ISLE Computational Lexicon Working Group

APPENDIX

Resources for spoken language and multimodal lexica in multilingual contexts

Dafydd Gibbon

August 2002 (Version: November 24, 2002)

Contents

1 Introduction

1.1 Background considerations

The main contexts in which spoken language and multimodal lexica are relevant for multilingual contexts are in localisable speech technology systems (in automatic speech recognition and speech synthesis), in speech-to-speech translation, and in other less striking contexts such as the provision of pronunciation information for each language in bilingual or multilingual dictionaries.

The difficult part of multilingual lexicon development for spoken language lies in the coordination of the corpus vocabularies for the languages concerned.

First, spoken language system development uses relatively small corpora of transcriptions (perhaps up to several hundred thousand words, yielding a lexicon of several tens of thousands of words, depending on application type). These corpora are expensive and very labour-intensive to make, with real-time factors of between 50 and 500 to transcribe and annotate, again depending on specification. A mere hour of speech, say 20 pages of relatively close-typed transcription, would therefore take between about 1 and 12 weeks to process as a resource for lexicon acquisition, depending on the lexicon requirements specification.

Second, in a speech-to-speech translation lexicon the scenario constrained corpus lexicon requirement which invariably has to be met for spoken language system lexica is over-strained by the need to translate from a given corpus lexicon into a translation-generated lexicon in the target language which is by definition not a corpus lexicon, and to process these lexical entries in the target language. The need to process items which are not in the corpus lexicon but need to be accounted for quasi-compositionally is known as the *Out Of Vocabulary* (OOV) problem; this problem is compounded by the translation situation.

The path towards a solution to these problems is more general than these specific examples might suggest. Consequently, the present contribution presents a preliminary clarification and systematisation of resources for spoken language lexica with a view to developing standards and resources in this area, and builds on a number of previously collated sources of information. The basic sources are [9], [12].

1.2 Spoken language lexicography

A central problem in discussing resources for spoken language lexica emerges from the fact that there is no unified notion of lexicography for spoken language, and therefore no relatively homogeneous guild of lexicographers as there is for written language. Many disciplines, independently of each other, manufacture spoken language lexica. The reason for this lexicographic inhomogeneity lies in the wide range of uses for which lexical information on spoken language is required, some of which are listed here:

- General lexica:
  - Transcription of pronunciation information as a data category in written language lexica,
  - Pronunciation lexica (orthographic wordlists with phonemic transcriptions).
  - Rhyming lexica,
  - Wordlists, glossaries, and lexica for unwritten languages;
- Machine lexica for human use:

  - Transcription and audio output for pronunciation in hyperlexica,
  - Audio and video concordances (wordlists, pre-compiled or generated on-the-fly) mapped to time-stamps in audio and video recording files);

- Machine lexica for written language systems:
  - Transcriptions of pronunciation for the pronunciation data category,
  - Spell check algorithms with pronunciation constraints,
  - "Phonetic search" wordlists with functions defining phonetic similarity via algorithms such as *Levenshtein distance* or *soundex*;

- Machine lexica for spoken language systems (currently always scenario constrained and corpus based):
  - orthography-pronunciation mapping for text-to-speech lexica in speech synthesis,
  - lexical search and mapping to orthography from output of decoder component in speech recognition in conjunction with a language model,
  - translation lexica for speech-to-speech translation;
  - resource lexica for generating optimised lexica;
  - stochastic language model (essentially a wordlist with statistical constraints on contexts of cooccurrence, with n-gram, regular grammar or context-free grammar structures.

This heterogeneity makes it somewhat difficult to integrate the requirements for spoken language lexical resources into the generic ISLE framework without considerable backtracking into the basics of computational spoken language lexicography.

1.3 Overview

The following sections deal with spoken languge and multimodal lexica, types of lexical information, Transcrption in corpora and lexica, corpus and lexicon annotation, and formal prerequisites for spoken language lexicon implementation, followed by conclusion and prospects.

2 Spoken language and multimodal lexica

We may start with a simple informal definition, assuming understanding of the basic concepts; the definition will be expanded below:

Lexicon: *a 4-column table of the non-compositional signs of a language and their category, constituent, semantic, and surface properties.*

Spoken languages differ from written language lexica, in structure, content and use (in the following discussion, a broader view of spoken language system lexica will be taken than is usual in the spoken langugage technologies).

In structure, spoken language lexica, particularly those constructed for use in spoken language systems, differ from written language lexica in several ways. The most important way is perhaps the need to link lexical entries, via time-stamps, to occurrences in corpora, whether for the training of statistical decoder models or for the construction of audio concordances.

In content, spoken language lexica require information of varying detail about pronunciation, and about the use of lexical items in dialogue contexts and differing pragmatic situations, as well as statistical information.

In use, spoken language lexica have a different deployment spectrum from written language lexica, particularly in the spoken language technologies, as already outlined.

Relatively few of the world's approximately 7000 languages are written languages, and construction of lexica for purely spoken, i.e. unwritten languages is perhaps the major single task in descriptive linguistics, A large part of the task is taken up with representing segmental (phonemic) and suprasegmental (prosodic) pronunciation information in the lexicon, and with mapping this to more detailed phonetic representations of pronunciation in actual utterances. In addition, spoken dialogue contexts require the differentiation of different semantic and pragmatic vocabulary fields for representation in the lexicon. Evidently, unwritten languages are not intrinsically less complex than written languages: they are supported by complex oral traditions with orally transmitted legal and religious systems, and sophisticated orature (oral "literature").

It is now widely recognised that spoken language is multimodal and not restricted to the acoustic-auditory modality, implying that spoken language systems have to consider "body language" components, including the

1. gestural (movements of head, face and limbs),
2. postural (configuration of body), and
3. proximal (interlocutor distance)

components of communication, as well as the more well-known locutionary components (though the latter are presumably the most complex by orders of magnitude). Until recently, these components have been investigated separately in different disciplines, from choreography through anthropological linguistics to the study of the complex sign languages used by acoustically handicapped.

Indirect confirmation of this generalisation of the definition of speech from the acoustic-auditory modality to multimodal communication is provided by the numerous contributions on multimodal spech in the recent events in the LREC and EUROSPEECH conference series; see also [8].

The lexical information required for multimodal speech will therefore be accommodated in the model developed here.

2.1 A note on spoken language genres

Spoken language lexica for system use are almost invariably scenario constrained corpus lexica, while spoken language lexica for direct human use are invariably general language lexica. Scenario constraints correspond largely to the criteria used in the traditional characterisation of registers, genres and sublanguages. The range of these spoken language registers, genres and sublanguages is wide, and beyond the scope of this study; it will be sufficient for present purposes to refer to previous traditional studies of genre, register and sublanguages, cf. [6].

In the contexts of the speech sciences, psychology, and spoken language engineering, the most commonly used genre of spoken language is *read speech*. In anthropological and descriptive linguistics, as well as in ethnolomethodology, conversation analysis and discourse analysis, spontaneous types of dialogue are focussed on. Increasingly, this is also the case for the spoken language technologies, under the influence of development requirements of producing interfaces for natural human-machine interaction.

2.2 Basic terminology for spoken language lexica

The basic terminology used in this contribution follows and systematises the usage in previous work in this field, and as far as possible is kept compatible with the work of the ISLE Computational Lexicon Working Group.

Corpus: a quadruple $< metadata, corpusdata, annotations, corpuslexicon >$.

Modality: a pair of human output (motor gesture) and input (sensory) channels such as $< acoustic, auditory >$ (e.g. speech), $< gestural, visual >$ (e.g. gesturing, signing), $< gestural, auditory >$ (with gestures transduced into sound, e.g. hand-clapping), $< gestural, tactile >$ (e.g. shoulder-slapping, kissing). In phonetics, speech is also commonly regarded as a specific kind of $< gesture, auditory >$ modality in which the gestures are restricted to the vocal tract and transduced into sound. Analogously, orthography is a $< gesture, visual >$ modality in which gestures are transduced into stored traces (inscriptions).

Submodality: An autonomously organised stream of intonation which modulates a modality (required for representing parallel streams of information such as prosody).

Corpus data: collection of (generally digital) audio/video/sensor signal recordings and/or transcriptions of spoken language utterances or hardcopy/scans and/or discrete electronic versions of written language inscriptions.

Transcription: a symbolic representation of corpus data.

Annotation: the enhancement of

- a transcription by time-stamps pointing to boundaries or segments in corpus data recordings (labelling); formally, a pair $< label, timestamp >$, where $timestamp$ can be a $point$ or an $interval$, the $interval$ generally being represented by a pair of $point$ timestamps $< point_i, point_{i+1} >$
- written language corpus data by a function mapping descriptive categories into boundaries or segments in the corpus data (tagging, tree-banking).

Corpus lexicon: a set of lexical items (words, idioms) induced from the corpus by the following functions,

- sorting,
- removing duplicates,
- (optionally) stemming, i.e. removing affixes,

- (optionally) lemmatising, i.e. extracting stems as lemmata.

and mapped into a set of types of lexical information.

Standard lexicon model: a model of a lexicon as a table with atomic cell entries.

Lexical entry: a row in a standard lexicon model representing a vector of lexical information of different types (corresponding to the traditional “lexicon article”).

Lexical data category: a column in a standard lexicon model representing a type of lexical information contained in lexical entries.

Microstructure: a vector of data categories comprising the types of lexical representation represented in lexical entries.

Spoken Language Reference Microstructure (ISLE-SLRM) model: a recursively structured vector of data categories (ignoring housekeeping and entry metadata):

Standard model:
$< STRUCTURE, INTERPRETATION >$

Standard model components:
$STRUCTURE$ is a pair $< CATEGORY, PARTS >$ and $INTERPRETATION$ is a pair $< MEANING, SURFACE >$. $SURFACE$ is, in turn, a pair $< MODALITY_{phonetic}, MODALITY_{visual} >$

Additionally, a category for corpus frequency information of different types (isolated frequency, frequency in various contexts such as digrams) is needed.

The grouping of data categories is derived from the ILEX model [7] and are closely related to the data category specifications of the ISLE Computational Lexicon Working Group, but extended for application to spoken and multimodal lexica.

The $CATEGORY$ attribute, in a spoken language lexicon, is very often a statistical function relating co-occurring neighbours in a corpus, but it may also be a function from the lexicon into the corpus which effectively defines a (pre-compiled or on-the-fly) concordance. If the corpus is purely textual the concordance is conventional. However, if the pointers are time-stamps relating to an audio or video signal, the concordance is a multi-media concordance (audio and/or video concordance) for human use or statistical system training,

For ease of comparison a full version of the ISLE-SLRM model is given as a feature structure in Figure 1.

This schema stands for a family of reduced or expanded microstructures in practical instantiations of the model, which depend on actual requirements in specific applications, and represent special cases of the ISLE-SLRM reference model.

For example, the model proposed by Bell & Bird [1] for lexicon metadata definition is a triple corresponding to a restricted instantiation of the triple $< STRUCTURE, MEANING, SURFACE >$.

And a descriptive linguistic glossary would be adequately modelled by the pair $< meaning, surface >$, with $meaning$ modelled by a gloss in the description language, and $surface$ modelled by a phonemic transcription in the source language.

A lexicon in a theoretical linguistic framework, on the other hand, such as the HPSG paradigm, requires a much fuller spelling out of the $STRUCTURE$ and $MEANING$ attributes.

$$\begin{bmatrix} \text{STRUCTURE =} & \begin{bmatrix} \text{CATEGORY =} & \text{...} \\ \text{PARTS =} & \text{...} \end{bmatrix} \\ \text{INTERPRETATION =} & \begin{bmatrix} \text{MEANING =} & \text{...} & \\ \text{SURFACE =} & \begin{bmatrix} \text{MODALITY}_1 \text{ =} & \text{...} \\ \text{... =} & \text{...} \\ \text{MODALITY}_i \text{ =} & \text{...} \\ \text{... =} & \text{...} \\ \text{MODALITY}_n \text{ =} & \text{...} \end{bmatrix} \end{bmatrix} \\ \text{FREQUENCY =} & \text{...} \end{bmatrix}$$

Figure 1: *Grouped representation of microstructure vector in the ISLE-SLRM model as an attribute-value structure. The modalities are* $orthography$, $phonetic$, $gestural$, *etc. (see text). The PARTS correspond to the Daughters attribute of HPSG-like grammars.*

Macrostructure: an ordering relation over the rows in a lexical relation table as an search-oriented data structure optimising operation, such as an alphabetic sorting by the orthography, a hyponymy-based tree relation induced over over senses of lexical entries, as in a thesaurus,

Mesostructure: a classification of lexical entries, and an ordering over these classes as a generalisation operation, induced from data category values of lexical entries, and represented as grammatical information in the front matter of a lexicon, or as an inheritance hierarchy in a formal lexicon.

Semasiological lexicon: a lexicon with a macrostructure optimised for search as a function from stems to sense representations.

Onomasiological lexicon: a lexicon with a macrostructure optimised for search as a function from sense representations to surface forms.

3 Types of lexical information

3.1 Generic lexical information

A generic model of syntactic and semantic types of lexical information for spoken language will for the most part correspond to the ISLE CLWG model for written language, and therefore need not be spelled out in this section.

3.2 Spoken language specific lexical information

The following are the most important types of lexical information which differ either gradually or categorically from the types of lexical information included in written language lexica:

1. Pronunciation representation in human readable lexica:
    - (a) non-standard adapted orthographies,
    - (b) International Phonetic Alphabet,
    - (c) Alphabets similar to the IPA or encoding the IPA in typewriter-friendly ASCII codes, e.g. the SAMPA alphabet [8].
2. Pronunciation representation in machine readable lexica (this systematisation goes beyond conventional descriptions in the disciplines concerned and may appear unconventional in these contexts):
    - (a) Statistical characterisation of components of pronunciation such as phonemes, diphones, disyllables and larger units by means of stochastic models such as Hidden Markov Models (HMMs), Neural Networks (NNs), Bayesian Networks;
    - (b) Statistical characterisation of distributional properties of functional items such as morphemes and words by means of stochastic models of cooccurrence in corpora (digram, n-gram models, regular grammars (equivalently: finite state automata), context-free grammars).
3. Gesture (as captured in current gesture transcription systems):
    - (a) HamNoSys: a gesture representation for sign languages,
    - (b) FORM: a gesture representation for automatic gesture recognition,
    - (c) CoGest: a linguistically motivated representation of gesture forms and functions.
4. Spoken language specific units corresponding to and generalising the notion of Part of Speech (POS), particularly
    - (a) Interjections,
    - (b) Clitic unit formation,
    - (c) Functional units,
    - (d) Stochastic Regular & CF Language Models.

4 Transcription in corpora and lexica

The issues of corpus and lexicon transcription, including prosodic transcription, were dealt with in detail in the SAM and EAGLES I and II projects [9], [8], and need only be summarised here.

4.1 Transcription in corpus representation

The kind of transcription used in corpus representations is highly variable and dependent on scenario constraints. The following tentative scale of transcriptions is proposed for corpus representations:

1-tier transcription: Minimally, a phonemic transcription or an orthographic transcription which is regularly related to pronunciation. The latter is perhaps the most widely used form of transcription, both in corpora for spoken language systems, as well as in written language corpora if alphabetic or syllabic orthographies are used. Logographic orthographies are unsuitable unless a function which maps them into a some pronunciation notation is provided.

2-tier transcription: Preferably, for minority language corpora and especially corpora of endangered language data aligned parallel transcriptions of forms (minimal transcription) and of functional categories (e.g. glosses in a standard description language such as English, French, Russian, Spanish, ...). This kind of transcription is generally referred to as an *interlinear gloss* in descriptive linguistics.

n**-tier (multi-tier) transcription:** In addition to a 2-tier transcription further aligned tiers with other lexically relevant information, such as:

- non-pronunciation-friendly orthographies,
- prosodic categories,
- morphosyntactic categories,
- semantic categories,
- pragmatic categories.

Any or all of the data categories in the ISLE CLWG microstructure definition can be expected in an annotation tier, in addition to corpus-specific tiers.

4.2 Transcription in lexical representation

In addition to any orthographic representation, a transcription system for the instantiation of a Data Category in the microstructure of a lexicon is by definition a *lexical* or *phonological underlying* transcription, not a *phonetic* transcription, which has the function of capturing phonetic detail of the pronunciation of sounds in context. For lexical transcription, the following levels are appropriate, depending on the typology of the language (no distinction will be made here between *prosodic* and *suprasegmental*, as distinctions made in the literature are generally idiosyncratic to a particular methodology):

Segmental: Depending on the degree of abstraction in the phonetic and phonological analysis involved in word identification, the following levels may be used:

1. Morphophonological or phonotypic: phoneme variation in uniquely identifiable morphological contexts are not marked but generated by rule; e.g. German orthography is morphophonemic Mond /mo:nt/ ‘moon’ - Monde /mo:ndə/ ‘moons’. For a monolingual lexicon meant for native speakers, who have internalised the syllabification and final devoicing rules, or computational models which have explicit syllabification and final devoicing rules, no distinction is necessary,
2. Phonemic: phonetic variation between the alloophones of phonemes within a word (such as the approximant and vocalised allophones of /r/ or the aspiration of voiceless plosives in English), without consideration of word-internal morphological information,
3. Phonetic: variations in phonemic or phonetic detail which may be required for capturing phonostylistic variation of pronunciation in different styles and registers.

The mesostructure of a spoken language lexicon (the front matter of a book lexicon, the rules or inheritance hierarchies of a formal lexicon) contains generalisations to supplement the explicit information in lexical entries:

1. *Morphophonemic rules* for specifying phoneme variation in morphophonemic contexts,
2. *Allophonic rules* for specifying allophone variation in phonemic contexts,
3. *Phonetic detail* rules for specifying dialectal and sociolectal variation,
4. *Supralexical rules* (traditionally known as ”postlexical rules” in Generative Phonology and its descendants) for specifying pronunciation variation in larger syntactic and phonostylistic contexts.

Suprasegmental (prosodic): In languages which have lexical prosody such as word stress or tone, with various distinctive and morphological functions, provision must be made for representing this. The main kinds of lexical prosody are the following:

stress, as in Dutch, English, German;

pitch accent, as in Japanese;

tonal accent, as in Swedish;

register tone, as in very many African languages;

contour tone, as in very many South East Asian languages.

The main lexical and morphological functions of lexical prosody are, illustrated with German examples (for clarity: stress indicated by upper case, standard initial upper case characters not used):

distinctive: *TEnor* ‘tenor’ (singer) - *teNOR* ‘tenor’ (gist)

inflectional: *DOKtor* ‘doctor’ - *dokTORen* ‘doctors’

derivational: *TElefon* ‘telephone’ (noun) - *telefoNIEren* ‘telephone’ (verb)

compounding: *Übersetzen* ‘cross over’ (verb), *überSETZen* ’translate’ (verb)

These functions occur with all types of lexical prosodic forms. Further relatively standardised information about prosodic transcription, with examples from 20 languages, can be found in [10].

4.3 Implementation of transcriptions

4.3.1 Criterial levels

The following sections pertain mainly to the implementation of lexica for human use, and not primarily to the implementation of system lexica, which are very product-specific.

In implementing transcription systems, at least the following four-way level distinction is needed:

1. *transcription category* (e.g. a voiceless alveolar plosive, a high tone),
2. *transcription id* (e.g. the code numbers in the Esling coding of the IPA, or the SAMPA ASCII encoding of the International Phonetic Alphabet, cf. [8], or in the Unicode conventions),
3. *transcription symbol* (e.g. /t/ for a voiceless alveolar plosive phoneme, ´ for a high tone),
4. *transcription font* (e.g. the actual glyph design for particular symbols) with properties such as *serif* or *non-serif.*

4.3.2 Transcription category

The level of transcription category is rather stable for segmental transcriptions, and is detailed in the Handbook of the International Phonetic Association [11] and in practically any leading phonetics textbook such as [5].

For prosodic transcriptions this is not the case. The IPA provides a set of categories for prosodic transcription, but other sets are in use.

The variation is too great to be detailed here, but several examples can be found in [10].

4.3.3 Transcription id

The level of transcription id has not completely stabilised, as there are a number of codings in current use, as already mentioned. It may be supposed that the Unicode conventions will be adopted as soon as adequate rendering engines for Unicode become available, which is currently not the case. For this reason, other encodings are commonly used.

4.3.4 Transcription symbol

At the level of symbol choice, the situation is rather similar, since the most symbols in current use have been fixed by the International Phonetic Association since the late 19th century. There is still some variation in different traditions, however, for example between /y/ in one US phonetic tradition, and /j/ in European phonetic traditions, for a palatal approximant, Although this distinction is rather trivial, nevertheless it can lead to misunderstandings. Consequently the recommendations of the International Phonetic Association are preferred.

In spoken language technologies and in general lexicographic practice conventions such as accent diacritics over words for stress or tones, or capitalisation of stressed syllables, are found.

4.3.5 Transcription fonts

There are many fonts available for encoding segmental symbols, from common orthographic fonts to specific implementations of the IPA symbols. Font implementations are very platform-specific, and easy font conversions are not possible.

Simplifying away from a number of issues which are not immediately relevant for present purposes, there are two basic font technologies which affect the usability and interoperability of fonts, word processor outline font technology such as TTF, and the Metafont glyph function technology used by TEX systems, mainly under UNIX.

The most well-known word processor IPA fonts implementations are:

- IPAkiel, based on the IPA chart as defined at the Kiel Convention in 1993.
- SIL (Summer School of Linguistics) fonts, of which there are many, perhaps the most well known being the Doulos series.
- Lucida, which maps into Unicode, and will possibly be the preferred font for this purpose when full Unicode rendering engines are available (current word processors have extensive but still partial Unicode implementations).

For the WYSIWYG oriented fonts, even with similar implementations, the functions which map ids to symbols and their glyphs, and which map keyboard combinations to ids and thence to symbols and their glyphs) vary greatly from one font implementation to another.

The most well-known implementations of IPA fonts for LATEX are:

- WSUIPA from Washington State University,
- TIPA from University of Tokyo.

5 Corpus and lexicon annotation

5.1 A definition re-visited

As defined in the glossary, an annotation in the context of spoken language context is:

Annotation: the enhancement of

- a transcription by time-stamps pointing to boundaries or segments in corpus data recordings (labelling); formally, a pair $< label, timestamp >$, where $timestamp$ can be a $point$ or an $interval$, the $interval$ generally being represented by a pair of $point$ timestamps $< point_i, point_{i+1} >$

The theoretical foundations of this concept of annotation are due to event logic based phonologies, as in the Event Phonology of Bird & Klein, cf. [2]. The Time Map theory of Carson-Berndsen [4] introduced extensions to phonetic levels and applications to spoken language processing, with finite state models for the event logic theory.

Annotation of spoken language data has been extensively dealt with, also in previous European project work, cf. [9] and [8]. Currently the most influential approach is by Bird & Liberman [3], who generalised the notion of annotation by means of the construct *Annotation Graph* in order to encompass known kinds of signal and text annotation.

Strictly speaking, the inclusion relation

$$Transcription \subset Annotation$$

holds, since maximally the beginning and end of a transcription are informally synchronised with the beginning and end of the speech signal, though this is not a very useful idea as the synchronisation is too fuzzy in general to be machine processable, and in any given case a signal recording may not actually exist.

5.2 Acquisition of spoken language lexical information

Work in the spoken language technologies, and modern hyperlexicon applications such as audio and video concordances, presuppose the availability of carefully annotated spoken language data. Many tools for automatically, semi-automatically and manually annotating audio signal data are available. Of primary importance are the manual tools, since the final quality criterion for accuracy (not necessarily consistency!) of annotations is the human annotator.

There are many proprietary and locally developed and used tools for audio annotation, and not many at all for video annotation. These developments are not available for standardisation, and have not developed into state of the art freeware or open source tools, and consequently, they will not be considered here.

The most widespread tools for this purpose are currently the following:

1. Praat, a comprehensive freeware toolset for annotation and experimentation in phonetics and speech technology, developed since the early 1990s at the University of Amsterdam phonetics laboratory by Paul Boersma and David Weenink.

2. esps/waves+ (”Xwaves”), a proprietary library and GUI developed by Entropic in Cambridge, UK, and not maintained or generally available since its purchase by Microsoft Corp. a number of years ago (though a change in this policy is apparently under discussion).
3. TASX, an open source workbench for video and audio annotation developed by Jan-Torsten Milde at Universität Bielefeld and implemented in Java.

Currently the most widely used audio tool, in research, development, resource creation and teaching environments is Praat.

In comparison with lexicon acquisition work on written language corpora, spoken language lexicon acquisition is relatively impoverished. Although the automatic construction of stochastic models for speech recognition is very highly developed, this is not true of the analysis of collocations, vocabulary fields and related activities characteristic of corpus linguistics and of natural language processing in general.

In addition to these activities, three areas of generalisation over corpora are on the horizon, represented by a scattering of studies, and it may be expected that these will increase in importance as spoken language input and output devices become more widely used and adaptability and portability requirements increase in importance:

1. Resource adaptation in spoken language technology, for example for permitting a generic speech recognition or speech synthesis application to be used by a wider range of users.
2. Annotation graph collation by fuzzy operations over near-simultaneous points and overlapping intervals in time.
3. Syntagmatic hierarchy induction in order to create phonotactic, morphotactic and phrasal grammars from data.
4. Paradigmatic induction of class hierarchies over lexical items for use in compact and robust inheritance hierarchies.

Effectively, these are Machine Learning (ML) applications which are gradually being transferred from other areas of language processing to spoken language, and will not only make the re-usability and interoperability of spoken language resources more feasible but will also enable resources to be related to and benefit from theoretical linguistics, and vice versa.

Finally, it may be noted that “lexicon acquisition” in general relates to the instantiation of pre-defined microstructures from corpus data. The notion may be generalised, however, to the process of defining lexicon microstructures by means of generalisation and disjunctive abbreviation procedures (e.g. for alternatives in lexical fields).

6 Formal prerequisites for spoken language lexicon implementation

6.1 Towards a generic lexical model for spoken language

Traditional lexicogaphy is very much a practical art, and many features of this art have been transferred to computational lexicography.

The concepts of microstructure (especially the ISLE-SLRM model of lexicon microstructure), macrostructure and mesostructure were introduced earlier as a basic generic starting point for defining a wide range of varieties of spoken language lexicon more precisely, and for integrating the notion of spoken language lexicon into the generic approach adopted by the ISLE Computational Lexicon Working Group.

Summarising the discussion of previous sections, the structure of the basic reference lexicon is thus a triple:

$$< microstructure, \ macrostructure, \ mesostructure >$$

The microstructure, which corresponds to the notion of *lexicon model* in most previous computational lexicographic work, defines the structure of the vector which represents specific lexical entries (LI means ”lexical information” and TLI means ”type of lexical information”, usually expressed as attribute-value pairs):

$$< LI_{TLI_1}, \ \ldots \ , \ LI_{TLI_n} >$$

In this reference model, microstructures are intended to be fully inflated, with no disjunctions or substructures of the kind to be found in conventional alphabetic semasiological dictionaries. The reason for this is to clarify the fact that disjunctions and tree-structured lexical entries implicitly express generalisations over more primitive structures.

The generic ISLE-SLRM model is mapped on to specific lexica by means of the following generalisation operations which define mesostructures and the macrostructures of the specific lexica:

1. A grouping operation for local alternatives in pronunciation in spelling data categories which are unrelated to other data categories (analogous to the grouping of readings or senses). A simple example is the /aɪðə- iðə/ pronunciation variation of English *either*.

2. A distributed disjunction operation (due to Krieger & Nerbonne), for example for relating linked local alternatives in pronunciation which are which are linked to alternatives in spelling (or other data categories), or in specifying morphological syncretisms. This operation is usually represented in lexicography by postulating separate lemmata when the disjunctions include syntactic or semantic categories. However, since each primitive lexical entry in the ISLE-SLRM model is inherently separate from all others, lemma groupings are by default left over after the groupings of alternatives and distributed disjunction operations have been applied.

3. Abstraction of distributed disjunctions into a type or default class hierarchy.

4. Definition of macrostructure ordering relations over lexical entries based on relations between the fields of lexical entries. Macrostructures which differ from the basic table structure are generally application specific, and defined for optimisation in lexical search, whether by human or machine. Macrostructures therefore, contrast with other generalisations, which may be said to be *declarative*, in that they have a *procedural* motivation. Examples of macrostructure orderings are:

- Onomasiological ordering: Tree graph induced over sense terms and representing a hyponym-hyperonym taxonomy or a meronymy (as in a thesaurus).
- Semasiological ordering: Alphabetical ordering over the orthography field (classically, ”the dictionary”).
- Unnamed ordering: Alphabetical ordering over reversed pronunciation fields (rhyming lexica).
- Rank ordering for defining idiom dictionaries as against word dictionaries or morpheme dictionaries, etc.
- Selection by sublanguage field for technical and terms and other collocationally restricted items for specialised sublangage dictionaries.
- Speech recognition ordering: over representations of pronunciation.
- Speech synthesis ordering (text-to-speech): as semasiological ordering, except in general with highly reduced microstructures.
- Speech synthesis ordering (concept-to-speech): as onomasiological ordering, but in principle with no intervening orthographic representation.

In traditional lexicography, macrostructure orderings such as these define different specific book lexica. In the context of lexical databases, these macrostructure ordering relations are implemented as database views. In hyperlexica, the macrostructures are defined as alternative superimposed hyperlink structures.

Any resource archive format will need to be at least “virtually” re-constitutable into the ISLE-SLRM format in order to be able to map this on to the different microstructures required in speech technology and human readable dictionary publication.

7 Conclusions and prospects

7.1 Integration into multilingual contexts

It has been demonstrated that by returning to some basic concepts the apparent heterogeneity in spoken language lexicography can be reduced to a fairly straightforward model, although the realisation of spoken language lexica themselves is anything but simple. The ISLE-SLRM (Spoken Language Reference Microstructure) model, like generalised models of lexical entries used in some theoretical linguistic frameworks, has the property that notions of headword and lemma are not basic, but the result of the application of operations of macrostructure optimisation and mesostructural generalisation.

The results of procedure followed in this contribution are thus similar in spirit to the results outlined by Bell & Bird [1], except that

- they based their study on an examination of 55 printed lexica, whereas this contribution is based on varied practical experience in manufacturing complex spoken language lexica,
- their abstract data model is more restricted, and turns out to be a special case of the ISLE-SLRM model,
- they do not systematically consider macrostructural optimisations or mesostructural generalisations of the lexicon (though one of their main points is that lexica are extremely inhomogeneious).

There has been no explicit discussion of metadata in relation to the ISLE-SLRM model because of the need to clarify requirements and design issues before moving on to specific metadata proposals for implementing archives and dissemination portals. However, the ISLE-SLRM model has been designed in such a way as to assist metadata design, and it is proposed that the ISLE-SLRM model, with the associated concepts of macrostructural optimisation operations and mesostructural generalisation operations is a suitable foundation for the definition of lexical metadata for spoken language lexica.

7.2 Realisation with XML technologies

The formal character of an attribute-value structure is common to linguistic feature structures, to the ISLE-SLRM model, and to XML tree structures, facilitating portability of structures from one methodology to another.

In the context of spoken language lexicography, a number of theoretical problems remain with XML, though in one way or another the structures discussed here may be represented in XML. The following points are relevant for the generation of the varieties of macrostructure required for different kinds of spoken language lexicon:

1. XML has no well-defined formal semantics beyond the assignment of tree-graphs.
2. For other abstract data structures, *ad hoc* definitions are required.
3. A specific example of an *ad hoc* solution to a well-known problem is the case of tables (familiar from the LaTeX and HTML table models): if the XML tree is row-based, there is no well-defined concept of column; the column is stipulated in an informal semantics for the tree. The same holds

vice versa. If it is accepted that XML is specifiable by context-free rules, the proof of this is obvious: a table has the structure $a^n b^n c^n$..., which is clearly not context-free but context-sensitive (in fact, an indexed language). An actual implementation for the purpose of system use in spoken language technologies (or browser construction) must take this into consideration.

4. The issue becomes correspondingly more complex with recursive tables.

5. Another specific example of *ad hoc* solutions, from the point of view of the syntax of XML, is the enhancement of XML with pointers. The internal syntax of pointers is tree-structured, but their semantics is that of variables with arbitrary values over positions in documents, which may be used to construct structures of arbitrary complexity.

6. it is unlikely that the transformations required by spoken language system technologies will be amenable to the procedural components of the XML technologies such as XSLT.

These points pertain to the projection of the ISLE-SLRM and will no doubt be resolved when an adequate semantics for XML is available. Currently it is sufficient to note them for future discussion and to illustrate the necessity for pragmatic strategies of operationalising computational lexica for spoken language.

References

[1] John Bell and Steven Bird. A preliminary study of the structure of lexicon entries. LDC, Philadelphia, 2000.

[2] Steven Bird. *Computational Phonology: A Constraint-Based Approach*. Cambridge University Press, Cambridge, 1995.

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

[4] Julie Carson-Berndsen. *Time Map Phonology: Finite State Models and Event Logics in Speech Recognition.* Kluwer Academic Publishers, Dordrecht, 1998.

[5] John Clark and Colin Yallop. *Introduction To Phonetics And Phonology, 2nd edn.* Blackwell, Oxford, 1995.

[6] Penelope Eckert and John R. Rickford, editors. *Style and Sociolinguistic Variation.* Cambridge University Press, Cambridge, 2002.

[7] Dafydd Gibbon. Compositionality in the English inheritance lexicon: English nouns. In Leila Behrens and Dietmar Zaefferer, editors, *The Lexicon in Focus*, pages 145–185. Lang, Frankfurt a. Main, 2001.

[8] Dafydd Gibbon, Inge Mertins, and Roger Moore. *Handbook of Multimodal and Spoken Dialogue Systems: Resources, Terminology and Product Evaluation.* Kluwer Academic Publishers, Dordrecht, 2000.

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

[10] Daniel Hirst and Albert Di Cristo, editors. *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, Cambridge, 1998.

[11] I.P.A. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Blackwell, Oxford, 1999.

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