task-specific training data. Although in the beginning the error rate on new data is likely to be rather high, this speech data can be used to re-train a recognition system. If carried out in an iterative manner, the speech data base for the new domain can be cumulatively extended over time *without* direct manual transcription. This approach has been investigated in (18; 22; 23; 36; 39).

In order to give an idea of the influence of the amount of training data on system performance, Table 3 shows the performance of a 10xRealTime American English BN system for different amounts of manually annotated training data. The language model News.Com.Cap is trained on large text corpora, and results from the interpolation of individual language models trained on newspaper and newswires tests (790M words), commercially produced transcripts and closed-captions predating the test epoch (240M words). The word error is seen to rapidly decrease initially, with only a relatively small improvement above 30 hours of usable data. However, there is substantial information available in the language models. Table 4 summarizes supervised training results using substantially less language model training material. The second entry is for a language model estimated only on the newpaper texts (790M words), whereas for the remaining two language models were estimated on only 30 M words of texts (the last 2 months of 1997) and 1.8 M words (texts from December 26-31, 1997). It can be seen that the language model training texts have a large influence on the system performance, and even 30 M words is relatively small for the broadcast news transcription task.

	Raw Acoustic training data		
Language model	*200 hours*	*1.5 hours*	*10 min*
News.Com.Cap, 65k	18.0	33.3	53.1
News, 65k	20.9	36.1	55.6
30 M words, 60k	24.1	40.8	60.2
1.8 M words, 40k	28.8	46.9	65.3

Table 4: Supervised acoustic model training: Reference word error rates (%) on the 1999 evaluation test data with varying amounts of manually annotated acoustic training data and a language model trained on 1.8 M and 30 M words of news texts from 1997.

Raw Acoustic training data		*WER (%)*
bootstrap models	10 min manual	65.3
1 (6 shows)	4 h	54.1
2 (+12 shows)	12 h	47.7
3 (+23 shows)	27 h	43.7
4 (+44 shows)	53 h	41.4
5 (+60 shows)	103 h	39.2
6 (+58 shows)	135 h	37.4

Table 5: Unsupervised acoustic model training: Word error rate (%) on the 1999 evaluation test data with varying amounts of automatically transcribed acoustic training data and a language model trained on 1.8 M words of news texts from 1997.

The basic idea of light supervision is to use a speech recognizer to automatically transcribe unannotated data, thus generating "approximate" labeled training data. By iteratively increasing the amount of training data, more accurate acoustic models are obtained, which can then be used to transcribe another set of unannotated data. The manual work is considerably reduced, not only in generating the annotated corpus but also during the training procedure, since it is no longer necessary to extend the pronunciation lexicon to cover all words and word fragments occurring in the training data. In (22) it was found that somewhat comparable acoustic models could be estimated on 400 hours automatically annotated data from the TDT-2 corpus and 150 hours of carefully annotated data.

The effects of reducing the amount of supervision are summarized in Table 5. The first observation that can be made, is that even using a recognizer with an initial word error of 65% the procedure is converging properly by training acoustic models on automatically labeled data. This is even more surprising since the only supervision is via a language model trained on a small amount of text data predating the raw acoustic audio data. As the amount of automatically transcribed acoustic data is successively doubled, there are consistent reductions in the word error rate. While these error rates are still quite high compared to supervised training, retranscribing the same data (36) can be expected to reduce the word error rate further. (Recall that even with supervised acoustic model training trained on 200 hours of raw data the word error rate is 28.8% with this language model.)

4.3. Unsupervised Cross-Task Adaptation

An incremental unsupervised adaptation scheme was investigated for cross-task adaptation from the broadcast news task to the ATIS task (26). In this system-in-loop adaptation scheme, a first subset of the training data is automatically transcribed using the generic system. The acoustic and linguistic models of the generic system are then adapted with these automatically annotated data and the resulting models are used to transcribe another portion of the training data. One obvious use of this scheme is for online model adaptation in a dialog system.

Using about one-third (15 hours) of the ATIS training corpus transcribed with a BN system to adapt both the acoustic and language models, the word error rate is reduced from 20.8% to 6.9%. Transcribing the remaining data, and readapting the models reduces the word error to 5.5% (which can be compared to 4.7% for a task-specific system). Contrastive experiments have shown that this gain is somewhat equally split between adaptation of the acoustic and language models.

4.4. Cross Language Portability

The same basic idea was used to develop BN acoustic models for the Portuguese language for which substantially less manually transcribed data are available. RTP and INESC, partners in the Alert project (http:alert.uni-duisburg.de) provided 5 hours of manually annotated data from 11 different news programs. Two of the programs (82 minutes) were reserved for testing purposes (JornalTarde_20_04_00 and 24Horas_19_07_00). The remaining 3.5 hours of data were used for acoustic model training. The language model texts were obtained from the following sources: the Portuguese Newswire Text Corpus distributed by LDC (23M words from 1994-1998); Correio da Manha (1.6M words), Expresso (1.9M words from 2000-2001), and Jornal de Noticias (46M words, from 1996-2001), The recognition lexicon contains 64488 words. The pronunciations are generated by grapheme-to-phoneme rules, and use 39 phones.

Initial acoustic model trained on the 3.5 hours of available data were used to transcribe 30 hours of Portuguese TV broadcasts. These acoustic models had a word error rate of 42.6%. By training on the 30 hours of data using the automatic transcripts the word error was reduced to 39.1%. This preliminary experiment supports the feasibility of lightly supervised and unsupervised acoustic model training.

5. Conclusions

This paper has discussed the main issues in speech recognizer development and portability across languages and tasks. Today's most performant systems make use of statistical models, and therefore require large corpora for acoustic and language model training. However, acquiring these resources is both time-consuming, costly, and may be beyond the economic interest for many languages. Research is underway to reduce the need for manually annotated training data, thus reducing the human investment needed for system development when porting to another task or language. By eliminating the need for manual transcription, automated training can be applied to essentially unlimited quantities of task-specific training data.

The pronunciation lexicon still requires substantial manual effort for languages without straightfoward letter-to-sound correspondences, and to handle foreign words and proper names. For languages or dialects without a written form, the challenge is even greater, since important language modeling data are simply unavailable. Even if a transliterated form can be used, it is likely to be impractical to transcribe sufficient quantities of data for language model training.

In summary, our experience is that although general technologies and development strategies appear to port from one language to another, to obtain optimal performance language specificities must be taken into account. Efforts underway to improve the genericity of speech recognizers, and to reduce training costs will certainly help to enable the development of language technologies for minority languages and less economically promising applications.

Acknowledgments

This work has been partially financed by the European Commission and the French Ministry of Defense. I would like to thank the members of the Spoken Language Processing group at LIMSI who have participated to the research reported here and from whom I have borrowed experimental results.

REFERENCES

[1] *IEEE Workshop on Automatic Speech Recognition*, Special session on Multilingual Speech Recognition, Snowbird, Dec. 1995.

[2] *ICSLP'96*, Special session on "Multilingual Speech Processing," Philadelphia, PA, Oct. 1996.

[3] *Multi-Lingual Interoperabilty in Speech Technology*, RTO-NATO and ESCA ETRW, Leusden, Holland, Sept. 1999.

[4] M. Adda-Decker, "Towards Multilingual Inoperability in Speech Recognition," *Multi-Lingual Interoperabilty in Speech Technology*, RTO-NATO and ESCA ETRW, Leusden, Holland, 69-76, Sept. 1999.

[5] M. Adda-Decker, L. Lamel, "Pronunciation Variants Across Systems, Languages and Speaking Style," *Speech Communication*, "Special Issue on Pronunciation Variation Modeling", **29**(2-4): 83-98, Nov. 1999.

[6] M. Adda-Decker, G. Adda, L. Lamel, "Investigating text normalization and pronunciation variants for German broadcast transcription," *ICSLP'2000*, Beijing, China, Oct. 2000.

[7] L.R. Bahl, P. Brown, P. de Souza, R.L. Mercer, M. Picheny, "Acoustic Markov Models used in the Tangora Speech Recognition System," *ICASSP-88* **1**, pp. 497-500.

[8] C. Barras, E. Geoffrois, Z. Wu, M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, **33**(1-2): 5-22, Jan. 2001.

[9] N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, D. Giuliani, "From Broadcast News to Spontaneous Dialogue Transcription: Portability Issues," *ICASSP'01*, Salt Lake City, May 2001.

[10] L. Chen, L. Lamel, G. Adda, J.L. Gauvain, "Broadcast News Transcription in Mandarin," *ICSLP'2000*, Beijing, China, Oct. 2000.

[11] R. Cordoba, P. Woodland, M. Gales "Improved Cross-Task Recognition Using MMIE Training" *ICASSP'02*, Orlando, Fl, May 2002.

[12] D. Dahl, M. Bates *et al.*, "Expanding the Scope of the ATIS Task : The ATIS-3 Corpus," *ARPA Spoken Language Systems Technology Workshop*, Plainsboro, NJ, 3-8, 1994.

[13] C. Dugast, X. Aubert, R. Kneser, "The Philips Large Vocabulary Recognition System for American English, French, and German," *Eurospeech'95*, 197-200, Madrid, Sept. 1995.

[14] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, J. Sakai, S. Seneff, V. Zue, "Multilingual spoken language understanding in the MIT Voyager system," *Speech Communication*, **17**(1-2): 1-18, Aug. 1995.

[15] D. Giuliani, M. Federico, "Unsupervised Language and Acoustic Model Adaptation for Cross Domain Portability" *ISCA ITRW 2001 Adaptation Methods For Speech Recognition*, Sophia-Antipolis, France, Aug. 2001.

[16] F. Jelinek, "Statistical Methods for Speech Recognition," Cambirdge: MIT Press, 1997.

[17] Katz, S.M. 1987. "Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer". *IEEE Trans. Acoustics, Speech, and Signal Processing*. **ASSP-35**(3): 400-401.

[18] T. Kemp, A. Waibel, "Unsupervised Training of a Speech Recognizer: Recent Experiments," *ESCA Eurospeech'99*, Budapest, Hungary, **6**, 2725-2728, Sept. 1999.

[19] J. Köhler, "Language-adaptation of multilingual phone models for vocabulary independent speech recognition tasks," *ICASSP'98*, **I**, 417-420, Seattle, May 1998.

[20] J. Kunzmann, K. Choukri, E. Janke, A. Kiessling, K. Knill, L. Lamel, T. Schultz, S. Yamamoto, "Portability of ASR Technology to new Languages: multilinguality issues and speech/text resources," slides from the panel discussion at *IEEE ASRU'01*, Madonna di Campiglio, Dec. 2001. (http://www.cs.cmu.edu/t̃anja/Papers/asru2001.ppt)

[21] L. Lamel, M. Adda-Decker, J.L. Gauvain, G. Adda, Spoken Language Processing in a Multilingual Context," *ICSLP'96*, 2203-2206, Philadelphia, PA, Oct. 1996.

[22] L. Lamel, J.L. Gauvain, G. Adda, "Lightly Supervised and Unsupervised Acoustic Model Training," *Computer, Speech & Language*, Jan. 2002.

[23] L. Lamel, J.L. Gauvain, G. Adda, "Unsupervised Acoustic Model Training," *IEEE ICASSP'02*, Orlando, Fl, May 2002.

[24] L. Lamel, F. Lefevre, J.L. Gauvain, G. Adda, "Portability issues for speech recognition technologies," *HLT'2001*, 9-16, San Diego, March 2001.

[25] F. Lefevre, J.L. Gauvain, L. Lamel, "Improving Genericity for Task-Independent Speech Recognition," *EuroSpeech'01*, Aalborg, Sep. 2001.

[26] F. Lefevre, J.L. Gauvain, L. Lamel, "Genericity and Adaptability Issues for Task-Independent Speech Recognition," *ISCA ITRW 2001 Adaptation Methods For Speech Recognition*, Sophia-Antipolis, France, Aug. 2001.

[27] R.G. Leonard, "A Database for speaker-independent digit recognition," *ICASSP*, 1984.

[28] D.S. Pallett, J.G. Fiscus, *et al.* "1998 Broadcast News Benchmark Test Results," *DARPA Broadcast News Workshop*, 5-12, Herndon, VA, Feb. 1999.

[29] D.B. Paul, J.M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," *ICSLP'92*, Kobe, Nov. 1992.

[30] D. Povey, P. Woodland, "Improved Discriminative Training Techniques For Large Vocabulary Continuous Speech Recognition", *IEEE ICASSP'01*, Salt Lake City, May 2001.

[31] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, **77**(2): 257-286. Feb, 1989.

[32] L.R. Rabiner, B.H. Juang, "An Introduction to Hidden Markov Models. *IEEE Acoustics Speech and Signal Processing ASSP Magazine*, **ASSP-3**(1): 4-16, Jan. 1986.

[33] T. Schultz, A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, **35** (1-2): 31-51, Aug. 2001.

[34] F. Van Eynde, D. Gibbon, eds., *Lexicon Development for Speech and Language Processing*, Dordrecht: Kluwer, 2000.

[35] A. Waibel, P. Geutner, L. Mayfield Tomokiyo, T. Schultz, M. Woszczyna, "Multilinguality in Speech and Spoken Language Systems," *Proceedings of the IEEE*, Special issue on Spoken Language Processing, **88**(8): 1297-1313, Aug. 2000.

[36] F. Wessel, H. Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *ASRU'01*, Madonna di Campiglio, Italy, Dec. 2001.

[37] S. Young, M. Adda-Decker, X. Aubert, C. Dugast, J.L. Gauvain, D.J. Kershaw, L. Lamel, D.A. van Leeuwen, D. Pye, A.J. Robinson, H.J.M. Steeneken, P.C. Woodland, "Multilingual Large Vocabulary Speech Recognition: The European SQALE Project," *Computer Speech and Language*, **11**(1): 73-89, Jan. 1997.

[38] S. Young, G. Bloothooft, eds., "Corpus Based Methods in Language and Speech Processing," Dordrecht: Kluwer, 1997.

[39] G. Zavaliagkos, T. Colthurst, "Utilizing Untranscribed Training Data to Improve Performance," *DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, 301-305, Feb. 1998.

Seven Dimensions of Portability for Language Documentation and Description

Steven Bird*[*] and Gary Simons[†]

[*]Linguistic Data Consortium, University of Pennsylvania, 3615 Market Street, Philadelphia, PA 19104, USA
[†]SIL International, 7500 West Camp Wisdom Road, Dallas, TX 75236, USA

Abstract

The process of documenting and describing the world's languages is undergoing radical transformation with the rapid uptake of new digital technologies for capture, storage, annotation and dissemination. However, uncritical adoption of new tools and technologies is leading to resources that are difficult to reuse and which are less portable than the conventional printed resources they replace. We begin by reviewing current uses of software tools and digital technologies for language documentation and description. This sheds light on how digital language documentation and description are created and managed, leading to an analysis of seven portability problems under the following headings: content, format, discovery, access, citation, preservation and rights. After characterizing each problem we provide a series of value statements, and this provides the framework for a broad range of best practice recommendations.

1. Introduction

It is now easy to collect vast quantities of language documentation and description and store it in digital form. It is getting easier to transcribe the material and link it to linguistic descriptions. Yet how can we ensure that such material can be re-used by others, both now and into the future? While today's linguists can access documentation that is over 100 years old, much digital language documentation and description is unusable within a decade of its creation.

The fragility of digital records is amply demonstrated. For example, the interactive video disks created by the BBC Domesday Project are inaccessible just 15 years after their creation.[1] In the same way, linguists who are quick to embrace new technologies and create digital materials in the absence of archival formats and practices soon find themselves in technological quicksand.

The uncritical uptake of new tools and technologies is encouraged by sponsors who favor projects that promise to publish their data on the web with a search interface. However, these projects depend on technologies with life cycle of 3-5 years, and the resources they create usually do not outlive the project any longer than this.

This paper considers portability in the broadest sense: across different software and hardware platforms; across different scholarly communities (e.g. field linguistics, language technology); across different purposes (e.g. research, teaching, development); and across time. Portability is frequently treated as an issue for software, but here we will focus on data. In particular, we address portability for language documentation and description, and interpret these terms following Himmelmann:

> The aim of a language documentation is to provide a comprehensive record of the linguistic practices characteristic of a given speech community. Linguistic practices and traditions are manifest in two ways: (1) the observable linguistic behavior, manifest in everyday interaction between members of the speech community, and (2) the native speakers' metalinguistic knowledge, manifest in their ability to provide interpretations and systematizations for linguistic units and events. This definition of the aim of a language documentation differs fundamentally from the aim of language descriptions: a language description aims at the record of A LANGUAGE, with ”language” being understood as a system of abstract elements, constructions, and rules that constitute the invariant underlying structure of the utterances observable in a speech community. (Himmelmann, 1998, 166)

We adopt the cover term DATA to mean any information that documents or describes a language, such as a published monograph, a computer data file, or even a shoebox full of hand-written index cards. The information could range in content from unanalyzed sound recordings to fully transcribed and annotated texts to a complete descriptive grammar. Beyond data, we are be concerned with language resources more generally, including tools and advice. By TOOLS we mean computational resources that facilitate creating, viewing, querying, or otherwise using language data. Tools include software programs, along with the digital resources that they depend on such as fonts, stylesheets, and document type definitions. By ADVICE we mean any information about what data sources are reliable, what tools are appropriate in a given situation, and what practices to follow when creating new data (Bird and Simons, 2001).

This paper addresses seven dimensions of portability for digital language documentation and description, identifying problems, establishing core values, and proposing best practices. The paper begins with a survey of the tools and technologies (§2), leading to a discussion of the problems that arise with the resources created using these tools and technologies (§3). We identify seven kinds of portability problem, under the headings of content, format, discovery, access, citation, preservation and rights. Next we give statements about core values in digital language documentation and description, leading to a series of “value statements”, or requirements for best practices (§4), and followed up with collection of best practice recommendations (§5). The structure of the paper is designed to build consensus. For instance, readers who take issue with a best practice recommendation in §5 are encouraged to review the corresponding statement of values in §4 and either suggest a different practice which better implements the values, or else take issue with the value statement (then back up to the corresponding problem statement in §3, and so forth).

2. Tools and Technologies for Language Documentation and Description

Language documentation projects are increasing in their reliance on new digital technologies and software tools. This section contains a comprehensive survey of the range of practice, covering general purpose software, specialized tools, and digital technologies. Reviewing the available tools gives us a snapshot of how digital language documentation and description is created and managed, and provides a backdrop for our analysis of data portability problems.

2.1. General purpose tools

The most widespread practice in language documentation involves the use of office software. This software is readily available, often pre-installed, and familiar. Word processors have often been used as the primary storage for large lexical database, including a Yoruba lexicon with 30,000 entries split across 20 files. Frequently cited benefits are the WYSIWYG editing, the find/replace function, the possibility of cut-and-paste to create sublexicons, and the ease of publishing. Of course, a large fraction of the linguist's time is spent on maintaining consistency across multiple copies of the same data. Word processors have also been used for interlinear text, with three main approaches: fixed width fonts with hard spacing, manual setting of tabstops, and tables.[2] All methods require manual line-breaking, and significant labor if line width or point size are changed. Another kind of office software is the spreadsheet, which is often used for wordlists. Language documentation created using office software is normally stored in a secret proprietary format that is unsupported within 5-10 years. While other export formats are supported, they may loose some of the structure. For instance, part of speech may be distinguished in a lexical entry through the use of a particular font, and this information may be lost when the data is exported. Also, the portability of export formats may be compromised, by being laden with presentational markup.

A second class of general purpose software is the hypertext processors. Perhaps the first well-known application to language documentation was the original Macintosh hypercard stacks of *Sounds of the World's Languages* (Ladefoged and Maddieson, 1996). While it was easy to create a complex web of navigable pages, nothing could overcome the limitations of a vendor-specific hypertext language. More recently, the HTML standard and universal, free browsers have encouraged the creation of large amounts of hypertext for a variety of documentation types. For instance, we have interlinear text with HTML tables (e.g. Austin's Jiwarli fieldwork[3]), interlinear text with HTML frames (e.g. Culley's presentation of Apache texts[4]), HTML markup for lexicons, with hyperlinks from glossed examples and a thesaurus (e.g. Austin and Nathan's Gamilaraay lexicon[5]), gifs for representing IPA transcriptions (e.g. Bird's description of tone in Dschang[6]), and Javascript for image annotations (e.g. Poser's annotated photographs of gravestones engraved with Déné syllabics[7]). In all these cases, HTML is used as the primary storage format, not simply as a view on an underlying database. The intertwining of content and format makes this kind of language documentation difficult to maintain and re-use.

The third category of general purpose software is database packages. In the simplest case, the creator shares the database with others by requiring them to purchase the same package, and by shipping them a full dump of the database (e.g. the StressTyp database, which requires users to buy a copy of "4th Dimension"[8]). A more popular approach is to put the database on a web-server, and create a forms-based web interface that allows remote users to search the database without installing any software (e.g. the Comparative Bantu Online Lexical Database[9] and the Maliseet-Passamaquoddy Dictionary.[10]) Recently, some sites have started allowing database updates via the web (e.g. the Berkeley Interlinear Text Collector[11] and the Rosetta Project's site for uploading texts, wordlists and descriptions[12]).

2.2. Specialized tools

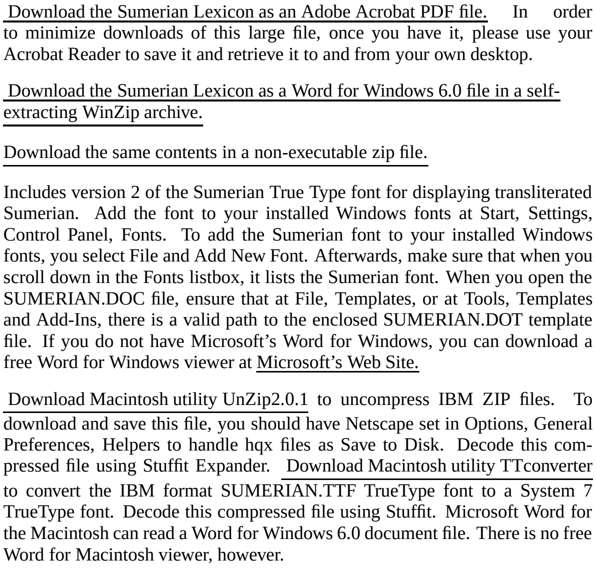
Over the last two decades, several dozen tools have been developed having specialized support for language documentation and description. We list a representative sample here; more can be found on SIL's page on *Linguistic Computing Resources*,[13] on the *Linguistic Exploration* page,[14] and on the *Linguistic Annotation* page.[15]

Tools for linguistic data management include Shoebox[16] and the Fieldworks Data Notebook.[17] Speech analysis tools include Praat[18] and SpeechAnalyzer.[19] Many specialized signal annotation tools have been developed, including CLAN,[20] EMU,[21] TableTrans, InterTrans, TreeTrans.[22] There are many orthographic transcription tools, including Transcriber[23] and MultiTrans.[24] There are morphological analysis tools, such as the Xerox finite state toolkit.[25] There are a wealth of concordance tools. Finally, some integrated multi-function systems have been created, such as LinguaLinks Linguistics Workshop.[26]

In order to do their specialized linguistic processing, each of these tools depends on some model of linguistic information. Time-aligned transcriptions, interlinear texts, syntax trees, lexicons, and so forth, all require suitable data structures and file formats. Given that most of these tools have been developed in isolation, they typically employ incompatible models and formats. For example, data created with an interlinear text tool cannot be subsequently annotated with syntactic information without losing the interlinear annotations. When interfaces and formats are open and documented, it is occasionally possible to cobble the tools together in support of a more complex need. However, the result is a series of increasingly baroque and decreasingly portable approximations to the desired solution. Computational support for language documentation and description is in disarray.

2.3. Digital technologies

A variety of digital technologies are now used in language documentation thanks to sharply declining hardware costs. These include technologies for digital signal capture (audio, video, physiological) and signal storage (hard disk, CD-R, DVD-R, minidisc). Software technologies are also playing an influential role as new standards are agreed. The most elementary and pervasive of these is the hyperlink, which makes it possible to connect linguistic descriptions to the underlying documentation (e.g. from an analytical transcription to a recording). Such links streamline the descriptive process; checking a transcription can be done with mouse clicks instead of digging out a tape or finding an informant. The ability to navigate from description to documentation also facilitates analysis and verification. Software technologies and standards have given rise to the internet which permits low-cost dissemination of language resources. Notably, it is portability problems with these tools and formats that prevents these basic digital technologies from having their full impact. The download instructions for the Sumerian lexicon[27] typify the problems (hyperlinks are underlined):



> Download the Sumerian Lexicon as an Adobe Acrobat PDF file. In order to minimize downloads of this large file, once you have it, please use your Acrobat Reader to save it and retrieve it to and from your own desktop.
>
> Download the Sumerian Lexicon as a Word for Windows 6.0 file in a self-extracting WinZip archive.
>
> Download the same contents in a non-executable zip file.
>
> Includes version 2 of the Sumerian True Type font for displaying transliterated Sumerian. Add the font to your installed Windows fonts at Start, Settings, Control Panel, Fonts. To add the Sumerian font to your installed Windows fonts, you select File and Add New Font. Afterwards, make sure that when you scroll down in the Fonts listbox, it lists the Sumerian font. When you open the SUMERIAN.DOC file, ensure that at File, Templates, or at Tools, Templates and Add-Ins, there is a valid path to the enclosed SUMERIAN.DOT template file. If you do not have Microsoft's Word for Windows, you can download a free Word for Windows viewer at Microsoft's Web Site.
>
> Download Macintosh utility UnZip2.0.1 to uncompress IBM ZIP files. To download and save this file, you should have Netscape set in Options, General Preferences, Helpers to handle hqx files as Save to Disk. Decode this compressed file using Stuffit Expander. Download Macintosh utility TTconverter to convert the IBM format SUMERIAN.TTF TrueType font to a System 7 TrueType font. Decode this compressed file using Stuffit. Microsoft Word for the Macintosh can read a Word for Windows 6.0 document file. There is no free Word for Macintosh viewer, however.

2.4. Digital Archives

Recently several digital archives of language documentation and description have sprung up, such as the Archive of the Indigenous Languages of Latin America,[28] and the Rosetta Project's Archive of 1000 Languages.[29] These exist alongside older archives which are in various stages of digitizing their holdings: the Archive of the Alaska Native Language Center,[30] the LACITO Linguistic Data Archive,[31] and the US National Anthropological Archives.[32] These archives and many others are surveyed on the *Language Archives* page.[33] Under the aegis of OLAC, the *Open Language Archives Community*,[34] the notion of language archive has been broadened to include archives of linguistic software, such as the Natural Language Software Registry[35]

These archives face many challenges, the most significant being the lack of funding. Other challenges may include: identifying, adapting and deploying digital archiving standards; setting up key operational functions such as offsite backup, migration to new digital formats and media over time, and the support of new access modes (e.g. search facilities) and delivery formats (e.g. streaming media); and obtaining the long-term support of a major institution to assure contributors and users that the materials will be available over the long term.

3. Seven Problems for Portability

With the rapid uptake of new digital technologies, many creators of language documentation and description are ignoring the question of portability, with the unfortunate consequence that the fruits of their labors are likely to be unusable within 5-10 years. In this section we discuss seven critical problems for the portability of this data.

3.1. Content

Many potential users of language data are interested in assimilating multiple descriptions of a single language to gain an understanding of the language which is as comprehensive as possible. Many users are interested in comparing the descriptions of different languages in order to apply insights from one analysis to another or to test a typological generalization. However, two descriptions may be difficult to compare or assimilate because they have used terminology differently, or because the documentation on which the descriptions are based is unavailable.

Language documentation and description of all types depends critically on technical vocabulary, and ambiguous terms compromise portability. For instance, the symbols used in phonetic transcription have variable interpretation depending on the descriptive tradition: “it is crucial to be aware of the background of the writer when interpreting an unexplained occurrence of [y]” (Pullum and Ladusaw, 1986, 168). In morphosyntax, the term “absolutive” can refer to one of the cases in an ergative language, or to the unpossessed form of a noun (in the Uto-Aztecan tradition) (Lewis et al., 2001, 151), and a correct interpretation of the term depends on an understanding of the linguistic context.

This terminological variability leads to problems for retrieval. Suppose that a linguist wanted to search the full-text content of a large collection of digital language data, in order to discover which other languages have relevant phenomena. Since there are no standard ontologies, the user will discover irrelevant documents (low precision) and will fail to discover relevant documents (low recall). In order to carry out a comprehensive search, the user must know all the ways in which a particular phenomena is described. Even once a set of descriptions are retrieved, it will generally not be possible to draw reliable comparisons between the descriptions of different languages.

The content of two descriptions may also be difficult to reconcile because it is not possible to verify them with respect to the language documentation that they cite. For example, when two descriptions of the same language provide different phonetic transcriptions of the same word, is this the result of a typographical error, a difference in transcription practice, or a genuine difference between two speech varieties? When two descriptions of different languages report that the segmental inventories of both languages contain a [k], what safe conclusions can be drawn about how similar the two sounds are? Since the underlying documentation is not available, such questions cannot be resolved, making it difficult to re-use the resources.

While the large-scale creation of digital language resources is a recent phenomenon, the language documentation community has been active since the 19th century, and much earlier in some instances. At risk of oversimplifying, a widespread practice over this extended period has been to collect wordlists and texts and to write descriptive grammars. With the arrival of new digital technologies it is easy to transfer the whole endeavor from paper to computer, and from tape recorder to hard disk, and to carry on just as before. Thus, new technologies simply provide a better way to generate the old kinds of resources. Of course this is a wasted opportunity, since the new technologies can also be used to create digital multimedia recordings of rich linguistic events. Such rich recordings often capture items which turn out to be useful in later linguistic analysis, and have immense intrinsic value as a record of cultural heritage for future generations. However, managing digital technologies in less controlled situations leads to many technical and logistical issues, and there are no guidelines for integrating new technologies into new documentary practices.

3.2. Format

Language data frequently ends up in a secret proprietary format using a non-standard character encoding. To use such data one must often purchase commercial software then install it on the same hardware and under the same operating system used by the creator of the data.

Other formats, while readable outside the tool that created them, remain non-portable when they are not explicitly documented. For example, the interpretation of the field names in Shoebox format may not be documented, or the documentation may become separated from the data file, making it difficult to guess what the different fields signify.

The developers of linguistic tools must frequently parse presentational formats. For example, the occurrence of `<b>[n]</b>` in a lexical entry might indicate that this is an entry for a noun. More difficult cases involve subtle context-dependencies. This presentational markup obscures the structure and interpretation of the linguistic content. Conversely, in the absence of suitable browsing and rendering tools, end-users must attempt to parse formats that were designed to be read only by machines.

3.3. Discovery

Digital language data is often presented as a physical or digital artefact with no external description. Like a book without a cover page or a binary file called `dict.dat`, one is forced to expend considerable effort to discover the subject matter and the nature of the content. Organized collections – such as the archive of a university linguistics department – may provide some metadescription, but it is likely to use a parochial format and idiosyncratic descriptors. If they are provided, key descriptors like *subject language* and *linguistic type* are usually given in free text rather than a controlled vocabulary, reducing precision and recall. As a consequence, discovering relevant language resources is extremely difficult, and depends primarily on word-of-mouth and queries posted to electronic mailing lists. Thus, new resource creation efforts may proceed in ignorance of prior and concurrent efforts, wasting scarce human resources.

In some cases, one may obtain a resource only to discover upon closer inspection that it is in an incompatible format. This is the flip-side of the discovery problem. Not only do we need to know that a resource exists, but also that it is relevant. When resources are inadequately described, it is difficult (and often impossible) to find a relevant resource, a huge impediment to portability.

3.4. Access

In the past, primary documentation was usually not disseminated. To listen to a field recording it was often necessary to visit the laboratory of the person who collected the materials, or to make special arrangements for the materials to be copied and posted. Digital publication on the web has alleviated this problem, although projects usually refrain from full dissemination by limiting access via a restrictive search interface. This means that only selected portions of the documentation can be downloaded, and that all access must use categories predefined by the provider. Moreover, these web forms only have a lifespan of 3-5 years, relying on ad hoc CGI scripts which may cease working when the interpreter or webserver are upgraded. Lack of full access means that materials are not portable. More generally, people have often conflated digital publication with web publication, and publish high-bandwidth materials on the web which would be more usable if published on CD or DVD.

Many language resources have applications beyond those envisaged by their creators. For instance, the Switchboard database (Godfrey et al., 1992), collected for the development of speaker-independent automatic speech recognition, has since been used for studies of intonation and disfluency. Often this redeployment is prevented through the choice of formats. For instance, publishing conversation transcripts in the Hub-4 SGML format does not facilitate their reuse in, say, conversational analysis. In other cases, redeployment is prevented by the choice of media. For instance, an endangered language dictionary published only on the web will not be accessible to speakers of that language who live in a village without electricity.

One further problem for access deserves mention here. It sometimes happens that an ostensibly available resource turns out not to be available after all. One may discover the resource because its creator cited it in a manuscript or an annual research report. Commonly, a linguist wants to derive recognition for the labor that went into creating primary language documentation, but does not want to make the materials available to others until deriving maximum personal benefit. Two tactics are to cite unresolved, non-specific intellectual property rights issues, and to repeatedly promise but to never finally deliver. Despite its many guises, this problem has two distinguishing features: someone draws attention to a resource in order to derive credit for it – "parading their riches" as Mark Liberman (pers. comm.) has aptly described it – and then applies undocumented or inconsistent restrictions to prevent access. The result may be frustration that a needed resource is withheld, leading to wasted effort or a frozen project, or to suspicion that the resource is defective and so must be protected by a smoke screen.

3.5. Citation

Research publications are normally required to provide full bibliographic citation of the materials used in conducting the research. Citation standards are high for conventional resources (such as other publications), but are much lower for language resources which are usually incorrectly cited, or not cited at all. This makes it difficult to find out what resource was used in conducting the research and, in the reverse direction, it is impossible to use a citation index to discover all the ways in which a given resource has been applied.

Often a language resource is available on the web, and it is convenient to have the uniform resource locater (URL) since this may offer the most efficient way to obtain the resource. However, URLs can fail as a persistent citation in two ways: they may simply break, or they may cease to reference the same item. URLs break when the resource is moved or when some piece of the supporting infrastructure, such as a database server, ceases to work. Even if a URL does not break, the item it references may be mutable, changing over time. Language resources published on the web are usually not versioned, and a third-party description of some item may cease to be valid if that item is changed. Publishing a digital artefact, such as a CD, with a unique identifier, such as an ISBN, avoids this problem.

Citation goes beyond bibliographic citation of a complete item. We may want to cite some component of a resource, such as a specific narrative or lexical entry. However, the format may not support durable citations to internal components. For instance, if a lexical entry is cited by a URL which incorporates its lemma, and if the spelling of the lemma is altered, then the URL will not track the change. In sum, language documentation and description is not portable if the incoming and outgoing links to related materials are fragile.

3.6. Preservation

The digital technologies used in language documentation and description greatly enhance our ability to create data while simultaneously compromising our ability to preserve it. Relative to paper copy which can survive for hundreds of years, digitized materials are evanescent because they use some combination of binary formats with undocumented character encodings saved on non-archival media and physically stored with no ongoing administration for backups and migration to new media. Presentational markup with HTML and interactive content with Javascript and specialized browser plugins require future browsers to be backwards-compatible. Furthermore, primary documentation may be embodied in the interactive behavior of the resource (e.g. the gloss of the text under the mouse may show up in the browser status line, using the Javascript "mouseover" effect). Consequently, digital resources – especially dynamic or interactive ones – often have a short lifespan, and typically become unusable 3-5 years after they are actively maintained.

3.7. Rights

A variety of individuals and institutions may have intellectual property vested in a language resource, and there is a complex terrain of legal, ethical and policy issues (Liberman, 2000). In spite of this, most digital language data is disseminated without identifying the copyright holder and without any license delimiting the range of acceptable uses of the material. Often people collect or redistribute materials, or create derived works without securing the necessary permissions. While this is often benign (e.g. when the resources are used for research purposes only), the researcher risks legal action, or having to restrict publication, or even having to destroy primary materials. To avoid any risk one must avoid using materials whose property rights are in doubt. In this way, the lack of documented rights restrict the portability of the language resource.

Sometimes resources are not made available on the web for fear that they will get into the wrong hands or be misused. However, this confuses medium with rights. The web supports secure data exchange between authenticated parties (through data encryption) and copyright statements together with licenses can be used to restrict uses. More sophisticated models for managing digital rights are emerging (Iannella, 2001). The application of these techniques to language resources is unexplored, and we are left with an all-or-nothing situation, in which the existence of any restriction prevents access across the board.

3.8. Special challenges for little-studied languages

Many of the problems reported above also apply to little-studied languages, though some are greatly exacerbated in this context. The small amount of existing work on the language and the concomitant lack of established documentary practices and conventions may lead to especially diverse nomenclature. Inconsistencies within or between language descriptions may be harder to resolve because of the lack of significant documentation, the limited access to speakers of the language, and the limited understanding of dialect variation. Open questions in one area of description (e.g. the inventory of vowel phonemes) may multiply the indeterminacies in another (e.g. for transcribed texts). More fundamentally, existing documentation and description may be virtually impossible to discover and access, owing to its fragmentary nature.

The acuteness of these portability problems for little-studied languages can be highlighted by comparison with well-studied languages. In English, published dictionaries and grammars exist to suit all conceivable tastes, and it therefore matters little (relatively speaking) if none of these resources is especially portable. However, when there is only one dictionary for the language, it must be pressed into a great range of services, and significant benefits will come from maximizing portability.

This concludes our discussion of portability problems arising from the way new tools and technologies are being used in language documentation and description. The rest of this paper responds to these problems, by laying out the core values that lead to requirements for best practices (§4) and by providing best practice recommendations (§5).

4. Value Statements

Best practice recommendations amount to a decision about which of several possible options is best. The notion of best always involves a value judgment. Therefore, before making our recommendations, we articulate the values which motivate our choices. Our use of “we” is meant to include the reader and the wider language resources community who share these values.

4.1. Content

TERMINOLOGY. We value the ability of users to identify the substantive similarities and differences between two resources. Thus the best practice is one that makes it easy to associate the comparable parts of unrelated resources.

ACCOUNTABILITY. We value the ability of researchers to verify language descriptions. Thus the best practice is one that provides the documentation that lies behind the description.

RICHNESS. We value the documentation of little-studied languages. Thus the best practice is one that establishes a record that is sufficiently broad in scope and rich in detail that future generations can experience and study the language, even when no speakers remain.

4.2. Format

OPENNESS. We value the ability of any potential user to make use of a language resource without needing to obtain unique or proprietary software. Thus the best practice is one that puts data into a format that is not proprietary.

DOCUMENTATION. We value the ability of potential users of a language resource to understand its internal structure and organization. Thus the best practice is one that puts data into a format that is documented.

MACHINE-READABLE. We value the ability of users of a language resource to write programs to process the resource. Thus the best practice is one that puts the resource into a well-defined format which can be submitted to automatic validation.

HUMAN-READABLE. We value the ability of users of a language resource to browse the content of the resource. Thus the best practice is one that provides a human-digestible version of a resource.

4.3. Discovery

EXISTENCE. We value the ability of any potential user of a language resource to learn of its existence. Thus the best practice is one that makes it easy for anyone to discover that a resource exists.

RELEVANCE. We value the ability of potential users of a language resource to judge its relevance without first having to obtain a copy. Thus the best practice is one that makes it easy for anyone to judge the relevance of a resource based on its metadescription.

4.4. Access

COMPLETE. We value the ability of any potential user of a language resource to access the complete resource, not just a limited interface to the resource. Thus the best practice is one that makes it easy for anyone to obtain the entire resource.

UNIMPEDED. We value the ability of any potential user of a language resource to follow a well-defined procedure to obtain a copy of the resource. Thus the best practice is one in which all available resources have a clearly documented method by which they may be obtained.

UNIVERSAL. We value the ability of potential users to access a language resource from whatever location they are in. Thus the best practice is one that makes it possible for users to access some version of the resource regardless of physical location and access to computational infrastructure.

4.5. Citation

CREDIT. We value the ability of researchers to be properly credited for the language resources they create. Thus the best practice is one that makes it easy for authors to correctly cite the resources they use.

PROVENANCE. We value the ability of potential users of a language resource to know the provenance of the resources it is based on. Thus the best practice is one that permits resource users to navigate a path of citations back to the primary linguistic documentation.

PERSISTENCE. We value the ability of language resource creators to endow their work with a permanent digital identifier which resolves to an instance of the resource. Thus the best practice is one that associates resources with persistent digital identifiers.

IMMUTABILITY. We value the ability of potential users to cite a language resource without that resource changing and invalidating the citation. Thus the best practice is one that makes it easy for authors to freeze and version their resources.

COMPONENTS. We value the ability of potential users to cite the component parts of a language resource. Thus the best practice is one that ensures each sub-item of a resource has a durable identifier.

4.6. Preservation

LONG-TERM. We value access to language resources over the very long term. Thus the best practice is one which ensures that language resources will still be usable many generations into the future.

COMPLETE. We value the ability of future users of a language resource to access the complete resource as experienced by contemporary users. Thus the best practice is one which preserves fragile aspects of a resource (such as dynamic and interactive content) in a durable form.

4.7. Rights

DOCUMENTATION. We value the ability of potential users of a language resource to know the restrictions on permissible uses of the resource. Thus the best practice is one that ensures that potential users know exactly what they are able to do with any available resource.

RESEARCH. We value the ability of potential users of a language resource to use it in personal scholarship and academic publication. Thus the best practice is one that ensures that the terms of use on resources do not hinder individual study and academic research.

5. Best Practice Recommendations

This section recommends best practices in support of the values set out in §4. We believe that the task of identifying and adopting best practices rests with the community, and we believe that OLAC, the *Open Language Archives Community*, provides the necessary infrastructure for identifying community-agreed best practices. Here, however, we shall attempt to give some broad guidelines to be fleshed out in more detail later, by ourselves and also, we hope, by other members of the language resources community.

5.1. Content

TERMINOLOGY. Map linguistic terminology and descriptive markup elements to a common ontology of linguistic terms. This applies to the obvious candidates such as morphosyntactic abbreviations and structural markup, but also to less obvious cases such as the phonological description of the symbols used in transcription. (NB vocabularies can be versioned and archived in an OLAC archive; archived descriptions cite their vocabularies using the `Relation` element.)

ACCOUNTABILITY. Provide the full documentation on which language descriptions are based. For example, where a narrative is transcribed, provide the primary recording (without segmenting it into multiple sound clips). Create time-aligned transcriptions to facilitate verification.

RICHNESS. Make rich records of rich interactions, especially in the case of endangered languages or genres. Document the "multimedia linguistic field methods" that were used. Provide theoretically neutral descriptions of a wide range of linguistic phenomena.

5.2. Format

OPENNESS. Store all language documentation and description in an open format. Prefer formats supported by multiple third-party software tools. NB some proprietary formats are open, e.g. Adobe Portable Document Format (PDF) and MPEG-1 Audio Layer 3 (MP3).

DOCUMENTATION. Provide all language documentation and description in a self-describing format (preferably XML). Provide detailed documentation of the structure and organization of the format. Encode the characters with Unicode. Try to avoid Private Use Area characters, but if they are used document them fully. Document any 8-bit character encodings. (OLAC will be providing detailed guidelines for documenting non-standard character encodings.)

MACHINE-READABLE. Use open standards such as XML and Unicode, along with Document Type Definitions (DTDs), XML Schemas and/or other definitions of well-formedness which can be verified automatically. Archive the format definition, giving each version its own unique identifier. When archiving data in a given format, reference the archived definition of that format. Avoid freeform editors for structured information (e.g. prefer Excel or Shoebox over Word for storing lexicons).

HUMAN-READABLE. Provide one or more human readable version of the material, using presentational markup (e.g. HTML) and/or other convenient formats. Proprietary formats are acceptable for delivery as long as the primary documentation is stored in a non-proprietary format.

N.B. Format is a critical area for the definition of best practices. We propose that recommendations in this area be organized by type (e.g. audio, image, text), possibly following the inventory of types identified in the Dublin Core metadata set.[36]

5.3. Discovery

EXISTENCE. List all language resources with an OLAC data provider. Any resource presented in HTML on the web should contain metadata with keywords and description for use by conventional search engines.

RELEVANCE. Follow the OLAC recommendations on best practice for metadescription, especially concerning language identification and linguistic data type. This will ensure the highest possibility of discovery by interested users in the OLAC union catalog hosted by Linguist.[37]

5.4. Access

COMPLETE. Publish complete primary documentation. Publish the documentation itself, and not just an interface to it, such as a web search form.

UNIMPEDED. Document all access methods and restrictions along with other metadescription. Document charges and expected delivery time.

UNIVERSAL. Make all resources accessible by any interested user. Publish digital resources using appropriate delivery media, e.g. web for small resources, and CD/DVD for large resources. Where appropriate, publish corresponding print versions, e.g. for the dictionary of a little-studied language.

5.5. Citation

CREDIT, PROVENANCE. Furnish complete bibliographic data for all language resources created. Provide complete citations for all language resources used. Document the relationship between resources in the metadescription (NB in the OLAC context, use the `Relation` element).

PERSISTENCE. Ensure that resources have a persistent identifier, such as an ISBN or a persistent URL (e.g. a Digital Object Identifier[38]). Ensure that at least one persistent identifier resolves to an instance of the resource or to detailed information about how to obtain the resource.

IMMUTABILITY. Provide fixed versions of a resource, either by publishing it on a read-only medium, and/or submitting it to an archive which ensures immutability. Distinguish multiple versions with a version number or date, and assign a distinct identifier to each version.

COMPONENTS. Provide a formal means by which the components of a resource may be uniquely identified. Take special care to avoid the possibility of ambiguity, such as arises when lemmas are used to identify lexical entries, and where multiple entries can have the same lemma.

5.6. Preservation

LONG-TERM. Commit all documentation and description to a digital archive which can credibly promise long-term preservation and access. Ensure that the archive satisfies the key requirements of a well-founded digital archive (e.g. implements digital archiving standards, provides offsite backup, migrates materials to new formats and media/devices over time, is committed to supporting new access modes and delivery formats, has long-term institutional support, and has an agreement with a national archive to take materials if the archive folds). Archive physical versions of the language documentation and description (e.g. printed versions of documents; any tapes from which online materials were created). Archive electronic documents using type 1 (scalable) fonts in preference to bitmap fonts.

COMPLETE. Ensure that all aspects of language documentation and description accessible today are accessible in future. Ensure that any documentary information conveyed via dynamic or interactive behaviors is preserved in a purely declarative form.

5.7. Rights

DOCUMENTATION. Ensure that the intellectual property rights relating to the resource are fully documented.

RESEARCH. Ensure that the resource may be used for research purposes.

6. Conclusion

Today, the community of scholars engaged in language documentation and description exists in a cross-over period between the paper-based era and the digital era. We are still working out how to preserve knowledge that is stored in digital form. During this transition period, we observe unparalleled confusion in the management of digital language documentation and description. A substantial fraction of the resources being created can only be re-used on the same software/hardware platform, within the same scholarly community, for the same purpose, and then only for a period of a few years. However, by adopting a range of best practices, this specter of chaos can be replaced with the promise of easy access to highly portable resources.

Using tools as our starting point, we described a diverse range of practices and discussed their negative implications for data portability along seven dimensions, leading to a collection of advice for how to create portable resources. These three categories, tools, data, and advice, are three pillars of the infrastructure provided by OLAC, the Open Language Archives Community (Bird and Simons, 2001). Our best practice recommendations are preliminary, and we hope they will be fleshed out by the community using the OLAC Process.[39]

We leave off where we began, namely with tools. It is our use of the new tools which have led to data portability problems. And it is only with new tools, supporting the kinds of best practices we recommend, which will address these problems. An archival format is useless unless there are tools for creating, managing and browsing the content stored in that format. Needless to say, no single organization has the resources to create the necessary tools, and no third party developing general-purpose office software will address the unique needs of the language documentation and description community. We need nothing short of an open source revolution, leading to new specialized tools based on shared data models for all of the basic linguistic types, and connected to portable data formats.

Acknowledgements

This research is supported by NSF Grant No. 9983258 “Linguistic Exploration” and Grant No. 9910603 “International Standards in Language Engineering.”

Notes

[1] http://www.observer.co.uk/uk_news/story/0,6903,661093,00.html

[2] http://www.linguistics.ucsb.edu/faculty/cumming/WordForLinguists/Interlinear.htm

[3] http://www.linguistics.unimelb.edu.au/research/projects/jiwarli/gloss.html

[4] http://etext.lib.virginia.edu/apache/ChiMesc2.html

[5] http://www3.aa.tufs.ac.jp/~austin/GAMIL.HTML

[6] http://www.ldc.upenn.edu/sb/fieldwork/

[7] http://www.cnc.bc.ca/yinkadene/dakinfo/dulktop.htm

[8] http://fonetiek-6.leidenuniv.nl/pil/stresstyp/stresstyp.html

[9] http://www.linguistics.berkeley.edu/CBOLD/

[10] http://ultratext.hil.unb.ca/Texts/Maliseet/dictionary/index.html

[11] http://ingush.berkeley.edu:7012/BITC.html

[12] http://www.rosettaproject.org:8080/live/

[13] http://www.sil.org/linguistics/computing.html

[14] http://www.ldc.upenn.edu/exploration/

[15] http://www.ldc.upenn.edu/annotation/

[16] http://www.sil.org/computing/shoebox/

[17] http://fieldworks.sil.org/

[18] http://fonsg3.hum.uva.nl/praat/

[19] http://www.sil.org/computing/speechtools/speechanalyzier.htm

[20] http://childes.psy.cmu.edu/

[21] http://www.shlrc.mq.edu.au/emu/

[22] http://sf.net/projects/agtk/

[23] http://www.etca.fr/CTA/gip/Projets/Transcriber/

[24] http://sf.net/projects/agtk/

[25] http://www.xrce.xerox.com/research/mltt/fst/

[26] http://www.sil.org/LinguaLinks/LingWksh.html

[27] http://www.sumerian.org/

[28] http://www.ailla.org/

[29] http://www.rosettaproject.org/

[30] http://www.uaf.edu/anlc/

[31] http://195.83.92.32/index.html.en

[32] http://www.nmnh.si.edu/naa/

[33] http://www.ldc.upenn.edu/exploration/archives.html

[34] http://www.language-archives.org/

[35] http://registry.dfki.de/

[36] http://dublincore.org/

[37] http://www.linguistlist.org/

[38] http://www.doi.org/

[39] http://www.language-archives.org/OLAC/process.html

7. References

Steven Bird and Gary Simons. 2001. The OLAC metadata set and controlled vocabularies. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*. http://arXiv.org/abs/cs/0105030.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: A telephone speech corpus for research and development. In *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing*, volume I, pages 517–20. http://www.ldc.upenn.edu/Catalog/LDC93S7.html.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36:161–195.

Renato Iannella. 2001. Digital rights management (DRM) architectures. *D-Lib Magazine*, 7(6), June.

Peter Ladefoged and Ian Maddieson. 1996. *The Sounds of the World’s Languages*. Cambridge, MA: Blackwell.

William Lewis, Scott Farrar, and D. Terence Langendoen. 2001. Building a knowledge base of morphosyntactic terminology. In Steven Bird, Peter Buneman, and Mark Liberman, editors, *Proceedings of the IRCS Workshop on Linguistic Databases*, pages 150–156. http://www.ldc.upenn.edu/annotation/database/.

Mark Liberman. 2000. Legal, ethical, and policy issues concerning the recording and publication of primary language materials. In Steven Bird and Gary Simons, editors, *Proceedings of the Workshop on Web-Based Language Documentation and Description*. http://www.ldc.upenn.edu/exploration/expl2000/papers/.

Geoffrey K. Pullum and William A. Ladusaw. 1986. *Phonetic Symbol Guide*. The University of Chicago Press.

Challenges and Opportunities in Portability of Human Language Technologies

Bojan Petek

Interactive Systems Laboratory
University of Ljubljana, Faculty of Natural Sciences and Engineering
Snežniška 5, 1000 Ljubljana, Slovenia
Bojan.Petek@Uni-Lj.si

Abstract

Availability of language resources (LR) is a decisive element that influences the vital issues such as linguistic and cultural identity and use of a particular language in information society. Specifically, natural interactivity in information age relies on the existence of mature Human Language Technologies (HLT) that need substantial amount of appropriate LR to be developed. Additional challenge is that research addressing portability issues in HLT is still in its infancy. In perspective, it is reasonably to expect that advances in construction of the multilingual LR [Bird, 2001] and insights into the portability issues of HLT [Lamel, 2002] could potentially lower the digital divide and increase the visibility of a much larger pool of languages than experienced today. In order to achieve this challenging goal, it is proposed to initiate an international network of excellence (NoE) on HLT portability that would complement the already established activities on HLT resources. Such NoE could actively contribute to and advise the national HLT efforts aiming to achieve the grand goal of a non-exclusive information society.

1. Introduction

Several research programs have already focused towards the next generation of intelligent conversational interfaces. Their fundamental goal is to create speech-enabled multi-modal systems that scale gracefully across modalities. Such interfaces typically include speech, graphics, gesture, and computer vision. They are capable of supporting complex conversational interaction comparable to the human-human natural interactivity.

The natural interactive systems integrate spoken language dialogue systems, multimodal communication systems, and web-based data handling tools. The long-term goal of computer-mediated natural interactivity is to transform the present computer systems to become transparent in communication tasks and to support similar communication patterns as those experienced in usual interpersonal communication.

From the theoretical point of view, traditional Human-Computer Interaction (HCI) model has recently evolved towards the enhanced Human-Human Computer Interaction (HHCI) model that also includes the information and communication technologies. The HHCI model positions the computer system as a networked facilitator of information access and sharing. Typical applications include video conferencing or distributed multimedia information systems.

In the following, this paper aims to reflect the on-going research efforts by providing an overview of the challenges and opportunities addressed by the EU HLT projects and the US DARPA programs with relevance to the portability issues in HLT.

2. EU HLT Projects

At the time of writing about 50 EU HLT projects are detailed on the HLTCentral, the gateway to speech and language technology opportunities. Some projects with high relevance to the focus of this paper (ie, portability issues in human language technologies) are overviewed in order to outline the challenges (ie, project objectives) and opportunities (ie, expected outcome and innovation perspectives) these projects describe at the HLTCentral.

2.1. CORETEX

Improving Core Speech Recognition Technology, [wwwCORETEX]

The CORETEX project aims to improve the core speech recognition technologies. This 3 year EU project started in April 2000. The project consortium includes RWTH Aachen (Germany), University of Cambridge (UK), Istituto Trentino di Cultura ITC - IRST (Italy), and Centre National de la Recherche Scientifique – CNRS (France). The proposed work is motivated by observation that the current commercial speech recognition systems perform fairly well for a limited number of tasks and languages. On the other hand, these systems are very difficult to adapt to new domains, languages, and/or changing acoustic environmental conditions. The main obstacle in efficient porting of speech applications to new tasks, languages, or new environments is a requirement for substantial investment of time, money, and expertise.

Therefore, the overall project objective is to devise generic speech recognition technologies that perform well in a task independent way. List of the CORETEX project objectives mentioned on the web is the following:

- To develop generic speech recognition technologies for a wide range of tasks with minimum domain dependencies.
- To devise methods for a rapid portability to new languages with a limited amount of training data.
- To research techniques for producing enriched symbolic speech transcription for higher level symbolic processing.
- To improve language models and provide automatic pronunciation generation.
- To integrate the methods into showcases and validate them in relevant applications.
- To propose an evaluation framework and define objective measures to assess improvements.
- To disseminate the CORETEX research results and to facilitate contact with the interested users in order to widely exploit the project results.

The project is expected to provide significant insights into how to develop conceptually new HLT that is *generic, adaptable and portable***.** Generic design of technology is analyzed by evaluating a system trained on one corpus and tested on another one. Aim of this research is to assess performance degradation under the non-optimal conditions with respect to the training and testing conditions, and in the context of new languages. Initial project phase has already defined objective evaluation criteria and measures, including common test suites and the protocol.

In summary, opportunities from the CORETEX project are an improved HLT that are less sensitive to the environmental and linguistic factors as well as efficiently portable to many languages. Evaluation and demonstration frameworks were already proposed and serve to analyze the progress on the project. Detailed descriptions of the project achievements are given in the CORETEX Annual Reports for the years 2000 and 2001.

2.2. ENABLER

*European National Activities for Basic Language Resources***,** [wwwENABLER]

The ENABLER project aims to improve collaboration activities that provide national language resources in Europe. Researchers, industry, and service providers identified the LR to be a critical issue in national HLT programs and that these efforts need to be supported by appropriate national funding. The LR are of central importance to any kind of HLT-based infrastructure. They are also of vital importance in the development of HLT applications and products, thereby fundamental for the overall industrial growth. Availability of the adequate LR for as many languages as possible is of paramount importance in the HLT development for a non-exclusive multilingual information society.

This 22-month project started in November 2001. The consortium includes Università di Pisa (Italy), Institute for Language and Speech Processing - ILSP (Greece), European Language Resources Distribution Agency – ELDA (France), Center for Sprogteknologi – CST (Denmark), as well as members from Belgium, Czech Republic, Germany, Portugal, Spain, Sweden, and the Netherlands.

The ENABLER project goals are:

- To strengthen the current network of national initiatives, creating links among them, thereby providing a regular, updated, structured and public repository of organizational and technical information.
- To provide an official and general coordination forum for exchange of information, data, best practices, sharing of tools, multilateral and bilateral co-operation on specific issues.
- To gradually enlarge the existing network by identifying representatives of national initiatives.
- To promote synergies across national activities, to enhance the compatibility and interoperability of the results, thereby facilitating efficient transfer of technologies between languages.
- To maintain compatibility across various national LR.
- To increase visibility and strategic impact of the national activities.
- To provide a forum for discussion of innovative research issues and to propose medium- and long-term research priorities.
- To provide a forum to assess industry needs and to formulate common medium and long-term priorities.
- To promote exchange of tools, specifications, validation protocols produced by the national projects.
- To create an EU center for the harmonization of metadata description of speech, text, multimedia and multi-modal LR.
- To promote industrial exploitation of LR.
- To contribute to the internationally agreed cooperative framework for the provision of LR.

In perspective, ENABLER will contribute to the natural interactivity by providing multimodal LR, and to the multilinguality by fostering harmonization of national LR.

2.3. FAME

*Facilitating Agent for Multicultural Exchange***,** [wwwFAME]

New information technology tools for human-human communication integrate speech understanding, computer vision and dialog modeling and enable communication between people from different cultures who use different languages. The FAME project aims to address the problem of integrating multiple communication modes, such as vision, speech and object manipulation. Communication support is provided by the integration of physical and virtual worlds in multi-cultural communication and problem solving. The major identified project challenges are in automatic perception of human action and in understanding of free dialog between the people from different cultures.

Consortium of Universität Karlsruhe - Interactive Systems Labs (Germany), Institut National Polytechnique de Grenoble - Laboratoire GRAVIR-IMAG (France), Université Joseph Fourier - Laboratoire CLIPS (France), Istituto Trentino di Cultura - ITC-IRST (Italy), Universitat Politècnica de Catalunya (Spain), Sony International - Europe (Germany), and Applied Technologies on Language and Speech (Spain) envisions to construct an information butler that will demonstrate the context of awareness in the problem solving scenario. This goal will be achieved by integration of computer vision, speech understanding and dialog modeling. The demonstration prototype in form of an enhanced computer human-to-human communication model will be developed for the 2004 Barcelona Cultural Fair.

2.4. HOPE/EUROMAP3

*HLT Opportunity Promotion in Europe***,** [wwwHOPE]

This project aims to accelerate the rate of technology transfer from the research to the market. The project contains 11 National Focal Points (NFPs) from Austria, Belgium/Netherlands, Bulgaria, Denmark, Finland, France, Germany, Greece, Italy, Spain and UK. The Bulgarian, French and UK partners joined the project in October 2001. Each NFP will build on skills and expertise from the previous HLT awareness-raising actions. It will strive to achieve the following objectives:

- To increase the number of projects that deliver market-ready results.
- To accelerate awareness of benefits of the HLT systems, services and applications within the user sectors, policy makers and national administrations.
- To increase the number of state-of-the-art technology developers participating in the research projects.
- To improve the relevance of project targets, technology supplier and user needs.
- To improve the match between the HLT design, supplier and end user expectations.
- To enable user partnerships for beta testing, demonstration and other market application activities.

In perspective, the project also aims to include the EU accession countries. HOPE is a 36-month project and started in February 2000.

2.5. ISLE-HLT

International Standards for Language Engineering, [wwwISLE-HLT]

The ISLE-HLT is the most recent initiative of the *Expert Advisory Group for Language Engineering Standards* (EAGLES, [wwwEAGLES]). This 36-month project started in January 2000. Consortium consists of Consorzio Pisa Ricerche (Italy), University of Southern Denmark, Institute Dalle Molle pour les Etudes Sémantiques et Cognitives (Switzerland), Center for Sprogteknologi (Denmark), University of Pennsylvania - Computer and Information Science (USA), University of Pennsylvania - Linguistic Data Consortium (USA), New York University - Computer Science Department (USA), and University of Southern California - Information Sciences Institute (USA).

The overall project aim is to develop HLT standards within a global (EU-US) international collaboration and continuing the success of EAGLES by developing, disseminating and promoting de facto standards and guidelines for the HLT language resources, tools and products. The policy of the EAGLES/ISLE is to closely interact with academia and industry, users and providers, funding bodies and research organisations. The project objectives are put on the following three areas judged to be of a long-term significance:

- *Multilingual computational lexicons*. Initial work in this area presented survey of bi- and multilingual lexicons covering publishers' dictionaries. Next, specification of the Multilingual Isle Lexical Entry (MILE) was made. This involved work on complex Italian-English word-pairs, better understanding of word sense representation and cross-language linkages, extraction and classification of sense indicators, and development of a prototype tool to manage MILE-based lexicons. The ISLE also contributed to recommendations for MILE bilingual dictionary entries. A prototype tool for management of computational lexicons conforming to ISLE recommendations was developed.
- *Natural interaction and multimodality (NIMM).* This work extended the previous EAGLES work on textual and spoken language resources. Surveys were done on resources, annotation schemes and tools, as well as on metadata descriptions and tools. A prototype tool was developed for NIMM data annotation. XML schemas were developed that handle ISLE metadata descriptions. Editing and browsing tools were devised using these descriptions, including across distributed resources. Future work is concentrated on producing draft guidelines for best practice in the areas covered by the project, and in refining and documenting the tools and resources intended to help users in applying the guidelines.
- *Evaluation of the HLT systems.* This work focuses on methods and metrics for Machine Translation (MT). User feedback was collected within three international workshops. This led towards a refined version of the ISLE evaluation framework.

2.6. NESPOLE!

Negotiating through Spoken Language in E-commerce, [wwwNESPOLE!]

The NESPOLE! project aims to integrate speech-to-speech translation in eCommerce and eService environments by extrapolating from the results of the large research projects (C-STAR and Verbmobil). This EU project started in January 2000 and has a duration of 30 months. Consortium includes ITC - IRST, Centro per la Ricerca Scientifica e Tecnologica (Italy), Universität Karlsruhe (Germany), Carnegie Mellon University (USA), Université Joseph Fourier (France), Aethra (Italy), and Azienda per la Promozione Turistica del Trentino (Italy). It uses standard communication protocols that allow for seamless integration of the multilinguality with the existing videoconferencing software.

NESPOLE! aims to understand issues related to the ability of people communicating ideas, concepts, thoughts and to solve problems in a collaborative framework. It also includes non-verbal communication facilities in the form of multimedia presentations, shared collaborative work spaces, multimodal interactivity and manipulation of objects. These facilities allow for sharing text, graphics, audio, video, therefore providing an improved interpersonal communication. The languages addressed in the NESPOLE! project are Italian, English, German and French.

NESPOLE! identifies the following dimensions that should allow construction of the effective eCommerce and eBusiness environments

- *Robustness*: ability to cope with distractions of spontaneous speech (interruptions, corrections, repetitions, false starts).

- *Scalability*: ability to ensure an adequate level of system performance when the number of users increases.
- *Cross-domain portability*: defined as an easy and cost-effective porting of a speech-to-speech translation system to a new domain.
- *Multimedia and multimodal support*: facilitates the close integration of, and interaction between, speech-based communication and visual cues and content.

The NESPOLE! project envisions to build three different speech to-speech translation systems, including

- A system for tourism applications, embedding multimedia features.
- A system for tourism with a larger coverage of the domain, richer interaction modalities, more sophisticated multimedia support. This should demonstrate the progress on the scalability issue.
- A system for an advanced multilingual help desk. This system should highlight the results concerning the cross-domain portability.

These demonstration systems will support the multilingual negotiations between a tourist service provider and a customer aiming to organize eg, her or his holidays. Portability is addressed by porting the developed system consisting of a video help-desk for technical support, troubleshooting and repair to a different domain.

2.7. ORIENTEL

Multilingual access to interactive communication services for the Mediterranean and the Middle East, [wwwORIENTEL]

The ORIENTEL project explores potential of the multilingual communication services for Mediterranean and the Middle East. Emphasis is put on the mobile applications that are on rise globally. Neither resources nor sufficient expertise are currently available to cope with the linguistic research challenges of the area and the problems posed for Automatic Speech Recognition technology.

The project started in June 2001 with the consortium of Philips Speech Processing (Germany), European Language Resources Distribution Agency (France), IBM Deutschland (Germany), Knowledge (Greece), Natural Speech Communication (Israel), Siemens (Germany), Universitat Politecnica de Catalunya (Spain), and Lucent Technologies Network Systems (UK).

Main objectives of the ORIENTEL project are to:

- Outline survey analysis of markets, technologies, languages and users of mobile communication.
- Gain fundamental knowledge about linguistic structure of the target languages.
- Develop strategies and standards for phonetic and orthographic transcriptions.
- Collect 23 speech databases to support mobile communication applications.
- Research for language, dialect and foreign accent adaptation techniques.
- Develop demonstrator applications.

The project outcome will therefore significantly contribute to the spoken language resources distributed by the ELRA/ELDA.

3. US DARPA Projects

The US DARPA supports a large pool of projects under the Translingual Information Detection, Extraction and Summarization (TIDES) umbrella [wwwTIDES]. These research projects also address the core issues in portability of HLT mentioned above.

4. Conclusions

This paper reflected some of the on-going research and development efforts towards the challenges and opportunities in portability of HLT. It advocates for the view that *every* language of the world contributes to the cultural richness of the information society. This vision should also be applied when the HLT support need to be developed for a small-market or non-prevalent languages [Ostler, 1999]. Furthermore, research in portability issues of HLT for prevalent languages has recently shown that a system developed with a bigger set of languages may exhibit better performance than a system trained with a large set of the target language task-specific data.

Since *every* language is constantly changing while adapting to the influences brought by globalization and increased human mobility, it is reasonably to expect that the state-of-the-art performance in HLT could only be achieved when the HLT development phase included a grand pool of languages instead of only a particular one. Additionally, robust HLT needs to be adaptive to the user and the task involved.

In conclusion, research in portability issues of HLT should be encouraged and strengthened. This could be achieved by forming a network of excellence on HLT portability under the forthcoming 6th EU framework program. Since HLT portability is a very important, difficult and challenging research problem, such NoE should include all interested major players in the field, as many national HLT entities as possible, as well as researchers concerned with the non-prevalent languages (eg., the ISCA SALTMIL SIG) [wwwSALTMIL].

5. References

Bird, S. (2001). Annotation graphs in theory and practice, http://media.nis.sdu.dk/elsnet/annotationgraphs/ . *9th ELSNET European Summer School on Language and Speech Communication, Text and Speech Corpora.* Lectures are accessible at http://media.nis.sdu.dk/elsnet/

Lamel, L. (2002). Some Issues in Speech Recognizer Portability. *This Proceedings*, pp. 14-22.

Ostler, N (1999). Does Size Matter? Language Technology and the Smaller Language. *ELRA Newsletter, 4(2): pp. 3-5.* Paris. (ISSN 1026-8200)

URL list accessed in April 2002:

[wwwCORETEX] http://coretex.itc.it/

[wwwEAGLES] http://www.ilc.pi.cnr.it/EAGLES/ home.html

[wwwEANBLER] http://www.HLTCentral.org/projects/ENABLER/

[wwwFAME] http://www.HLTCentral.org/projects/ FAME/

[wwwHLTCentral] http://www.hltcentral.org/

[wwwHOPE] http://www.HLTCentral.org/projects/HOPE/

[wwwISLE] http://www.HLTCentral.org/projects/ISLE-HLT/

[wwwNESPOLE] http://nespole.itc.it/

[wwwORIENTEL] http://www.orientel.org/

[wwwSALTMIL] http://isl.ntftex.uni-lj.si/SALTMIL/

[wwwTIDES] http://www.darpa.mil/ipto/research/tides/

The Atlantis Observatory: Resources Available on the Internet to Serve Speakers and Learners of Minority Languages

Salvador Climent[*]
Miquel Strubell[*]
Marta Torres[*]
Glyn Williams[**]

*Universitat Oberta de Catalunya (UOC) / Internet Interdisciplinary Institute (IN3)
scliment@uoc.edu / mstrubell@uoc.edu / mtorresv@uoc.edu

**Foundation for European Research, Wales
g.williams@bangor.ac.uk

Abstract

The ATLANTIS Project (Academic Training, Languages and New Technologies in the Information Society) and its outcome, The Atlantis Observatory, are presented. The project's website (www.uoc.edu/in3/atlantis) brings together totally updated information on digital tools and resources available for Lesser-Used Languages of the European Union in a searchable database. The structure and classification of the database is explained and some preliminary results are also offered.

1. Introduction

Globalisation and the development and spread of digital technology in the Information Society provide excellent opportunities for creating spaces and tools for the use of many smaller languages. But the degrees of enterprise and know-how on which to draw from within the linguistic community vary, largely as a function of the size of that community. So, unless special support is given to such communities, there is a real danger that networks will develop only in larger languages, and particularly the hegemonic languages in the respective States, in rapidly growing areas such as the Internet. Thus the smaller linguistic communities, and especially those whose language is not that of the State, need to have at their disposal both products that can satisfy new demands, and platforms which will allow them to share initiatives with partners whose languages face a similar challenge.

In this framework, The ATLANTIS Project was aimed to create a virtual network that facilitates regular contact among individuals from all European Union lesser-used languages (LUL) to share knowledge on digital tools and resources available for such linguistic communities.

In the following section, the background and the main goals of the project are presented; then in section 3 we acknowledge the languages that are the subject of study. Sections 4 and 5 are devoted to describing the main areas to be analysed and their structuring and presentation in the database. Section 6 presents some preliminary results and reports. The paper ends with some concluding remarks.

2. The ATLANTIS Project. Baseline and objectives

The ATLANTIS Project, *Academic Training, Languages and New Technologies in the Information Society*, (funded by the EU under the terms of contract nº 2001 – 0265 / 001 – 001 EDU – MLCME) has been carried conjunctly by the Internet Interdisciplinary Institute (IN3) of the Universitat Oberta de Catalunya (Open University of Catalonia, UOC), the Foundation for the European Research University of Wales and the Onderzoeks Centruum voor Meertaligheid (Multilingualism Research Centre) of the Katholieke Universiteit Brussel (Dutch Language Catholic University of Brussels).

It leads on naturally from the Euromosaic report (Euromosaic, 1996), a study of the minority language groups of the European Union (EU) in order to ascertain their current situation by reference to their potential for production and reproduction, and the difficulties which they encounter in doing so. The Euromosaic report highlighted the shift in thinking about the value of diversity for economic deployment and European integration. It argued that language is a central component of diversity, and that if diversity is the cornerstone of innovative development, then attention must be given to sustaining the existing pool of diversity within the EU.

Now focusing on one of the various social and institutional aspects whereby a language group produces and reproduces itself –digital technology in the IS–, The ATLANTIS Project was designed to accomplish the following main objectives:

a. Bring together totally updated information on digital tools and resources available for Lesser-Used Languages.

b. Place the results on a new website –The Atlantis Observatory: www.uoc.edu/in3/atlantis/ – that will consist of a searchable database of the resources detected and thus duly classified.

c. Draw up a final report that will underline areas, projects and technology which, in the view of the participants, offer greatest potential for multiplying effects from one language group to another.

It must be noticed that these aims go along to a great extent with the general aims of the SALTMIL SIG (Special Interest Group on Speech and language Technology for Minority Languages) –promotion of research, development and education in the area of Human Language Technologies for less prevalent languages. Nadeu et al. (2001) specifically point that *"the vision of the SALTMIL SIG is that sharing of information and the forming of a network of researchers is important to begin with. It is hoped that this networking will form the seed-bed out of which more substantial projects will grow"*.

3. Languages targeted

The languages included in this study are all the autochthonous languages in the European Union which are not one of the eleven official EU languages –therefore, those minority languages which are EU official on account of being the official language in a neighbouring State are not included. In a few cases (such as Albanian or Slovene), though the language is official in a neighbouring State, it has been included because that State has not yet joined the enlarged Union.

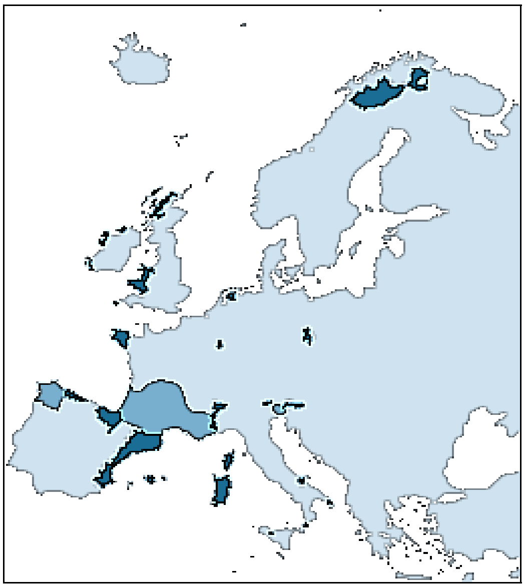


Fig. 1: Location of the languages targeted by the project

Therefore, languages targeted are the following (see Fig. 1):

1. Albanian (as spoken in Italy)
2. Asturian (Spain)
3. Basque (Spain, France)
4. Breton (France)
5. Catalan (Spain, France, Italy)
6. Cornish (UK)
7. Corsican (France)
8. Franco-provençal (Italy)
9. Frisian (Netherlands)
10. Friulian (Italy)
11. Gaelic (UK)
12. Galician (Spain)
13. Irish (UK, Ireland)
14. Ladin (Italy)
15. Luxembourgish (Luxembourg)
16. Occitan (France)
17. Sami (Finland, Sweden)
18. Sardinian (Italy)
19. Slovene (Austria, Italy)
20. Sorbian (Germany)
21. Welsh (UK)

4. Work package categories

Information from all EU LUL has been gathered in six parallel work packages:

1. Learning Platforms in LUL
2. Human Language Technology Developments
3. Information and Communication Technology: Regional Plans, computer software and Internet tools
4. Cultural Digital Resources and Linguistic Diversity
5. Convergence and LUL Broadcasting
6. Electronic Publishing and LUL

In order to do that, each partner took charge of a group of linguistic communities and distributed a comprehensive questionnaire to as many researchers, professionals and academic specialists they could contact. Those informants were also requested to circulate the questionnaire among other specialists in their fields. For those few linguistic communities where feedback resulted to be scarce, the partner in charge committed itself to gather information.

Each of the six work package categories is now described in more detail.

4.1. Learning Platforms in Lesser-used Languages

On-line learning offers cost-saving contexts for small dispersed populations and can thus be of considerable value for numerous language groups. In this section, information has been gathered on the extent to which LUL groups are incorporated into on-line learning platforms being developed in each of the European regions studied which have a LUL group. All levels of educational delivery have been studied, as well as the various associated training programmes. The information on the selected sites and products will allow potential users to see how knowledge resources are being made available in the LUL.

4.2. Human Language Technology Developments

Human language technology for lesser-used languages is the basis for much further development. The goals of e-mail, web page translation or discussion group translation require the appropriate technology for the language pairs that involve the LUL and the state language. Before this is possible, however, the basic requirements of such development have to be available: electronic corpora, dictionaries, spell checkers, grammars etc. These developments are expected to focus in on-line learning, administration and electronic publishing.

4.3. Information and Communication Technology (ICT): Regional Plans, computer software and Internet tools

Information and Communication Technologies are advancing fast. The extent to which Regional Authorities are addressing the issue, the importance they attach to the availability of tools in the relevant language, and the range of existing computer software and internet tools in each language, are the subjects of the category.

4.4. Cultural Digital Resources and Linguistic Diversity

The development, storage and accessing of digital resources in the context of the emerging Digital Economy requires the creation of Media Asset Management Systems. The extent to which this is proceeding within each region is an object of study. The development of appropriate resource locators allow such materials to be available not merely for industrial development based on the New Media sector, but also for on-line learning developments which, increasingly, will rely on digital resources. The EU's e-Content initiative is highly relevant to these developments.

4.5. Digital Convergence and Lesser-used Language Broadcasting

Many lesser-used language groups have their own audiovisual broadcasting media. The transition from solely analogue broadcasting, to the inclusion of digital systems, which a limited number of minority language communities have already embarked upon, opens up the potential of convergence. More and more audiovisual products are being made, and even shot, in a digital format. This is relevant to some learning developments and user-friendly platforms that encourage interactivity and can increase the potential of digital democracy.

4.6. Electronic Publishing and Lesser-used Languages

Electronic publishing in most LUL is already underway, if only, as happens in some cases, only through LUL web sites. The scope for low cost newspaper and journal publication has greatly expanded thanks to the web. Data has been gathered about the progress of such developments for all the language groups.

5. Structure of the database interface

The database has been organized according to the categories described above and several corresponding subcategories in a way that users can perform searches by language, by (sub)category or by any possible cross grouping of languages and/or (sub)categories.

The first category, *Learning Platforms in LUL*, is arranged for products and resources around two main axis: level (primary, secondary, tertiary and adult education) and area (language, science, mathematics and arts & social science). Moreover, users can search for online educational projects organised in two categories: (i) for learning and information purposes, and (ii) leisure oriented (games, etc.).

Due to its complexity, the *Human Language Technologies* (HLT) package is the one that has undergone a richer and stricter organisation. It has been tailored according to Sarasola (2000) levels and categories, which acknowledge the phases a minority language should follow to incrementally develop its HLT capabilities.

Sarasola's five phases have been simplified within ATLANTIS to the following three: (i) Foundations, (ii) Tools and Resources for Application Development, and (iii) Advanced Tools and Applications. Each one of such level-categories is divided in several field subcategories – such as Lexicon, Speech, Corpus, etc. These, at their turn, subdivide in types of tools, resources or applications – such as Database, Parser, Integrated System and the so.

Foundations is detached in three subcategories: Corpus (raw text), Lexicon and Morphology (raw lists, description of phenomena, different kinds of machine-readable dictionaries) and Speech (collections of recordings, descriptions).

The *Tools and Resources* category is in turn organised around five standard levels (Corpus, Lexicon and Morphology, Syntax, Semantics and Speech) each one including several tool subcategories (such as different kinds of parsers and knowledge bases), plus an Integration of Tools and Resources level.

Last, in *Advanced Tools and Applications* the following subcategories apply: Authoring Aids (spell, grammar and style checkers), Translation (Machine Translation and integrated Computer Assisted Translation environments), Information Retrieval and Extraction systems and advanced tools, Speech (synthesis, recognition, dialog systems) and Language Learning environments.

The third main category of the database, *Information and Communication Technology–Regional Plans, computer software and Internet tools*, is searchable by two subcategories: Regional Plans, and Software and Internet Tools.

The fourth, *Cultural Digital Resources and Linguistic Diversity*, is organised around seven kinds of media or resources: TV stations, Radio stations, Libraries, Museums, Music, Voice recordings and Other.

Last, the two remaining main categories, *Convergence and Broadcasting* (in fact, Radio or TV digitised) and *Electronic Publishing*, are not subdivided.

Every search in the database returns as output the list of matching items with the following information:

- Name of the product, a link to the URL of the product, name of the organization which has developed or is the owner of the tool, resource or application.

- A record with basic information about the tool, resource or application and the set of ATLANTIS categories under which it has been classified.

6. Preliminary results and reports.

At the moment we are writing this paper most of the data are still being gathered, studied and classified in order to produce six final per-category reports and the final overall report of the project. Nevertheless, we can already offer preliminary summary reports for the following languages: Breton, Friulian, Irish, Scots Gaelic, Slovene and Welsh (**§6.1** below); and Asturian, Basque, Catalan, Corsican, Galician, Occitan and Sardinian (**§6.2**). Such groupings simply correspond to work packages as distributed to the Atlantis Project research centres.

With respect to data figures, we can only show now as being reliable the total number of entries for the languages of the second group.

6.1. Breton, Friulian, Irish, Scots Gaelic, Slovene, and Welsh

For such languages, one can state the points that are detailed below.

6.1.1. Learning Platforms in Lesser-used Languages

All states are developing connectivity and establishing ICT (Information & Communication Technology) as a basis for its educational system. Those states that do acknowledge the relevance of minority languages for learning do not necessarily develop the tools and materials required for this to operate. However, this does not guarantee development. In Italy, the frontier agreement with Slovenia means that many of the developments for the Slovene language group await developments in Slovenia. In Austria on the other hand, the same language group does have the advantage of a concerted effort to develop supporting materials for the limited amount of teaching in Slovene. The main problem here is the tendency to interpret the legal requirement liberally, which means that the service is not very effective. In Scotland, connectivity is available but the developing of materials and the use of ICT is left to each individual learning enterprise and there is little central support. The situation is similar in Ireland by reference to Irish. In Brittany on the other hand the state makes virtually no provision for Breton medium education and therefore the limited amount of on-line learning that is available is the consequence of private initiative. The best situation appears to be in Wales where institutions responsible for developing on-line learning in Welsh match connectivity. This supported by the fact that the local authorities as learning providers are obliged to have their language plans confirmed by the Welsh Language Board. Also, the National Assembly for Wales, which has the sole responsibility for education in Wales, is devoted to developing a bilingual nation. Friulian lacks any support of this nature.

6.1.2. Human Language Technology Developments

Again, the situation is highly variable. In Wales, there have been certain developments but these have yet to developing machine translation and voice recognition capacity even with Welsh/English language pair. This is partly because the issue is driven by the translation agenda, which has become a powerful lobby rather than by economic needs. In both Austria and Italy the developments depends entirely on Slovenia, which is one of the few states in Europe that has not developed full capacity. In Ireland the picture is broadly similar to that in Wales whereas in Scots Gaelic has a limited presence even though dictionaries, corpora and grammars have been developed. In Brittany much of the initiative is the result of private efforts and is limited to on-line dictionaries, grammar checkers, etc. It is clear that this area requires considerable investment, usually by private commercial enterprises. Friulian also lacks any development other than limited private initiatives.

6.1.3. Information and Communication Technology: Regional Plans, computer software and Internet tools

Not all regions have such plans. Thus in Ireland there is little such coherent development even though the new initiatives in the West are developing plans which, between them, can be said to constitute regional technology plans. However, things are in their infancy and the failure of large companies to extend broadband to these areas is holding things back. Little is happening by reference to language in these areas but the awareness of the need to do so is high. In Scotland, such plans are in the hands of the Scottish Parliament and the Highlands & Islands Enterprise. The latter has responsibility for Gaelic but its plans make little reference to ICT and Gaelic. In Friulian and the Slovene border areas regional development is limited to European Regional Development Fund initiatives and there is little reference to language in such plans. The same can be said of Carinthia (in Austria) where the plans which are developed are relatively sophisticated but have little of relevance for the Slovene language group. Wales was one of the first to develop a Regional Technology Plan under the RISI programme of the EU. This has been superseded by the Cymru ar Lein' initiative. While there is a strong awareness of the need to incorporate Welsh development awaits the ability to incorporate the language into economic development writ large. In Brittany, the technological features of regional development make no reference to Breton.

6.1.4. Cultural Digital Resources and Linguistic Diversity

We must realise that we are in the beginning of any development of the Digital Value Chain. Thus far, it is unlikely that there is a regional DVC anywhere. Nonetheless, there are early developments. The Cymru'n Creu project in Wales is developing at least one end of the DVC. The exploitation end is emerging but is not articulated with the content end. The content end requires considerable investment whereas the production end does not. It is likely that Ireland will eventually develop one but is moving slowly in this direction at present. Scotland is in the same situation as Wales with SCRAN being an important innovative venture. SCRAN (Scottish Cultural Resources Access Network) was set up by museums, libraries and archives to create multimedia, manage digital IPR and provide educational access.

It is less likely that the other regions will be DVC regions and may well emerge as either content regions or production regions, more likely the later. This is largely because regional resources are housed in the capital region of the state so that initiatives will derive from that location on a state-wide basis. This does not preclude the emergence of regional eContent economies but it is less likely than in the historic regions with strong political autonomy. Carinthia is digitising some resources and there are multimedia companies capable of exploiting these but it is very limited. It is even less so among the Slovenes in Italy and also in Friulian. Brittany is in a similar situation. Whether the DVC regions focus on minority language digital resources depends on two things:

i. The extent to which they appreciate that diversity is driver of the Digital Economy and markets will be structured by language and not by states.

ii. The specific drive to incorporate minority languages into the New Economy.

6.1.5 Convergence and Lesser-used Language Broadcasting

This is also a matter of regional and central policy. The two language groups in north-eastern Italy will be hampered by the limited amount of exposure to media for the languages and the centralized nature of the broadcasting framework. However as costs plummet and deregulation takes hold, it will be possible to develop private initiatives. Carinthia also has a limited regional broadcasting presence and less so for Slovene. The Slovene language groups will, in all likelihood, benefit from deregulation and the entry of Slovenia into the EU, which will create a more integrated digital broadcasting region. Ireland is hampered by the size of its population and the dependence on terrestrial cabling which tends to be expensive. The main providers have recently pulled out and the state system is being partly privatised. Thus, development is hindered. Its minority language service will involve transformation of existing analogue services. Brittany has started developing a strong regional broadcasting capacity in the minority language and this will benefit from digitisation and the opportunities afforded by convergence.

6.1.6. Electronic Publishing and Lesser-used Languages

Electronic publishing is easier to conceive of partly because orthodox publishing in the minority language already exists and partly because of the relatively low cost. In all likelihood, this will be a parallel venture involving both orthodox publishing and electronic publishing existing side by side. The interesting developments involve exploiting convergence. This is already happening in Wales using Welsh where the main newspaper and the BBC are cooperating and also by reference to the community newspapers which are linked to the BBC's web service for the Welsh diaspora. As costs fall regional broadcasting and publishing will converge and will become far more localized. The publishing houses in Carinthia are also developing electronic Slovene language services. The Slovene newspaper in Italy is also available on-line but further developments are limited. Friulian has a limited development, as does the Gaelic language group in Scotland. Ireland's developments are also in a rudimentary state.

6.2. Asturian, Basque, Catalan, Corsican, Galician, Occitan and Sardinian.

For such languages we have collected and processed the following number of entries:

Asturian	50
Basque	408
Catalan	400
Corsican	50
Galician	225
Occitan	100
Sardinian	50

Some preliminary conclusions for this group are detailed below.

6.2.1 Learning Platforms in Lesser-used Languages

In this field we find a number of resources for the teaching/learning of languages on-line at different levels and for different target groups. Some are multilingual in nature, and a number are simple websites for adults, such as World Language Resources (which caters for Basque, Catalan, Galician, Sardinian, Corsican), Tandem Agency, etc. Monolingual language courses, grammars and lexicons are often offered by private individuals keen on disseminating their language on the net. For instance, an on-line Occitan course at

http://occitanet.free.fr/cors/intro.htm. Institutional support for developing and/or disseminating language educational products can be observed in most of these languages, such as *A Palabra Herdada. Curso de Galego*, promoted by the Dirección Xeral de Política Lingüística of Galicia.

Other educational projects are not for teaching the language, but rather use it as the medium of instruction. "*Recursos educativos para ciencias naturais*" in Galician (http://www.galego21.org/ciberlingua/recur.htm) is a good example of well-sorted links to available resources of this nature, but there are not many. Another is aimed at primary school education: *CD ikastola.net*. Nearly all material aimed at primary education is for Catalan or Basque: nothing has been detected in Occitan, Sardinian or Corsican, and a few tools are in Galician or Asturian. The same can also be said about secondary education, though in this case more Galician products have been found. At university level most Catalan universities offer on-line language courses both for non-native learners and for native speakers improving their literacy skills. Other tertiary level sites offer information on literature, e.g. Biblioteca d'Autores Asturianos at http://www.araz.net/escritores/ or philosophy resources in Galician, http://filosofia.00go.com/.

These are nearly all single products unrelated to digital educational platforms as such. Others cover leisure products which range from digital games such as *Trivial Pursuit euskaraz eta on line!* in Basque, at http://www.argia.com/tribiala.htm, to distribution lists and newsgroups in Occitan (soc.culture.Occitan, or the forum at http://www.oest-gasconha.com/listadif.php3).

There are, however a number of digital learning platforms. These are to be found in the virtual campuses of many universities such as the Universitat Oberta de Catalunya, which offers a range of degree courses, both undergraduate and postgraduate, in Catalan (http://www.uoc.edu/). The universities involved are virtually all Catalan (including, of course, Valencian universities) or Basque.

6.2.2. Human Language Technology Developments

Much of the digital work in the "Foundations" section has been done on Basque, by a wide range of organisations, many of which publicly supported. But all the languages studied do have at least some work in this area. Projects include untagged corpora, speech recordings and mono- or bi-lingual dictionaries. One example of an oral archive is the *Archivu Oral de la Llingua Asturiana*, http://www.asturies.org/asturianu/archoral/. In the "Tools and Resources for Application" section, Basque, Catalan and Galician seem to be the most productive. As regards taggers and tagged corpora, most of the work appears to have been done on Catalan and, to a lesser extent, Galician. Lexical and speech databases can be found in and for Catalan, Basque, Galician and for Corsican: www.ac-corse.fr/expos_autres/webdlc2/webdlc/Acceuil.html. No such developments have been found for Asturian, Occitan or Sardinian. Several terminology research centres offer resources on the Internet including Termcat (for Catalan) and UZEI (for Basque). In the field of Lexical-Semantic knowledge bases, WordNets have been developed for Catalan and Basque, and a Galician version is being developed.

Moving now to "Advanced tools and applications" we have found Authoring aids (spelling correctors and, just for Catalan, a grammar and style checker) for the following languages: Catalan, Galician and Basque. An Asturian product is about to be launched. Bilingual machine-translation systems have been developed for Basque, Galician and Catalan, such as the Basque-Spanish tool developed by the Basque government: http://www1.euskadi.net/hizt_3000/. Speech tools include those developed by Telefónica for speech recognition and synthesis for Catalan, Basque and Galician (alongside Spanish). Philips has developed the continuous-speech recognition tool *Free Speech* for Catalan only. No other languages in our group seem to have similar tools. Linguistic information retrieval and extraction tools have been located for Catalan and Galician. Web crawlers have been developed to manage Basque, Catalan and Galician.

6.2.3 Information and Communication Technology: Regional Plans, computer software and Internet tools

Regional ICT plans of greatly varying scope and objectives have been found at least for the following: Catalan (Catalonia, Balearic Islands), Basque and Sardinia. They also vary in the importance they attach to language in the plan. The Regional Development Plan (http://dursi.gencat.es/ca/de/pla_estrategic.htm) developed by the government of Catalonia does include specific projects related to the Catalan language.

As to software developed for the internet, languages such as Basque, Catalan and Galician have developed versions of the most widely used tools, such as several operating systems, the music file manager WinAmp (available also in Asturian) and web crawlers such as Netscape. Softcatalà, Softkat (for Basque) and Proxecto Xis-Galego 21 are three organizations devoted primarily to this work, as well as to developing new software. Very little has been found for Corsican, Occitan and Sardinian.

6.2.4. Cultural Digital Resources and Linguistic Diversity

In this section the digital treatment of libraries is interesting. Several cases have national libraries with limited digital services. All languages studied have at least websites reproducing literary and/or academic texts. Some libraries are fully digital: *Biblioteca Joan-Lluís Vives* is a good example of a resource containing digital versions of important Catalan literary texts. *InterRomania* has literary texts in Sardinian, Catalan and Corsican.

The small subsection on Museums is devoted strictly to those which offer digital resources related to the area, or from its own stocks, on the Internet.

The “voice” subsection contains a heterogeneous collection of resources relating to Basque, Galician, Sardinian and Corsican, from recited toponyms to full-scale digital archives.

The musical resources are plentiful. Significantly, each language studied has at least one website offering such recordings, with MP3, given the ease with which digitisation is possible. Thus the range of resources is enormous: from versions of original recordings on record. Some are sung, others are instrumental, and they range from traditional music to rock.

The section also includes a wide variety of other digital resources related to each culture: from photography to the visual arts, cartoons, catalogues of films made and/or dubbed (in Catalan). A high quality Galician multimedia resource centre is housed at http://www.culturagalega.org

6.2.5. Convergence and Lesser-used Language Broadcasting

The situation for radio and for television is somewhat different. Radio stations are available on the Internet in most of the languages, except Asturian. Al least 11 Catalan radio stations broadcast (live or stored) on the Internet, as do several of Basque stations. As regards television, digital satellite television is available in Catalan, Basque and Galician, whereas the picture for Internet TV is different: Catalan is not available, whereas Galician and Basque, and even Occitan, broadcasts are available. France 3 Corse has uploaded some of its programmes, a few of which (local news programmes) are in Corsican: http://www.france3.fr/semiStatic/382-1250-NIL-NIL.html. France 3 promises the same for Occitan and Catalan.

6.2.6 Electronic Publishing and Lesser-used Languages

The section is very rich in quantity and variety. The “Academia de la Llingua Asturiana” has fully four digital journals of a cultural and literary nature. Indeed, every language has similar journals. There are also other journals of a non-cultural nature, such as the Basque cooperative movement Eroski’s journal *Consumer*, which is published in Galician, Catalan (http://revista.consumer.es/web/ca/) and Basque as well as Spanish. http://codigocero.com/ is a Galician journal designed as a portal which offers information and news about new technologies.

Other linguistic products which have been regarded as electronic publications include multimedia encyclopædias, dictionaries (including interesting combinations such as Occitan-Basque), vocabularies and grammars. A CD-Rom on lesser-used languages, Lingua+, can also be viewed via the Internet. Publishers and/or sellers of electronic books include Basque houses, and the Catalan http://www.llibres.com, most of whose sales are still printed books. Electronic short stories in Corsican are sold through http://www.ac-corse.fr/fole2/fole.htm, where a demo can be viewed.

In dealing with daily newspapers a distinction has to be made between printed newspapers which also have an electronic edition, and strictly electronic dailies. In both cases a considerable investment is needed. Among the former there are many examples. Including http://www.egunkaria.com/ in Basque, and http://www.avui.com and http://www.diaridebalears.com/, among others, in Catalan. Among the latter we find Vieiros-Hoxe http://www.vieiros.com/ in Galician, and http://www.diaridebarcelona.com, which is run by the Barcelona city council, in Catalan.

Several languages have regular news services. Good examples are http://www.vilaweb.com, which operates in Catalan and uses mostly links with other electronic dailies, http://www.asturies.com/ in Asturian.

7. Concluding remarks

The Observatory has concluded its development phase, and has fulfilled one of its main aims: to bring together into its database a wide variety of initiatives relating to the new digital age. It is to be hoped that many of these will spur others into similar initiatives, hopefully working synergistically: as we stated at the outset, it is hoped that it will facilitate “regular contact among individuals from all European Union lesser-used languages to share knowledge on digital tools and resources available for such linguistic communities”.

What remains to be done, as we write this paper, is to determine to what extent each of these communities is well placed to enter the Digital Economy. Perhaps it is too ambitious, or pretentious, to imagine that we can pinpoint, for each community, which obstacles may appear in its drive towards the Digital Economy, as lack of basic tools, weak levels of networking, etc. Were this to be feasible, the Observatory could act as a useful reference point for planners.

At a less ambitious level, we are confident that users of the database will point out the inaccuracies and help us to continuously update the information it contains.

8. References

Euromosaic (1996) Euromosaic: The production and reproduction of the minority language groups in the European Union. www.uoc.edu/euromosaic.

Nadeu C., D. Ó Cróinín, B. Petek, K. Sarasola, B. Williams (2001) ISCA SALTMIL SIG: Speech and Language Technology for Minority Languages. In *Proceedings of EUROSPEECH 2001*. Alborg, Denmark

Sarasola K. (2000). Strategic priorities for the development of language technology in minority languages. In *Proceedings of Workshop on "Developing language resources for minority languages: re-usability and strategic priorities"*. Second International Conference on Language Resources and Evaluation (LREC 2000) Athens, Greece

Towards the definition of a basic toolkit for HLT

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

IXA Group
Dept. of Computer Languages and Systems
University of the Basque Country, 649 P. K.,
E-20080 Donostia, Basque Country

KSarasola@si.ehu.es

Abstract

This paper intends to be an initial proposal to promote research and development in language independent tools. The definition of a basic HLT toolkit is vital to allow the development of lesser-used languages. Which kind of public HLT products could be integrated, at the moment, in a basic toolkit portable for any language? We try to answer this question by examining the fifty items registered in the Natural Language Software Registry as language independent tools. We propose a toolkit having standard representation of data and develop a strategy for the integration, in a common framework, of the NLP tools.

1. Introduction

SALTMIL, the ISCA SIG (International Speech Communication Association Special Interest Group) on Speech and Language Technology for Minority Languages, has the overall aim of promoting research and development in the field of speech and language technology for lesser-used languages. Actually, its main activity is providing a channel of communication between researchers by means of workshops and the discussion list. The members of SALTMIL, we often wonder how to promote research and development in a more active way. In this paper we would like to propose a medium term project to accomplish that goal: the definition of a basic toolkit for HLT. Of course, this toolkit should be designed following the basic principles of reusability and portability[1]. So, the adoption of common standards and procedures will help to minimise costs and workload in research. This way will be beneficial for any kind of language (and vital for lesser-used languages), and would define a new collaboration-space for researchers working with different languages.

The real challenge is, however, how to define a basic toolkit for HLT? In this paper we will not resolve this problem, but we want to lay some foundations to address it. First, we will try to collect an initial list of present tools and applications that are portable (usable) for different languages:

- How many of the present HLT tools and applications are portable?
- How many of them are free for academic and public uses?
- Is there any tool for any of main basic applications? or… Is there any application with no accessible tool?

In this way, by recognizing which are the most basic tools, we propose four phases as a general strategy to follow in the processing of any language. Therefore, tools considered in the first phase will be taken as more basic than the later ones.

The paper is organized as follows: Section 2 proposes a strategy to develop language technology for language, grouping linguistic resources, tools and applications in four different phases. Section 3 examines the programs registered by the Natural Language Software Registry (NLSR) in order to determine the present proportion between portable and not-portable HLT products. Section 4 proposes a standard representation of linguistic data; it is a method we use in IXA Group in order to allow the integration between different tools in the same HLT framework; the standard representation would be fundamental for any possible basic toolkit. Finally, some concluding remarks are included.

2. Recognizing basic tools and their preference

We present here an open proposal for making progress in Human Language Technology. This proposal is based on the fifteen years experience of the IXA Group with the automatic processing of Basque. Anyway, the steps here proposed do not correspond exactly with those observed in the history of the processing of English, it is due to the high capacity and computational power of present computers allows arranging problems in a different way. We must remark that our work has been centered on the processing of written language and that we do not have any reliable experience on spoken language. However, in this proposal some general steps on speech technology have included.

Language foundations and research are essential to create any tool or application; but in the same way tools and applications will be very helpful in research and improving language foundations. Therefore, these three levels (language foundations, tools and applications) have

[1] Main themes chosen for the last two ISCA SALTMIL SIG workshops were "*Re-usability and strategic priorities*" (Athens 2000) and "*Portability Issues in Human Language Technologies*" (Gran Canaria 2002).

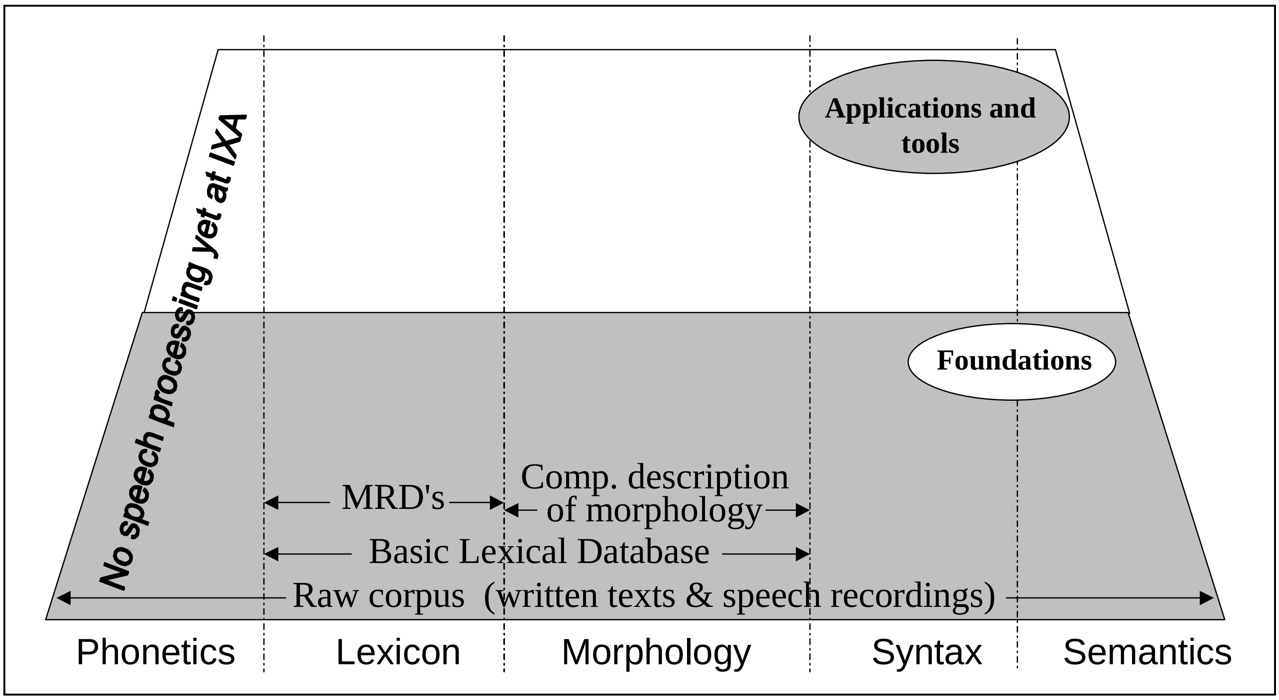

Applications and tools
Foundations
No speech processing yet at IXA
MRD's
Comp. description of morphology
Basic Lexical Database
Raw corpus (written texts & speech recordings)
Phonetics
Lexicon
Morphology
Syntax
Semantics

Figure 1. First phase: Foundations.

to be incrementally developed in a parallel and coordinated way in order to get the best benefit from them. Taking this into account, we propose four phases as a general strategy to follow in the processing of the language.

Initial phase: Foundations (see Figure 1).

- Corpus I. Collection of raw text without any tagging mark.
- Lexical database I. The first version could be simply a list of lemmas and affixes.
- Machine-readable dictionaries.
- Morphological description.
- Speech corpus I.
- Description of phonemes.

Second phase: Basic tools and applications.

- Statistical tools for the treatment of corpus.
- Morphological analyzer/generator.
- Lemmatizer/tagger.
- Spelling checker and corrector (although in morphologically simple languages a word list could be enough).
- Speech processing at word level.
- Corpus II. Word-forms are tagged with their part of speech and lemma.

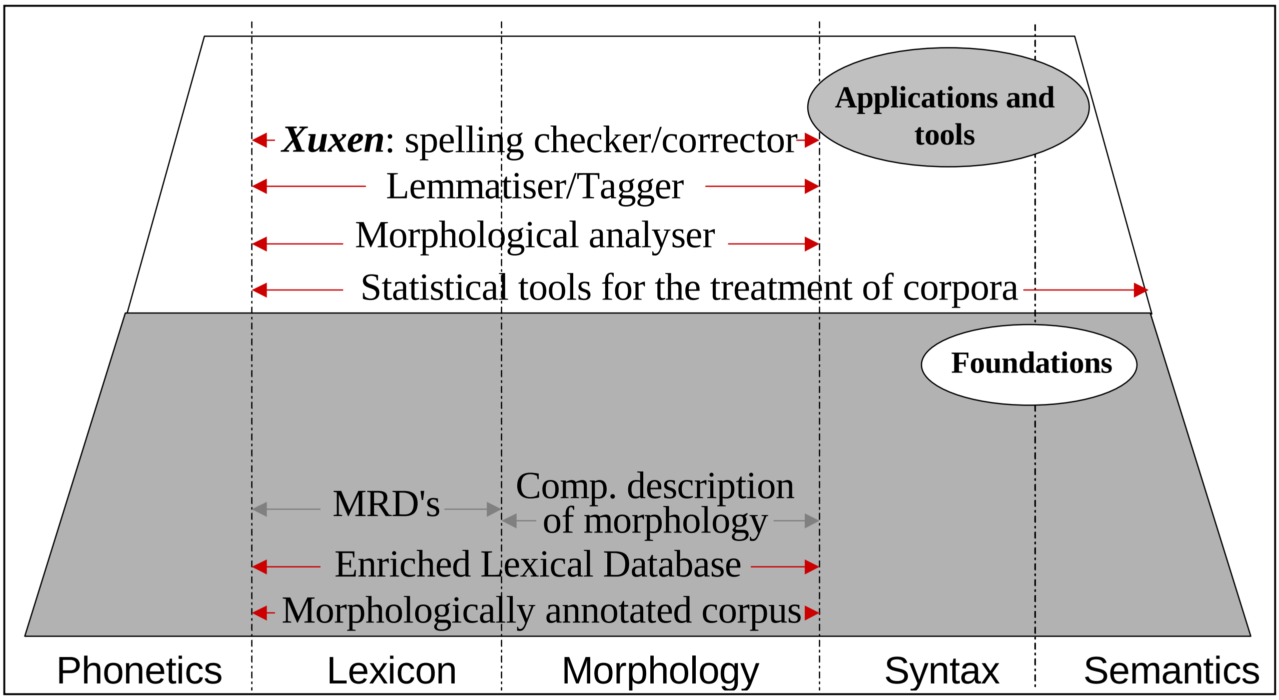

Applications and tools
Xuxen: spelling checker/corrector
Lemmatiser/Tagger
Morphological analyser
Statistical tools for the treatment of corpora
Foundations
MRD's
Comp. description of morphology
Enriched Lexical Database
Morphologically annotated corpus
Phonetics
Lexicon
Morphology
Syntax
Semantics

Figure 2. Second phase: Basic tools and application.

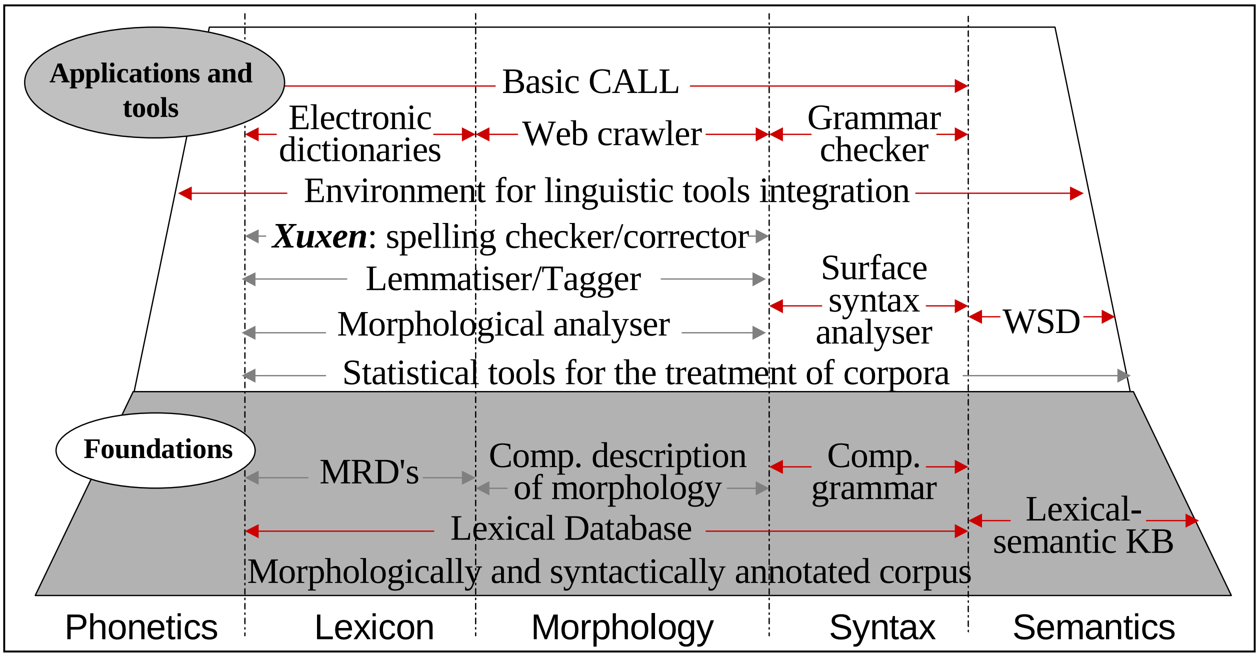


Figure 3. Third phase: advanced tools and applications.

- Lexical database II. Lexical support for the construction of general applications, including part of speech and morphological information.

Third phase: Advanced tools and applications.

- An environment for tool integration. For example, following the lines defined by TEI using XML. Section 4 describes this proposal.
- Web crawler. A traditional search machine that integrates lemmatization and language identification.
- Surface syntax.
- Corpus III. Syntactically tagged text.
- Grammar and style checkers.
- Structured versions of dictionaries. They allow enhanced functionality not available for printed or raw electronic versions.
- Lexical database III. The previous version is enriched with multiword lexical units.
- Integration of dictionaries in text editors.
- Lexical-semantic knowledge base. Creation of a concept taxonomy (e.g.: Wordnet).
- Word-sense disambiguation.
- Speech processing at sentence level.
- Basic Computer Aided Language Learning (CALL) systems.

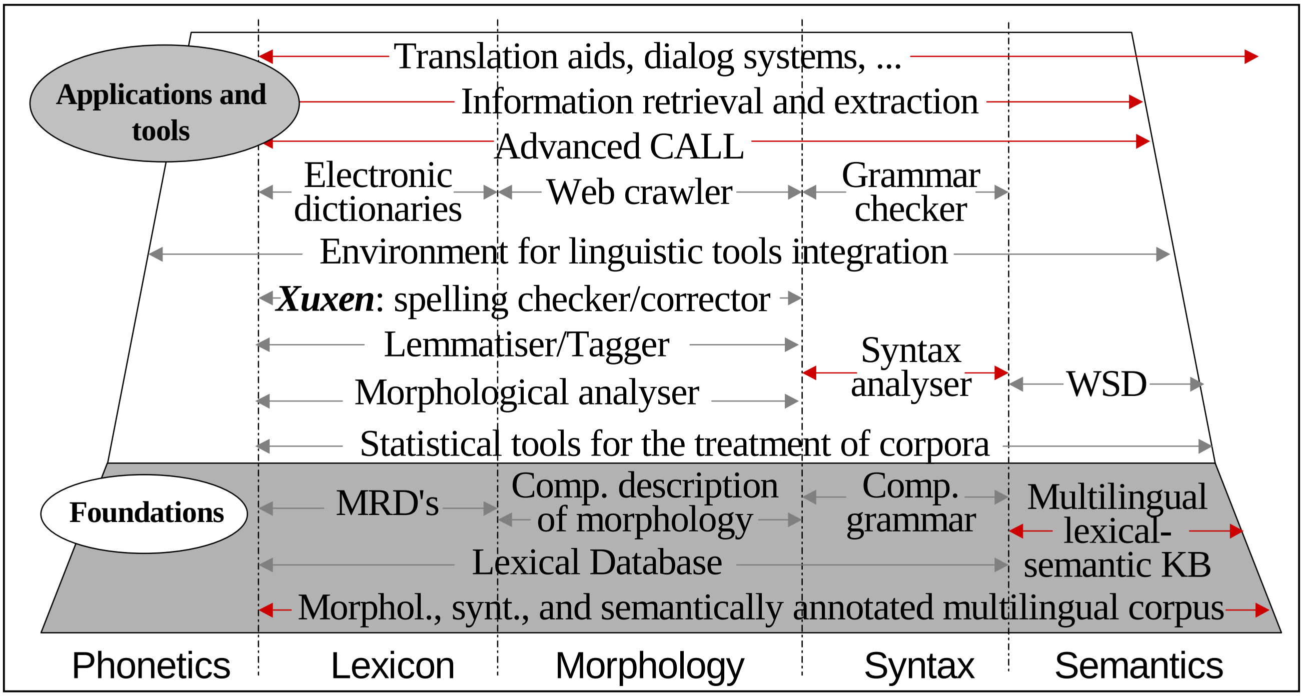
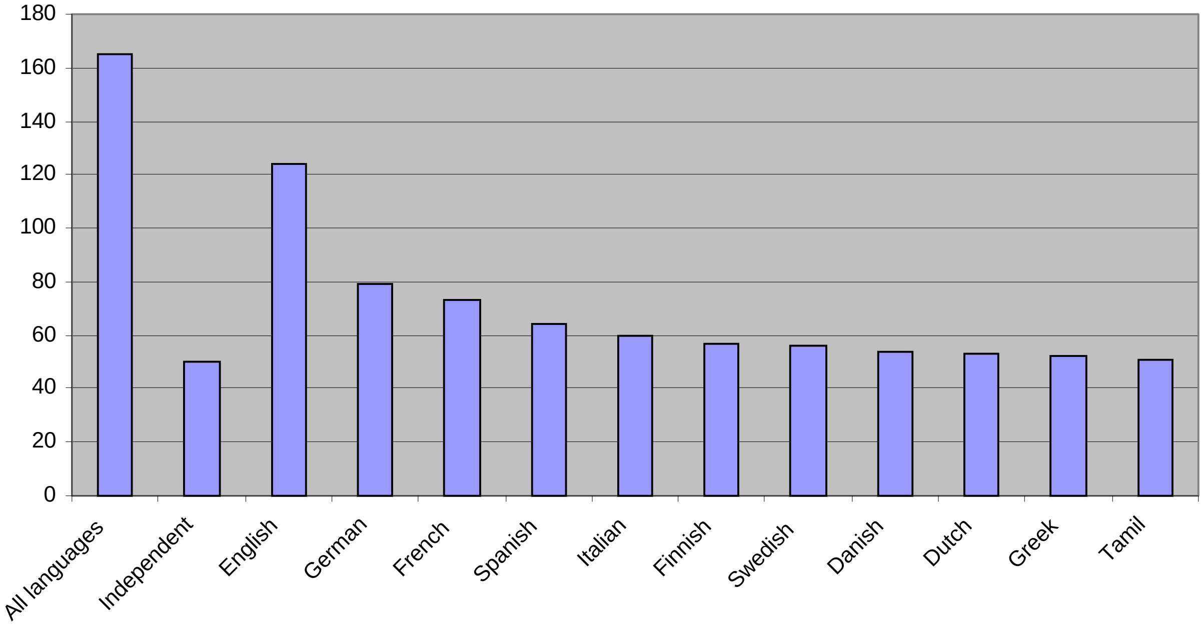

Applications and tools
Translation aids, dialog systems, ...
Information retrieval and extraction
Advanced CALL
Electronic dictionaries
Web crawler
Grammar checker
Environment for linguistic tools integration
Xuxen: spelling checker/corrector
Lemmatiser/Tagger
Morphological analyser
Syntax analyser
WSD
Statistical tools for the treatment of corpora
Foundations
MRD's
Comp. description of morphology
Comp. grammar
Multilingual lexical-semantic KB
Lexical Database
Morphol., synt., and semantically annotated multilingual corpus
Phonetics
Lexicon
Morphology
Syntax
Semantics

Figure 4. Fourth phase: Multilingualism and general applications..



Fourth phase: Multilingualism and general applications.

- Information retrieval and extraction.
- Translation aids. Integrated use of multiple on-line dictionaries, translation of noun phrases and simple sentences.
- Corpus IV. Semantically tagged text after word-sense disambiguation.
- Knowledge base on multilingual lexico-semantic relations and its applications.
- Dialog systems.

Now that we have started working on the fourth phase, every foundation, tool and application developed in the previous phases is of great importance to face new problems.

3. Present portable HLT products

Which is the start point at the present? Which kind of public HLT products could be integrated, at the moment, in a basic toolkit portable for any language?

With the aim of looking for data to answer to those questions, we examined the programs registered in the Natural Language Software Registry[2] (NLSR), an initiative of the Computational Linguistics Association (CL) and hosted at DFKI in Saarbrücken. The NLSR concentrates on listing HLT software, but it does not exclude the listing of linguistic resources (corpus, monolingual and multilingual lexicon). Other institutions, such as ELRA/ELDA or the Linguistic Data Consortium, provide listings of such resources. However, looking for portable products, to be precise, looking for products usable for multiple languages, the NLSR result sufficient because, actually, all linguistic resources are related to particular languages and so, they are not significant in this search. Of course, there are other HLT tools that have not been submitted to the NLSR, but we think that examine this database is a good start point.

3.1. Present proportion between portable and not-portable HLT products

First of all, we looked for how many of the present HLT tools and applications support different languages. This task was not very difficult because the system allows queries with a particular value for the slot named *Supported language(s)*. Figure 5 shows that a) the all amount of programs registered is 167; b) 50 of them (30%) has been declared to be language independent; c) of course, English is the language that support most of the programs. 125 support English (75%), that means that only 42 systems have been defined for the remaining 24 languages defined in NLRS; d) German, French, Spanish and Italian are the next languages an they are supported only by 79, 73, 64 and 60 respectively; and e) other languages are supported by those fifty defined as language independent and, occasionally, by a few other programs, for example 51 hits for Tamil. Those data reveals evident the significance of portability in Natural Language Software.

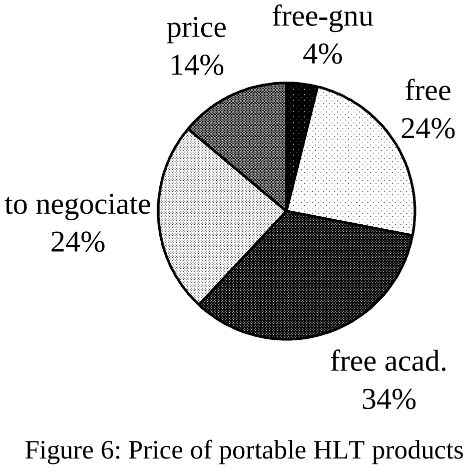


Figure 6: Price of portable HLT products

[2] http://registry.dfki.de

3.2. Price of portable HLT products

How many of the portable HLT products are free for academic and commercial uses? Among the fifty products they are 14 programs that free for any use (two of them, Zdatr and the speech synthesizer MBROLA, are distributed under the GNU Public Public License). Other 17 systems are free for academic uses. The price of 12 systems is defined as "to negotiate" even for academic uses. And finally 7 systems has a fixed price stated from $129 to $799; their average price is $546.

3.3. Distribution of portable products between HLT sections

Is there any portable tool for all the main basic sections in HLT? Or… is there any application with no accessible tools? Table 1 shows the distribution by sections of language independent software in NLSR. Similar data is shown for products that support English. We remark the following points: a) the number of products for the last four sections is not enough to be considered: b) the distribution of language independent products is similar to that of the total amount of products; c) there is any system in every section; d) the percentage of language independent products is considerable higher in Spoken Language and in NLP Development Aid.

Section	Total	Indep.	% indep.	Eng.	% Eng.
Total	167	50	0,30	125	0,75
Annotation	15	4	0,27	13	0,87
Written lang.	122	28	0,23	90	0,74
Spoken language	31	15	0,48	23	0,74
NLP development Aid	41	16	0,39	31	0,76
Lang. Resources	23	6	0,26	18	0,78
Multimedia	2	1	0,50	1	0,50
Multimodality	5	1	0,20	4	0,80
Evaluation	4	3	0,75	4	1,00

Table 1: Distribution of software by HLT sections

And now let's consider the distribution of NSLR products taking into account the kind of linguistic knowledge they manage. The kinds of knowledge to be considered are those referred in the previous section plus special points for NLP frameworks than includes facilities for lexical, morphology, syntax or speech. There is not any program to deal with dictionaries (creation of structured versions of dictionaries or integration of them in other applications), nor for semantics.

3.3.1. Corpus

Product	Description	Price
Alembic Workbench	a multi-lingual corpus annotation development tool	free
Bigram Statistics Package	Bigram analysis software	free
emdros	text database engine for linguistic analysis and research	free
PWA	Word Aligner	free acad.
SRILM -- SRI Language Modeling Toolkit	Statistical language modeling toolkit	free acad.
Entropizer 1.1	A toolbox for sequential analysis	to negotiate

Table 2: NLSR language independent products for corpus

3.3.2. Morphology

Product	Description	Price
PC-KIMMO	Two-level morphological analyzer	free acad.
TnT - Statistical Part-of-Speech Tagging	a statistical part-of-speech tagging for german, english and languages that delimit words with space	free acad.

Table 3: NLSR language independent product for morphology

3.3.3. Lexical databases

Product	Description	Price
DATR	A formalism for lexical knowledge representation	free
Xerox TermOnLine	Xerox TermOnLine is a terminology database sharing tool	to negotiate
Xerox TermOrganizer	Xerox TermOrganizer is a terminology database management system.	to negotiate

Table 4: NLSR language independent product for lexical databases

3.3.4. Speech

Product	Description	Price
IVANS: The Interactive Voice ANalysis System	Voice analysis, voice quality rating, voice/client data management	$749
CSRE - Computerized Speech Research Environment	speech analysis, editing, synthesis and processing system	$750
The OroNasal System	Nasalance measurement, analysis of oral and nasal airflow/energy in speech	$799
CSLU Toolkit	a comprehensive suite of tools to enable exploration, learning, and research into speech and human-computer interaction	free acad.
CSL -- Computerized Speech Lab	speech acquisition, analysis and playback	to negotiate
Signalyze(tm)	Interactive program for speech/signal analysis (runs only on Macintosh)	$350
TFR: The Time-Frequency Representation System	a comprehensive speech/signal analysis, editing and processing system	$599
Multi-Speech	a comprehensive speech recording, analysis, feedback, and measurement software program	to negotiate
WinPitch, WinPitch II	Speech analysis and annotation	to negotiate
ProTrain	speech analysis and speech production training system	$349
Praat	a research, publication, and productivity tool for phoneticians	free acad.
MBROLA	a speech synthesizer based on the concatenation of diphones	free-GNU
EULER	a freely available, easy-to-use, and easy-to-extend, generic multilingual TTS	to negotiate

Table 5: NLSR language independent product for speech

3.3.5. Syntax

Product	Description	Price
ASDParser and ASDEditor	Parser and editor for Augmented Syntax Diagram grammars, implemented in Java.	free
XLFG	Syntactic analysis using the LFG formalism	free
AGFL Grammar Work Lab	Formalism and tools for context free grammars	free acad.
CUF	constraint-based grammar formalism	free acad.
GULP -- Graph Unification Logic Programming	an extension of Prolog for unification-based grammar	free acad.
LexGram	development and processing of categorial grammars	free acad.

Table 6: NLSR language independent product for syntax

3.3.6. NLP framework

Product	Description	Price
Alembic	an end-to-end multi-lingual natural language processing system	free
The Quipu Grok Library	a library of Java components for performing many different NLP tasks	free
PAGE: A Platfrom for Advanced Grammar Engineering.	System for linguistic analysis and test of linguistic theory (HPSG, FUG, PATR-II). Can be used as part of a deep NLP system or as part of a speech system (a special version is used in Verbmobil).	to negotiate
TDL---Type Description Language	System for linguistic analysis and test of linguistic theory (HPSG, FUG, PATR-II). Can be used as part of a deep NLP system or as part of a speech system (a special version is used in Verbmobil).	to negotiate
QDATR	An implementation of the DATR formalism	free acad.
Kura	Kura is a system for the analysis and presentation of linguistic data such as interlinear texts.	free
Zdatr	Zdatr is a standardised DATR implementation in ANSI C	free-GNU

Table 7: NLSR language independent product for NLP frameworks

3.3.7. Applications

Product	Description	Price
BETSY - Bayesian Essay Test Scoring sYstem	Free Windows based text classifier/essay scorer	free acad.
Flag	Terminology, style and language checking	to negotiate
Universal Translator Deluxe	An omni-directional translation system	$129
Onix	High performance information retrieval engine	to negotiate
Brevity	Document summarization toolkit	to negotiate

Table 8: NLSR language independent product for applications

4. A standard representation for linguistic data using TEI conformant feature structures

The standard representation of linguistic data in order to allow the integration between different tools in the same HLT framework will be fundamental for any possible basic toolkit. In this section we present as a proposal the strategy used for the integration, in a common framework, of the NLP tools developed for Basque during the last twelve years (Artola et al.; 2000). The documents used as input and output of the different tools contain TEI-conformant feature structures (FS) coded in SGML[3]. These FSs describe the linguistic information that is exchanged among the integrated analysis tools.

The tools integrated until now are a lexical database, a tokenizer, a wide-coverage morphosyntactic analyzer, a general purpose tagger/lemmatizer, and a syntactic parser.

Due to the complexity of the information to be exchanged among the different tools, FSs are used to represent it. Feature structures are coded following the TEI's DTD for FSs, and Feature System Declaration (FSD) descriptions have been thoroughly defined.

The use of SGML for encoding the I/O streams flowing between programs forces us to formally describe the mark-up, and provides software to check that this mark-up holds invariantly in an annotated corpus.

A library of Abstract Data Types representing the objects needed for the communication between the tools has been designed and implemented. It offers the necessary operations to get the information from an SGML document containing FSs, and to produce the corresponding output according to a well-defined FSD.

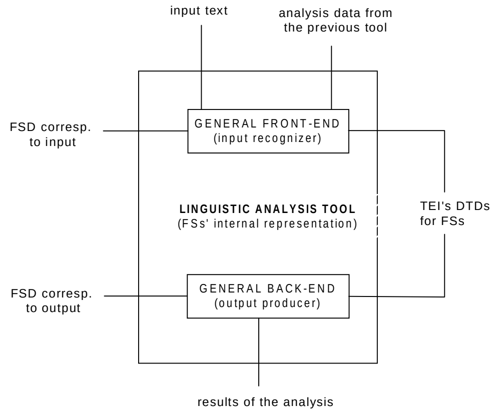

input text
analysis data from the previous tool
FSD corresp. to input
GENERAL FRONT-END (input recognizer)
LINGUISTIC ANALYSIS TOOL (FSs' internal representation)
TEI's DTDs for FSs
FSD corresp. to output
GENERAL BACK-END (output producer)
results of the analysis

Figure 7. Schematic view of a linguistic analysis tool with its general front-end and back-end.

The use of SGML as an I/O stream format between programs has, in our opinion, the following advantages:

a) It is a well-defined standard for the representation of structured texts that provides a formal framework for the internal processing.
b) It provides widely recognized facilities for the exchange of data: given the DTD, it is easy to process any conformant document.
c) It forces us to formally define the input and the output of the tools used for the linguistic analysis of the text.
d) It facilitates the future integration of new tools into the analysis chain.
e) Pieces of software are available for checking the syntactic correctness of the documents, information

[3] All the references to SGML in this section could be replaced by references to XML.

retrieval, modifications, filtering, and so on. It makes it easy to generate the information in different formats (for processing, printing, screen-displaying, publishing in the web, or translating into other languages).

f) Finally, it allows us to store different analysis sets (segmentations, complete morphosyntactic analyses, lemmatization results, and so on) linked to a tokenized piece of text, in which any particular analysis FS will not have to be repeated.

.

5. Conclusions

If we want HLT to be of help for more than 6000 languages in the world, and not a new source of discrimination between them, the portability of HLT software is a crucial feature. Looking for language independent software in the Natural Software Registry, we saw that only 30% of the tools has been so declared; that 62% of those language independent programs are at least academic free and that they are quite homogeneously distributed among the different sections of HLT and among the kinds of knowledge they manage.

As many problems would arise when trying to coordinate several of those language independent programs, we present as a proposal the strategy used for the integration, in a common framework, of the NLP tools developed for Basque. Feature structures are used to represent linguistic information, and feature structures are coded following the TEI's DTD for FSs, and Feature System Declaration descriptions (FSD) have been thoroughly defined.

Worldwide international organizations that work for the development of culture and education should promote the definition and creation of a basic toolkit for HLT available for as many languages as possible. ISCA SALTMIL SIG should coordinate researchers and those organisations to initiate such project.

References

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Soroa A.. A Proposal for the Integration of NLP Tools using SGML-tagged documents. In *Proc. of the Second Int. Conf. on Language Resources and Evaluation*. Athens (Greece). 2000

Petek B. "Funding for research into human language technologies for less prevalent languages" Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.

Sarasola K. "Strategic priorities for the development of language technology in minority languages". Proceedings of Workshop on "Developing language resources for minority languages: re-useability and strategic priorities". Second International Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece.

Ubiquitous multilingual corpus management in computational fieldwork

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Postfach 100131
D–33501 Bielefeld
Germany
gibbon@spectrum.uni-bielefeld.de

Abstract

The present application addresses the issue of portability in the context of linguistic fieldwork, both in the sense of platform interoperability and in the sense of ultra-mobility. A three-level networked architecture, the UbiCorpus model, for information gathering in the field is described: (1) a Resource Archive server layer, (2) a Data Processing application layer, and (3) a new Corpus Pilot layer designed to support specific fieldwork sessions under adverse conditions, for on-site questionnaire presentation and metadata editing.

1. Goals

In linguistic fieldwork,[1] conceptually the initial stage in any language documentation procedure, the issue of portability is important in two senses: first, the sense of platform interoperability and second, in the sense of ultra-mobility. This issue is addressed by the present application. A three-level networked architecture, the UbiCorpus model, for information gathering in the field is described: (1) a Resource Archive server layer, typically non-mobile, and distributed; (2) a Data Processing application layer, typically a local laptop or desktop; and (3) a new Corpus Pilot layer, designed to support specific fieldwork sessions under adverse conditions with questionnaire presentation and metadata editing, and typically, it is suggested, implemented on a handheld PDA. The UbiCorpus model is based on extensive fieldwork experience, mainly in West Africa. The Corpus Pilot layer is described in detail.

Owing to severe financial and platform resource limitations in practical linguistic fieldwork situations, the general development strategy is to use available freeware or open source components as far as possible, and to augment these with custom applications which are distributed as freeware for initial testing, and subsequently published as as open source software.

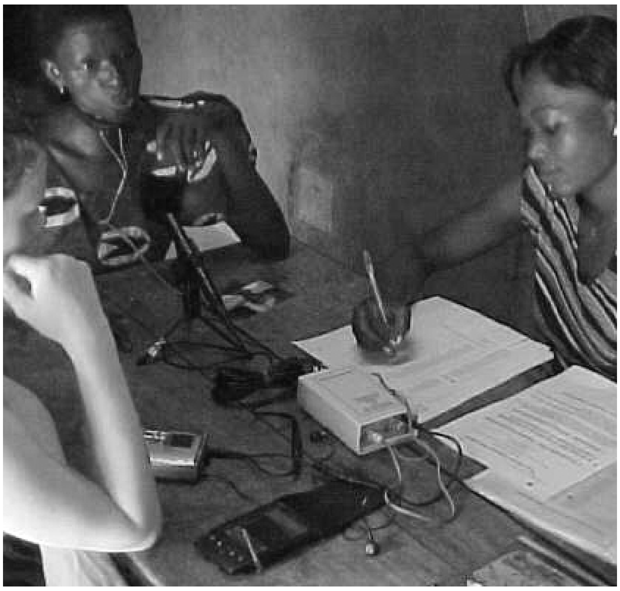


Figure 1: Questionnaire-based interview on Anyi syntax with Kouamé Ama Bié by Sophie Salffner & Sandrine Adouakou in Adaou, Ivory Coast (equipment: field laryngograph, DAT, Palm, pen & paper).

2. Requirements specification

Relatively recently, issues of corpus standards and resources as developed in the field of speech technology (Gibbon et al., 1997; Gibbon et al., 2000; Bird and Liberman, 2001) have been extended to fieldwork corpora in linguistics, ethnography, and related sciences, and specific issues such as the role of metadata in resource archiving and reusability have come to the fore, adding to the complexity of the documentation task facing the fieldworker. The present application area is computational support for this fieldwork documentation task within an integrated fieldwork resource environment. This concern is on the one hand more comprehensive than the currently popular issues of annotation-based data enhancement and web-based resource dissemination, and on the other hand orthogonal to these expensive technologies in that an effective but inexpensive practical new “low end high tech” technique for grass roots applications in geographically inaccessible areas is introduced.

From the perspective of field linguistics, language documentation traditionally consists in the main of field notes, an outline of the situation of the language, transcriptions, and generally including a sketch grammar consisting of basic phonology, morphology, and grammar, together with a lexicon containing glosses and examples and perhaps a thesaurus. The prompt materials for eliciting this kind of documentation are mainly systematic linguistic and ethnographic questionnaires, and the media for production of the documentation are generally office-oriented software such

[1] Grateful acknowledgements to Sandrine Adouakou, Firmin Ahoua, Doris Bleiching, Bruce Connell, Eddi Gbery, Ulrike Gut, Ben Hell, Sophie Salffner, Thorsten Trippel and Eno-Abasi Urua for discussion of problems addressed in this contribution.

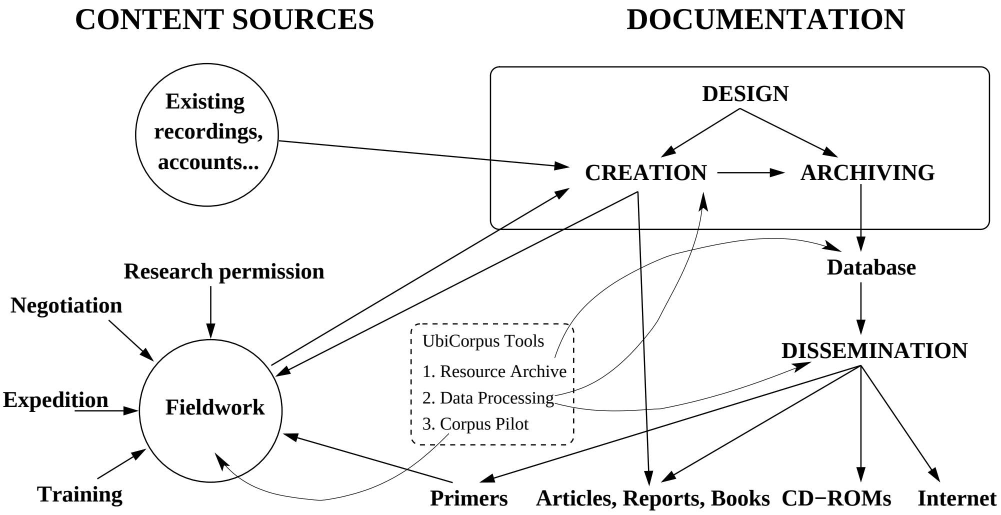


Figure 2: Language documentation logistics model.

as word processors (MS-Word etc.), DBMS (Access, FilemakerPro etc.), and spreadsheets (Excel, etc., also used for database entry). The guiding objectives of this concept of documentation are applications in the production of translations, terminologies, and alphabetisation materials.

The UbiCorpus model is designed to support this kind of fieldwork in the following main respects:

1. questionnaire presentation (either by database or in free format, as a plain text editor or with special formatting and rendering, for example by means of an IPA font),

2. transcription (either plain ASCII such as X-SAMPA, or in an IPA font),

3. metadata input.

One of the main advantages of the model is that when implemented on a modern palmtop device it provides a convenient, efficient and — important for many applications — inconspicuous method for the frequently neglected task of systematic on-site metadata logging.

However, the scope of the model is more general, and supports both the documentation of spoken language corpora in general, and further corpus processing in the form of the development of structured computational lexica (van Eynde and Gibbon, 2000) and computationally supported grammar testing. The UbiCorpus model is embedded in a comprehensive documentation model which covers not only the fieldwork activity itself, but the environment of preparation, archiving and application in which fieldwork is embedded (cf. Figure 2).

The first general operational requirement for the UbiCorpus model is portability. In the present context the term is systematically ambiguous:

- interoperability of applications on different OS and hardware platforms,

- compatibility of data formats through import and export filters for functionally equivalent or interfaced applications,

- ubiquity, i.e. time and place independent mobile deployment.

In the present context, the primary focus is on ubiquity, with interoperability and compatibility seen from this perspective.

Computational support for certain aspects of linguistic fieldwork has been available for many years, both for laptop-based data entry and initial analysis on the move or in isolated areas, and for desktop-based detailed descriptive work and document production (with increasing overlap between laptop and desktop functionalities). Software applications have been characteristically in the following areas:

- Lexical databases, either using general office DBMS such as FileMakerPro and MS-Access, or custom lexicon project software such as SIL's Shoebox; the latter also includes lexical support for textual glossing.

- Publication support such as DB export functions, fonts.

- Phonetic software, for signal analysis (e.g. general signal editors such as CoolEdit, or SIL's CECIL and signal analysis packages, or Praat) and for the symbol-signal time alignment (labelling) of digital recordings (e.g. Praat, Transcriber).

- Computational linguistic software for basic phonological, morphological and syntactic processing.

Some of this functionality (lexical databases, document production, computational linguistic processing) overlaps with the new Corpus Pilot layer, but this layer has the following characteristic additional fieldwork corpus acquisition functionality (Gibbon et al., 1997; Gibbon et al., 2000):

Pre-recording phase: planning of the overall corpus structure and contens, in particular design of corpus recording sessions, including the preparation of scenario descriptions, interview strategies, questionnaires, data prompts (for instance with prompt randomisation),

Recording phase: conduct of corpus recording sessions, including session management with the logging of metadata in a metadata editor and database, questionnaire consultation and data prompt presentation;

Post-recording phase: provision of recorded and logged data for archiving and processing, including metadata export, transcription, lexicon development, systematic sketch grammar support and document production.

3. Design: modules, interfaces

The language documentation model within which the UbiCorpus model is deployed is visualised in Figure 2; the documentation model was developed for project work in West Africa. The two components of the model with which the UbiCorpus tools are concerned are the *Creation* and *Archiving* component, and the *Fieldwork* information source. The latter is directly associated with the Corpus Pilot layer described below. The UbiCorpus model itself is visualised in Figure 3.

The three layers of the UbiCorpus model are characterised as follows:

Resource Archive (RA) layer

The bottom layer represents the archive database and the access and media dissemination functions associated with it. On the declarative side, a number of current language resource and documentation proposals may be assigned to the Resource Archive layer: a single resource database such as a corpus or a lexicon, a multiple resource database such as a browsable corpus or concordance system, a web portal constituting a large and systematic resource world, or an entire dissemination agency. On the procedural side, the Resource Archive layer provides search functions of various kinds, from standard browsing strategies to intelligent search and concordance construction, with token renderings of resources in any suitable media, whether entire corpora or lexica.

Data Processing (DP) layer

This is the layer which is familiar to the "ordinary working linguist". The data include paper fieldwork logbooks, transcriptions, sketch grammars and card index lexica; word processor and database versions of these; analog and digital audio and video recordings; time aligned digital annotations of recordings, and concordance or browsing software based on annotations; metadata catalogues for all of these Data Processing layer data types. Procedurally, the platforms and applications used at the Data Processing layer are very varied, though there is a tendency to go for platform independence and standardised data interchange formats. By using modern laptops, both the Resource Archive and Data Processing layers can be integrated into a single mobile environment.

Corpus Pilot (CP) layer

The top layer of the model represents the functionality which needs to be available in an actual fieldwork situation. This functionality can be very varied, and much — especially free format interviews and film recording — lies outside the range of systematic computational support. However, the following on-site support features can easily be covered:

1. metadata editor and database,
2. participant database for interviewee, interviewer etc.,
3. structured or free format questionnaire presentation.

Interfaces

The interfaces between these three layers, and modules within these layers, are defined mainly on the basis of generic ASCII formats, including XML annotated text, CSV database tables, and RTF formatted documents (including IPA font information). For the interface between a palmtop implementation of the Corpus Pilot layer and the Data Processing layer, conversion scripts are provided as required, in order to export palmtop database and text formats into the generic ASCII formats. Data transfer at the implementation level is via the usual synchronisation functions provided with handheld devices, or via scp, http, and ftp procotols for laptops, desktops and server.

4. Implementation: hybrid applications

Resource Archive (RA) layer

The server archive provides web portal access for the local and global linguistic communities, CD-ROM access for the local linguistic community, and analogue selections (in general, tape cassette, print media) for practical applications in the local user community. Currently, the leading models for the Resource Archive level are provided by the LDC and ELRA dissemination agencies; the E-MELD project is developing a general model for best practice in resource collation, and a meta–portal for flexible access to language resources. The local server currently used for initial database collation contains a number of specific search functionalities for corpus analysis, in particular an audio concordance (Gibbon and Trippel, 2002).

Data Processing (DP) layer

The classical environment for fieldwork data processing is a laptop, often a Mac, but also very frequently an Intel based device configured alternatively with Linux or MS based portable standard software. The kinds of application typically used are for basic corpus processing: Transcriber and Praat for transcription and annotation; Shoebox for lexical database development; MS Office or StarOffice for word processor, database and spreadsheet applications. These may be augmented with custom applications in Java (cf. the TASX engine (Milde and Gut, 2001)) and Perl (PAX audio concordance).

Corpus Pilot (CP) layer

The Corpus Pilot layer is implemented as custom-developed Palm compatible PDA applications. The rationale behind the use of the PalmOS based handhelds, as opposed to the use of a laptop, is based on the following considerations:

1. extremely inexpensive (in relation to other computational equipment),

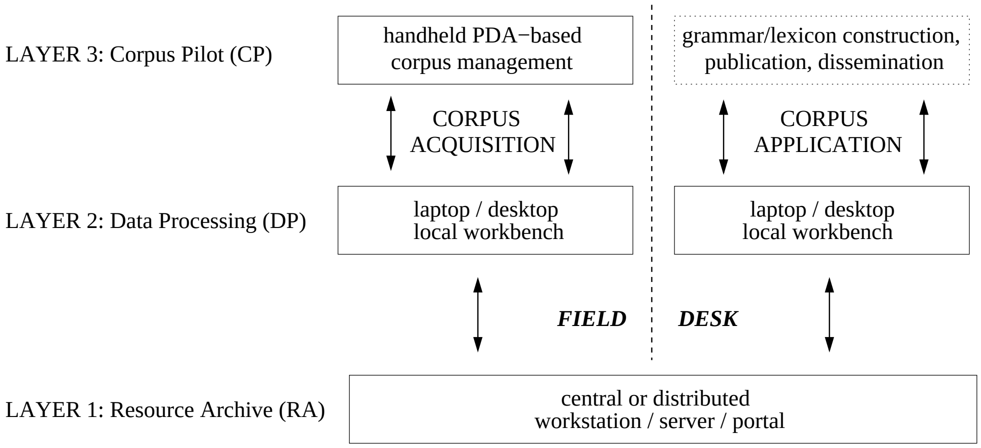

LAYER 3: Corpus Pilot (CP)
handheld PDA−based corpus management
grammar/lexicon construction, publication, dissemination
CORPUS ACQUISITION
CORPUS APPLICATION
LAYER 2: Data Processing (DP)
laptop / desktop local workbench
laptop / desktop local workbench
FIELD
DESK
LAYER 1: Resource Archive (RA)
central or distributed workstation / server / portal

Figure 3: The three layer UbiCorpus model.

2. ultra-lightweight (lighter than other standard portable fieldwork equipment such as field laryngograph, DAT recorder),

3. long operating cycle (with normal use, around 3 weeks on 2 AAA batteries or one charge), depending on model,

4. fast and highly ergonomic in use,

5. small and unobtrusive in the interview situation,

6. an integrated environment with other PDA functionalities such as calendar, diary, address and other databases, other custom applications in C and Scheme.

Networking

The three levels are networked by standard techniques: server-to-applications in general via TCP/IP-based protocols and mobile or landline telephone. The applications-to-acquisition via dedicated sychronisation software of the kind typically used to link handheld PDAs to desktop installations.

Use in the field

The satisfaction of these criteria points towards a high level of suitability for use in extreme fieldwork situations without power supplies, for instance in isolated outdoor locations (forest, village, etc.).

The functionality which has been included in the Corpus Pilot layer so far covers the following:

- Metadata editor and database for audio/video recordings, photos, paper notes, artefact cataloguing. This application is based on a widely used PalmOS DBMS application, HanDBase, which provides a wide range of input support facilities (popups, date picker, free format notes, etc.), as well as cross-table linking.

- Questionnaire administration. In general, free text format has been used for questionnaire administration, and responses have been recorded for later out-of-field processing. For some questionnaire types (e.g. demographic information), the HanDBase DBMS is used.

- Lexicon development tools. Three applications are used for lexical database input (excluding freely formatted notes):

  1. an Excel-compatible spreadsheet, QuickSheet, which permits export in either CSV or Excel format (Excel is widely used in field linguistics as a convenient input tool for lexical databases, because of the ease with which databases may be constructed and restructed, and because it has many database-like functions, as well as built-in arithmetic functions if required for corpus work),
  2. the HanDBase DBMS which is also used for the metadata editor database,
  3. an implementation of the DATR lexicon knowledge representation language in LispMe, a Scheme implementation for the PalmOS platform (this application is a more Data Processing layer oriented tool, but is included in the Corpus Pilot layer implementation suite for convenience).

- Transcription support. In general, transcription in X-SAMPA (Gibbon et al., 2000) is used, but if required, IPA fonts may be used with the WordSmith word processor for PalmOS devices; RTF import and export facilities are available.

- Statistics package for initial evaluations. This is also a more Data Processing layer application, but integrated into the Corpus Pilot layer; functions include all the measures used in basic experimental and corpus work (including random sorting, mean, median, standard deviation, standard error, as well as standard pairwise comparison measures).

- Context-free parser package for basic grammar development. This is another Data Processing layer application, which is integrated into the Corpus Pilot layer because of the convenience of the LispMe Scheme application in which the parser suite is implemented.

The metadata application has been selected for detailed description, because it is most immediately relevant to the issue of language resources.

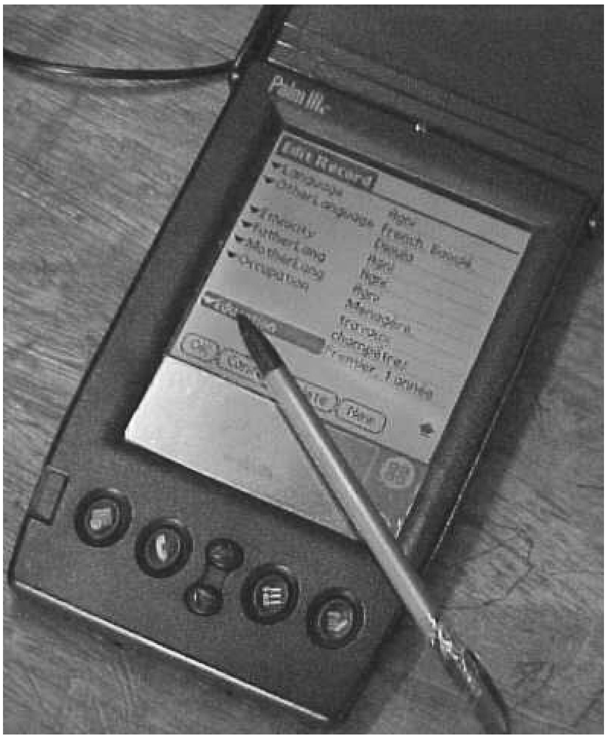


Figure 4: Palmtop metadata editor.

5. Metadata editor and database application

A metadata editor for audio/video recordings, photos, paper notes, artefact cataloguing was designed, based on a standard PalmOS relational database shell (HanDBase). The metadata editor provides a fast and inconspicuous input method for structured metadata for recordings and other field documentation, based on current work on metadata in the ISLE, E-MELD projects, and in the pilot phase of the DOBES project.

For the work in hand, standardised metadata specifications, such as the Dublin Core and IMDI sets, were taken into account. However, new resource types such as those which are characteristic of linguistic fieldwork demonstrate that the standards are still very much under development, since some of the standard metadata types are not relevant for the fieldwork data, and the fieldwork data types contain information not usually specified in metadata sets, but which are common in the characterisation of spoken language resource databases (Gibbon et al., 1997). In respect of the fieldwork resource type, it appears that it cannot be expected that a truly universal — or at least consensual — set of corpus metadata specifications will be developed in the near future, or perhaps at all, at a significant level of granularity. It may be possible to constrain the attribute list, though the existence of many different fieldwork questionnaire types belies this. However, the values of the attributes are in general unpredictable, entailing not only free string types but possibly unpredictable rendering types (e.g. different alphabets; scanned signatures of approval).

Indeed, it may be noted in passing that the expectation of fully standardising the entire metadata specification tends to reveal singularly little awareness of the potential of machine learning and text mining procedures for handling generalisation tasks of this kind. It may be predicted that such procedures will be applied in future not only to extensive resource data sets but also to increasingly extensive sets of metadata.

Table 1: Fieldwork metadata specifications.

Attribute	Type
RecordID:	string
LANGname(s):	popup: Agni,Agni; Ega
SILcode:	popup: ANY; DIE
Affiliation:	string
Lect:	string
Country:	popup: Côte d'Ivoire
ISO:	popup: CI
Continent:	popup: Africa; AmericaCentral; AmericaNorth; AmericaSouth; Asia; Australasia; Europe
LangNote:	longstring
SESSION:	popup: FieldIndoor; FieldOutdoor; Interview; Laboratory
SessionDate:	pick
SessionTime:	pick
SessionLocale:	string
Domain:	popup: Phonetics; Phonology; Morphology; Lexicon; Syntax; Text; Discourse; Gesture; Music; Situation
Genre:	Artefacts; Ceremony; Dialogue; ExperimentPerception; ExperimentProduction; History; Interview; Joke/riddle; Narrative; Questionnaire; Task
Part/Sex/Age:	string
Interviewers:	string
Recordist:	string
Media:	popup: Airflow; AnalogAudio; AnalogAV; AnalogStill; AnalogVideo; DigitalVideo; DigitalAudio; DigitalAV; DigitalStill; DigitalVideo; Laryngograph; Memory; Paper
Equipment:	longstring
SessionNote:	longstring

In consequence, the metadata specifications used in the UbiCorpus applications are deliberately opportunistic, in the sense that they are task-specific and freely extensible. A selection of attributes and values for the current fieldwork application are shown in Table 1. Metadata attributes concerned with the Resource Archive layer of archiving and property rights are omitted.

For current purposes, databases are exported in the attribute-value format shown below and converted into the TASX reference XML format (Milde and Gut, 2001). A specific example of the application of the metadata editor in the fieldwork session pictured in Figure 1 is shown in the exported record shown in Table 2.

The metadata editor and database application has been tested extensively in fieldwork on West African languages, and has proved to be an indispensable productivity tool, especially in difficult situations where very limited time is available.

6. Conclusion

Architectures using the first two levels, e.g. a server configuration and a laptop for use in the field, are very com-

Table 2: Fieldwork metadata example.

Attribute	Value
RecordID:	Agni2002a
LANGname(s):	Agni, Anyi
SILcode:	ANY
Affiliation:	Kwa/Tano
Lect:	Indni
Country:	Côte d'Ivoire
ISO:	CI
Continent:	Africa
LangNote:	
SESSION:	FieldIndoor
SessionDate:	11.3.02
SessionTime:	8:57
SessionLocale:	Adaou
Domain:	Syntax
Genre:	Questionnaire
Part/Sex/Age:	Kouamé Ama Bié f 35
Interviewers:	Adouakou
Recordist:	Salffner, Gibbon
Media:	Laryngograph
Equipment:	1) Audio: 2 channels, 1 laryngograph, r Sennheiser studio mike 2) Stills: Sony digital 3) Video: Panasonic digital (illustration of techniques)
SessionNote:	Adouakou phrases repeat

mon. However, in many situations the laptop concept is unsuitable because of heavy power requirements which are not available in many fieldwork locations. For these applications, the PalmOS based family constitutes the platform of choice because of minimal size and power requirements, permitting several weeks use on one charge or small battery. Although the PalmOS platform is obviously unsuitable for signal processing applications (such as time-aligned annotation) it is well-suited for logging, transcription and reference purposes.

The power of PDA miniature computing platforms as useful components of laboratory and office environments is often underestimated, and we demonstrate that a number of applications for which even a laptop is clumsy or unsuited for the developing field of computational ethnolinguistic fieldwork may be elegantly provided on the Palm PDA platform. The addition of a foldable keyboard further enhances the text handling capacity of the devices.

In the medium term, it will be possible to integrate the hybrid applications at the Corpus Pilot, Data Processing and Resource Archive levels into a corpus management environment which not only permits seamless dataflow and workflow, a goal already achieved, but also into a non-technical user-friendly prototype which may serve as the basis of a fieldwork management product implementation.

The UbiCorpus architecture has been used as the basic specification for different kinds of language documentation work in a variety of different projects. The Resource Archive layer was originally designed and implemented for web–based lexical database development in the VerbMobil project (Wahlster, 2000), funded by the German Federal Ministry of Education and Research (BMBF). The concept has been further developed theoretically and practically in connection with the projects *Theorie und Design multimodaler Lexika* funded by the German Research Council (DFG), *Enzyklopädie der Sprachen der Elfenbeinküste* funded by the German Academic Exchange Service (DAAD) and *Ega: a documentation model for an endangered Ivorian language* in the pilot phase of the DOBES funding programme of the Volkswagen Foundation.

In its local implementation, the current Resource Archive layer version also includes support for telecooperation and web-teaching. The Data Processing layer includes numerous applications which cannot be specified here. The Corpus Pilot layer as described in the present contribution has been informally but extensively field tested at a number of fieldwork locations, most recently in the framework of DAAD funded doctoral thesis work. It is planned to apply the field testing criteria defined in (Gibbon et al., 2000) to an extended implementation of the components of UbiCorpus model.

7. References

Steven Bird and Mark Liberman. 2001. A formal framework for linguistic annotation. *Speech Communication*, (33 (1,2)):23–60.

Dafydd Gibbon and Thorsten Trippel. 2002. Annotation driven concordancing: the pax toolkit. In *Proceedings of LREC 2002*. LREC.

Dafydd Gibbon, Roger Moore, and Richard Winski, editors. 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin.

Dafydd Gibbon, Inge Mertins, and Roger K. Moore, editors. 2000. *Handbook of Multimodal and Spoken Dialogue Systems, Resources, Terminology and Product Evaluation*. Kluwer Academic Publishers, Boston/Dordrecht/London.

Jan-Torsten Milde and Ulrike Gut. 2001. The TASX-engine: an XML-based corpus database for time aligned language data. In *Proceedings of the IRCS Workshop on Linguistic Databases*, Philadelphia. University of Pennsylvania.

Frank van Eynde and Dafydd Gibbon. 2000. *Lexicon Development for Speech and Language Processing*. Kluwer Academic Publishers, Dordrecht.

Wolfgang Wahlster, editor. 2000. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer Verlag.

A Theory of Portability

Hyo-Kyung Lee[*]

[*]Department of Computer Science
1304 W. Springfield Ave., Urbana, IL 61801, USA
h-lee5@cs.uiuc.edu

Abstract

To discuss the portability of human natural language technology, it is necessary to define the portability precisely first. If one claims that his or her language technology works for other languages, how can we verify such claim when every language has a different set of features, i.e. speech or text tagging system? This paper presents a view of protabilty as a function of a common representable set of features and argues that the development of such representation is critical in discussing portability issues.

1. Introduction

If you try a sentence boundary identification program[1] developed for English, you will easily notice that it does not work for other language such as Korean. However, the developers never mention that the program will not work for other languages. It is a very common practice among developers to ignore the portability issues in human language related technology because it is often targeted for only one language and assumed to work for that language. Yet, such ignorance is missing too many opportunities for the future success of the technology. If a technology that works for one language can be extended to another language with a minimal modification, such technology can be regarded as the most valuable technology in its potential considering the fact that there exists more than 6,000 languages in the world.

The main difficulty of applying a technology that works for one language into another is obviously due to the set of features that are unique to one specific language. The more the technology resorts to those features, the less it will succeed on other languages. In this regard, it is necessary to separate out those features and to concentrate on the common features that every language shares for maximal portability. Finding the common grounds for all languages is not an easy task but can be achieved by abstracting levels of language processing into hierarchy. In other words, there are different levels of sources that hinders the portability during human language processing and the portability problem should be discussed in as high level as possible.

This leads to the central question of this paper: if one claims that one's language technology works for other languages, how can we verify such claims when every language has a different set of features, i.e. speech or text tagging system and quantify its portability? The only way to determine it is to test how many features are translatable into the common feature sets which are similar to the interlingua in machine translation. This paper presents a view of portability as a function of a common representable set of features and argues that the development of such representation is critical in discussing portability issues.

2. Sources of Portability Problem

To identify the sources of portability problem in human language technology, it seems wise to illustrate it with the actual examples that might occur in statistical machine translation, namely sentence boundary identification for aligning sentences and word sense disambiguation for word selection. Here, we identify two categories in portability problems and formalize it. For the rest of the paper, we use a term *program* to denote a particular instance of human language technology.

2.1. Representational Problem

Almost every language has its own unique set of features. At the same time, some languages share many common features. For example, semantic or syntactic features like the notion of person's name and noun phrases are quite universal. On the other hand, honorifics used in Korean or Japanese language is hard to find in European languages. Such uniqueness of features is the major obstacle for the portability of human language processing program. For example, if you want to create a sentence-aligned corpus for statistical machine translation, the first step is to identify sentence boundaries. If one program uses the notion of capitalized word to determine whether the period is used for abbreviation or not, it won't work for a language, like Korean, that doesn't have any notion of the capital word in its writing system. For a program to be fully portable, it should avoid using such features.

Clearly, we can distinguish between two different feature sets which we will call *soft* and *hard* features. If some features are common in two or more languages, we call them soft features; otherwise hard features. Soft features are ubiquitous in the same families of languages and they are all functionally equivalent. One key observation is that features are independent from the surface forms of one particular language. For example, the same parsing program can be used to parse two different languages although grammatical notations of the languages are different as long as they can be mutually translated into the equivalent representation.

2.2. Functional Problem

Although two languages share the same soft features, not all functions consistently generate the desired outputs based on them. Let's assume that we want to disambiguate

[1] http://l2r.cs.uiuc.edu/~ cogcomp/cc-software.html

senses of English words based on the local context feature such as n-grams to find the corresponding Korean words in statistical machine translation(Ng and Lee, 1996). If you could achieve 90% of accuracy on the task with such method in English, it does not guarantee the same accuracy in Korean. The reason for such discrepancy can be attributed to the previous representational problem but the key issue that we want to emphasize here is only the performance aspect of the program. This is a separate dimension of portability problem which is related to the performance-wise consistency issue. A good portable program should perform well with the minimal variance and high accuracy across several languages. For example, if a machine translation system that performs very well on English-Korean translation fails on English-Japanese translation with the same soft set of features, we can say that such system has a functional problem in portability.

3. Theory of Portability

In this section, we present a functional view of portability in a more formal way by providing definitions and examples first and then theorems derived from them.

3.1. Definition of Portability

Definition 1 Features *are any properties of language that are used in a program P as inputs X and outputs Y. A* program *P is a collection of* functions *f.*

Example 1 *Period, question mark, and exclamation mark are features used in English for sentence boundaries.*

Definition 2 *Let's denote a set of all features for language L as $\chi(L)$.* Soft features *Z for a set of languages $L_1, \ldots, L_n$ are features s.t. $\forall z \in Z, z \in \bigcap_i^n \chi(L_i)$. All other features that are not soft features are* hard features *$Z' = \bigcup_i^n \chi(L_i) - Z$.*

Example 2 *Let $n = 2$ and L_1 is Korean and L_2 is English. Period, question mark, and exclamation marks are soft features used in both L_1 and $L2$ for sentence boundaries. The capitalization of words is a hard feature unique in L_2.*

Definition 3 *A family of languages are called σ-*similar *if $\frac{|Z|}{|Z|+|Z'|} = \sigma$ w.r.t. P.*

Example 3 *Let's assume that a sentence boundary identification program P uses only four features: period, question mark, exclamation mark, and a test value $\{0, 1\}$ for capitalization for the word that ends with a period. Again, if $n = 2$ and L_1 is Korean and L_2 is English, two languages are $\sigma(= 0.75)$-similar since $|Z| = 3$ and $|Z'| = 1$.*

Definition 4 *A function $f{:}\, X \rightarrow Y$ is called σ-*portable *for n σ-similar languages $L_1, \ldots, L_n$, if $X \subseteq Z$ and $Y \subseteq Z$ for a soft feature set Z.*

Example 4 *Let f be a classification function that uses the previous four features in Example 3 as input X and the boolean truth value $\{0, 1\}$ (to indicate a sentence boundary) as output Y. Then, f is $\sigma(= 0.80)$-portable.*

Definition 5 *A program P is σ-*portable *iff $\exists\sigma > 0$ among n languages and all f are σ-portable over $\sigma \cdot \sum_{i=1}^n |\chi(L_i)|$ soft features.*

Example 5 *Let P be a sentence identification program that has two functions f_1 and f_2. Let f_1 be the classification function in the previous example and f_2 be a boolean function that test the capitalization of words. Then, P is* not *σ-portable for English and Korean because f_2 uses hard feature. If P has only one function f_1, we can claim that P is σ-portable.*

Definition 6 *A function $f{:}\, X \rightarrow Y$ is called ϵ-*portable *over n languages $L_1, \ldots, L_n$, if $\forall i, j (1 \leq i, j \leq n), Pr(f_i \neq f_j) < \epsilon$*

Example 6 *Let f be a classification function in Example 4. Since Korean sentences do not use periods for abbreviation purposes, it is easy to see that $Pr(f_k = 1) > Pr(f_e = 1)$ when equal number of examples are represented with soft features. If the difference $Pr(f_k = 1) - Pr(f_e = 1)$ in such empirical performance of f over two languages is less than the predefined bound 0.02, we can say that f is $\epsilon(= 0.02)$-portable.*

Definition 7 *A program P is ϵ-portable if all functions $f \in P$ are ϵ-portable and generates the coherent output over n languages with the confidence at least $1 - \delta$ for some small ϵ, δ.*

Example 7 *Let P is a program that has only one f in Example 6. If f is tested on Japanese and also produced the result of $Pr(f_k = 1) - Pr(f_j = 1) < \epsilon$ and $Pr(f_e = 1) - Pr(f_j = 1) < \epsilon$ with at least $1 - \delta$ accuracy over many examples, we can say that P is $\epsilon(= 0.02)$-portable.*

3.2. Theory of Portability

Here, we introduce the notion of portability similar to the learnability notion in the learning theory (Valiant, 1984). The first theory is related to the representational problem.

Theorem 1 *Soft features are harder to obtain as the number of languages n increases.*

$$|Z_n| \geq |Z_{n+1}|$$

Corollary 1 *Hard features are easier to obtain as the number of languages n increases.*

$$|Z'_n| \leq |Z'_{n+1}|$$

Proof This is obvious from the definition of soft features. Since soft sets are extracted from the common feature sets, there are less features than previous n-th soft feature set unless all features are soft features in $n+1$-th language. On the other hand, hard features are obtained from the union of the feature sets and the size of them grows over n. ∎

Lemma 1 *σ is monotonically decreasing over the number of languages n increases.*

Now, let's look at the hardness of portability issue which is the main topic of this paper by combining two parameters — σ and ϵ. First, let's define probably approximately correct(PAC) portability as follows:

Definition 8 *A program P is* PAC-portable *if P is both σ-portable and ϵ-portable.*

If we apply the same technique used in (Haussler, 1988), we get the *PAC-portability bound* which is similar to the PAC-learnability bound as shown in Equation (1). It is adapted from (Mitchell, 1997) for illustration purpose.

$$m \geq \frac{1}{\epsilon}(\ln |H| + \ln(1/\delta)) \qquad (1)$$

We can replace m with $\binom{n}{2} = \frac{n(n+1)}{2}$ and $|H|$ with $2^{\sigma k}$ where $k = \sum_{i=1}^{n} |\chi(L_i)|$ by assuming that the number of different functions are only dependent on the size of soft set $|Z| = \sigma k$.

Theorem 2 *If a program P is PAC-portable with some small σ, ϵ and δ, the total number of portable languages n is bounded by:*

$$\frac{n(n+1)}{2} \geq \frac{1}{\epsilon}(\sigma k \ln 2 + \ln(1/\delta)) \qquad (2)$$

The following proof is essentially same as what Haussler (1988) showed.

Proof Let F be all the functions that use soft features in n languages. Clearly, there are at most $|F| = 2^{\sigma k}$ possible functions in P over σk soft features and $k = \sum_{i=1}^{n} |\chi(L_i)|$ is a constant for n languages. Let $f^1, f^2, \ldots, f^l$ be all functions in P such that each pair of functions (f_i^l, f_j^l) over two languages i, j have true error greater than ϵ. We need to consider the all pairs of functions over n languages and there exists $\binom{n}{2} = \frac{n(n+1)}{2}$ pairs for each f^l. We fail in ϵ-portability if and only if at least one of these f^l pairs fails. The probability that at least one of these will be consistent with all n languages is at most

$$l(1-\epsilon)^{\frac{n(n+1)}{2}}$$

And since $l \leq |F|$, this is at most $|F|(1-\epsilon)^{\frac{n(n+1)}{2}}$. Finally, we use an inequality that if $0 \leq x \leq 1$ then $(1-x) \leq e^{-x}$. Thus,

$$l(1-\epsilon)^{\frac{n(n+1)}{2}} \leq |F|(1-\epsilon)^{\frac{n(n+1)}{2}} \leq |F|e^{\epsilon\frac{n(n+1)}{2}}$$

upper bound holds.

We can use the above result to determine the number of languages required to reduce the portability failure below some δ.

$$|F|e^{\epsilon\frac{n(n+1)}{2}} \leq \delta$$

which means that we need

$$\frac{n(n+1)}{2} \geq \frac{1}{\epsilon}(\ln |F| + \ln(1/\delta))$$

By substituting $|F| = 2^{\sigma k}$, we get the

$$\frac{n(n+1)}{2} \geq \frac{1}{\epsilon}(\sigma k \ln 2 + \ln(1/\delta))$$

which is identical to Equation (2) ∎

This is correct because a program P can be regarded as portable for n languages as long as at least one function in F survives the portability test bounded by ϵ and δ. Consequently, it is easy to see that a program P that has many functions needs more languages to ensure portability.

4. Discussion

Although there has been a significant amount of computational linguistic research for major languages such as English for more than fifty years, the portability issue of natural language technology based on such research has not been studied until recently.

However, portability of technology is neither cheap to obtain nor trivial to implement according to our theory. From the functional perspective of language technology, the efforts of linguists can be described as finding good theories or rules that can generate both universal and local features for various languages. Likewise, one of the main reasons in the recent success of statistical natural language processing techniques(Manning and Schutze, 1999) can be found in its portability. Statistical approaches, unlike traditional symbolic approaches, are less dependent on language specific features. Our definition of portability demonstrates that as the number of soft features increases, the same technology is portable for more languages. If σ is fully dependent on n and decreases linearly, the technology is not portable. If one can find a good features that are not affected by n and a robust technology that depends only those features, then such technology can enjoy its maximum portability.

To claim portability of a technology, empirical justifications of its performance guarantee are also required over many languages and this is reflected in the parameter ϵ. What it suggests is that even if the same statistical method that uses the common features in many languages, the distribution of features could be dependent on each language and thus significantly different from others. Our theory clearly demonstrates that reducing portability error ϵ requires $O(\sqrt{n})$ languages to be verified with.

5. Conclusion

We presented a formal PAC framework for the functional view of portability. Although it is still a sketchy work, the main contribution of this work is to define portability in a formal way and show the relation among features and performance measures. Therefore, the development of good theories and rules that can work for as many languages as possible and the empirical application of them is critical in discussing the portability issues.

6. References

D. Haussler. 1988. Quantifying inductive bias: Ai learning algorithms and valiant's learning framework. *Artificial Intelligence*, (36):177–221.

C. Manning and H. Schutze. 1999. *Foundations of Statistical Natural Langauge Processing*. MIT Press.

T. Mitchell. 1997. *Machine Learning*. McGraw Hill.

H. T. Ng and H. B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of 34th Conference of the ACL*.

L. Valiant. 1984. A theory of the learnable. In *ACM Symposium on Theory of Computing*, pages 436–445.

A requirement analysis for an open set of human language technology tasks

Fredrik Olsson

Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
fredrik.olsson@sics.se

Abstract

This work presents a requirement analysis and a design proposal for a general architecture for a specified, yet open set of human language technology (HLT) tasks — the set chosen is dubbed *information refinement.* Apart from using information refinement as a means to focus the requirement analysis and accompanying design proposal, the analysis and proposal are based on a survey of a number of projects that have had great impact on the realisation of today's HLT architectures, as well as on the experiences gained from a long-term case study aiming at composing a general purpose tool-kit for Swedish. The analysis and design are currently used in an ongoing effort at SICS to implement an open and general architecture for information refinement.

1. Introduction

During the last few years, the need for general, reusable software for computational linguistics and human language technology (HLT) has become widely acknowledged by the research community as well as by the industry. Usually, the overall motivation of striving for reusable software is to shorten the way from the origin of an idea to its implementation in a prototype system. Utilising reusable software also means that the effort spent in building an HLT system is reduced, and thus, that personal labour can be focused on more important issues.

The aim of this paper is to present a requirement analysis and design proposal for a specified, yet open set of human language technology tasks — *information refinement* is introduced as constituting a set of related tasks intended to serve as a target for developing a general and open architecture, Kaba. The requirement analysis and design proposal presented in sections 7. and 8. are based on three parts: the notion of information refinement (Section 2.); a survey of a number of projects and software that have had great impact on how HLT software is constructed today (TIPSTER, CLE, ALEP, GATE, DARPA Communicator, and ATLAS presented in Section 3.); and on the experience gained from a case study on constructing a language processing tool-set for Swedish in a national project called SVENSK (Section 4.). See (Olsson, 2002) for an elaboration on the requirements specification and design of an open architecture for information refinement.

2. The notion of information refinement

By the term information refinement, the *process* is referred to in which text is handled with the aim of *accessing* the pieces of content that are relevant from a certain *perspective* (Olsson et al., 2001).

Information access is about providing people with different tools and methods for granting reliable and simple access to the information they need, ideally with awareness of task and context of the access situation. A system for information access is intended to see to an expressed information need. Such a need is not always static — the *process* of searching for information is a dynamic one in which the information need, sources of information, characteristics of the task, and the type of text involved may change during a search session.

Since different readers have different information needs, prerequisites, and attitudes, they also have different *perspectives* when reading one and the same text. When considering that there are different perspectives, it is natural to think of information access and refinement systems as something that should not (only) deliver texts in their entirety, but rather in some sense understand the contents of the text and tailor the information according to the reader's perspective.

Information exctraction, information retrieval and automatic summarisation are all examples of human language techniques that fall under the information refinement category. Current work concerning information refinement at SICS include protein name tagging (Eriksson et al., 2002), information access using mobile services (Hulth et al., 2001), and support of professionals in information seeking (Hansen and Järvelin, 2000).

3. Some important HLT projects

This section introduces some of the software and projects that have, or have had, impacts on the ways today's software for HLT is designed and implemented. The survey of the literature in the area is not exhaustive, but merely provides an overview of the aspects and features of some important projects.

3.1. The TIPSTER architecture

The TIPSTER project (Grishman et al., 1997) was a joint effort between a number of U.S. government agencies led by DARPA and funded by CIA, DARPA, and DoD in collaboration with NIST and SPAWAR. The project started in 1991 and ended due to lack of funding in 1998.

The main focus of TIPSTER was to improve document processing efficiency and cost effectiveness, and in doing that, technologies such as information retrieval, information extraction, and automatic text summarisation were of great interest. There were two primary goals of the TIPSTER project, the first of which was to provide developers and users with an architecture that allowed for information retrieval in several gigabytes of texts, and the second goal was to provide an environment for research in document detection and data extraction. However, by the time the project was discontinued, no fully implemented version of the TIPSTER architecture was produced.

3.2. CLE

SRI International's Cambridge Research Centre and Cambridge University's Computer Laboratory in 1985 suggested a UK-internal project developing a Core Language Engine (CLE), a domain independent system for translating English sentences into formal representations (Moore and Jones, 1985; Alshawi et al., 1992).

SRI's CLE built on a modular-staged design in which explicit intermediate levels of linguistic representation were used as an interface between successive phases of analysis. The CLE has been applied to a range of tasks, including machine translation and interfacing to a reasoning engine. Smith (1992) gives two examples of such systems; the LF-Prolog Query Evaluator and the Order Processing Exemplar (OPEX). The modular design also proved well suited for porting to other languages and the implementation was quite efficient. Thus, the project proved its purpose. However, even though the CLE system received considerable attention, it failed to spread in the community, the main reason being that it simply was too expensive to obtain it.

3.3. ALEP

The origin of the Advanced Language Engineering Platform (ALEP), the work on which started in 1991 and ended in 1995, was the issue of the lack of a general platform for research and development of large scale natural language processing systems (Simpkins, 1995; Bredenkamp et al., 1997). ALEP was an initiative of the Commission of the European Community (CEC) based on the experiences from the Eurotra and CLE projects.

ALEP was intended to function as a catalyst for speeding up the process of going from a research prototype of a system to a ready-to-ship product. The kind of users that ALEP first and foremost was targeted at were advanced experts, i.e., researchers in computational linguistics, possibly in conjunction with application developers. Simpkins (1995) expected that the openness of ALEP would attract users for research and development. Later, it turned out that this was not the case and ALEP never became widely spread.

3.4. GATE

Since the mid 90's, the General Architecture for Text Engineering (GATE) platform as reported on by, e.g., Cunningham (2000) is being developed at the University of Sheffield and funded by the U.K. Engineering and Physical Sciences Research Council (EPSRC). GATE provides a communication and control infrastructure for linking together language engineering software. It does not adhere to a particular linguistic theory, but is rather an architecture and a development environment designed to fit the needs of researchers and application developers. GATE, currently available as version 2.0, is free for non-commercial and research purposes.

GATE supports reuse of resources, data as well as algorithms, since it provides for well-defined application programmers interfaces (APIs). Once a module has been integrated in the system, it is very easy to combine it with already existing modules to form new systems. Each component integrated into GATE has a standard I/O interface, which conforms to a subset of the TIPSTER annotation model. The infrastructure of GATE provides several levels of integration, reflecting how closely a new module should be connected to the core system.

3.5. The DARPA Communicator

Currently, the MITRE Corporation is (under DARPA funding) developing the DARPA Communicator. The goal of the DARPA Communicator is to set the scene for the next generation of conversational, multi-modal, interfaces to distributed information to be used in, e.g., travel planning, that require information from different sources to be combined.

The reference DARPA Communicator architecture builds on MIT's Galaxy-II system (Polifroni and Seneff, 2000; Seneff et al., 1999; Seneff et al., 1998). Among its key features, the authors list the ability to control system integration using a scripting language: each script includes information about the active servers, a set of operations supported by the server, as well as a set of programs. An in-depth explanation of the program control is given by Seneff et al. (1999). Essentially, the Galaxy-II system builds on a central process, the Hub, which mediates information between a number of different servers. The Galaxy-II system supports a wide range of component types, e.g., language understanding and generation, speech recognition and synthesis, dialogue management, and context tracking (Goldschen and Loehr, 1999).

There is a freely available, public version of the core DARPA Communicator.

3.6. ATLAS

The Architecture and Tools for Linguistic Analysis Systems (ATLAS) project is conducted by NIST, MITRE and LDC (Bird et al., 2000). The main goal is to develop a general architecture for annotation of linguistic data, including a formal/logical data format, a set of APIs, a tool-set, and persistent storage.

Within the ATLAS project, the participants are mainly interested in creating a formal framework for constructing, maintaining, and searching in linguistic annotations. In some aspects, the ATLAS annotation set model seems very similar to the TIPSTER annotation scheme. Bird and Liberman (2000) say that there are several ways of translating a TIPSTER-style annotation to a corresponding ATLAS one. In the end, the ATLAS working group concludes that TIPSTER-like annotations are not appropriate for audio transcriptions, except for "cases where such transcriptions are immutable in principle", (Bird and Liberman, 2000).

4. A case study — SVENSK

The SVENSK project was a national effort funded by the former Swedish National Board for Industrial and Technical Development (Nutek) and SICS addressing the problem of reusing language engineering software, see e.g., (Eriksson and Gambäck, 1997; Gambäck and Olsson, 2000). The SVENSK project was divided into three phases, spanning the spring of 1996 to the end of 1999. The aim has been to develop a multi-purpose language processing system for Swedish based, where possible, on existing components. Rather than building a monolithic system attempting to meet the needs of both academia and industry, the project has created a general tool-box of reusable language processing components and resources, primarily targeted at teaching and research.

The re-usability of the language processing components in SVENSK system arises from having each component integrated into GATE.

Collecting and distributing algorithmic resources and making different programs inter-operate present a wide range of challenges, along several different dimensions outlined next.

4.1. Diplomatic challenges

Making language processing resources freely available and, in particular, re-usability of resources is really a very uncommon concept in the computational linguistic community. Possibly this also reflects another uncommon concept, that of experiment reproducibility. In most research areas the possibility for other researchers to reproduce an experiment is taken for granted. It is even considered as the very core of what is accepted as good research at all. Strangely enough, this is seldom the case in computer science in general and even more rare within computational linguistics, perhaps because of tradition or lack of interest.

4.2. Technical challenges

From the technical point of view, one major conclusion is that the difficulties of integrating language processing software never can be over-estimated. Even when using a liberal architecture such as GATE it is hard work making different pieces of software from different sources and built according to different programming traditions meet any kind of interface standard.

In a way, it is understandable that academia does not always put much effort in packaging and documenting their software, since their main purpose is not to sell and widely distribute it. More surprising and discouraging, however, is that some of the actors on the commercial scene do not document their systems in a proper manner, either. Far too often this has resulted in inconsistencies with the input and output of other modules.

4.3. Linguistic challenges

Of course, language engineering components differ with respect to such things as language coverage, processing accuracy and the types of tasks addressed. It is also the case that tasks can be carried out at various levels of proficiency. The trouble is that there is no quality control available neither to the tool-box developer nor to the end-user. If a large number of language processing components are to be integrated, they should first be categorised so that components with a great difference in, say, lexical coverage are not combined.

A familiar problem for all builders of language processing systems relates to the adaptation to new domains. When reusing resources built by others this becomes even more accentuated, especially if a language engineering resource is available only in the black-box form (and thus relates to the issues of the previous subsection).

5. General observations and experiences

Below are some broad conclusions — focal points — drawn from the previous and present chapters, of what should be considered when creating a general HLT architecture:

1. **An architecture should be general with respect to a class of tasks, not to an entire field of research** The issue of *how* general an architecture should be needs to be considered since a too general one tends to be hard to handle.

2. **Keep the software open** There are various dimensions along which software could be considered open: distributing and licensing it; keeping its source open and inviting other people to participate in developing it; and to achieve software that are easily adaptable to new domains and types of information.

3. **Allow for use of existing programs as well as for the creation of system-specific ones** The potential drawback in using existing, externally produced software concerns issues such as, e.g., maintenance, fixing bugs, and extending/updating resources such as lexica and ontologies. All these things rely on the external program being supported by its producer.

4. **Support maintenance of systems and the components making them up** Develop tools and methods to support maintenance of components and systems, both on the linguistic level, e.g., integrated machine learning methods for lexical acquisition and grammar induction, and on the software level, e.g., new file formats and operating systems.

6. Motivation for a new architecture

The motivation for building a new architecture is primarily due to the fact that when information refinement emerged as a research area at SICS, there was no single architecture which fulfilled the demands that SICS's projects made at the time. In particular, no one of the existing platforms granted us full access to the source code and full distributional rights of the code, something which would be of great interest to us since we wanted to be able to distribute the source code of future information refinement systems freely, and since the functionality of the tools used for information refinement will have to be tuned to each new information refinement task. The latter may include changes to, e.g., the way the tools interact with each other and with the user, as well as the kind of data they produce — such changes may be difficult to achieve unless the software architecture hosting the tools is accessible at the source code level.

The work on a new HLT architecture called Kaba is an ongoing effort which was initiated in 1999 by Kristofer Franzén and Jussi Karlgren at SICS. At first, Kaba was intended to constitute an information extraction system for Swedish. An attempt at porting an existing information extraction system from English to Swedish turned out to be cumbersome (Franzén, 1999). Along the above lines, the conclusion was reached that future research in information refinement at SICS would benefit from a research vehicle having been built on site. Since 1999, the research focus has shifted slightly from information extraction to the more general goal of information refinement, which makes the need for an open and general architecture even clearer.

7. Requirement analysis

Deciding on what requirements are relevant for a given project tends to be a top-down process, going from broad issues such as, e.g., that the software under development should be portable to new operating systems, to splitting the portability into more specific sub-requirements. Requirements analysis always asks the *what*-questions regarding the software, e.g.: *what* equipment constraints exist, and *what* functions are to be incorporated. The *how*-questions are issued in the design phase described in Section 8..

Kaba is intended to function as a tool for developers of information refinement systems, first and foremost for research systems, but also for prototypes for testing ideas within information refinement. Kaba will *not* be a fixed set of tools for creating ready-to-ship products.

A typical Kaba user is a computational linguist with programming skills. This person's role is to use Kaba for the creation of information refinement systems to be used further in research and prototyping.

7.1. Project constraints and external factors

To accomplish the portability of Kaba on the software level, a widely supported programming language, such as Java, has to be used throughout the development process to implement all parts of the architecture. Further, Kaba will require (and presuppose) a linguistic processor that performs basic linguistic analysis of the texts to be processed, e.g., part-of-speech tagging and some fundamental grammatical analysis. Most likely the processor will be the Swedish and the English Functional Dependency Grammars (FDG) from Conexor Oy, Helsinki, Finland (Tapanainen and Järvinen, 1997).

Kaba must be implemented using a technology and an environment that facilitates easy integration of in-house or third party software for linguistic analysis as well as basic computational facilities, e.g., for reading and writing various file formats.

7.2. The scope of the work

Figure 1 shows three different ways that an information refinement system based on Kaba can interface with its environment, and thus gives some notion of what a developer of such a system has to deal with. What differs between the three constellations is the kind of user the system is intended for. In Figure 1 A, the system interacts with an information provider of some sort, e.g., a web site, a database, or a mobile service, on the one hand, and a human user on the other.

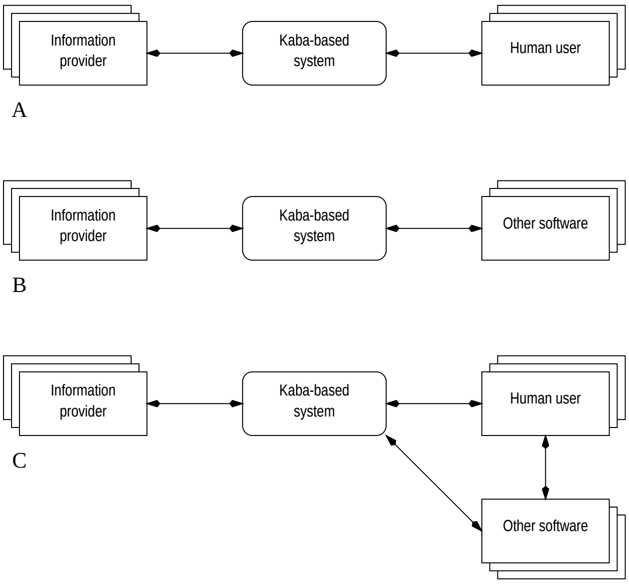


Figure 1: Characteristics of the environment of a Kaba-based system.

In Figure 1 B, the Kaba-based system communicates with the same kind of information provider as in Figure 1 A, but with another machine as counterpart instead of a human. The setup illustrates the case when a Kaba-based system is part of a larger system.

Finally, Figure 1 C, shows a configuration in which the system interacts with a human user as well as another machine.

7.3. The scope of the architecture

When starting to look at what a user may want to do with Kaba, it seems as a good idea to structure the requirements into what is commonly known as use cases (UC). Cockburn (1997) gives an overview of a method that deals with the identification and structuring of UCs. He defines a use case as being what happens when *actors* interact with a system to achieve a desired goal. An actor is an external entity (human or other software) that uses the system. In effect, UCs hold the functional requirements of a system in an easy-to-read format, and they represent the goal of an interaction between an actor and the system.

In total, 30 use cases have been identified for Kaba and seven of these constitute the top level of the use case hierarchy (Olsson, 2002):

UC 1: Develop an information refinement research and development prototype system.

UC 2: Evaluate an information refinement research and development prototype system.

UC 3: Port an existing system to a new domain or language.

UC 4: Document system.

UC 5: Maintain system.

UC 6: Create learning material or tutorial.

UC 7: Manage LR and PR components.

Use cases 1 and 7 each have several sub-goals which are illustrated in Figure 2 and Figure 3, respectively.

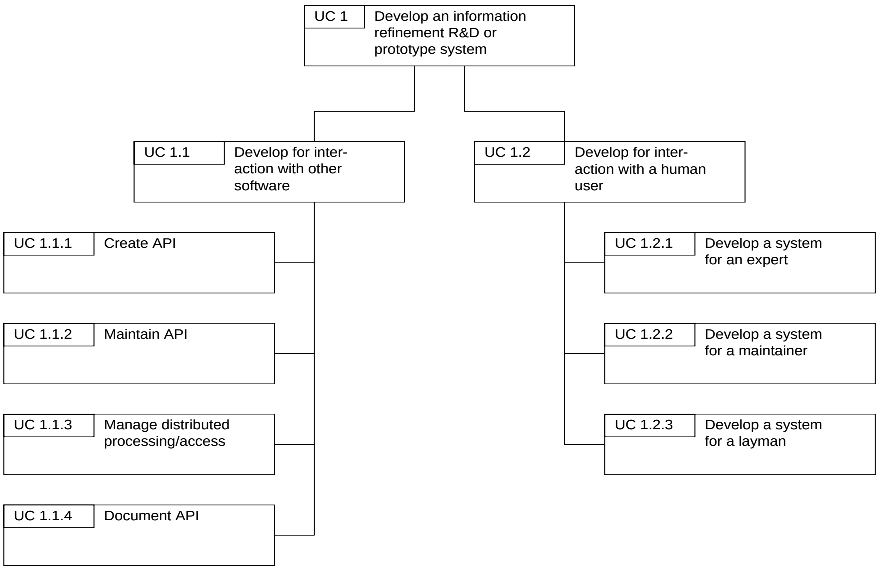

UC 1
Develop an information refinement R&D or prototype system
UC 1.1
Develop for interaction with other software
UC 1.2
Develop for interaction with a human user
UC 1.1.1
Create API
UC 1.1.2
Maintain API
UC 1.1.3
Manage distributed processing/access
UC 1.1.4
Document API
UC 1.2.1
Develop a system for an expert
UC 1.2.2
Develop a system for a maintainer
UC 1.2.3
Develop a system for a layman

Figure 2: Schematic view of use case 1 and its sub-goals.

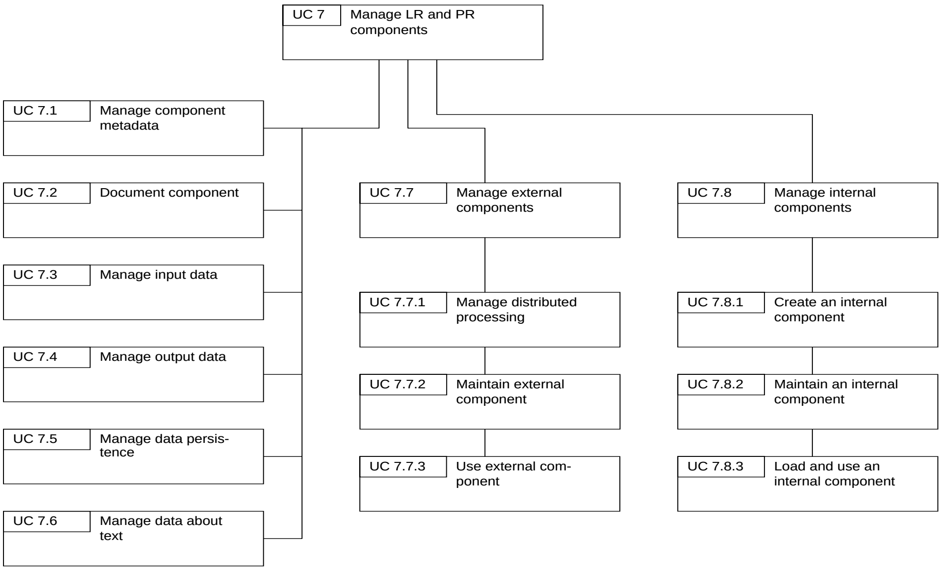

UC 7
Manage LR and PR components
UC 7.1
Manage component metadata
UC 7.2
Document component
UC 7.3
Manage input data
UC 7.4
Manage output data
UC 7.5
Manage data persistence
UC 7.6
Manage data about text
UC 7.7
Manage external components
UC 7.7.1
Manage distributed processing
UC 7.7.2
Maintain external component
UC 7.7.3
Use external component
UC 7.8
Manage internal components
UC 7.8.1
Create an internal component
UC 7.8.2
Maintain an internal component
UC 7.8.3
Load and use an internal component

Figure 3: Schematic view of use case 7 and its sub-goals.

8. Design proposal

The design proposal is intended to give a hint as to how the requirement analysis could be realised.

8.1. Component metadata

This section covers use case 7.1 (*Manage component metadata*). Metadata about both language resources (LR) and processing resources (PR) is needed for several reasons, the first of which is to allow the developer (and the future users of the system) to browse a collection of components to see what components there are in order to build an information refinement system utilising existing components. In the same manner, metadata can be used to identify shortcomings of existing components and act as a basis for requirements analysis and specification when new language processing components need to be constructed or when new language resources need to be developed.

There are several means by which metadata can be expressed, and it seems natural to convey such data in the same format as the components themselves are annotated or produce annotations about text. Thus, the system internal format of metadata should correspond to the internal format of the data about text as described in Section 8.3., while the external format of metadata should agree with the format for data persistence described in Section 8.4..

8.2. Input and output

This section deals with use cases 7.3 (*Manage input data*) and 7.4 (*Manage output data*). The Kaba information refinement development platform presupposes that some sort of linguistic analysis has been performed on the text to be processed by a Kaba-based system. Currently, the FDG for English and Swedish are intended to be used, but it should also be possible to use any TIPSTER compliant linguistic processing component.

On the output side, a Kaba-based system should be able to generate representations of the text it has processed in a format suitable to the user, regardless of whether the user is another computer program or a human.

8.3. System internal representation of annotated text

This section covers use case 7.6 (*Manage data about text*). Data about text can be expressed in various ways and the crucial point in all data representation is that it should facilitate rapid access to arbitrary pieces of information about the text. The representation formalism should allow for scaling up without causing the system's performance to drop.

While the format of the external and persistent data is like XML (see Section 8.4.), the internal representation is based on the TIPSTER annotation scheme. Although the two schemes are conceptually different the conversion between TIPSTER-style annotations and XML-based representations is quite straightforward.

8.4. Data persistence

This section deals with use case 7.5 (*Manage data persistence*). Data persistence is needed in order to provide Kaba with multiple-session capabilities, that is, to allow a user to work with the same source of information during several sessions and, in each session, having access to the results from the previous ones. The need for working in multiple sessions may occur, e.g., due to a system crash, for saving intermediate results, or simply because the user needs to interrupt the refinement process for other reasons.

The most suitable format is likely to be some instance of XML, partially because of the fact that it is becoming increasingly widespread in language engineering applications, and partially because there exist tools for manipulating and converting between different instantiations of XML.

8.5. Interacting with others

There are several aspects of interaction which have to be taken into account when designing an information refinement architecture like Kaba: (1) when a Kaba-based system is used by other software as a part of a larger system, (2) when a Kaba-based system utilises external components, both processing and data, as a part of an information refinement system, and (3) when a Kaba-based system needs to interact with human users.

Case (1) is reflected in use case 1.1.1 (*Create API*). In effect, what is required for a Kaba-based system to function in the context of a larger system, is a means for the developer of the larger system to have access to a restricted and well-defined set of the functionality in the Kaba-based system. Such access can be provided by means of a Java API.

Case (2) is addressed in use case 7.7.3 (*Use external component*) which concerns how to allow a Kaba-based system to use external components, i.e., components not primarily implemented for use within Kaba such as, for instance, part-of-speech taggers and ontologies. To allow Kaba to interact with external components, it is important that the components all look the same from Kaba's point of view. This means that the APIs that Kaba has to use to achieve this interaction have to be well defined and consistent.

Case (3) is addressed in use cases 1.2.1 (*Develop a system for an expert*), 1.2.2 (*Develop a system for a maintainer*), 1.2.3 (*Develop a system for layman*), all of which aim at facilitating interaction between different kinds of end-users and a Kaba-based system. Case (3) boils down to creating a connection between a tool or library for constructing GUIs, such as the Java Swing Classes (Topley, 1998), and Kaba.

8.6. Distributed processing

This section addresses use cases 1.1.3 (*Manage distributed processing/access*) and 7.7.1 (*Manage distributed processing*). In various settings, the parts making up a Kaba-based system need to be situated on different machines, connected by a network. One such setting occurs when some component, for example the one providing the initial linguistic analysis of input text, is available only for a particular operating system, while the rest of the system runs on another machine in the network. The different parts of the system then have to communicate using some protocol, e.g., SOAP.

8.7. Documentation and tutorials

This section addresses use cases 1.1.4 (*Document API*), 4 (*Document system*), 6 (*Create learning material or tutorial*), and 7.2 (*Document component*).

Kaba should come with incentives for developers, both of the Kaba architecture itself and of Kaba-based systems, to document their efforts. Such stimulus should be in the form of guide-lines and examples. There is a range of possible formats for documenting software systems, e.g., HTML and plain ASCII. It is also important that the guidelines are tied as little as possible to the chosen format. As for documenting the source code, existing tools such as Javadoc should be used.

Examples and tutorials should be encouraged by providing templates, example examples and tutorials to Kaba users and system developers.

8.8. Creating internal components

This section deals with use case 7.8.1 (*Create an internal component*). In Kaba, an internal component is one that is under the control of the developer in that it provides him with a more elaborate API than external components do. Typically, an internal component is created explicitly for use within a Kaba-based system.

A variant of the Common Pattern Specification Language (CPSL) called Kaba Pattern Specification Language (KPSL) will form the base formalism in which the functionality of the internal components will be expressed. CPSL is an effort by the TIPSTER working group that, unfortunately, has not been officially released. However, Appelt (1999) as well as Cunningham et al. (2000) present implementations of annotation engines based on CPSL. Essentially, a CPSL rule describes a finite state transducer for TIPSTER annotations.

It should be possible to construct internal components in several ways, for instance by hand-crafting rules using a graphical rule editor, or by breeding them using machine learning methods.

8.9. Loading and using internal components

This section deals with use case 7.8.3 (*Load and use an internal component*). Once the KPSL rules making up a component have been developed, they are turned into Java code by a KPSL rule compiler. Along with the compiler come Java classes that facilitate dynamic loading of compiled sets of rules. Thus, as long as the KPSL rules have been compiled to Java and the Kaba-based system knows where to find the components, there are means by which they can be dynamically loaded into the system at run time.

8.10. Maintenance

The fundamental question when it comes to maintenance of any software is *When is maintenance necessary for this piece of software in this particular setting?* and, in the context of information refinement systems, this calls for well-defined criteria that can be used to probe the system's performance with respect to the task it is supposed to accomplish, or the system's affordance with respect to the users' expectations as to what the system is really supposed to do.

8.10.1. Maintenance of external components

This section addresses use cases 1.1.2 (*Maintain API*) and 7.7.2 (*Maintain external component*). The cases are closely related in that communication between a Kaba-based system and other software will always take place via some kind of API. Thus, maintaining an external component is in many cases the same as maintaining the API that Kaba uses for communicating with that component.

8.10.2. Maintenance of internal components

This section deals with use case 7.8.2 (*Maintain an internal component*). Maintenance of internal components should be facilitated by a graphical interface for inspecting, editing, loading, executing, and evaluating KPSL rules with respect to some success criteria set up for the component. It should be possible to do all this using the same, or a similar, graphical interface as when creating internal components.

8.10.3. Maintenance of systems

This section deals with use case 5 (*Maintain system*) which involves all other kind of maintenance mentioned previously in this section, i.e., maintenance of component APIs and external components (Section 8.10.1.), as well as of internal components (Section 8.10.2.). In addition, maintenance of systems also involves taking care of the whole formed by the pieces, e.g., seeing to it that the documentation is up to date, installing new software when needed, and monitoring the system's performance on a regular basis. This should be supported in the same way as maintenance of external components is, e.g., by giving guidelines for how to integrate the documentation of the parts into a central repository, and collect information about availability of new components.

8.11. Providing support for porting systems to new domains

This section deals with use case 3 (*Port an existing system to a new domain or language*). While maintenance may accommodate correction of minor changes to a system, there will also be occasions when the shift of domain or information need is so different from that captured by an existing system that maintenance of the system or one of its components is not enough to compensate for it. In these cases, the question of whether to use an existing system or to create a new one from scratch arises. One of the issues of providing support for porting systems to new domains and needs should be to supply the developer with clues for deciding the answer to that question. If the answer is that an existing system could probably be altered (ported) to meet the new needs, then the follow-up question should be: *What parts of the existing system can be re-used, and to what degree do they need to be modified?* Again, Kaba should provide methods that makes answering this question easier.

8.12. Providing support for evaluation

This section addresses use case 2 (*Evaluate an information refinement R&D or prototype system*). Evaluation of information refinement systems is a crucial issue in several aspects. The basic support for evaluation of information refinement systems can be of two kinds: by providing linguistically annotated data that act as a key to the questions for which a system is to be evaluated, or by providing tools that presuppose the presence of an answer-key for comparing data structures and calculating measurements of performance. Both kinds of support are necessary. In the former case, machine learning methods are often used as an aid in obtaining the correctly annotated corpora constituting the answer-key. In the latter case, the comparison of data structures should yield values in an appropriate metric, e.g., precision and recall, depending on the features that are evaluated.

9. Conclusions

When developing a general tool or architecture, it is possible to focus the technical and linguistic efforts in several ways. The most obvious one is to formulate and maintain an explicit goal regarding the kind of tasks that programs developed within the general architecture at hand should cope with. By obtaining and focusing on the goal at an early stage in the development of the open architecture, one can avoid ending up with a definition and design of a far too general system: when it comes to generality for language engineering, it should be with respect to a class of tasks, rather than to the field as such.

Acknowledgements

Lars Borin, Björn Gambäck, Kristofer Franzén, Jussi Karlgren, Preben Hansen, Gunnar Eriksson, Mikael Eriksson, Anette Hulth, Mark Tierney, Anna Jonsson.

10. References

Hiyan Alshawi, David Carter, Jan van Eijck, Björn Gambäck, Robert C. Moore, Douglas B. Moran, Fernando

C. N. Pereira, Stephen G. Pulman, Manny Rayner, and Arnold G. Smith. 1992. *The Core Language Engine*. MIT Press, Cambridge, Massachusetts, March.

Douglas E. Appelt, 1999. *The Complete TextPro Reference Manual*, June.

Steven Bird and Mark Liberman. 2000. A Formal Framework for Linguistic Annotation. *Speech Communication*, 33(1,2):23–60.

Steven Bird, David Day, John Garofolo, John Henderson, Christophe Laprun, and Mark Liberman. 2000. ATLAS: A flexible and Extensible Architecture for Linguistic Annotation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 1699–1706, Athens, Greece, June.

Andrew Bredenkamp, Thierry Declerck, Frederik Fouvry, Bradley Music, and Axel Theofilidis. 1997. Linguistic Engineering using ALEP. In R. Mitkov and N. Nicolov, editors, *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, pages 92–97, Tzigov Chark, Bulgaria, September.

Alistair Cockburn. 1997. Structuring Use Cases with Goals. *Journal of Object-Oriented Programming*, Sep-Oct and Nov-Dec.

Hamish Cunningham, Diana Maynard, and Valentin Tablan. 2000. JAPE: A Java Annotation Patterns Engine. Technical Report CS–00–10, University of Sheffield, Department of Computer Science, Sheffield, UK. Second Edition.

Hamish Cunningham. 2000. *Software Architecture for Language Engineering*. Ph.D. thesis, University of Sheffield, UK.

Mikael Eriksson and Björn Gambäck. 1997. SVENSK: A Toolbox of Swedish Language Processing Resources. In R. Mitkov and N. Nicolov, editors, *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing*, pages 336–341, Tzigov Chark, Bulgaria, September.

Gunnar Eriksson, Kristofer Franzén, Fredrik Olsson, Lars Asker, and Per Lidén. 2002. Exploiting Syntax when Detecting Protein Names in Text. In *Proceedings of Workshop on Natural Language Processing in Biomedical Applications*, Nicosia, Cyprus, March.

Kristofer Franzén. 1999. Adapting an English Information Extraction System to Swedish. In *Proceedings of the 12th Nordic Conference of Computational Linguistics*, pages 57–65, Norwegian University of Science and Technology, Trondheim, Norway, December.

Björn Gambäck and Fredrik Olsson. 2000. Experiences of Language Engineering Algorithm Reuse. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, volume 1, pages 161–166, Athens, Greece, May. ELRA.

Alan Goldschen and Dan Loehr. 1999. The role of the DARPA Communicator Architecture as a Human Computer Interface for Distributed Simulations. In *Spring Simulation Interoperability Workshop*, Orlando, Florida, USA, March. Simulation Interoperability Standards Organization (SISO).

Ralph Grishman, Ted Dunning, Jamie Callan, Bill Caid, Jim Cowie, Louise Guthrie, Jerry Hobbs, Paul Jacobs, Matt Mettler, Bill Ogden, Bev Schwartz, Ira Sider, and Ralph Weischedel, 1997. *TIPSTER Text Phase II Architecture Design. Version 2.3*. New York, New York, January.

Preben Hansen and Kalervo Järvelin. 2000. The Information Seeking and Retrieval Process at the Swedish Patent and Registration Office. Moving from Lab-based to Real Life Work-task Environment. In *Proceedings of the ACM-SIGIR 2000 Workshop on Patent Retrieval*, pages pp. 43–53, Athens, Greece, July 28.

Anette Hulth, Fredrik Olsson, and Mark Tierney. 2001. Exploring Key Phrases for Browsing an Online News Feed in a Mobile Context. In *Proceedings of Management of uncertainty and imprecision in multimedia information systems*, Toulouse, France, September. A workshop held in conjunction with the Sixth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU-2001).

Robert C. Moore and Karen Sparck Jones. 1985. A research programme in natural language processing. CRC technical report, SRI International, Cambridge, England.

Fredrik Olsson, Preben Hansen, Kristofer Franzén, and Jussi Karlgren. 2001. Information Access and Refinement — A Research Theme. *ERCIM News*, 46, July.

Fredrik Olsson. 2002. *Requirements and Design Considerations for an Open and General Architecture for Information Refinement*. Licentiate of philosophy thesis, Department of Linguistics, Uppsala University, Uppsala, March. Available at `http://www.sics.se/~fredriko/lic`.

Joseph Polifroni and Stephanie Seneff. 2000. Galaxy-II as an Architecture for Spoken Dialogue Evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, May. ELRA.

Stephanie Seneff, Ed Hurley, Raymond Lau, Christine Pao, Philipp Schmid, and Victor Zue. 1998. Galaxy–II: A Reference Architecture for Conversational System Development. In *Proceedings of the 5th International Conference on Spoken Language Processing*, volume 3, pages 931–934, Sydney, Australia, December.

Stephanie Seneff, Raymond Lau, and Joseph Polifroni. 1999. Organization, Communication, and Control in the Galaxy-II Conversational System. In *Proceedings of Eurospeech 99*, Budapest, Hungary, September.

Neil K. Simpkins. 1995. ALEP — An Open Architecture for Language Engineering. Technical report, Cray Systems, 151 rue des Muguets, L-2167 Luxembourg.

Arnold Smith. 1992. The CLE in Application Development. In Hiyan Alshawi, editor, *The Core Language Engine*, chapter 12, pages 235–250. MIT Press, Cambridge, Massachusetts, USA, March.

Pasi Tapanainen and Timo Järvinen. 1997. A Non-Projective Dependency Parser. In *Proceedings of the 5th Conference of Applied Natural Language Processing*, Washington, D.C. USA, April. ACL.

Kim Topley. 1998. *Core — Java Foundation Classes*. Prentice Hall PTR Core Series. Prentice Hall.

Taking Advantage of Spanish Speech Resources to Improve Catalan Acoustic HMMs

Jaume Padrell and José B. Mariño

TALP Research Center
Universitat Politècnica de Catalunya, Barcelona 08034, Spain
{jaume,canton}@talp.upc.es

Abstract

At TALP, we are working on speech recognition of official languages in Catalonia, i.e. Spanish and Catalan. These two languages share approximately 80 % of their allophones. The speech databases that we have available to train HMMs in Catalan have a smaller size than the Spanish databases. This difference of size of training databases results in poorer phonetic unit models for Catalan than for Spanish. The Catalan database size is not enough to allow correct training of more complex models like triphones. The aim of this work is to find segments in Spanish databases that, used in conjunction to the Catalan utterances to train the HMM models, get an improvement of the speech recognition rate for the Catalan language. To make this selection, the following information is used: the distance between the HMM which are trained separately in Spanish and Catalan, and the phonetic attributes of every allophone. A contextual acoustic unit, the demiphone, and a state tying approach are used. This tying is done by tree clustering, using the phonetic attributes of the units and the distances between the HMM states. Different tests have been carried out by using different percentage of tied states in training simultaneously in Catalan and Spanish. In this way, Catalan models are obtained that give generally better results than the models trained only with the Catalan utterances. However, we observe from one of the tests that, when the number of gaussians is increased, that improvement becomes a loss of performance. Currently, we are working on the inclusion of additional labels to avoid that tree clustering puts in the same pool phoneme realizations that are too much different.

1. Introduction

It is not strange to develop in a same laboratory speech recognition systems for different languages. In the TALP we work in official languages of Catalonia, the Spanish and the Catalan. These two languages share approximately 80% of their allophones (both come from the Latin and share geographic space).

The available speech databases to train the HMMs in Catalan have an smaller size that the available databases in Spanish.

This difference in the dimensions of the training databases is one of the causes by which poorer estimations of the Catalan phonetic units are obtained than in Spanish. Thus, whereas in Spanish the units with what we obtain higher recognition rates are triphonemes or demiphones, in Catalan the best results are obtained modeling allophones (Pachès, 1999), 3 states CDHMM with 32 Gaussian for state, since the database size is not sufficient to allow correct train of more complex models like triphones.

In other works (Mariño et al., 2000b), for bilingual recognition systems with these two languages that should work with limited resources (memory, time, etc.), a set of bilingual HMMs has been created (modeling demiphones) that share some models for both languages. These models, trained with utterances from both languages, obtain a lower recognition rate than their respective monolingual models, but the degradation is not significant.

These last recognition results suggests the possibility of a carefully selection of some utterances from the Spanish databases might the Catalan acoustic HMMs.

(Bonaventura et al., 1997; Wheatley et al., 1994) already suggested the idea to train phoneme models for a language using other languages and implemented a number of different metrics for measuring similarities among cross-language phonetic models. (Bub et al., 1997) considered this task as a question of model adaptation and (Imperl and Horvat, 1999) already used context-depending phonetic units (triphones) in multilingual models.

All these works discus the difficult to select the utterances segments to train the shared models. As a framework to do this selection, in this paper we present some preliminary results using demiphones (Mariño et al., 2000a) as context-depending phonetic units and clustering algorithms that are usually employed to train units that appear little in the training corpus. So, the aim is to use these clustering algorithms to relate contextual units from different languages.

The paper is organized as follows. Section 2. describes the work methodology, section 3. gives some preliminary results and section 4. presents conclusions and future work.

2. Procedure

The procedure that we followed has been: 1) to choose an acoustic unit inventory for both languages, 2) to choose algorithms to select and to tie acoustic units and, 3) to train and to evaluate units for the Catalan with a different percentages of units also trained with the Spanish utterances.

2.1. Spanish and Catalan Allophones Inventory

The allophone transcription is made by different softwares (Saga for the Spanish and Segre for the Catalan) developed in TALP research center. These programs use rules to turn the orthographic text to strings of allophone coded in SAMPA notation.

The transcriptor Saga uses the rules described in (Llisterri and Mariño, 1993) to obtain the phonetic transcription. The program Segre uses extern rules developed in the UAB.

The program Segre transcribes the Catalan sentences using 34 different allophones. In table 1 are shown these allophones with the attributes[1] that we associated to them. They are used to indicate common characteristics between the units for tree-clustering. These attributes can have phonetic meaning (voiced, manner and point of articulation) or not (for example speaker gender), however, in the present case all attributes have a phonetic meaning.

Al	Attributes	L
a	vowels, open, central, voiced	C,ES
e	vowels, mid_close, front, voiced	C,ES
E	vowels, mid_open, front, voiced	C
i	vowels, close, front, voiced	C,ES
o	vowels, mid_close, back, voiced, rounded	C,ES
O	vowels, mid_open, back, voiced, rounded	C
u	vowels, close, back, voiced, rounded	C,ES
@	vowels, schwa, central, voiced, unrounded	C
j	glides, palatal, semivowel, voiced, close, front	C,ES
w	glides, labial_velar, approximant, voiced, close, back	C,ES
uw	glides, voiced, close, back, rounded	C
p	consonants, bilabial, plosive, voiceless, stop	C,ES
t	consonants, dental, plosive, voiceless, stop	C,ES
k	consonants, velar, plosive, voiceless, stop	C,ES
b	consonants, bilabial, plosive, voiced	C,ES
d	consonants, dental, plosive, voiced	C,ES
g	consonants, velar, plosive, voiced	C,ES
B	consonants, bilabial, approximant, voiced	C,ES
D	consonants, dental, approximant, voiced	C,ES
G	consonants, velar, approximant, voiced	C,ES
f	consonants, labiodental, fricative, voiceless	C,ES
s	consonants, alveolar, fricative, voiceless	C,ES
z	consonants, alveolar, fricative, voiced	C,ES
x	consonants, velar, fricative, voiceless	ES
jj	consonants, palatal, approximant, voiced	ES
T	consonants, interdental, fricative, voiceless	ES
tS	consonants, palatal, affricate, voiceless, mid_palatal	ES
S	consonants, palatal, fricative, voiceless, mid_palatal	C
Z	consonants, palatal, fricative, voiced, mid_palatal	C
y	consonants, palatal, approximant, voiced	C
l	consonants, alveolar, lateral, voiced, liquid, back	C,ES
L	consonants, palatal, lateral, voiced	C,ES
m	consonants, bilabial, nasal, voiced	C,ES
n	consonants, alveolar, nasal, voiced	C,ES
N	consonants, velar, nasal, voiced	C,ES
J	consonants, palatal, nasal, voiced	C,ES
r	consonants, alveolar, tap, voiced, rothics, liquid	C,ES
rr	consonants, alveolar, trill, voiced, rothics, vibrate	C,ES
R	Alveolar, Voiced, Rothics, vibrate	ES

Table 1: Allophone list (All.) that the program Segre (C) and Saga (ES) uses (in SAMPA notation) and attributes that are assigned to each unit.

The program Saga provides the 32 allophones for Spanish language. They are also shown in table 1. These inventory were used in the Spanish SpeechDat database (Moreno, 1997) design.

Between the 32 selected allophones to represent Spanish (ES) and the 34 to represent the Catalan (C), there are 27 allophones (C,ES) that share the same SAMPA notation, for example the vowels /a/, /e/, /i/, /o/ and /u/.

2.2. Shared Training and HMM Distance Measure

In an initial step, we trained units separately for each language. In a second step, we re-estimated the Catalan units using also utterances in Spanish. To do this, we tied the Catalan HMM states and the Spanish ones by tree-clustering (Young et al., 1999) with the separately trained HMMs values and the allophones attributes from table 1.

We sort the HMM states that have been tied in both languages by distances between their values. This will later help us to decide which units are finally shared in the experiments, when we only leave tied a ρ percentage of these HMM states.

The distance measure between two HMMs that we use is described in (Young et al., 1999). This measure is based on the sum of the probabilities that the averages that characterize HMM_1 belong to HMM_2 and vice versa. The probability is evaluated logarithmically so, in the case that HMM_1 and HMM_2 are the same model, we obtain a distance zero among them.

So, we re-estimate the HMMs of both languages jointly. The Catalan HMMs with the Catalan utterances and the Spanish HMMs with Spanish utterances. However, there is a certain ρ percentage of HMMs states shared (or tied) between both languages and therefore, that are trained simultaneously in Catalan and Spanish.

3. Experiments and Results

HTK (Young et al., 1999) software is used to train HMMs and to carry out the speech recognition experiments. A classic parametrization of four characteristics has been used, three of dimension 12 and one of dimension 2, trained respectively with a Mel-Cepstrum with mean subtraction, its first differential, its second differential and, in joint form, second and third energy differential. CDHMM are used to model the acoustic units.

3.1. Spanish and Catalan Speech Databases

3.1.1. Training Data

In order to estimate the HMMs two sets of utterances are used (in both cases the automatic transcription is done without considering coarticulation between words):

- Catalan corpus is formed by 3,981 sentences from the SpeechDat Catalan database (Hernando and Nadeu, 1999), with 639 different speakers (Catalan Eastern dialect) and 171,443 allophones according to the Segre transcription (6.5 hours of speech for training).

- Spanish corpus is formed by 4,951 sentences from 976 different speakers (Spanish speakers from Catalonia) from the SpeechDat Spanish database (Moreno, 1997). This sentences set is formed by 242,813 allophones according to the program Saga (also more than 6 hours of speech for training). This corpus represents a fourth database size than it is available for Spanish.

[1] They were designed, in addition to the TALP members, by the Laboratory of Phonetics from the UAB and by Sílvia Llach from the UG.

<table>
<tr><td>3</td><td colspan="9">@ ESa a ESo o O ESe e E</td><td colspan="2">ESi i</td><td colspan="2">ESu u</td></tr>
<tr><td>4</td><td colspan="6">@ ESa a ESo o O</td><td colspan="3">ESe e E</td><td colspan="2">ESi i</td><td colspan="2">ESu u</td></tr>
<tr><td>5</td><td colspan="3">@ ESa a</td><td colspan="3">ESe e E</td><td colspan="2">ESi i</td><td colspan="3">ESo o O</td><td colspan="2">ESu u</td></tr>
<tr><td>8</td><td>@</td><td colspan="2">ESa a</td><td colspan="2">ESe e</td><td>E</td><td colspan="2">ESi i</td><td colspan="2">ESo o</td><td>O</td><td colspan="2">ESu u</td></tr>
<tr><td>13</td><td>@</td><td>ESa</td><td>a</td><td>ESe</td><td>e</td><td>E</td><td>ESi</td><td>i</td><td>ESo</td><td>o</td><td>O</td><td>ESu</td><td>u</td></tr>
</table>

Table 2: Vowels clustering by smaller between models distances according to the final maximum number of groups (first column). The Spanish allophones are distinguished from the Catalans by the prefix ES added to its SAMPA representation.

In this paper, we focus our experiments on the ρ percentage of HMMs states tied between both languages. Future research will be addressed to the size database ratios.

Between Catalan and Spanish models we notice 27 allophones that have the same (SAMPA) representation. These are the 94.43% total number of allophones in our Spanish training corpus and the 76.95% in the Catalan one (the difference is mainly due to the allophone schwa /@/ that does not exist in Spanish and has 16.90% frequency of apparition in our Catalan database). On the other hand, the most frequent allophone in our Spanish database (/a/ with 13.04% frequency) is also in Catalan database.

3.1.2. Evaluation Data

The evaluation tests have been carried out with a database with locality names (2,633 sentences with 232 different names with length from one to five words by sentence) and with a people names database (2,956 sentences with 510 different names). All these sentences come from the Vocatel database (Nadeu et al., 1997).

3.2. Allophone Clustering

Allophone models have been trained separately for each language (CDHMM of 3 states with 4 Gaussian for state). These models objective is twofold: first, to do a preliminary analysis of distances that there are between models from both languages and, second, like a departure point for demiphone units training.

To study the correlation between SAMPA representation and HMM distance we clustered the 13 models that represent the vowels set for both languages into 8 models, matching the models that are at smaller distance. In table 2 is shown the clusters that are obtained according to their final maximum number (3, 4, 5 and 8) that are requested to the clustering algorithm. We obtain that each vowel joins with whom shares symbol SAMPA.

The table 3 shows similar experiment but clustering HMM states independently. The clusters are ordered from less distance between HMM states to more. It can be seen that the similarity between models depends on the HMM state.

We created models simultaneously training the allophones that share its SAMPA representation. The experiments gave recognition rates poorer than with the allophones trained only with Catalan utterances. Probably this is due to many shared allophones occur in different allophone contexts in both languages.

In above mentioned work (Mariño et al., 2000b), where demiphones were used for a bilingual recognition system, the degradation was not significant. In order to approach the different allophone context problem we also use the demiphone as acoustic unit, so that the context tied can be better controlled.

Cl. Order	Left State	Middle State	Right State
1	a ESa	a ESa	a ESa
2	i ESi	e ESe	e ESe
3	e E	e ESe E	o ESo
4	@ a ESa	O o	@ a ESa
5	e E ESe	i ESi	i ESi

Table 3: Clustering order depending on distance between HMM states followed by Catalan and Spanish (with prefix ES) vowels.

3.3. Demiphone Clustering

These demiphones which were trained simultaneously in both languages were chosen by tree clustering, using the allophone attributes (table 1) and the distance between the HMMs (we use the tree clustering described in (Young et al., 1999)).

In Catalan, after tree-clustering, are used 1,092 demiphones modeled by CDHMM of 2 states with 1 Gaussian for state. We also use a model for silence and one for the speaker noise, both of 3 states and 1 Gaussian for state. Following the same procedure, in Spanish 852 CDHMM for demiphones are obtained, plus one for silence and one for the speaker noise.

The analysis of the clusters that are obtained tying by trees is complex. First, we obtained different clusters depending on if we tried simultaneously to cluster all states that form a model or make clusters by state. Second, some of the clusters had that we would name phonetic explanations, but others were inexplicable from this point of view.

It is difficult to evaluate which tying improve the recognition in Catalan and which not. Preliminary experiments with our databases seems indicate that tying between some vowels (for example, /e/ /E/ and /ESe/, or /o/ /O/ and /ESo/) worsen the speech recognition in Catalan language.

Several tests have been done operating only on the tying ρ percentage allowed between demiphones pre-tied by tree-clustering with both languages.

In order to have baseline models for the evaluation demiphones CDHMM with only the Catalan utterances have been trained (it is the case of $\rho = 0.00\%$). In the table 4 are shown the different recognition rates that were obtained. In the first column it is indicated ρ, the states percentage for a total HMM sates set of $(2 * 1,092)$ states that were tied in the training and which, therefore, were trained simultaneously both languages.

ρ(%)	Corr. Names (%)	Corr. Localities (%)
1 Gaussian for state		
0.00	71.28 [69.76,72.80]	85.26 [84.02,86.50]
12.34	71.96 [70.44,73.48]	86.75 [85.51,88.00]
24.95	71.96 [70.44,73.48]	86.06 [84.82,87.30]
34.01	72.63 [71.11,74.15]	85.72 [84.48,87.00]
4 Gaussian for state		
0.00	75.51 [74.00,77.03]	89.21 [88.14,90.28]
34.01	76.52 [75.00,78.04]	87.20 [85.96,88.44]

Table 4: Recognition rates for Catalan sentences depending on the ρ percentage of states trained simultaneously. Between parenthesis there are the probabilities margin with a level of significance of 95 %.

One of the main causes for database people names had worse recognition rate than the site names is that many names are only different by last allophone (due to the gender; for example Francesc for male and Francesc**a** for female) and, in addition, are shorter.

In the results of the table 4 Catalan models with one gaussian for state are obtained that give generally better results using a percentage of bilingual states than the models trained only with the Catalan utterances. However, when we increased the number of Gaussians the recognition improvement becomes a loss of performance for localities database experiment.

4. Conclusion and Future Work

In this paper, we described a method to take advantage of Spanish language speech resources to improve Catalan language acoustic HMMs to speech recognition. We used language as an attribute in the clustering algorithm and CDHMM modeling demiphones. They allowed a better control over the tied allophone context between languages. Further research is need to improve the phonetic transcription and the attributes of these units, for example distinguishing units that at the moment have same symbol SAMPA and, to experiment other types of distances between pdfs, for example the Hellinger distance (Settimi et al., 1999). Our next step will be to carry out experiments increasing the size of the Spanish speech databases and to carry out recognition tests with other tasks, observing the amount of used Spanish material in the training and the test, not only the shared states percentage. Once developed this tying procedure it will be interesting to extend it to other languages that have poor speech databases resources. In our center similar works between dialects of Spanish are being made (Nogueiras et al., 2002).

5. Acknowledgments

This research was supported by CICYT of Spanish government under contract TIC2000-1005-C03-01. The authors would like to thank Climent Nadeu due to his help and encouragement. The authors also wish to acknowledge the access to the VOCATEL database allowed by Telefónica I+D.

6. References

P. Bonaventura, F. Gallocchio, and G. Micca. 1997. Multilingual Speech Recognition for Flexible Vocabularies. In *European Conference on Speech Communication and Technology*, pages 355–358, Rhodes.

U. Bub, J. Köhler, and B. Imperl. 1997. In-Service Adaptation of Multilingual Hidden-Markov-Models . In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1451–54, Münich.

J. Hernando and C. Nadeu. 1999. SpeechDat. Catalan Database for the Fixed Telephone Network. Technical report TALP center, Universitat Politècnica de Catalunya.

B. Imperl and B. Horvat. 1999. The Clustering Algorithm for the Definition of Multilingual Set of Context Dependent Speech Models. In *European Conference on Speech Communication and Technology*, pages 887–90, Budapest.

J. Llisterri and J. B. Mariño. 1993. Spanish Adaptation of SAMPA and Automatic Phonetic Transcription. Technical report, ESPRIT Project 6819, SAM-A/UPC/001/V1, London.

J.B. Mariño, A. Nogueiras, P. Pachès, and A. Bonafonte. 2000a. The Demiphone: an Efficient Contextual Subword Unit for Continuous Speech Recognition. *Speech Communication*, 32(3):187–197.

J.B. Mariño, J. Padrell, A. Moreno, and C. Nadeu. 2000b. Monolingual and Bilingual Spanish-Catalan Speech Recognizers Developed from Speechdat Databases. In *Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities, LREC*, pages 57–61, Athens.

A. Moreno. 1997. SpeechDat Spanish Database for Fixed Telephone Network. Technical report, SpeechDat Project LE2-4001.

C. Nadeu, J. Padrell, and A. Febrer. 1997. Diseño de la Base de Datos Vestel y Preparación de la Captura. Technical report, Projecte VOCATEL (Telefónica I+D), Universitat Politècnica de Catalunya.

A. Nogueiras, M. Caballero, and A. Moreno. 2002. Multi-Dialectal Spanish Speech Recognition. In *International Conference on Acoustics, Speech, and Signal Processing*, Orlando.

P. Pachès. 1999. *Improved modelling for robust speech recognition*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Raffaella Settimi, J.Q. Smith, and Ali S. Gargoum. 1999. Approximate Learning in Complex Dynamic Bayesian Networks. In *Fifteenth Conference on Uncertainty in Artificial Intelligence*, Stockholm.

B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy. 1994. An Evaluation of Cross-Language Adaptation for Rapid HMM Development in a New Language. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 237–40, Adelaine.

S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. 1999. *The HTK Book, Version 2.2*. Entropic.

Portability Issues of Text Alignment Techniques

António Ribeiro*, Gabriel Lopes* and João Mexia[+]

Universidade Nova de Lisboa
Faculty of Sciences and Technology, Department of *Informatics/[+]Mathematics
Quinta da Torre, Monte da Caparica, P–2829–516 Caparica, Portugal
{ambar, gpl}@di.fct.unl.pt

Abstract

Much of the work on parallel texts alignment tries to push the boundaries of language independence as far as possible. This has been a trend since the first approaches on sentence alignment in the early 1990s. In this paper we discuss portability issues of parallel texts alignment techniques. How language independent can they be? We examine several alignment techniques proposed by previous authors, discuss how far they went with language independent methodologies, why some authors decided to add linguistic knowledge to their systems and what improvements they attained by doing it. We will also discuss some methodologies and the problems faced by systems which aim at extracting Translation Equivalents from aligned parallel texts.

1. Introduction

Text alignment techniques aim at identifying automatically correspondences between *parallel texts*, i.e. either correspondences between text segments, words or even sequences of characters. Parallel texts are sets of texts which are translations of each other in different languages, like the proceedings of the European Parliament, which is published in the eleven official languages[1] of the European Union – the Official Journal of the European Communities –, or the proceedings of the Canadian Parliament which is published in both English and French – the Canadian Hansards.

Much of the work on parallel texts alignment tries to push the boundaries of language independence as far as possible, i.e. by not using language specific knowledge for the alignment process. This has been a trend since the first approaches on sentence alignment in the early 1990s (Kay and Röscheisen, 1993; Brown *et al.*, 1991; Gale and Church, 1991). Still, some authors have resorted to adding some linguistic knowledge in order to improve the alignment results, either by adding short bilingual dictionaries to bootstrap the alignment process (Wu, 1994; Melamed, 1999) or by using word similarity measures to find similar words automatically (Simard *et al.*, 1992; Melamed, 1999).

Many authors have tried not to feed linguistic knowledge to their alignment systems, in particular, short bilingual dictionaries, since this makes them easily language dependent and, consequently, hardly portable to other languages. Also, those dictionaries may be incomplete and outdated. In addition, they usually do not provide all the possible word variants due to possible language inflection. Moreover, linguistic knowledge may be expensive to get, may require much time to compile, may be hard to get especially for minority languages or languages for which there are not much linguistic resources available (as in African languages).

In this paper we discuss portability issues of parallel texts alignment techniques. How language independent can they be? We examine several alignment techniques proposed by previous authors, discuss how far these authors have gone with language independent methodologies, why some authors decided to add linguistic knowledge to their systems and what improvements they attained by doing it. We will also discuss some methodologies and the problems faced by systems which aim at extracting Translation Equivalents from aligned parallel texts.

This paper is organised as follows: the next section gives a brief overview of what parallel texts alignment is. Section 3 provides some evidence on the amount of lexical cues available in European languages. Section 4 describes previous work on alignment techniques developed by some authors, both on sentence and word level, and discusses the strategies they used regarding language independence. Section 5 describes how language independent the extraction of Translation Equivalents can be. Finally, section 6 presents the conclusions and some future work.

2. Parallel Texts Alignment

Today, it has become quite common to find parallel texts virtually everywhere from translations of books in bookshops, to consumer products information in supermarkets, instructions manuals in the industry, multilingual portals in the Internet, and it has even become trendy to find parallel versions of songs in English and Spanish. In all these parallel texts, one can notice a continuum in the 'degree of non-parallelness' from legislative texts and instructions manuals, which tend to be very faithful to the originals, to translations of books or lyrics of songs, which leave more freedom and creativity to the translator.

Text alignment techniques aim at identifying automatically correspondences between those parallel texts. Once they are aligned it is possible to start using them for various purposes. For example, an immediate application is the production of bilingual concordances. Bilingual concordances are particularly useful for the preparation of commercial bilingual dictionaries, for translators and even for foreign language learners. They allow the examination of the way specific words or terms are translated into another language, providing simultaneously part of the context in which they appear.

Furthermore, they can also be used to build Bilingual Dictionaries, Bilingual Terminology Databanks, Translation Memories, to name but a few immediate applications.

[1] Danish (da), Dutch (nl), English (en), Finnish (fi), French (fr), German (de), Greek (el), Italian (it), Portuguese (pt), Spanish (es) and Swedish (sv).

This data can then be included in Machine Translation systems, Computer Assisted Translation tools, Cross-Language Information Retrieval systems or Lexicographers workbenches.

The lexical cues found in parallel texts have been quite used. They can be identical tokens in two texts (numbers, proper names, punctuation marks), similar words (cognates, like *Comissão* and *Commission*, in Portuguese and English, respectively) or known translations (like *data* and *fecha*, in Portuguese and Spanish, respectively). These tokens, called *anchors* (from Kay and Röscheisen, 1993, p. 128), allow correspondences between the texts, and help the alignment system to keep track of the evolution of text and to avoid straying away from the correct alignment.

3. Sharing Words

Several authors have used lexical cues as potential anchors for alignment. In fact, the number of identical tokens available in parallel texts should not be underestimated.

According to the results reported in Ribeiro *et al.* (2000), almost 15% of the 'vocabulary' (different tokens) found in their texts from the Official Journal of the European Communities was found to be the same in its various official languages with respect to the Portuguese text (this number also includes names, numbers and punctuation). They used a sample of parallel texts from three sources: records of the Written Questions to the European Commission, records of Debates in the European Parliament and Judgements of The Court of Justice of the European Communities. Table 1 gives an overview of the equal vocabulary size across the ten language pairs (see footnote 1 for the abbreviations):

	Sub-Corpus			
Pair	Written Questions	Debates	Judgements	Average
pt-da	1.2k (17%)	2.0k (10%)	0.2k (19%)	1.9k (11%)
pt-de	1.0k (15%)	1.9k (10%)	0.2k (19%)	1.8k (10%)
pt-el	1.0k (15%)	1.5k (8%)	0.1k (18%)	1.5k (9%)
pt-en	1.3k (19%)	2.2k (11%)	0.2k (20%)	2.1k (12%)
pt-es	2.5k (38%)	6.5k (32%)	0.3k (36%)	6.0k (33%)
pt-fi	---	---	0.2k (19%)	0.2k (19%)
pt-fr	1.3k (19%)	2.3k (11%)	0.2k (22%)	2.1k (12%)
pt-it	1.4k (22%)	3.0k (15%)	0.2k (25%)	2.8k (16%)
pt-nl	1.2k (17%)	2.0k (10%)	0.1k (19%)	1.9k (11%)
pt-sv	---	---	0.2k (19%)	0.2k (19%)
Average	1.3k (20%)	2.7k (13%)	0.2k (22%)	2.5k (14%)

Table 1: Average size of common vocabulary per pair of parallel texts in thousands.

Table 1 also shows the average percentages with respect to the size of the vocabulary found in Portuguese parallel texts are in brackets.

For example, an average of 2500 tokens were found to be exactly the same for the Written Questions parallel texts in Portuguese and Spanish (pt-es). This corresponds to an average of 38%, i.e. 38% of the vocabulary found in the Portuguese Written Questions parallel texts was equal to the Spanish vocabulary.

In the case of close languages such as Portuguese and Spanish, the average rate rises to more than 30%; for the opposite reason, it drops to about 10% for the pair Portuguese–German. Furthermore, the number of occurrences of these shared vocabulary tokens in the parallel texts (see Table 2) reaches an average of almost 50% in parallel texts in Portuguese and Spanish. For Portuguese and German parallel texts, this number is about 20% on average.

	Sub-Corpus			
Pair	Written Questions	Debates	Judgements	Average
pt-da	18.3k (32%)	103.6k (25%)	1.5k (33%)	92.5k (26%)
pt-de	15.0k (27%)	80.7k (19%)	1.4k (31%)	72.2k (20%)
pt-el	16.4k (29%)	66.7k (16%)	1.4k (31%)	60.1k (18%)
pt-en	17.8k (31%)	100.5k (24%)	1.4k (30%)	89.8k (25%)
pt-es	29.7k (52%)	192.5k (46%)	2.4k (52%)	171.4k (47%)
pt-fi	---	---	1.3k (30%)	1.3k (30%)
pt-fr	22.8k (40%)	106.3k (26%)	1.9k (41%)	95.5k (27%)
pt-it	20.3k (35%)	96.7k (23%)	1.8k (38%)	86.7k (25%)
pt-nl	19.8k (35%)	106.0k (25%)	1.6k (35%)	94.8k (26%)
pt-sv	---	---	1.3k (29%)	1.3k (29%)
Average	20.0k (35%)	106.6k (25%)	1.6k (35%)	95.4k (27%)

Table 2: Average number of common tokens per pair of parallel texts in thousands

Table 2 also shows the average percentages of common tokens with respect to the number of tokens of the Portuguese parallel text are in brackets. For example, about 1400 tokens are were found to be equal in both Greek and Portuguese for the Judgements parallel texts; this covers 31% of the total number of tokens of the Portuguese Judgements parallel texts.

This is a wealthy source of lexical cues for parallel texts alignment that should not be left unused.

Homographs, as a naive and particular form of cognates, are likely translations, which makes them potential reliable anchors. For example, *Portugal* is written like this in several European languages, which makes it a potential anchor for alignment.

These anchors end up being mainly numbers and names. Here are a few examples of anchors from a parallel text in English and Portuguese: *2002* (numbers, dates), *ASEAN* (acronyms), *Patten* (proper names), *China* (names of countries), *Manila* (names of cities), *apartheid* (foreign words), *Ltd* (abbreviations), *habitats* (Latin words), *ferry* (common words), *global* (common vocabulary).

4. Alignment Techniques

Some alignment techniques establish correspondences between sentences – *sentence alignment* – where as other techniques try to provide more fine-grained alignments by establishing correspondences between words – *word alignment*. The next section will describe some sentence alignment techniques. Section 4.2 describes word alignment techniques.

4.1. Sentence Alignment

Back in the early days of alignment, in the 1990s, sentences were set as the basic units for alignment. Each text was viewed as a sequence of sentences and alignment algorithms attempted at making correspondences between the sentences in the parallel texts.

The method proposed by Kay and Röscheisen (1993) assumed that for sentences in a translation to correspond, the words in them must also correspond. Two words were considered to have similar distributions if they tended to co-occur in the tentatively aligned sentences. In this case, if their measure of similarity was above a threshold, it would mean they were translations and, finally, sentences were aligned if the number of words associating them was greater than an empirically defined threshold.

In other alternative approaches, less knowledge based, sentences were aligned as long as they had a proportional number of words (Brown *et al*., 1991) or characters (Gale and Church, 1991). They started from the fact that long sentences tend to have long translations and, conversely, short sentences tend to have short translations. This correlation was the basis for their statistical models. Brown *et al.* (1991, p. 175) remarked that the error rate was slightly reduced from 3.2% to 2.3% when using some linguistic knowledge like the time stamps, question numbers and author names found in the parallel texts. This confirmed that sentences could be aligned just by looking at sentence lengths measured in number of tokens and that extra linguistic knowledge did not improve the results significantly.

Although none of these algorithms depend on some word similarity measure as in later work (e.g. Simard *et al*., 1992), these algorithms tended to break down when sentence boundaries were not clearly marked.. This means full stops would have to be clearly interpreted as sentence boundaries markers. However, they are not safe markers of sentence boundaries.

Gale and Church (1991, p. 179) reported that only 53% of the full stops found in the Wall Street Journal were used to mark sentence boundaries. Full stops may be part of abbreviations (*Dr. A. Bromley*) or numbers (*1.3%*), they are not usually found in headlines (*Tyre production*), they may not even exist because they were not added, or they were either lost or were mistaken for noise in the early days when electronic versions of parallel texts were still rare and texts needed to be scanned.

Wu (1994) also aligned English–Chinese sentences with proportional lengths. He also began by applying a method similar to the one used by Gale and Church (1991) and reported results not much worse than those expected by this algorithm. Still, he claimed sentence alignment precision over 96% when the method incorporated a seed bilingual lexicon of words commonly found in the texts to be aligned (e.g. names of months, like *December* and its equivalent in Chinese 十二月). So, again Wu's work confirmed that the use of lexical cues would be beneficial for alignment.

4.2. Word Alignment

If word alignment is the main goal, alignment algorithms must be more 'careful' in order to avoid wrong word correspondences. This is a much more fine-grained alignment since it is no longer done at sentence level but at word level. In contrast with sentence alignment algorithms which permit a margin of tolerance for occasional wrong word matches, at word level, the sentence is no longer a 'safety net'. Consequently, the penalty on wrong word matches becomes much higher.

By adding some lexical information, Church (1993) showed that alignment of parallel text segments was possible by exploiting orthographic *cognates* instead of sentence delimiters. He used the rule of equal 4-grams in order to find 'cognate' (similar) sequences of characters in the parallel texts, i.e. sequences of four characters which are equal in the texts. This is a good strategy for languages which share lexical similarities like languages which share a character set.

The idea of exploiting *cognates* for alignment had been proposed one year earlier in a paper by Simard *et al.* (1992). According to the Longman Dictionary of Applied–Linguistics, a cognate is "a word in one language which is similar in form and meaning to a word in another language because both languages are related" (Richards *et al*., 1985, p. 43). For example, the words *Parliament* and *Parlement*, in English and French respectively, are cognates. They are similar in form and have the same meaning. When two words have the same or similar forms in two languages but have different meanings in each language, they are called false cognates or false friends (Richards *et al*., 1985, p. 103). For example, the English word *library* and the French word *librairie* are false cognates (Melamed, 1999, p. 114): *library* is translated as *bibliothèque* in French and, conversely, *librairie* as *bookstore* in English.

Simard *et al.* (1992) used a simple rule to check whether two words were cognates. They considered two words as cognates if their first four characters were identical (Simard *et al*., 1992, p. 71), as is the case of *Parliament* and *Parlement*. This simple heuristic proved to be quite useful, providing a greater number of lexical cues for alignment though it has some shortcomings. According to this rule, the English word *government* and the French word *gouvernement* are not cognates. Also, *conservative* and *conseil* ('council'), in English and French respectively, are wrongly considered as cognates (Melamed, 1999, p. 113). The rule is sensitive to variations in the first four letters and it does not distinguish different word endings.

In fact, both the rule proposed by Simard *et al.* (1992) and the one used by Church (1993) are two variants of Approximate String Matching Techniques. The former technique corresponds to *truncation*, where only the *n* first characters are considered. The latter technique resembles *n*-gram matching, which determines the similarity of two words by the number of common *n*-grams. A technique developed by Adamson and Boreham (1974) uses contiguous bigrams and base their word similarity score on the coefficient of Dice to compare the number of common bigrams between two words and the number of bigrams of each individual word.

McEnery and Oakes (1995) tried to improve the definition of cognates by comparing the truncation technique, the number of shared bigrams in two words with a score based on the coefficient of Dice and using dynamic programming. In experiments they performed comparing English and French vocabulary, they found that the bigram matching technique precision was 97% using a threshold of 0.9, and 81% for a similarity score between 0.8 and 0.9; the truncation technique precision was 97.5% for a length of eight characters and 68.5% for a length of six characters.

The word alignment approaches just described are not appropriate for pairs of languages for which it is not possible to find some common cues. In order to overcome this problem, Melamed (1999, p. 113) also suggests the use of *phonetic cognates* especially for languages with different alphabets. Phonetic cognates are words which are phonetically similar though written differently or in different scripts, like 'program' /pr græm/ and 'プログラム' /puroguramu/ in English and Japanese. This increases the number of cues available for alignment.

The requirement for clear sentence boundaries was dropped in Fung and Church (1994) on a case-study for English-Chinese. It was the first time alignment procedures were being tested on texts between non-Latin languages and without finding sentence boundaries. Each parallel text was split into K pieces and word correspondences were identified by analysing their distribution across those pieces. In particular, a binary vector of occurrences with size K (hence, the K-vec) would record the occurrence of a word in each of the pieces. Should the word occur in the *i*-th piece of the text, then the *i*-th position of the vector would be set to ‘1’. Next, in order to find whether two words corresponded, their respective K-vecs were compared. In this way, it was possible to build a rough estimate of a bilingual lexicon. This would feed the alignment algorithm of Church (1993), where each occurrence of two translations would become a dot in the graph.

This method was extended in Fung and McKeown (1994). It was also based on the extraction of a small bilingual dictionary based on words with *similar distributions* in the parallel texts. However, instead of K-vecs, which stored the occurrences of words in each of the K pieces of a text, Fung and McKeown (1994) used vectors that stored the distances between consecutive occurrences of a word (DK-vec’s). For example, if a word appeared at offsets (2380, 2390, 2463, 2565, ...), then the corresponding distances vector would be (10, 73, 102, ...). Should an English word and a Chinese word have distance vectors with a similarity above a threshold, then those two words would be used as potential anchors for alignment. Later, in Fung and McKeown (1997), rather than using only single words, the algorithm extracted terms to compile the list of reliable pairs of translations, using specific syntactic patterns. However, this made it become language dependent.

Melamed (1999) also used orthographic cognates. Moreover, he used lists of stop words to avoid matching of closed-class words (like articles and prepositions) which tended to generate much noise, which requires some linguistic knowledge to be hand-coded into the system. In order to measure word similarity, he defined the Longest Common Sub-sequence Ratio as follows:

$$Ratio(w_1, w_2) = \frac{Length(LongestCommonSub\text{-}Sequence(w_1, w_2))}{Max(Length(w_1), Length(w_2))}$$

where w_1 and w_2 are the two words to be compared (Melamed, 1999, p. 113).

This measure compares the length of the longest common sub-sequence of characters with the length of the longest token. For the previous example, the ratio is 10 (the length of *government*) over 12 (the length of *gouvernement*) whereas the ratio is just 6 over 12 for *conservative* and *conseil* (council). This measure tends to favour long sequences similar to the longest word and to penalise sequences which are too short compared to a long word. However, for this very same reason, it fails to consider *gouvernement* and *governo* in French and Portuguese as cognates because *governo* is a shorter word. Their ratio is also 6 over 12.

For alignment purposes, Melamed (1999) selects all pairs of words which have a ratio above a certain threshold, heuristically selected. However, this becomes a language dependent value. Still, this comparison measure seems to provide better results than the one first proposed by Simard *et al.* (1992) but it is not based on a statistically supported study.

Danielsson and Mühlenbock (2000) aim at aligning cognates starting from aligned sentences in two quite similar languages: Norwegian and Swedish. The ‘fuzzy match’ of two words is “calculated as the number of matching consonants[,] allowing for one mismatched character” (Danielsson and Mühlenbock, 2000, p. 162). For example, the Norwegian word *plutselig* (suddenly) and the Swedish word *plötsligt* would be matched through *pltslg*: all consonants match except for the ‘t’. However, *bakspeilet* (rear-view mirror) and *backspegeln*, in Norwegian and Swedish respectively, would not match because four consonants are not shared ‘c’, ‘g’, ‘n’ and ‘t’. This strategy resembles the technique developed by Pollock and Zamora (1984, p. 359) whereby words are coded using the first letter of the word, the remaining unique consonants in order of occurrence and, finally, the unique vowels also in order of occurrence – the *skeleton key*. For example, *plutselig* would be coded as *pltsguei* and *plötsligt* as *pltsgöi* where the sequence of consonants is equal.

Choueka *et al.* (2000) present an alignment algorithm for English and Hebrew, a highly inflected language with a different alphabet, a complex morphology and flexible word order. For example, *and since I saw him* is translated into a single Hebrew word (Choueka *et al.*, 2000, p. 74): åëùøàéúéå /ukhshereitiv/. First, texts were lemmatised, i.e. each word was reduced to its basic form as found in a dictionary entry (e.g. *saw* to *see*). This is clearly a language dependent task though it is quite difficult to solve for highly inflected languages. Also high frequency words were removed using a list of stop words. Next, parallel texts were aligned using the methodology of Fung and McKeown (1997). The lemmatisation step increases the chances of finding similar words in the aligned parallel texts in order to compile automatically a bilingual dictionary by analysing their distribution across the aligned parallel texts; otherwise, non-lemmatised words would just become too rare which would make it difficult to find trustworthy Translation Equivalents due to data sparseness.

In contrast with previous approaches, Ribeiro *et al.* (2001b) consider two words to have a high level of ‘cognateness’, if they share a typical sequence of characters that is common to that particular pair of

Language Pairs	pt-da	pt-de	pt-el	pt-en	pt-es	pt-fi	pt-fr	pt-it	pt-nl	pt-sv	Average
punctuation	60%	63%	66%	60%	46%	68%	59%	60%	63%	74%	62%
numbers	16%	20%	20%	21%	11%	18%	16%	12%	19%	20%	17%
names	17%	8%	7%	10%	8%	7%	6%	5%	9%	1%	8%
common words	3%	3%	6%	7%	33%	6%	14%	18%	9%	3%	10%
others	4%	6%	1%	2%	2%	2%	5%	5%	1%	1%	3%
Total	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 3: Percentages of types of tokens used for alignment in the alignment algorithm developed by Ribeiro *et al.* (2001).

languages. The typical sequences of characters can be extracted by statistical data analysis of contiguous and non-contiguous sequences of characters, based on the notion of ‘textual unit’ association. They were able to find typical sequences which lie in the beginning of words such as •*Comis*, for *Comissão* and *Comisión* in Portuguese and Spanish, lie in the middle of words as in *f_rma* which matches both *information* and *informação* in English and Portuguese respectively, cross word boundaries as *i_re•ci* for the Portuguese–French pair as in *livre•circulação–libre•circulation* (‘free movement’), which made it quite adequate for the pairs of languages which use words written in the same character set. It is up to the alignment algorithm proper to confirm whether words are cognates depending on their position in the text.

This particular experiment used the Judgements sub-corpora, in three language pairs: Portuguese–English, Portuguese–French and Portuguese–Spanish. The size of the parallel texts for each language pair amounted to about 150k characters (about 30k tokens). Table 4 shows the number of typical sequences of characters extracted from each parallel text.

Language Pair	Number of Sequences
Portuguese–English	677
Portuguese–Spanish	1137
Portuguese–French	877

Table 4: Number of typical sequences of characters for each pair of languages.

Interestingly, Table 4 also confirms language similarity. Bearing in mind that Portuguese and Spanish are two quite close languages, it does not come as a surprise to see that this pair shares more typical sequences of characters than any of the other. French comes next for its closeness as a Romance language and English comes last confirming that Portuguese and English are more distant languages.

Table 3 presents an analysis of a sample of aligned parallel texts, using the previously mentioned methodology though just using equal tokens. The table shows that punctuation marks are indeed good cues for alignment. On average, more than 60% of the tokens used for correspondence points are punctuation marks. This confirms the success of early approaches that started by using sentences as the basic alignment unit and exploiting full stops for sentence alignment. It shows that the number of common words used as correspondence points is higher for similar pairs of languages like Portuguese–French, Portuguese–Italian and Portuguese–Spanish than for other pairs. It shows that, on average, 10% of the tokens used as correspondence points are common words and that 17% are numbers, which makes it more than a quarter of all tokens used as correspondence points.

5. Translation Equivalents

The extraction of Translation Equivalents is one of the most important tasks for building either Translation Memories or Bilingual Dictionaries. Translations databanks are useful language resources either for Machine Translation, Cross-Language Information Retrieval or even for human translators themselves.

Aligned parallel texts are ideal sources to extract Translation Equivalents for they provide the correspondences between the original text and their translations in other languages made by professional translators. They allow the examination of the way specific words or terms are translated into other languages. Aligned parallel texts can reduce the amount of effort necessary to build Translation Databanks.

The key issue in the extraction of Translation Equivalents is to find a correlation between co-occurrences of terms in parallel texts. In general, if two terms co-occur often in aligned text segments, then they are likely to be *equivalent*. The alignment of parallel texts splits the texts into small aligned text segments and reduces the number of words that must be checked for co-occurrence. In order to identify Translation Equivalents, their *distribution similarity* must be analysed in those aligned segments.

However, the larger the aligned text segments, the more difficult it gets to extract Translation Equivalents for more alternative translations become possible and, consequently, the search space becomes larger and with fewer evidences. This may be the case for distant languages where fewer cues may be available for alignment. As a result, the number of Translation Equivalents which can be more reliably extracted gets more reduced.

Nonetheless, the few Equivalents extracted can be subsequently fed back into the alignment system to improve the alignment proper, reduce the size of aligned text segments, and extract more Translation Equivalents in an iterative and unsupervised way. Even though it should only be possible to extract a small bilingual lexicon as in Fung and McKeown (1994), it can be quite helpful to bootstrap a more fine-grained alignment as Wu (1994) has shown.

6. Conclusions and Future Work

The exploitation of lexical cues for parallel text alignment is indeed quite helpful for alignment methods based on lexical information found in the texts. The more lexical information shared between a pair of languages, the more candidate correspondence points for alignment can be generated. As a result, this leads eventually to a more fine-grained alignment beyond the sentence level as in the early 1990s. Language similarity should be seen as bonus for alignment.

Language independent approaches are quite dear in multilingual regions where the possibility of using a single methodology to handle different languages increases portability and greatly reduces the amount of human effort. Ideally, an alignment algorithm should be completely language independent: character set independent; no previous linguistic knowledge, either from machine-readable bilingual dictionaries or hand coded seed bilingual translation lexicons; no lemmatised and/or tagged texts; no requirement for the detection of sentence boundaries.

However, as described in section 4, previous alignment approaches have often resorted to making use of sentence boundaries, lexical cues available in the parallel texts and even to hand-coding some linguistic knowledge through small bilingual lexicons and building list of stop words. This increases the number of potentially reliable anchors for alignment and increases the chances of having more accurate alignments of parallel texts.

Nevertheless, it is wise to make good use of the lexical cues available in parallel texts. The larger the overlap between common lexical cues between two languages, the higher the number of potential anchors for alignment. Eventually, this means that the average size of aligned parallel texts gets smaller for non-sentence based alignment algorithms. The extraction of Translation Equivalent becomes more reliable and easier since there may be fewer alternative translations to choose from.

Consequently, when it comes to more distant languages like Portuguese and Chinese, where the number of lexical cues available is more reduced, the number of Translation Equivalents extracted is usually more reduced (Ribeiro *et al*., 2001a). However, it is still possible to extract some Translation Equivalents reliably in order to re-feed the alignment algorithm. Indeed, for distant languages, previous authors (Wu, 1994; Melamed, 1999) have resorted to building a small bilingual lexicon to bootstrap the alignment algorithm.

We believe that it is possible to extract some Translation Equivalents in a 'self-enriching' process instead of feeding an alignment system manually with either hand-coded bilingual lexical information or incomplete machine readable dictionaries.

By re-feeding the extracted Translation Equivalents back into the aligner it is possible to increase the number of candidate correspondence points for new lexical cues become available for the generation of correspondence points. The more candidate correspondence points, the more fine-grained the alignment and the better are the extracted equivalents. This means that the alignment precision may be improved, i.e. more correspondences may be established between words or phrases.

As the example of Choueka *et al*. (2000) has shown, it becomes more difficult to get cues for highly inflected languages where words can suffer major changes. Still, it would be interesting to test whether it should be possible to automatically lemmatise texts either by using a strategy similar to the one presented in Kay and Röscheisen (1993) whereby common suffixes and prefixes of words were automatically identified in a language independent fashion, or by extracting automatically character patterns using a methodology similar to the developed in Ribeiro *et al*. (2001b).

All in all, most work on alignment has been carried out on a wide range of 'popular' languages, most of them on English and French, but also including other Western European languages, Arabic, Chinese, Hebrew, Japanese and even some Korean. It would be quite interesting to test alignment algorithms on radically different languages to check for their degree of language independence.

7. References

Adamson, G. and Boreham, J. (1974). The Use of an Association Measure Based on Character Structure to Identify Semantically Related Pairs of Words and Document Titles. *Information Storage and Retrieval*, 10, 253–260.

Brown, P., Lai, J. and Mercer, R. (1991). Aligning Sentences in Parallel Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (pp. 169–176), Berkeley, California, USA.

Choueka, Y., Conley, E. and Dagan, I. (2000). A Comprehensive Bilingual Word Alignment System: Application to Disparate Languages – English and Hebrew. In J. Véronis (ed.), *Parallel Text Processing: Alignment and Use of Translation Corpora* (pp. 69–96). Dordrecht, The Netherlands: Kluwer Academic Publisher.

Church, K. (1993). Char_align: A Program for Aligning Parallel Texts at the Character Level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics* (pp. 1–8). Columbus, Ohio, USA.

Danielsson, P. and Mühlenbock, K. (2000). Small but Efficient: The Misconception of High-Frequency Words in Scandinavian Translation. In J. White (ed.), *Envisioning Machine Translation in the Information Future – Proceedings of the 4th Conference of the Association for Machine Translation in the Americas, AMTA 2000 – Lecture Notes in Artificial Intelligence*, 1934, pp. 158–168. Berlin, Germany: Springer-Verlag.

Fung, P. and Church, K. (1994). K-vec: A New Approach for Aligning Parallel Texts". In *Proceedings of the 15th International Conference on Computational Linguistics – Coling'94* (pp. 1096–1102), Kyoto, Japan.

Fung, P. and McKeown, K. (1997). A Technical Word- and Term-Translation Aid Using Noisy Parallel Corpora across Language Groups. In *Machine Translation*, 12(1–2), 53–87. Dordrecht, The Netherlands: Kluwer Academic Publisher.

Fung, P. and McKeown, K. (1994). Aligning Noisy Parallel Corpora across Language Groups: Word Pair Feature Matching by Dynamic Time Warping. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas* (pp. 81–88). Columbia, Maryland, USA.

Gale, W. and Church, K. (1991). A Program for Aligning Sentences in Bilingual Corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (pp. 177–184), Berkeley, California, USA (short version). Also (1993) *Computational Linguistics*, 19 (1), 75–102 (long version).

Kay, M. and Röscheisen, M. (1993). Text-Translation Alignment. *Computational Linguistics*, 19 (1), 121–142.

Melamed, I. (1999). Bitext Maps and Alignment via Pattern Recognition. *Computational Linguistics*, 25 (1), 107–130.

McEnery, A. and Oakes, M. (1995). Sentence and Word Alignment in the CRATER Project: Methods and Assessment. In S. Warwick-Armstrong (ed.), *Proceedings of the SIGDAT Workshop "From Texts to Tags: Issues in Multilingual Language Analysis"* (pp. 77–86). Dublin, Ireland.

Pollock, J. and Zamora, A. (1984). Automatic Spelling Correction in Scientific and Scholarly Text. *Communications of the Association for Computing Machinery (ACM)*, 27 (4), 358–368. ACM Press.

Simard, M., Foster, G. and Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. In *Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation TMI-92* (pp. 67–81), Montréal, Canada.

Simard, M. and Plamondon, P. (1998). Bilingual Sentence Alignment: Balancing Robustness and Accuracy. *Machine Translation*, 13 (1), 59–80. Dordrecht, The Netherlands: Kluwer Academic Publisher.

Ribeiro, A., Lopes, G. and Mexia, J. (2000). Using Confidence Bands for Parallel Texts Alignment. In *Proceedings of the 38th Conference of the Association for Computational Linguistics* (pp. 432–439). Hong Kong, China.

Ribeiro, A., Lopes, G. and Mexia, J. (2001a). Extracting Translation Equivalents from Portuguese-Chinese Parallel Texts. In *Proceedings of Asialex 2001 – The Second International Congress of the Asian Association for Lexicography (Asialex)* (pp. 225–230). Seoul, South Korea.

Ribeiro, A., Dias, G., Lopes, G. and Mexia, J. (2001b). Cognates Alignment. In B. Maegaard (ed.), *Proceedings of the Machine Translation Summit VIII – MT Summit VIII – Machine Translation in the Information Age* (pp. 287–292). Santiago de Compostela, Spain.

Richards, J., Platt, J. and Weber, H. (1985). *Longman Dictionary of Applied Linguistics*. London, United Kingdom: Longman.

Wu, D. (1994). Aligning a Parallel English–Chinese Corpus Statistically with Lexical Criteria. In *Proceedings of the 32nd Annual Conference of the Association for Computational Linguistics* (pp. 80–87), Las Cruces, New Mexico, USA.

SPE based selection of context dependent units for speech recognition

Matjaž Rodman♦, Bojan Petek and Tom Brøndsted*

Interactive System Laboratory
Faculty of Natural Sciences and Engineering
University of Ljubljana
Snežniška 5, 1000 Ljubljana, Slovenia
matjaz.rodman@ntftex.uni-lj.si, bojan.petek@uni-lj.si

*Center for PersonKommunikation (CPK)
Institute of Electronic Systems
Aalborg University
Niels Jernes Vej 12, 9220 Aalborg, Denmark
tb@cpk.auc.dk

Abstract

Decision tree-based approach is a well known and frequently used method for tying states of the context dependent phone models since it is able to provide good models for contexts not encountered in the training data. In contrast to the other approaches, this method allows us to include expert linguistic knowledge into the system. Our research focused on the inclusion of standard generative theory by Chomsky & Halle (1968), called the SPE theory (the Sound Pattern of English), into the decision tree building process as expert linguistic knowledge. Our attempt was to "merge" the SpeechDat2 SAMPA label set, used for English and Slovenian languages, with the SPE. We created all possible natural groups of phones (SAMPA segments defined by a set of binary phonological features) for both languages and included them into a set of questions used in the process of creating the decision trees. Based on the decision tree constructed this way, we created an English and Slovenian speech recognition systems and tested both of them. Compared with the reference speech recognition system (Lindberg et al., 2000; Johansen et al., 2000) we got some promising results that encouraged us to continue this work and to perform further testing.

1. Introduction

Much of the phonetic variation in natural speech is due to contextual effects. In order to be able to accurately model variations in natural speech a careful choice of the units represented by each model is required. In large-vocabulary speech recognition systems, modelling of vocabulary words by subword units (phonemes or units derived from phonemes) is mandatory. For example, triphone models have been one of the most successful context dependent units because of their ability to model well the co-articulation effect. Yet if we create distinct models for all possible contexts, the number of models becomes very high. In practical applications of building speech recognition systems, there is often a conflicting desire to have a large number of models and model parameters in order to achieve high accuracy, whilst at the same time having limited and uneven training data in form of labelled utterances of a particular language (Young et al., 2000). In the case of triphone context dependent models, tying of HMM states gives us a possible solution of how to overcome this problem.

In our work we analysed the influence of the decision tree method on the acoustic modelling. We also analysed parameters that influence the decision tree building process and tested the proposed method based on the theory of naturalness (the theory that phonological segments cluster into "natural groups" defined by universal features), (Chomsky et al., 1968). We first examined this issue within the Slovenian language and then also addressed its portability to other languages.

2. Decision tree

When building large vocabulary cross-word triphone systems, unseen triphones are unavoidable. A limitation of the data-driven clustering procedure is that it does not deal with triphones for which there are no examples in the training data. Decision tree based approach gives us a possibility to include expert linguistic knowledge into a procedure of creating acoustic models. This methodology provides appropriate models also for contexts that are not seen in the training data. Therefore, decision trees are used in speech recognition with large numbers of context dependent HMMs, to provide models for contexts not seen in the training data. Sharing data at the model level may not be the most appropriate method for models composed of distinct states (Odell, 1995). Sharing distributions at the state level allows for finer distinctions to be made between the models by allowing left and right contexts to be modelled separately.

2.1. Decision tree building process

A phonetic decision tree is a binary tree in which a yes/no phonetic question is attached to each node (Young et al., 2000). Initially, all states in a given item list (typically a specific phone state position) are placed at the root node of a tree. Depending on each answer, every node is successively split and this continues until the states have trickled down to leaf-nodes. All states in the same leaf node are then tied and trained from the same data.

The question at each node is chosen to (locally) maximise the likelihood of the training data (using a log likelihood criterion) and gives the best split of the node. This process is repeated until the increase in log likelihood falls below the specified threshold. As a final stage, the decrease in log likelihood is calculated for merging terminal nodes, which belong to different parents. Any pair of nodes for which this decrease is less than the threshold used to stop splitting is then merged (Young et al., 2000). The algorithm for building a decision tree is summarised in figure 1.

♦ Socrates/Erasmus exchange student under the multilateral agreement UL D-IV-1/99-JM/Kc.

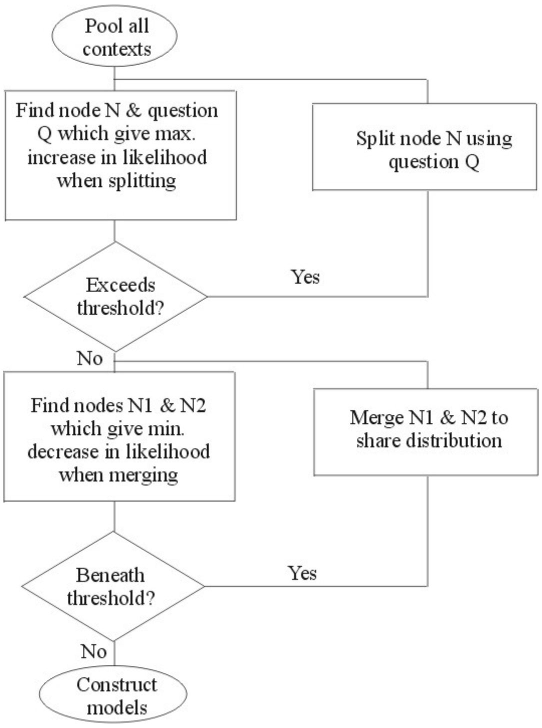

Pool all contexts
Find node N & question Q which give max. increase in likelihood when splitting
Split node N using question Q
Exceeds threshold?
Yes
No
Find nodes N1 & N2 which give min. decrease in likelihood when merging
Merge N1 & N2 to share distribution
Beneath threshold?
Yes
No
Construct models

Figure 1. Algorithm for constructing decision tree (Odell, 1995)

Questions asked in the decision tree have a form:

QS "L_SL_Nasal" { m-,n-*, N-* }*

As an example, the command above defines the question "Is the left context a nasal?" where the group of nasals is represented by *{m-*, n-*,N-*}*. Only a finite set of questions can be used to divide each node. So questions have to be defined in a way that all possible natural groups of phonological segments are stated. That allows the incorporation of expert linguistic knowledge needed to predict contextual similarity when little or no data is available in order to determine which contexts are acoustically similar.

Decision tree building process has two stop criteria that determine how deep the tree will be. The first one is increase in the log likelihood that has to be achieved if node was split. In HTK (Young et al., 2000) it is defined with the command TB. The second one is the minimal occupation count that determines how many training data each node has to have. In HTK it is defined with the command RO.

3. SPE theory

Distinctive feature theory was introduced first by R. Jakobsen. He set up twelve universal inherent feature classes. Chomsky and Halle took over Jakobsens idea and defined 22 universal feature classes, which according to the standard SPE theory are sufficient for analysing expression segments of any language into distinctive oppositions.

The idea of natural phonetic groups is based on the so-called Sound pattern of English theory, "SPE", of Chomsky & Halle (1968). By this theory an inventory of expression segments can be described in terms of a hierarchical tree structure where upper nodes represent major class features (like +/- vocalic, +/- consonantal) and lower nodes cavity features, manner of articulation etc., and terminal nodes represent phones. A phonetic representation of an utterance in a given language has by this theory the form of a two-dimensional matrix in which the rows are labelled by features of universal phonetics; the columns stand for the consecutive segments of the utterance generated; and the entries in the matrix determine the binary value (+/-) of each segment with respect to the universal features (Chomsky et al., 1968). A set of phonological segments ("phonemes") sharing the same feature matrix and unequivocally defined by this matrix form a natural group. There are more degrees of naturalness. The SPE theory claims that one group is more natural than the other if the number of features defining it is smaller. The main natural groups (vowels, consonants, semi-vowels) are separated just by different values in major class features. Specific groups (e.g. back-vowels, plosives, nasals, labials) are defined by further features in the matrix and are consequently "less natural". Groups of segments that cannot be defined by a feature matrix are not natural (e.g., the pseudo group: k, a, m, h).

3.1. The use of SPE on SpeechDat2 databases

The starting point of our distinctive features composition can be described as follows:

- We intended to use the SPE as a generally accepted standard theory of phonology and with as few modifications as possible.
- Most notably, we have tried to utilise the Chomsky & Halle decomposition of English segments (1968) as directly as possible.
- Finally, we have attempted to make as few changes to the SpeechDat2 label set as possible.

Hence, our starting point can be paraphrased as attempt to "merge" the SAMPA label set used in SpeechDat2 database with the SPE.

The SPE sets up a total number of twenty-two feature classes, which according to the standard theory are sufficient for analysing expression segments (phonemes) of any language into distinctive oppositions. For a distinctive feature composition of the segments of a specific language, not all 22 feature classes are utilised. For instance, the SPE-description of English segments (Chomsky et al., 1968) makes references only to 13 feature classes. The remaining 9 classes may be regarded as redundant or "irrelevant" to English.

The set of 15 features was sufficiant to represent the set of Slovenian and English SAMPA symbols used in the SpeechDat2 database by the standard SPE theory. In general, we tried to preserve the original distinctive features used in the SPE. We had to, however, make some changes. In short, we replaced the feature vocalic with sonorant and syllabic, and added a feature front (Brøndsted, 1998). The feature +/- front is not within the set of 22 universal binary features defined in the SPE. However, the feature is needed additionally to +/-back because the SAMPA symbols include segments of a dubious phonological state, only specifiable with reference to three places of articulation: [-back, +front], [-back, -front], and [+back, -front].

3.2. Major Class Features

In standard generative phonology, the major class features sonorant, syllabic and consonantal are used to classify phonological segments into five major groups: vowels, non-syllabic liquids/nasals, syllabic liquids/nasals, glides, and obstruents. However, as the SAMPA segments defined for English and Slovenian do not include syllabic liquids/nasals, this in our case resulted in only four major groups (cf. table 1).

	Sonorant	Syllabic	Consonantal
Vowels	+	+	-
Glides	+	-	-
Syllabic Liquids and Nasals	+	+	+
Non-Syllabic Liquids and Nasals	+	-	+
Obstruents	-	-	+

Table 1: The main natural groups represented by major class features

3.3. The use of SPE on the Slovenian SpeechDat2 database

To create a distinctive feature composition table of the Slovenian SAMPA symbols used in SpeechDat2 we had to modify the phonetic transcriptions. In total, SpeechDat2 uses 46 SAMPA symbols in the Slovenian transcriptions. However, according to (Šuštaršič, 1999; Toporišič 2000) Slovenian only has 29 phonemes. Thus, 17 symbols must be considered allophonic variants. These allophones include certain composite pseudo segments (t_n, d_n, p_n, b_n, t_l, d_l) used along with the normal polyphonematic transcriptions (t n, d n ... etc.) in a way that appeared non-systematic to us. Consequently, we decided to change phonetic transcriptions in the database according to the following seven rules:

- Change string "t_n n" with two symbols "t n"
- Change string "d_n n" with two symbols "d n"
- Change string "p_n n" with two symbols "p n"
- Change string "b_n n" with two symbols "b n"
- Change string "t_l l" with two symbols "t l"
- Change string "d_l l" with two symbols "d l"
- Change symbol "W" with symbol "w"

This reduced the set of segments from 46 to 39. The resulting distinctive feature composition table of the Slovenian vowel and consonantal segments is shown in tables 2 and 3.

	i	i:	e	e:	E	E:	a	a:	u	u:	o	o:	O	O:	@	@r
Sonor.	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Conson.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
High	+	+	-	-	-	-	-	-	+	+	-	-	-	-	-	-
Back	-	-	-	-	-	-	+	+	+	+	+	+	+	+	-	-
Front	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-
Low	-	-	-	-	+	+	+	+	-	-	-	-	+	+	-	+
Round	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-
Tense	-	+	-	+	-	+	-	+	-	+	-	+	-	+	-	-
Anterior																
Coronal																
Voice																
Cont.																
Nasal																
Strident																

Table 2: Distinctive feature composition of Slovenian vowel segments

	b	d	g	p	t	k	dZ	ts	tS	s	S	z	Z	f	v
Sonor.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Conson	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
High	-	-	+	-	-	+	+	-	+	-	+	-	+	-	-
Back															
Front															
Low															
Round															
Tense															
Anterior	+	+	-	+	+	-	-	+	-	+	-	+	-	+	+
Coronal	-	+	-	-	+	-	+	+	+	+	+	+	+	-	-
Voice	+	+	+	-	-	-	+	-	-	-	-	+	+	-	+
Cont	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Strident	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+

	w	j	x	r	l	m	n	N
Sonor	+	+	+	+	+	+	+	+
Syllabic	-	-	-	-	-	-	-	-
Conson.	-	-	-	+	+	+	+	+
High	+	+	-	-	-	-	-	+
Back	+	-	-	-	-			
Front	-	+	-	-	-			
Low	-	-	+	-	-	-	-	-
Round	+	-						
Tense	-	-						
Anterior	-	-	-	-	+	+	+	-
Coronal	-	-	-	+	+	-	+	-
Voice			-	+	+	+	+	+
Cont.			+	+	+	-	-	-
Nasal			-	-	-	+	+	+
Strident			-	-	-	-	-	-

Table 3: Distinctive feature composition of Slovenian consonantal segments

3.4. The use of SPE on the English SpeechDat2

Similarly we had to modify the transcriptions of the English SpeechDat2 database. The major problem was the monophonematic representation of diphtongs (as single phones). In SPE theory there are no phonological features differentiating diphthongs from monofthongs. This theory handles diphthongs with certain appropriate diphthongisation rules applied to the underlying representations (Chomsky et al., 1968). In order to provide a level of description conforming to the underlying representation presupposed by the SPE, the diphtongs were re-written according to the 8 rules:

- Change symbol “eI” with phones “e” and “j”
- Change symbol “aI” with phones “{” and “j”
- Change symbol “OI” with phones “Q” and “j”
- Change symbol “@U” with phones “@” and “w”
- Change symbol “aU” with phones “{” and “w”
- Change symbol “I@” with phones “I” and “@”
- Change symbol “e@” with phones “e” and “@”
- Change symbol “U@” with phones “U” and “@”

The resulting distinctive feature composition of the English vowels and consonants are presented in tables 4 and 5.

	i:	u:	3:	O:	A:	I	U	e	{	Q	V	@
Sonor.	+	+	+	+	+	+	+	+	+	+	+	+
Syllabic	+	+	+	+	+	+	+	+	+	+	+	+
Conson.	-	-	-	-	-	-	-	-	-	-	-	-
High	+	+	-	-	-	+	+	-	-	-	-	-
Back	-	+	-	+	+	-	+	-	-	+	+	-
Front	+	-	-	-	-	+	-	+	+	-	-	-
Low	-	-	-	+	+	-	-	-	+	-	-	-
Round	-	+	+	+	-	-	+	-	-	+	-	-
Tense	+	+	+	+	-	-	-	-	-	-	-	-
Anterior												
Coronal												
Voice												
Cont.												
Nasal												
Strident												

Table 4: Distinctive feature composition of English vowel segments

3.5. Definition of natural groups

During the process of creating the decision tree, groups of phones are used to define questions that may be used in each node of the decision tree. This is the most important stage in the entire model-building procedure where expert phonological knowledge can be included (another one is the prior stage, where the actual set of phones to be used for segmentation and classification of the acoustic signal is established). For that reason, groups of phones for five languages - among these both Slovenian and English - were defined as a part of the COST 249 project. As the languages partly use the same phonemic label set (SAMPA), the groups are reuseable across languages. Slovenian contributes with 45 groups and English with 17 groups. During the process of creating the decision tree, two questions are created from every group defined. One is about the left context and the other about the right one. On the basis of these definitions we created English and Slovenian reference recognition systems.

Our main goal was to create another two systems for both languages that would have phone groups defined on the basis of the SPE theory. Therefore we automatically generated all natural groups of phones from the distinctive feature compositions table set up for the two languages. This resulted in 174 natural groups for Slovenian and 171 for English. The groups were used to create the set of all possible questions to be included in the process of building the experimental SPE-based speech recognition systems.

	b	d	g	p	t	k	dZ	tS	s	S	z	Z	f	T	v	D
Sonor.	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Syllabic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Conson	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
High	-	-	+	-	-	+	+	+	-	+	-	+	-	-	-	-
Back																
Front																
Low																
Round																
Tense																
Anterior	+	+	-	+	+	-	-	-	+	-	+	-	+	+	+	+
Coronal	-	+	-	-	+	-	+	+	+	+	+	+	-	+	-	+
Voice	+	+	+	-	-	-	+	-	-	-	+	+	-	-	+	+
Cont	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Strident	-	-	-	-	-	-	+	+	+	+	+	+	+	-	+	-

	w	j	h	r	l	m	n	N
Sonor	+	+	+	+	+	+	+	+
Syllabic	-	-	-	-	-	-	-	-
Conson.	-	-	-	+	+	+	+	+
High	+	+	-	-	-	-	-	+
Back	+	-	-	-	-			
Front	-	+	-	-	-			
Low	-	-	+	-	-	-	-	-
Round	+	-						
Tense								
Anterior	-	-	-	-	+	+	+	-
Coronal	-	-	-	+	+	-	+	-
Voice			-	+	+	+	+	+
Cont.			+	+	+	-	-	-
Nasal			-	-	-	+	+	+
Strident			-	-	-	-	-	-

Table 5: Distinctive feature composition of English consonantal segments

4. Importance of the order of questions for “unseen” contexts

We hypothesised a case of why it would be not advisable to create questions that would include all possible combinations of phonemes (including "unnatural" groups) and leave to the decision tree building process to chose the best ones by it’s own criteria. This way the decision tree building process would pick up only the important questions (likely involving only "natural" groups) and leave out the irrelevant ones. The idea emerged because of the explanation in the HTK documentation considering the problem of how to build questions for a decision tree: “There is no harm in creating extra unnecessary questions, because those which are determined to be irrelevant to the data will be ignored” (Young et al., 2000). That would yield us the optimal decision tree for this particular system without including any linguistic knowledge. By this definition also the order of the questions in the file that HTK uses for creating a decision tree should have no effect on the structure of the decision tree. But already the first experiment showed us that the order of questions in this file *does* matter.

When we changed the order of questions in the file also the structure of decision tree has changed. Considering how questions are chosen in the process of building decision tree, we got a possible explanation for this change. For example let's suppose that we in the process of deciding how to cluster the centre state of the phone */m/*. Let's assume that we have data only for the triphones *a-m+**, *b-m+**, *c-m+** and *d-m+** where * means any context. Suppose further that we have defined the questions *QS "L_context1" {a-*, b-*, x-*}* and *QS "L_context2" {a-*, b-*}* where the first one is a superset of the first one (including also the left context '*x*'). The log-likelihood can only be calculated for data that is available for training. Therefore these two questions would cause the same increase in log likelihood if they were used for splitting the node because the left context *x-** does not appear in the training data. So if *L_context1* was used, the middle state of the model with the left context *x* would be trained from the same data as middle states of the models with left contexts *a* and *b*! Likewise, if *L_context2* was used, the middle state of the model with the left context *x* would be trained from the same data as the middle states of the models with left contexts *c* and *d* so from different data as in the first case. Both situations are presented in figure 2. Increase in log-likelihood would be the same in both cases. Therefore, only the order of questions in the file where questions are defined or the procedure that defines which question to use, if more questions give the same increase in log likelihood, would decide from which data model with left context *x* was trained. This means that for the models with contexts not seen in the training data (like *x* here) the decision from which data they'll be trained would depend on the order of questions.

From this we concluded that the phone groups that are later transformed into questions must not be defined without linguistic knowledge, because of the classification of contexts not appearing in the training data.

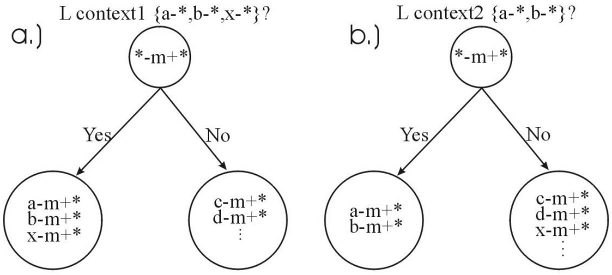


Figure 2. Effect of the order of questions on decision tree

5. Experimental methodology

The main scripts for training and testing acoustic models were implemented as Perl scripts invoking HTK. They were the outcome of the COST 249 project and intended to be used on the SpeechDat2 databases (Lindberg, 2000; Johansen, 2000) and are an extended version of the tutorial example in the HTK Book (Young et al., 2000). They can all be found on the Refrec homepage at

http://www.telenor.no/fou/prosjekter/taletek/refrec/

On this web page we can also find descriptions of standard tests and results of comparative tests done on many SpeechDat2 databases. We used hidden Markov models (HMM) having the 3-state left-right topology. We built triphone models and increased the number of Gaussian mixtures per state sequentially to 32.

5.1. The reference speech recognition systems

For building reference recognition systems we defined questions used in decision tree from groups of phones that were created as a part of the COST 249 project. For the English system we had 17 groups and for Slovenian 45 groups. During the training of acoustic models, data from labelled pronunciations of 800 speakers were used, while the data of the remaining 200 speakers was used as a test set.

The choice of good threshold values is important for the decision tree building process and requires some experimentation in practice. We therefore decided to experiment with the threshold set with the HTK RO command. This threshold determines how many training data each leaf in the decision tree must have. We built one Slovenian system with the threshold set to 100 and two English systems with thresholds set to 100 and 350, respectively (we named them sl-ref100, en-ref100 and en-ref350).

5.2. Speech recognition system with groups based on the SPE theory

In order to evaluate the effect of including the SPE theory into the decision tree building process we built five additional systems – three Slovenian and two English ones. For the model training we used the modified phonetic transcriptions as described in sec. 3.3 and 3.4. We automatically generated all natural phonetic groups from the distinctive feature compositions tables for both languages. From these groups, questions were generated that were used in the process of building decision trees for the two languages. Because of the modified phonetic transcriptions (less phones were used) and the modification of the broad classes, the number of leaves in the decision tree also changed and with that the distribution of the training data. In attempt to alter the amount of training data, we changed the threshold set with the HTK RO command for Slovenian systems from 100 to 267 and 350 and for English to 350 and 477. In this way we got five systems named sl-spe100, sl-spe267, sl-spe350, en-spe350 and en-spe477.

6. Speech recognition results

Six standard tests defined in the framework of the SpeechDat project (Johansen, 2000) were used on all reference and SPE based systems. These tests had the self-explanatory names: Yes/No test, Digits test, Connected Digits test, Application Words test, City Names test and Phonetic Rich Words test. In all tests but one (Connected Digits), each spoken test utterance consists of only one word. Therefore the word error rate (WER) is equal to the sentence error rate (SER) in these cases. Best results of tests done on all systems are given in table 6 and 7.

From these tables it can be observed that the SPE based systems performed either better or at least as good as the reference systems for both languages. The only exception was the Application Words test on the Slovenian systems. We should also take into consideration that Yes/No, Digits and Connected Digits tests only applied to a small part of the decision tree. Specifically, the vocabulary in these tests is very limited and only a small number of triphones are therefore used.

	sl-ref100	sl-spe100	sl-spe267	sl-spe350
Yes/no	0,63	0,63	0,63	0,63
Digits	3,85	3,85	3,85	3,30
Con. Digits	4,12	3,91	3,95	3,98
App. Words	3,20	3,38	3,74	3,38
City Names	7,65	8,16	7,14	7,14
Ph. R. Words	17,62	17,36	15,93	15,51

	en-ref100	en-spe350	en-spe350	en-spe477
Yes/no	0,00	0,00	0,00	0,00
Digits	3,98	3,98	2,84	2,84
Con. Digits	5,42	5,51	4,22	4,33
App. Words	3,53	3,72	3,72	3,53
City Names	6,21	6,21	7,91	6,21
Ph. R. Words	36,83	35,01	32,68	31,56

Table 6: Lowest WER achieved by the Slovenian and English speech recognition systems in all six tests

	sl-ref100	sl-spe100	sl-spe267	sl-spe350
Con. Digits	15,75	14,56	14,56	14,32

	en-ref100	en-ref350	en-spe350	en-spe477
Con. Digits	30,72	30,92	24,50	25,10

Table 7: Lowest SER achieved by the Slovenian and English speech recognition systems in Connected Digits test

Without doubt, the most reliable evaluation of the SPE based concept can be taken from the Phonetic Rich Words test, employing the largest vocabulary (1491 words for Slovenian and 3043 for English) and more than 710 utterances. This test involves a very big part of the decision tree. This test also gave us the biggest decrease of the WER when comparing the SPE based concepts with the reference systems. The results achieved on the English systems had even bigger impact on the WER. The difference in WER of the best reference system and the best SPE based system is for Slovenian 1,85% and for the English 3,45%. Also the SER achieved with the SPE based systems in the Connected Digits test is better than the one achieved with the reference systems. The impact is again much bigger for English.

7. Conclusions

Within bounds of our experimental set-up we observed an advantage to include the SPE theory as an expert linguistic knowledge into the speech recognition systems. In general we got better results with the SPE-based processing for the English systems than for Slovenian ones. Several possible reasons can be referenced for such behaviour. One is probably the definition of phone groups for the reference systems. There were 45 phone groups defined in the Slovenian reference system while only 17 in the corresponding English one. Therefore, the increase in the number of natural groups resulting from the inclusion of the SPE theory had bigger impact on the English systems than on the Slovenian ones. Another possible reason is the presence of noise. Pronunciations in the Slovenian database were recorded in much higher presence of noise than the English ones. This could potentially have reduced the distinctive ability of some of the features used in the SPE theory.

One possible reason for achieving much better WER for the Phonetic Rich Words test and SER for the Connected Digits test with the English SPE based systems could be the fact that the English reference systems had much bigger error rates than the Slovenian ones. The lowest WER in the Phonetic Rich Words test achieved by the Slovenian reference system was 17,62% whereas it in case of English was 36,83%. The same was observed for the SER in the Connected Digits test (English reference system: 30,72%, Slovenian reference system: 15,75%).

From our experiments, we also concluded that groups of phones should never include actual "unnatural groups" and leave it to the decision tree building process to disregard them in favour of the more natural groups. That would present no significant problem to the classification of triphones that do appear in training data but would lead to the incorrect classification of triphones with contexts that do not appear in the training set.

Based on the experimental evidence we have shown that the creation of the natural groups of phonemes by the SPE theory could effectively be used in defining phone groups for the multilingual speech recognition system including multilingual triphone Markov models. When porting the HLT technology to a new target language, this provides us a promising alternative to the more widespread approach of using the union of phone group definitions from all languages (Zgank et al., 2001).

8. References

Brøndsted, T., 1998. A SPE based Distinctive Feature Composition of the CMU Label Set in the TIMIT Database. *Technical Report IR 98-1001,* Center for PersonKommunikation, Institute of Electronic Systems, Aalborg University

Chomsky, Noam, and Halle, Morris, 1968. *The Sound Pattern of English.* Harper & Row, Publishers New York, Evanston, and London.

Johansen, F.T., N. Warakagoda, B. Lindberg, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, 2000. The COST 249 SpeechDat multilingual reference recogniser. *Paper for XL-DB.*

Lindberg, B., F.T. Johansen, N. Warakagoda, G. Lehtinen, Z. Kačič, A. Žgank, K. Elenius, and G. Salvi, 2000. A Noise Robust Multilingual Reference Recognizer Based on SpeechDat(II). *In Proc. ICSLP, International Conference on Spoken Language Processing,* Beijing,

Odell, J.J., 1995. *The Use of Context in Large Vocabulary Speech Recognition.* Dissertation submitted to the University of Cambridge for the degree of Doctor of Philosophy. Queens' College.

Šuštaršič, R., S. Komar, and B. Petek, 1999. *Illustrations of the IPA: Slovene.* Handbook of the International Phonetic Association: A Guide to the Use of the

International Phonetic Alphabet. Cambridge University Press, 135-139.

Toporišič, Jože, 2000. *Slovenska slovnica*. Maribor: Založba Obzorja.

Žgank, A., B. Imperl, F.T. Johansen, Z. Kačič, and B. Horvat. 2001. Crosslingual Speech Recognition with Multilingual Acoustic Models Based on Agglomerative and Tree-Based Triphone Clustering. *In Proc. EUROSPEECH, European Conference on Speech Communication and Technology*. Aalborg.

Young, Steve, Kershaw, Dan, Odell, Julian, Ollason, Dave, Valtchev, Vatcho, and Woodland, Phil. 2000. *The HTK Book (for HTK Version 3.0).* Cambridge: Entropic Cambridge Research Laboratory.

VIPTerm

The Virtual Terminology Information Point for the Dutch Language.

A Supranational project on terminology documentation and resources.

Prof. Dr. Frieda Steurs

Lessius Hogeschool
Dept. of Translators and Interpreters
Sint Andriesstraat 2
B-2000 Antwerp
e-mail : Frieda.Steurs@lessius-ho.be

1. VIPTerm

In 2001, the "Nederlandse Taalunie"(NTU) (Dutch Language Union) initiated a project to set up a virtual informationpoint for terminology (VIPTerm). This project can be considered as the Dutch part of the information and documentation requirements as stipulated by the TDCNet project.

The TDCNet project (European Terminology Documentation Centre Network) is an EU funded project(MLIS 4000 TDCNet 24264/0) with the main objective to create a virtual terminology directory in the form of a logical and physical network of terminology information and documentation centres in Europe.

Within this framework, both bibliographical data (data collections, literature, theses, etc.) and factual data (organisations, software, events, experts and training institutes) will be exchanged between national and regional information centres and compiled in an international terminology directory.

To reach this aim, the NTU set up a project to compile the data for both the Netherlands and Dutch speaking Belgium (Flanders).

VIPTerm (short for Virtual Terminology Information Point) will fullfil the function of a terminology institute, providing a documentation service and information point for users from different backgrounds.

The VIPTerm will also be designed to take up a function in the organisation of the terminologyfield and the networking in this field (a.o. by means of exchange, e.g. through a mailing list). This type of networking and fieldsupport is not the type of task the Nederlandse Taalunie wants to take care of through its own services. Tasks like these will have to be taken up by fieldorganisations, such as NL-TERM, the Dutch terminology association.

The main focus in the structure of this portal is to create an inventory and informationbase of organisations, events, activities , etc. that support a terminology policy for the Dutch language area. Next to this, in the future also the management, maintenance and distribution of electronic resources for Dutch, which is an important task of the NTU (cfr. Euromap and the Dutch Platform for language and speech research : TST (taal en spraaktechnologie), can be organised through this VIPTerm portal.

Both input and output formats were considered carefully in this project. By input format we refer to the actual database format that can be used to register the data.

The ISIS software has been known for some time as the central archiving system used by Unesco and other important datacollectors. The Winisis is a menu-driven generalised information storage and retrieval system designed specifically for the computational management of text-oriented data. Compared to ISIS, WinISIS has a Windows GUI.

The output format, on the other hand, has special requirements as well. We need to make the database available through a number of information points, a.o. the special interactive website of the NTU ('Taalunieversum'), the ETIS portal to TDCNet, and websites of other user groups such as NL-TERM, the Dutch terminology association.

In the output structure, the webportal, the following basic categories will be taken into consideration :

- general information and history (short outline of the agents in the field : NTU; Coterm, NL-TERM, and the policy concerning terminology)
- publications, literature
- termcollections
- events/projects

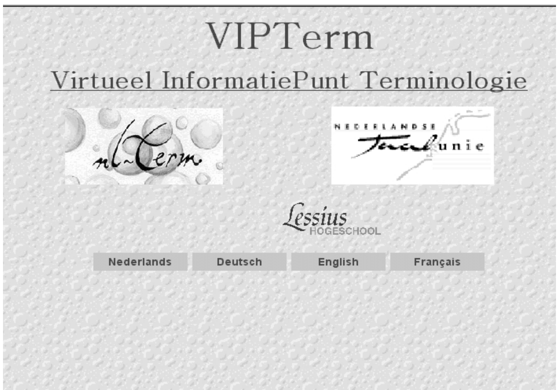
- training and education

- standardisation
- language technology and terminology management tools
- neoterm
- novelties
- international links

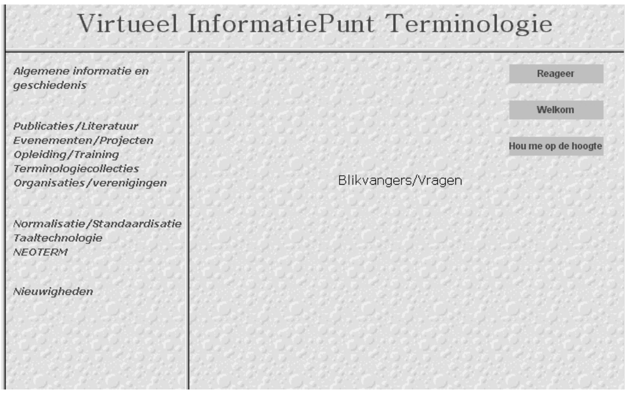
The project and pilotdesign was discussed thoroughly with our colleagues from the Fachhochschule Köln, who develop a similar project called DTP (Deutsches Terminologie Portal). We thoroughly investigated the classification of data (taking into account the existing classifications and TeDif format). We agreed on the principle to structure our portals in a similar way, so as to avoid unnessecary confusion.
Our sites will also be compatible with the ETIS server , as it was recently reprogrammed by the Union Latine.
For Dutch, an active exchange will be organised between the VIPTerm portal and ETIS, providing the data on Dutch terminology for the European level.

Once the analysis and study phase has been concluded, and advice has been collected through Coterm-experts and NL-TERM board members, we will build a sample portal site for this pilot project.
It will then be evaluated thouroughly and tested among a limited group of users.
If the final outcome of this evaluation is positive, then the VIPTerm project will be continued and will be organised on a more permanent basis.

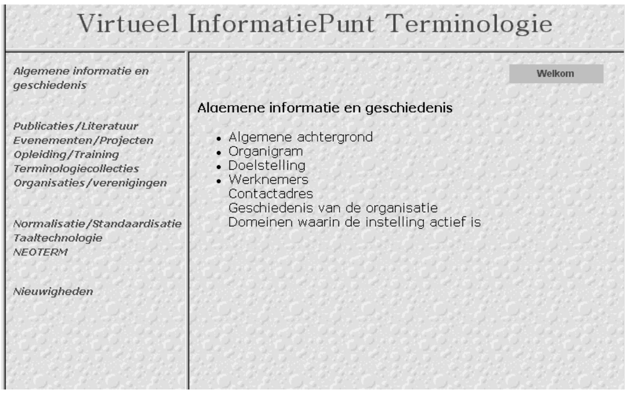
1.1. Prototype


VIPTerm

Virtueel InformatiePunt Terminologie

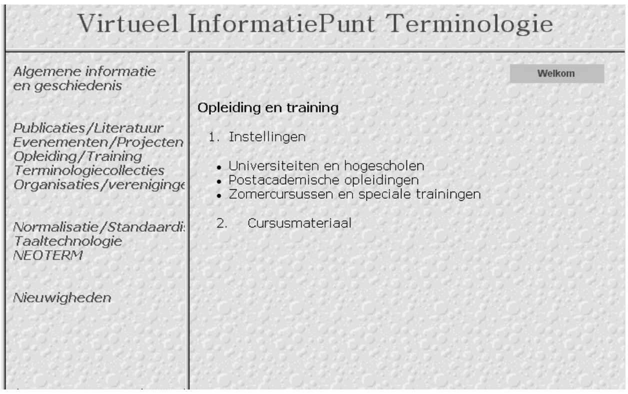

Virtueel InformatiePunt Terminologie

Blikvangers/Vragen


Virtueel InformatiePunt Terminologie

Algemene informatie en geschiedenis

- Algemene achtergrond
- Organigram
- Doelstelling
- Werknemers
  Contactadres
  Geschiedenis van de organisatie
  Domeinen waarin de instelling actief is

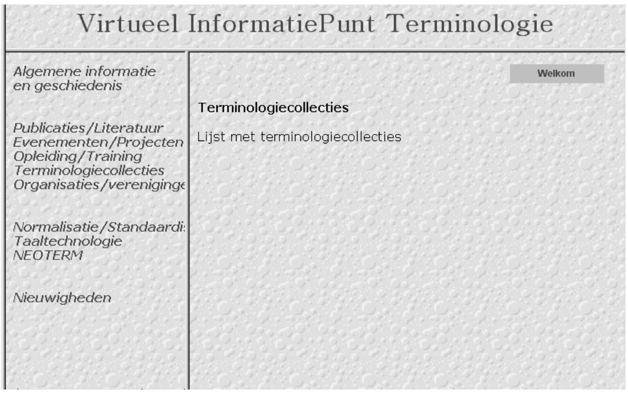

Virtueel InformatiePunt Terminologie

Opleiding en training

1. Instellingen

- Universiteiten en hogescholen
- Postacademische opleidingen
- Zomercursussen en speciale trainingen

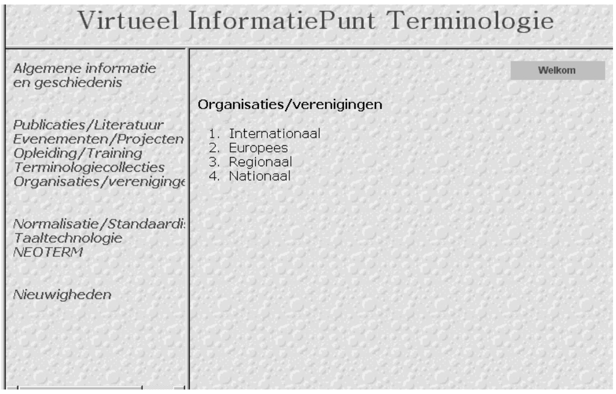
2. Cursusmateriaal


Virtueel InformatiePunt Terminologie

Terminologiecollecties

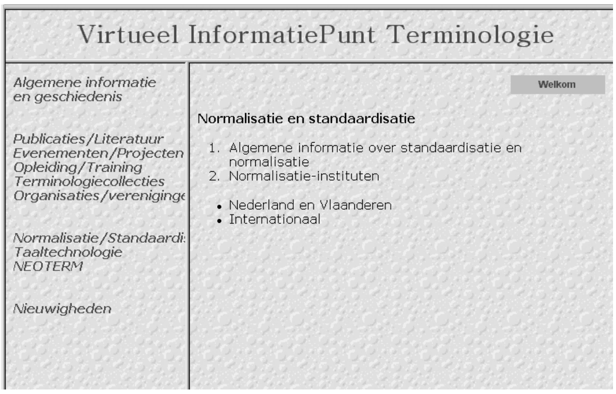
Lijst met terminologiecollecties

If the VIPTerm project and analogous projects such as DTP (Deutsches Terminologie Portal) are succesful and can be continued, a larger European platform for terminology is within reach and terminology awareness among experts, professionals and users will grow.



Organisaties/verenigingen

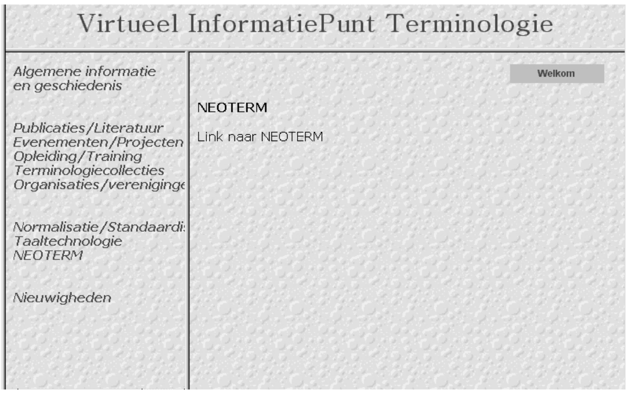
1. Internationaal
2. Europees
3. Regionaal
4. Nationaal



Normalisatie en standaardisatie

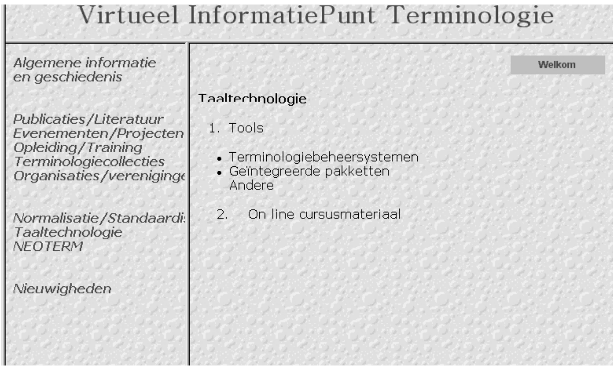
1. Algemene informatie over standaardisatie en normalisatie
2. Normalisatie-instituten

- Nederland en Vlaanderen
- Internationaal



NEOTERM

Link naar NEOTERM



Taaltechnologie

1. Tools

- Terminologiebeheersystemen
- Geïntegreerde pakketten
  Andere

2. On line cursusmateriaal