

Papers from the Fourth International Conference on Sinology

HUMAN LANGUAGE RESOURCES AND  
LINGUISTIC TYPOLOGY

Human Language Resources:  
Their Role in Research, Development  
and Application

**Dafydd Gibbon**

Universität Bielefeld and COCOSDA

Academia Sinica

# Human Language Resources: Their Role in Research, Development and Application

Dafydd Gibbon

Universität Bielefeld and COCODA

A key issue in the language sciences and technologies is the provision of an adequate clear and computer accessible empirical foundation for research, description and development, with suitable high quality corpora, corpus analyses, lexica, grammars and software tools. Documentary Linguistics has emerged as a new branch of applied linguistics in the past 15 years, concerned with language documentation and maintenance, especially for endangered languages. In the Human Language Technologies, 'Human Language Resources' for text, speech and gesture are the basis for applications such as internet information dissemination and retrieval, machine translation, automatic speech recognition and synthesis.

The present contribution surveys the motivations for developing Human Language Resources gathered on a 'fair play' basis, argues for the need for data and corpus analyses based on a comprehensive semiotic architecture, and finally provides a linguistic perspective on practical ways in which to use computational methods to use Human Language Resources, ranging from syllable grammars and the analysis of lexical tone to lexicography, the documentation of legacy text manuscripts and to the computational modelling of tone sandhi.

**Keywords: documentary linguistics, human language technology, corpus linguistics, lexicography, phonetics**

## 1. The empirical base of language sciences and technologies

### 1.1 Human Language Resources and Language Documentation

Our word processors, internet software and mobile phones would not work without *Human Language Resources (HLR)*. The entire World-Wide Web is built on natural text and markup text resources, with other media embedded into the text. And our linguistic models and theories, lexicons and grammars would not be possible without one form or another of *Language Documentation (LD)*. These two concepts, *HLR* and *LD*, are technological and linguistic sides of the same coin: they express programmes for collating authentic text and speech data, lexicons and grammars, and the methods, formats and tools which are needed for collecting and processing the text and speech data. With this background in mind, the present study falls into two sections: first, a discussion of speech and text resources in the context of scientific and technological development; second, a series of case studies based on projects which the author has conducted.

All the disciplines which deal with language — whether literary or linguistic, whether Sinology or English Studies, whether machine translation or internet search — are to greater or lesser extents dependent on *HLR* and *LD*. In this overview I will concentrate more on the technical side, *HLR*, since *LD* is increasingly adopting the technical methodologies of *HLR*, but I will often combine references to the two as *HLR/LD*. This contribution to the *Sinology Congress 2012* is not conceived primarily as a classic academic study of a particular empirical or formal problem, but as a strategic overview and as an information source for the methodologies required in modern empirically founded research in the language sciences and technologies.

Language and speech data are, fortunately, ubiquitous: they are found in libraries, bookshops, CD stores, lexicons, grammars, word processors, web sites,

tweets, blogs, novels, dramas and poems, and of course by direct interviewing of language users. This wide range of data sources is required on the one hand for corpus-based theoretical and descriptive linguistics and for the speech and text technologies such as information retrieval, speech interfaces for computers, and machine translation. But on the other hand these data sources are required for many other sciences. Indeed the ‘close reading’ study of literary texts is actually far more corpus-based than most varieties of linguistics. Theology and jurisprudence, like many other disciplines, are also dependent on large collections of texts for their work, and the media industry is dependent on language and speech technologies for accessing large collections of speech and text, as well as video media. The need for extensive speech and text data from sources such as those listed above originated in two fields. The first source, concerned with *LD*, is *Documentary Linguistics*, which deals with the documentation and description of the languages of the world, particularly endangered languages, for scientific reasons as well as for the documentation of language and culture heritage. This avenue of text and speech data collation and processing is the *LD* direction.

The second source, *HLR*, emerged mainly from the *Human Language Technologies*, i.e. speech and text technology, which require large quantities of data for statistical analysis in order to create working models for creating text products such as large dictionaries, and speech products such as dictation and reading software. Gradually theoretical and descriptive directions in linguistics are making use of high quality *HLR/LD* information from these two sources in order to underpin the traditional linguistic methodology of deduction of hypotheses and testing with intuitively produced or observed opportunistic data.

In a nutshell: modern methodologies in the language sciences, both in general and theoretical linguistics, as well as the speech and text technologies, require extensive data in order to induce and validate their working models and

explanatory theories. The requirement for extensive data — ‘big data’ — marks a shift from traditional intuition based and deductive methodologies in linguistics to quantitative methods and inductive ‘close-reading’ corpus-based methodologies.

The field of *HLR/LD* has emerged as a general methodological field. Therefore discussion need not be restricted to specific languages or language families, whether Sino-Tibetan languages, Turkic languages, Indo-European languages, Niger-Congo languages or any others. On the contrary: scholars working in these different language domains can profit from and share each other’s results. I will accordingly discuss specific *HLR/LD* examples from different sources and with different aims, in which I have been involved:

**Language structure:** the visualization of the complexity of syllable structure constraints in Mandarin.

**Speech:** the phonetic modelling of spoken language for technological purposes, with reference to the Tibeto-Burman language *Kuki-Thadou* (ISO 639-3 *tcz*), in cooperation with a linguist who is a native speaker.

**Text:** the documentation of the heritage of a short-lived variety of written language, the Nigerian ‘spirit language’ isolate *Medefaidrin* (no ISO 639-3 classification) in cooperation with a local linguist and the local community.

**Lexicography:** the provision of *CESAF* tools for automatizing the acquisition and processing of lexicon data from language corpora, here the Eastern Turkic language *Uyghur* (ISO 639-3 *uig*), for use in human language technologies.

After this I will outline specific technical aspects of creating resources for selected language and speech technologies, lexicon creation and speech synthesis and conclude with brief indications of institutions involved in text and speech resource creation. Finally I offer a commented list of books for further reading, with recommendations for internet search.

## 1.2 The inductive turn: the need for data

The concept — and the problem — of *Natural Resources* has become one of the permanent issues in politics, media, business and science. The concept of *Human Resources* is one which is at the centre of any work programme, whether a national economy, a business, or a scientific project. The concept of *Financial Resources* is one which concerns most people most of the time. But *Human Language Resources* and *Language Documentation* represent concepts which are perhaps unfamiliar.

The term *Human Language Resources* originates in the more technical areas of speech technology, text technology and computational corpus linguistics, the *Human Language Technologies*, while *Language Documentation* originates in the more traditional area of field linguistics and the linguistic description of the languages of the world. As already noted, the domain of these terms covers the data and methods which we need in order to do our job as empirical language scientists and technologists, whether Sinologists, Anglicists, literary scholars, linguists or engineers in the speech and language technologies.

The thesis underlying these twin methodological disciplines is that the need for *HLR/LD* is universal in respect to the empirical description and modelling of any language in any of the language disciplines. A definition of the domain of these two terms will be a useful reference point. I propose the following definition as being very close to consensual in the field:

Both *Human Language Resources (HLR)* and *Language Documentation (LD)* provide the foundation for empirical and theoretical language studies and technological applications, while differing in focussing on applications versus descriptions, and consist of

- (a) *corpora* of text and speech data,
- (b) *tools* for efficiently processing, storing and communicating these corpora,
- (c) *information* such as transcriptions, glosses, lexicons and grammars derived from the corpora by means of tools.

The *HLLR/LD* domains have arisen because of increasing awareness not only that there is a need for more inductive approaches to language description and modelling, but also that text and speech, and the principles underlying them, are too varied and too complex for the individual human mind to capture all their subtleties of form in the form of deductive rules. This ‘inductive turn’ in the study of speech and text assigns generalizations such as ‘principles’, ‘parameters’, ‘rules’ a secondary status: they are output of a wide range of inductive methodologies whose input is extensive, authentic, observed data and whose methodology is to some extent rule-based but largely quantitative and statistical.

It has turned out, for example, that automatic translation based on statistical evaluation of very large quantities of parallel or comparable corpora and of existing human translations provides a more effective empirical basis than collections of translation rules, however detailed. Naturally, machine translation goes wrong, as anyone can test using the translation machines on the internet, because adequate contexts are missing, but then human translation also goes wrong when an adequate context is missing. Anyone can translate simple texts and spoken utterances, but not everyone can translate technical texts or literature, first, because we lack the contexts for wide ranges of language activities, and second, because the world is categorized differently in the vocabulary and the grammar of different languages.

Another area where extensive *HLR/LD* are required is automatic information retrieval and language understanding, closely related to machine translation: the more data for training the system, the better the result. I do not know if anyone has quantified the amount of data to which a child is exposed in the first seven years of life, by which time his command of the language has more or less stabilized — it may well be of the same order of magnitude as that which would be required for the development and statistical training of modern speech and text technology systems.

In speech technology it is impossible to capture all the subtle variations in individual utterances by means of rules, and so for automatic speech recognition, as used in dictation software and in menu control, the systems have to be trained with statistical methods, using very large amounts of data. The same applies to speech synthesis, which is used in many contexts from public announcements to satellite navigation software in vehicles and to reading software for people with sight impairments.

As already indicated in the introduction, *HLR/LD* apply to any discipline in which language studies are foremost, such as Sinology, English Studies, Comparative Literature Studies, Oral Literature Studies, Cultural Studies, Linguistics, Phonetics, Speech Technology. Indeed, as we see in this Congress, there is increasing cooperation between disciplines such as these in the area of *HLR/LD*, as well as increasing cooperation between scholars working on entirely different languages, with the common goal of providing high quality empirical foundations for their work.

### **1.3 Deployment zones for *HLR/LD***

We find *HLR/LD* information in many places, though they may not be



traditionally referred to in such terms. Libraries, for example, embody *HLR/LD*: they contain corpora of data in the form of texts and recordings; they provide generic conceptual tools such as cataloguing systems for describing and for searching these data; they provide specific ‘metadata’ information in the form of catalogue entries. The Internet has emerged as the largest and most easily accessible source of *HLR/LD* which has ever been created, though a wild one, for which methods of taming are constantly being developed.

From the point of view of the literary scholar, perhaps also of the philosopher, perhaps it sounds a little odd to characterize books as ‘data’. Colleagues in Departments of Literature all over the world may shudder at the ‘naive’ notion that the works of fiction, of poetry, of drama which they study with hermeneutic methods are a corpus of ‘data’ or ‘resources’ — these terms could call up controversial notions of positivism in the philosophy of science, or perhaps more straightforward associations with applied sciences such as engineering and economics. Nevertheless, without works of literature in their concrete manifestations, and without collections of these works in libraries, the study of literature would be poor indeed. So I will risk the scorn of some literary colleagues by using a rather general concept of *HLR/LD* repository. For colleagues who work in rhetoric, or stylistics, or in empirical literary studies, and who study the forms of literary language, this usage will seem uncontroversial. And I will risk the scorn of some linguistic colleagues by insisting on the prime role of *HLR/LD* in the language sciences, but will attempt to appease them by conceding that our minds are potential Human Language Resources — but not *HLR/LD* of a type which is sharable independently of the individual researcher unless recorded as a corpus. The tendency of post-structuralist generative linguistics from the 1960s to the 1990s to assign corpus data a secondary role and to concentrate on modelling intuited data is even more extremely hermeneutic than the interpretative

methodologies of literary studies and conversation analysis: the data are not ‘given’ — the original Latin meaning of the term — independently of the observer who discovers them, but constructed intuitively. In varieties of linguistics which study corpora of speech and text, whether qualitatively or quantitatively, the need for *HLR/LD* will seem uncontroversial.

The first disciplines concerned with *HLR/LD*, though not as massively as in modern times, were *translation* — whether on the Rosetta Stone, or on paper and parchment — and lexicography. From the 19<sup>th</sup> to the 21<sup>st</sup> centuries, lexicography in Western Europe gradually turned from being primarily the art of the lexicographer into a modern digital technology based on corpora of many millions of words of text and transcriptions of spoken language. Translation proceeded at the same time in a similar direction with the aid of databases as translation memories and machine learning techniques.

The key to this development lies in the increasing importance of extensive *sharable data* for empirical language studies, the need for which has been generally recognized since the 18th century in connection with the development of empiricism and detailed studies of the natural world (though sporadically recognized in many cultures long before that).

In the day-to-day use of *HLR/LD* there are also practical considerations, as in the following criterial properties of *HLR/LD* which are constantly under discussion in the field:

1. **Independent.** The data are ontologically independent of the observer by virtue of being recorded on storage media and identically accessible to many observers. This is not true of the intuited opportunistic data which are often used in linguistics, typically: “Can you say...?”.

2. **Reusable.** The data are usable many times, perhaps by many analysts, and not just once by an individual scholar.
3. **Sustainable.** The data are not *ad hoc* individual observations but have an extensive lifetime. This is an obvious condition for endangered languages — if sustainable forms of data are not available, then if the language dies, the research on it dies. For this purpose, the notion of *sustainable repository* is crucial. In general, the structure of funded projects does not provide for sustainable repositories. These are a supraregional, national or international responsibility.
4. **Searchable.** This condition ensures that the data which are relevant for a given research purpose can be found. The condition also implies that the data are provided with adequate descriptive ‘metadata’, i.e. catalogue data.
5. **Interoperable.** The resources must be usable in a variety of environments, for example on different computer systems. What is true of cars must be true of *HLR/LD*.
6. **Standardized.** The role of a standard format, category system or tool, is to facilitate interoperability and exchange of data. This applies whether the standard is proprietary (such as the *de facto* norms of WAV files in the audio domain, or PDF and DOC files in the text domain) or official (such as HTML norms). Sometimes a crippling effect of over-standardization on innovation and creativity is feared, but this not generally justified: standards are, in the medium and long terms, mutable.
7. **Trustworthy.** The *validity* of *HLR/LD* must be checkable, the method of its creation must be *available* and *replicable*, the *HLR/LD* must be *accessible*, and storage of *HLR/LD* must be *reliable*.

All of these criteria are, from the scientific point of view, language and culture independent, and also discipline independent in regard to the scientific and technological language disciplines. The criteria also apply to data from other kinds of communication such as conversational gestures, and the sign languages used by the speech and hearing impaired.

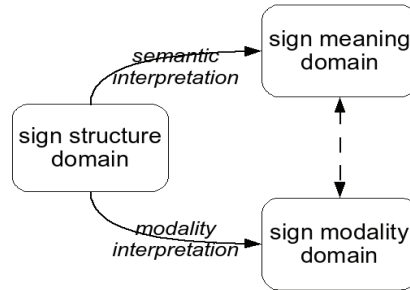
In the following sections I discuss scenarios for the creation of *Human Language Resources*, taking a broad perspective in view of the immense number of varieties of human languages and then the case studies mentioned previously.

## **2. Scenarios for human language resource creation**

### **2.1 The architecture of language**

In a sense, *HLR/LD* are primarily concerned with the *forms* of texts and speech rather than their *functions*, i.e. their *meaning* and their *purpose* — no functions without forms. But of course interest goes beyond the forms of texts and speech.

Figure 1 shows a model of the minimal requirement for a unit of documentation in a human language resource: the language sign, a representation of observable forms of text or speech, a representation of the structure of each form (whether simple or complex, paradigmatic or syntagmatic), and a representation of its meaning.



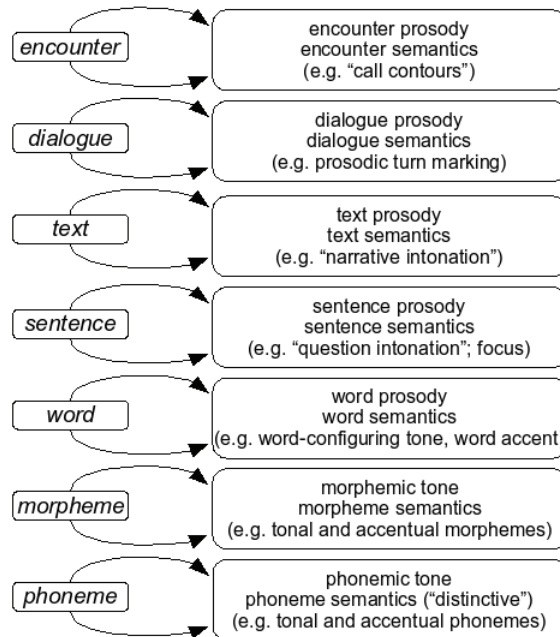
**Figure 1:** The units of language documentation: signs

The triadic structure is more adequate than the common dyadic structure of *significans* and *significatum* in many areas of linguistics, or in distinctions between the *logical structure* and the *rendering structure* of representations on the World Wide Web. The core of the sign, often interpreted as a mental representation, is abstract: the simple or complex *structure* of the sign. The structure of the sign has two interpretations in terms of the real world within which the sign user is located. The first interpretation is the *modality* interpretation, either acoustic (phonetic interpretation) or visual (as in gesture or in writing). The second interpretation is the *semantic* interpretation in terms of abstract concepts or objects in the environment.

But the triadic sign representation is insufficient in two ways. First, the triadic sign representation is too general. More detail, as shown in Figure 2, the *Rank Interpretation Model (RIM)* of language architecture, is needed in order to gain an understanding of the breadth and depth of HLR/LD.

Second, the triadic sign representation is incomplete: there is no representation of the situational semantic or pragmatic context. The situational semantics and pragmatics of these forms — meanings in context and the background knowledge and strategies of their users — can only be documented indirectly

and less explicitly than the textual and spoken forms of language. The situational semantic and pragmatic functionality of text and speech in use can be partially documented by use of additional textual characterization and by video recordings, however, and this practice is increasingly in use, with a growing interest in systematic transcription and annotation practices for communicative movement and gesture as well as speech and text.

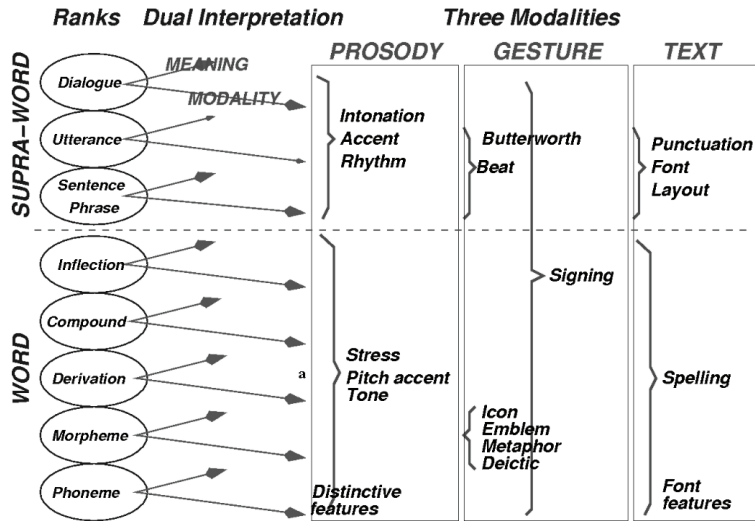


**Figure 2:** The Rank Interpretation Model (RIM) of the architecture of language

Signs are highly complex, though at each level there is a need for modality and semantic interpretations. Figure 2 is related to and differs from conventional models of language in many ways. Essentially, it combines the notion of the *rank* of language sign units of different sizes, a characteristic of functionalist

views of language, with a notion of *semantic interpretation* taken originally from formal logic and *phonetic interpretation* from linguistics. At each rank, from the phoneme through morpheme, word and sentence to the dialogue contribution and the encounter (and many sublevels between these), the units thus receive a *semantic interpretation* and a *modality interpretation*. The usual modality interpretation in linguistics is *phonetic interpretation*, but there is increasing interest in the documentation of writing in different scripts (including handwriting) conversational gesture and also of sign languages, which are comparable in complexity to spoken languages, into language documentation and human language resources. This extension of the Rank Interpretation Model is shown in Figure 3.

The purpose of proposing this integrated *Multimodal Rank Interpretation Model (MRIM)* here is to underline the fact that *HLR/LD* are not only about the properties of phonemes, words and sentences, but about the roles of all these items in a structured rank hierarchy, about the modality interpretation of these items in terms of speech, gesture and writing at different rank levels, and about the semantic interpretation of these signs at different rank levels. The inter-modality correspondence tendencies of intonation, accent and rhythm with punctuation, font choice and layout at the supra-word level are obvious. Perhaps less obvious is the inter-modality correspondence of prosodies such as accents and intonations with the accent-like and intonation-like rising and falling and rhythmical movements of hands, head, facial features. The correspondence of iconically demonstrating or deictic pointing gestures with words is more obvious.



**Figure 3:** Multimodal Rank Interpretation Model (MRIM)

## 2.2 Varieties of text and speech

We have seen rapid developments in the *HLR/LD* area under the influence of modern communication technologies: in all languages, when SMS text messages are used, the limitations — but also the potential — of the technology enforce changes in the use of language. The development of writing technologies 3000 years ago enforced not only changes in the use of language but changes in language itself. There are many scenarios for the use of language in text and speech, and for many reasons we want to document and harness these resources, for heritage preservation, for the advancement of education in the languages of the world, and for deployment with new technologies. It is this practical need for actual usage which prompts the use of ‘text’ and ‘speech’, rather than ‘language’.

The word ‘language’ itself has a wide range of meanings, across the enormous number of languages, dialects, speech varieties, written registers with



which we communicate naturally as humans, and the specializations of written registers of language which we use in mathematics, logic, the formal sciences and technology, for description, programming computers, for describing knitting patterns and chess games, for transcribing the pronunciation of languages, and for technical communication. The word ‘language’ refers, also, to the gesture sign languages of the visually and vocally impaired. The word ‘language’ is also used for the communication systems of birds, apes, whales and other animals. And the word ‘language’ is used metaphorically for many other things in nature which are said to ‘speak’ to us.

In these broad senses, there are uncountably many languages and varieties of languages. But for *HLR* documentation — as for any other activity — we only have limited means, which compete with the means required by other activities. Resource scarcity means rationing, selection, the setting of priorities. There is a ‘magic number’ which is often cited for the number of human languages which have developed until the present day: the *Ethnologue*<sup>1</sup> catalogue of the world’s languages lists 6909 languages and regional variants; this catalogue underlies the international language name codes standardized in ISO 639-3, and is based on the activities of multitudes of linguists around the world. Aleksandr Kibrik once characterized language from this point of view as *the spoken communication of a village*. One might even say, from the point of view of migrating populations, *the spoken communication used in a family*.

The number 6909 is a simplification, in relation to the varieties outlined in the present discussion, and it is already an enormous reduction in the number of language varieties in the broad sense of the term, but still a rather large, though not astronomical number.

---

<sup>1</sup> URLs: <[www.ethnologue.com](http://www.ethnologue.com)>, <[www.ethnologue.org](http://www.ethnologue.org)>.

So we need additional, more pragmatic criteria for selecting languages for creating human language resources. For *HLR*, in a large technology company the answer to selection lies in the standard languages of the world and other languages spoken by large populations, which promise a market for technology. But for *LD* in the humanities, the criterion is different: the priority lies with the languages which are on the verge of disappearing, the ‘endangered languages’. Endangerment-based criteria for selecting languages for resource-creation are used by funding agencies; the main criterion (apart from dwindling community size) is whether an affluent linguist or group of linguists is interested in these languages.

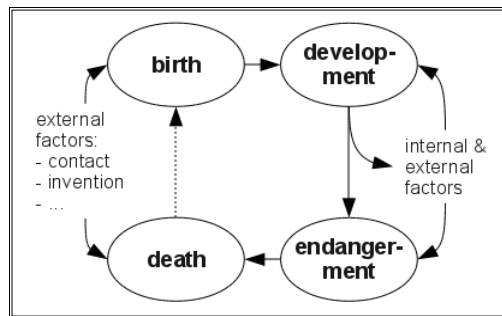
The reasons for these pragmatic criteria are obvious: there are economic priorities — we have limited means. For this reason, a group of linguists in West Africa and I have formulated more detailed criteria for both *HLR* and *LD*, and formulated this as the *CESAF* list of *HLR/LD* criterial properties:

- ♣ **Comprehensive:** all levels of linguistic analysis, relative to the task.
- ♣ **Efficient:** maximal automatization of data acquisition and processing.
- ♣ **State of the art:** data structures, algorithms; web applications, mobile intranet servers.
- ♣ **Affordable:** browser, any desktop/laptop/tablet/cellphone.
- ♣ **Fair:** open source payback to data donors.

In the following discussion I will deal with both the heritage type of motivation and the technological type of motivation for selecting languages for resource creation, and provide examples, concentrating on the following areas: the cycle of language birth-development-endangerment-death.

### 2.3 The cycle of language birth, development, endangerment, death

Behind the many language varieties which have already been outlined is an ever-changing cycle of language change: language *birth* (e.g. pidgins, creoles, artificial languages), language *change* (and ultimately *birth*) by medium and long term social and internal processes, language *endangerment* and language *death*, and sometimes also language *re-birth* as in the case of *Ivrit* (Modern Hebrew) in Israel, or, to a marginal extent, *Kernewek* (Cornish) in the United Kingdom.



**Figure 4:** The cycle of language birth – development – endangerment – death

To clarify the background I propose a model of cycle of language *birth* – *development* – *endangerment* – *death*, shown in Figure 4, and discuss a number of details: the first decision in creating human language resources is to decide for which phase in this cycle *HLR/LD* are to be provided — a technology company would concentrate on different phases than a cultural history project.

*Language birth* itself has many facets: the slow evolution of human languages from unknown sources around 200,000 years ago with the evolution of *homo sapiens sapiens*, the splitting of languages into new languages both within Africa and after the out-of-Africa migration 70,000 years ago, and again

after the last Ice Age 10,000 years ago and the development of agriculture, metal processing and large urban civilizations. Later facets of language birth and development include the mutation of colonial languages into regional languages, pidgins and creoles, from the first empires in Europe and Asia 5000 years ago to several Indo-European languages and to Chinese in the modern age. And ‘natural’ languages may be invented: *Esperanto* is the most famous example, based on the structural typology and vocabulary of European languages, and later in this paper a different kind of ‘invented language’ will be described. Sometimes older phases of languages which have developed into new languages are retained, mainly in written form, in formal and ritualized contexts such as religious language: Classical Chinese, Sanskrit, Greek, Latin, Old Church Slavonic, Classical Arabic, with legacy documents with special requirements for manuscript documentation.

*Language development* is part of a many-faceted process of language change. On the one hand, it involves the standardization of languages for sociolinguistic, cultural and political reasons, in which factors which influence the stability of societies are subject to the control of academies, education and media. On the other hand, it involves the de-standardization of languages which were once standards, as in the cases of Classical Chinese, Sanskrit, Greek, Latin, Classical Arabic, which have already been mentioned, but also in many other cases in which the language in its standard form has disappeared, such as the East Germanic Language Gothic, or the languages of South and Central American civilizations.

A different development of this kind is where dialects become standards and then give way to other dialects: this has the case with the dialects of Britain since the Middle Ages, where the creation of a standard was pushed by Norman French colonists, Norman French itself being a variety of mediaeval French which had been influenced by Scandinavian colonists. German is another case in

point: the dominance of South-West Germany in the Middle Ages gave way over the centuries to the dominance of East German varieties under the influence of Luther's Bible translation, and later still to the dominance of North German dialects under the influence of Prussia. We can see the grain of truth behind Max Weinreich's joking definition of a standard language: *A language is a dialect with an army and a navy*. Evidently, this is the kind of language variety preferred on political and macro-economic grounds in societies with a strong need for a unified identity.

The point is that languages are not defined by mutual comprehensibility alone, but by political identity, and both of these can change rapidly. The very different dialects of Austria, Switzerland and Germany are referred to for historical political reasons as 'German'. In contrast, the languages Dutch and Flemish, on the other hand, are linguistically very similar, both to each other and, less so, to German, but have different political histories and are treated as different languages. Both are very similar to the neighbouring German dialects of the Rhineland, which in turn are very different from many other German dialects. Evidently the motivation for defining a language is political as well as linguistic.

In principle, though with major historical differences, these considerations also apply to the development of Chinese through the centuries and under different political entities in various parts of Eastern Asia and in the Chinese diaspora.

*Language endangerment* is a function of language development: the languages of some communities thrive as the communities themselves thrive and influence others, while the languages of other communities decline as the communities decline and are influenced by others. The immediate source of language endangerment is when languages are no longer transmitted by parents

to their children. But this complex process has reasons. The reasons may be migration and diaspora, due to climate change such as is currently going on in many parts of the world. Or, in the case of a language like Yiddish, endangerment and at least partial extinction may be due to persecution and destruction of the community. Or the reasons may be due to political repression, of which there are many examples in all continents. The most common reason, indirectly linked to these other reasons, is economic survival and progress. My first contact with this was in a taxi in Dublin in the 1980s: I asked the taxi driver if he spoke Irish. He replied “Of course!”. I then asked if his children spoke Irish. His reply: “Of course not! I forbid them to speak it. They’ll never get rich if they speak Irish!” And in fact my own history — an English-speaking descendant of Welsh and English families — is coloured by some of these reasons.

*Language death* is the consequence of extreme language endangerment, via a moribund stage in which very few speakers of the language remain, and is easily defined: there are no more users of the language, and potential insights into human cognition and culture which the language might have provided have disappeared forever.

## **2.4 Preliminary conclusion: the variety of language resources**

The preceding discussion has ranged widely over the different varieties of text and speech for which *HLR/LD* are collected. First, the complexity of language signs themselves was outlined with reference to the *Multimodal Rank Interpretation Model* of the architecture of communication in text, speech and gesture. Then varieties of text and speech were sketched, and finally the cycle of language birth, development, endangerment and death was outlined. Each of these criteria enters into decisions about what kinds of *HLR/LD* to create.

### 3. Case studies

The following discussions of aspects of HLR/LD creation are based on projects and bilateral communications with scholars of different linguistic interests, but with the goal of developing high quality empirical data with appropriate methodologies for use in various linguistic and technological applications. Resources are being collected on a grand scale for many languages, for technological applications in large communities to language documentation of endangered languages. But resources are not just data: they also include enhancements of data with appropriate tools. The following case studies illustrate selected resource enhancement projects for written and spoken language.

#### 3.1 Speech resource visualization: Mandarin syllable complexity

Large quantities of text and speech resources are available for Mandarin Chinese, and their deployment in many areas of applied linguistics and speech technology is extensive. In particular, syllables in Mandarin and their properties such as tone are well researched and the subject of continuing research. I propose a technique here for adding to existing knowledge, using a visualization strategy which is related to the ‘big data’ visualization and sonification techniques which are being developed currently in order to gain insights into the structure of data prior to further analysis.

The structure of Mandarin syllables is often portrayed in the form of a ‘pinyin table’, as in Table 1 (He 2004). The Table visualizes a binary syntagmatic relation between the initials (onset consonants) and finals (vowels and vowel-nasal sequences, i.e. rhymes). The relation licences existing lexical syllables associated with characters. The table contains 399 characters; other tables with slightly different numbers of syllables may be found, e.g. Aidaoguangci (a nickname)





	b	p	m	f	d	t	n	l	g	k	h	j	q	x	zh	ch	sh	r	z	c	s		
<b>in</b>	+	+	+				+	+				+	+	+									+
<b>iang</b>							+	+				+	+	+									+
<b>ing</b>	+	+	+		+	+	+	+				+	+	+									+
<b>iong</b>												+	+	+									+
<b>u</b>	+	+	+	+	+	+	+	+	+	+	+				+	+	+	+	+	+	+	+	+
<b>ua</b>									+	+	+				+		+						+
<b>uo</b>					+	+	+	+	+	+	+				+	+	+	+	+	+	+	+	+
<b>uai</b>									+	+	+				+	+	+						+
<b>uei</b>					+	+			+	+	+				+	+	+	+	+	+	+	+	+
<b>uan</b>					+	+	+	+	+	+	+				+	+	+	+	+	+	+	+	+
<b>uen</b>					+	+		+	+	+	+				+	+	+	+	+	+	+	+	+
<b>uang</b>									+	+	+				+	+	+						+
<b>ueng</b>																							+
<b>ü</b>							+	+				+	+	+									+
<b>üe</b>							+	+				+	+	+									+
<b>üan</b>												+	+	+									+
<b>ün</b>												+	+	+									+

Examination of the syllable finals shows that they are syntagmatically complex, so their internal cooccurrence constraints are not expressed in the table. It is important to know these constraints for a number of reasons, which will be detailed below. There are a number of analytic steps which can be taken in order to bring out the constraint patterning:

1. Introduction of a glottal stop consonant as initial for syllables which the Table notes as being vowel-initial. The motivation for this is that vowel and vowel-nasal syllables start with a glottal stop when spoken in isolation.

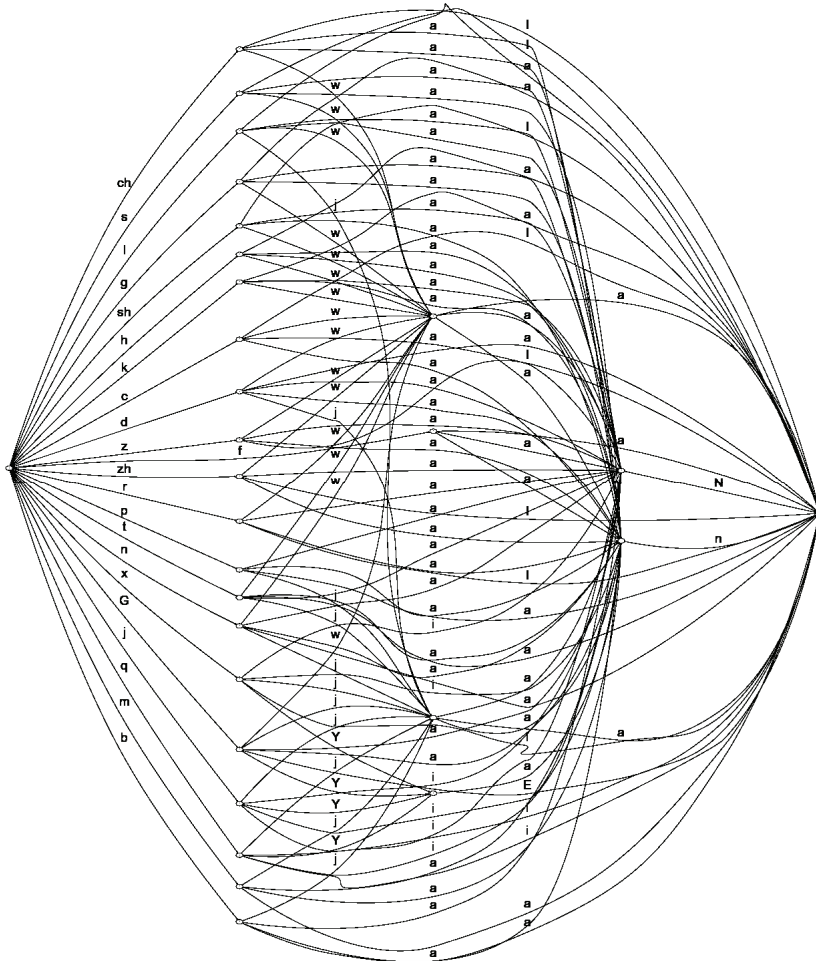


In Figure 5 automatically generated visualizations of two models of Mandarin syllables are shown, using left-right directed graphs. In each graph, there are differing numbers of parallel edges between pairs of nodes. These are pruned, in order to make the structure clearer, with only one edge representing the set of parallel edges in each case; the important issue here is the structure, not the specific symbol inventories on the edges. The names of the nodes are not significant. The graphs are deliberately rendered in such a way as to emphasize structure at the expense of detail; the underlying structures are highly detailed, however.

The graphs were constructed from the table in a four-stage process: first, the nodes and edges were created, based on Table 1, and, second, extended with a node insertion algorithm on the basis of the additional segmentations listed above (implemented in *Python*). Third, the graph was visualized using the *Graphviz* graph construction and optimization package. Fourth, a graph traversal algorithm (also implemented in *Python*) was used to generate the entire syllable set for testing purposes by starting at the leftmost node, traversing the edges to the next node on the right, picking up the edge label symbol on the way, and concatenating these symbols until the final node on the right is reached.

On the left in Figure 5 is a straightforward model of the binary relation expressed in the ‘pinyin table’, with no further analysis of syllable finals; the only nodes intervening between the initial and the terminal node are those which link the initial and final edges, the long vertical series of nodes in the centre. The model on the left captures exactly the syllables defined in the Table. The model on the right in Figure 5 expresses the constraints for approximants as separate segments. The series of three nodes for the approximants can be seen between the initial-final linking nodes and the terminal node. The explicit handling of approximants leads to an enriched understanding of internal syllable constraints.

The extended network with the interpolated nodes and edges for approximants overgeneralizes a little, i.e. produces syllables which are not in the basic inventory of 399 syllables defined by the Table.



**Figure 6:** Visualization of Mandarin syllables with explicit incorporation of constraints on approximants and nasals as independent segments

When the final nasals [n], [ŋ] are treated as separate segments, a more complex constraint network shown in Figure 6 is induced from the basic resource data in the Table. The nodes linking the vowels to the nasals appear as a vertical series of two, immediately to the left of the terminal node. The graph also overgeneralizes slightly, taking over the same overgeneralizations introduced by the treatment of approximants. Adding the nasals did not increase the number of overgeneralizations.

It is legitimate to ask what are the actual advantages of such visualizations. These advantages are both theoretical and practical. On the theoretical side, they permit a more sophisticated appreciation of the typological properties of the language, supporting comparison of dialects and the study of language history. On the practical side an enhancement of basic resource data is needed when theories of sound production and perception and of the sequentiality of speech acquisition (with child and adult learning) and speech loss (due to accident or illness) are examined. Also on the practical side, a detailed understanding of sound patterns is a prerequisite for the design of corpus markup for the speech-corpus-based technologies of speech recognition and speech synthesis. In these contexts, the graphs can be also interpreted as finite automata (state machines) for symbolic or stochastic processing purposes.

### **3.2 Spoken language resources: developing tone models for Kuki-Thadou**

The long-term intention of providing *HLR/LD* for the Tibeto-Burman language Kuki-Thadou (Thadou) is to provide an empirical foundation not only for traditional applications in education, but also to examine the problems which need to be handled in providing *HLR/LD* for linguistics and speech technology

in a Sino-Tibetan language, here specifically a Tibeto-Burman language on the India-Myanmar border. The particular features of interest are phonemic tone and its phonetic correlates.

The typical *HLR/LD* information collection procedure encompasses lexicon compilation, and simultaneous phonemic analysis, supplemented by tonemic analysis. The phonemes of Thadou, as elicited by these methods, are shown in Table 2.

**Table 2:** Thadou phonology and tonology

left: consonants; top right: vowels; bottom right: tones

	bilabial	labio-dental	alveolar	palatal	velar	laryngeal		Front	Central	Back	
<b>plosive</b>	p	ph b	t	th d	k	kh g	ʔ	<b>close</b>	i	u	
<b>nasal</b>		m		n		ŋ		<b>mid</b>	e	ə o	
<b>fricative</b>			v	s z		[x]	h	<b>open</b>		a	
<b>apic. affr.</b>				ts					<b>Gibbon</b>	<b>Hyman</b>	<b>Gloss</b>
<b>lat. appr.</b>				l				<b>sá</b>	H	HL	animal
<b>lat. fric.</b>				ɬ				<b>sǎ</b>	LH	LH (H)	hot
<b>appr.</b>	w				j			<b>sà</b>	L	L	thick

The starting point for the analysis is the classic method of ‘ear phonetics’. But for speech technology, in particular speech synthesis, a more precise model is required, of the fundamental frequency (F0) of tones in isolation (citation contexts) and of tones in sequence, and this is supplemented and supported by instrumental measurements and perceptual tests. For application purposes, whether in language teaching or in speech technology, it is important to document each phoneme in context.

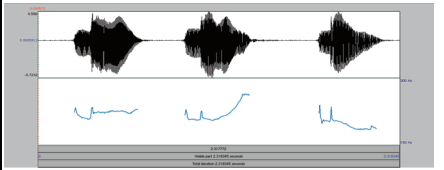
Among the consonants, unlike the different two-way distinctions between plosive types in Chinese or English, Thadou has a three-way distinction between unvoiced, aspirated and voiced plosives, presumably an areal phenomenon related to the neighbouring Indo-Aryan languages which have such distinctions.

The property of phonemic tone is shared by Thadou with other Sino-Tibetan languages. As with segmental phonemes, standard vocabulary and minimal contrast methods are used to determine the set of tones. As with segmental phonemes, varying analytic methodologies and varying theoretical assumptions can sometimes lead to slightly different inventories or characterizations of the tones, as in the difference between the Hyman and Gibbon tone inventories in Table 2.

The next step, at least for the purposes of *HLR/LD* in speech technology, is to determine the physical correlates of the tones. In medical applications, physiological correlates are investigated, and in the more common speech technology applications the acoustic correlates are examined. The usual procedure is to use software for examining the waveform and pitch traces of utterances, to annotate the occurrences of tones in the utterance, and to extract the acoustic measurements of frequency and duration associated with these tones (cf. Table 3).

**Table 3:** Thadou tone model with basic descriptive statistics

Tone	N	min	max	mean	sd	offset	slope
H	18	200	244	221	0.29	221	-0.03
LH	17	215	237	220	7.07	209	1.3
L	18	192	213	203	6.3	215	-1.31



The three tones of Thadou, spoken in isolation, show clear phonetic properties. The H (high) and L (low) tones are clearly different, with barely

overlapping ranges and clearly different means of 221 Hz and 203 Hz respectively; the small, non-overlapping standard deviations show, in relation to these means, that the difference in frequency is significant. The LH (low-high) contour tone is not so different from the H tone, but the shape is very different, as shown by the considerable positive slope of 1.3, in relation to the marginally negative slope of -0.03 shown by the H tone and the clearly negative slope of -1.31 shown by the L tone.

These are measurements on citation forms in isolation. Tones in context may behave rather differently. Figure 7 shows an example of tone spreading in an authentic utterance, and in a model of this tone spreading in a speech synthesis example of this utterance.

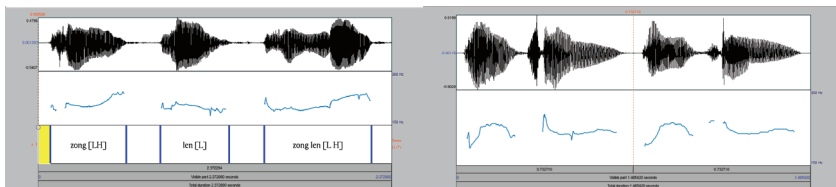


Figure 1: Thadou tone spreading (left). Tone original and synthesis (right).

The application of *HLR/LD* procedures such as these provides a documentary foundation for further linguistic description and technological application.

### 3.3 Written language resources: documenting Medefaidrin texts

Customarily one thinks of written language as books, newspapers, the internet. But in *HLR*, for many purposes handwriting and manuscript documents are equally important. Examples in technology are scanning with optical character recognition, the recognition of handwriting gestures on touchscreens, and in



forensics and historical philology writer identification. In heritage documentation, oftentimes available data are in manuscript form.

In the case of manuscripts, which are particularly important in the study of local and minority languages of many kinds, especially endangered languages, there are many more stages of documentation on the physical level, comparable in many ways with speech, than with electronically available written text. Manuscripts may be ‘noisy’, i.e. faded, smudged, stained; they may be subject to ‘fast writing’ phenomena of assimilation and reduction, like ‘fast speech’ phenomena; the media may be subject to physical degradation such as discolouring and decomposition through dryness or damp. The first stages of documentation are, therefore, as with speech, segmentation, transcription and annotation, together with vocabulary collation and analysis and contrastive study of the characters, punctuation marks and layout conventions of the text, and the grammatical conventions, which may differ from the conventions of spoken language.

Figure 8 shows documentary fragments pertaining to the case of the Nigerian ‘spirit language’ Medefaidrin.

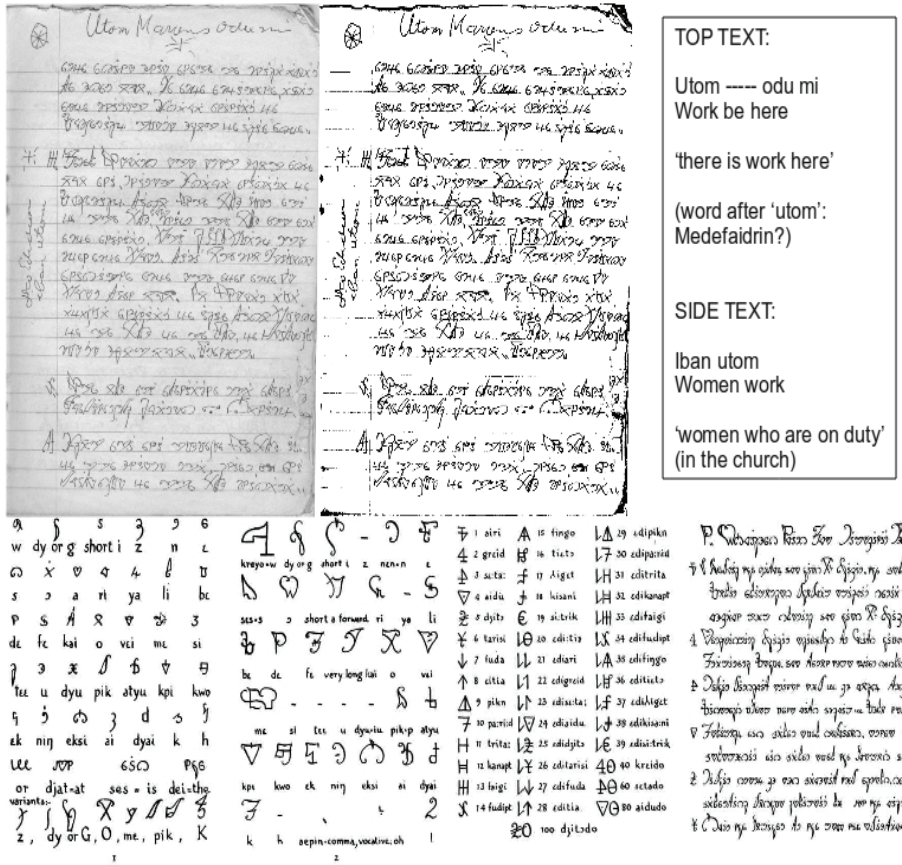


Figure 8: Samples of phases in Medefaidrin 'spirit language' manuscript analysis

Medefaidrin is spoken and written by the Oberi Okaime religious community near the Calabar River in South Eastern Nigeria. The language is said to be a 'spirit language' which was revealed to the originator of the language by divine inspiration (Gibbon et al. 2010).

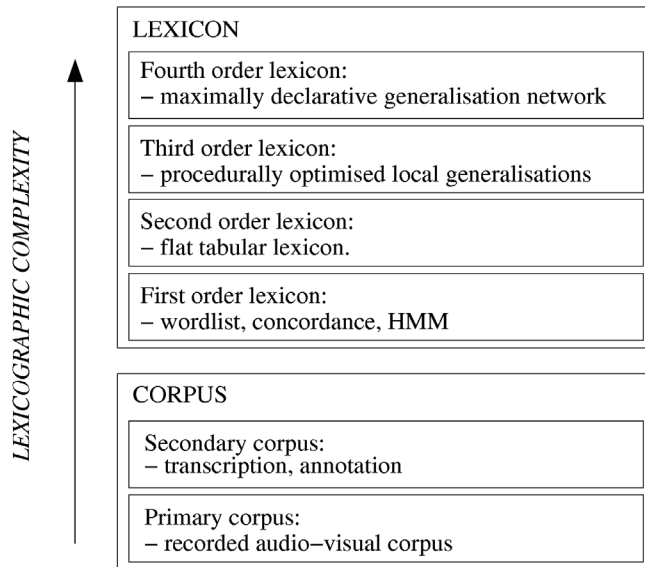
A number of phases of Medefaidrin documentation at different levels of analysis are shown (top left clockwise to bottom left): the original scanned pages of a notebook; the same page, optically enhanced for readability; metadata about

the marginalia of the page; number system inventory; alphabet inventory. The phases are somewhat analogous to the phonetic analysis of speech signals: noisy signals are filtered to extract the essential information, then the signals are tokenized and inventarized. The scanning and filtering destroys some important background information, however, which is only of indirect interest to the linguist, but may be of immense interest to the historian: the state of preservation of the document, and its age — only the original document can provide direct evidence for this information, though circumstantial evidence and historical reconstruction may contribute.

The divine inspiration of Medefaidrin is not entirely immune to influence from the contemporary prevailing colonial language of missionaries in the area, English, as individual characters in the script, the number system and text layout show. An interesting feature of number formation is, however, the vigesimal (base 20) system, as inspection of the number system in Figure 8 shows, of which only vestiges in English ('score', 20; biblical 'three score', 60) and French ('quatre-vingt', 80).

### **3.4 Lexicography: prelexical resources for Uyghur**

One central area of *Human Language Resources* — perhaps the central area — is the lexicon or dictionary, of which there are many forms. The lexicographic data acquisition and product creation process has often been seen as an art, but there is a logic to the process, and tools are employed at each stage. The stages of lexicon acquisition, each stage involving abstraction and generalization from corpus data (perhaps with lateral input from other dictionaries) is illustrated in Figure 9.



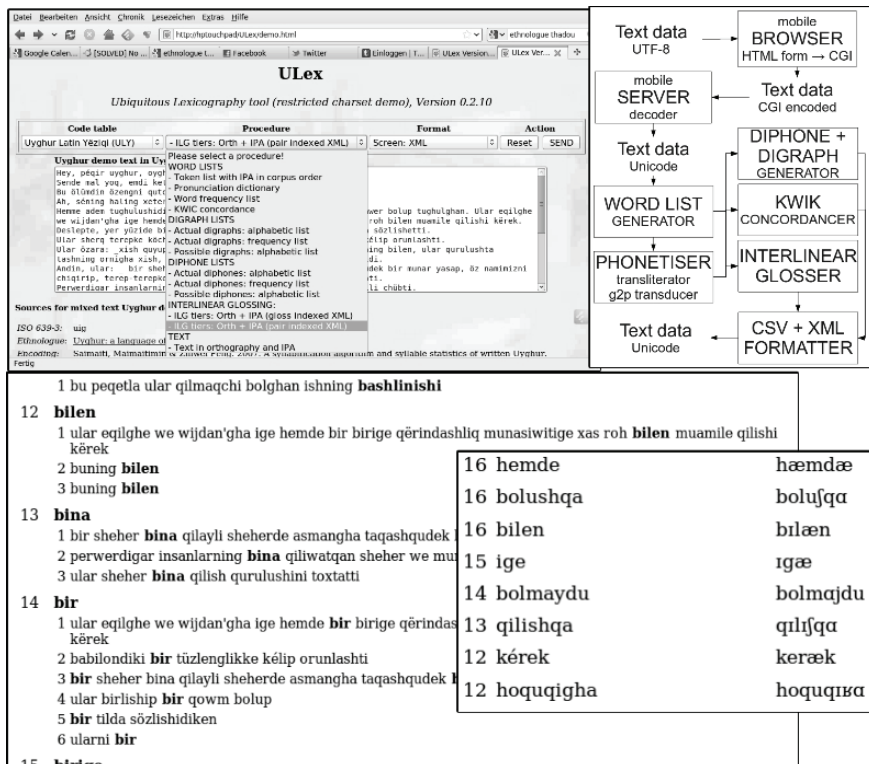
**Figure 9:** Hierarchical Lexicon Generalization Model

The core resources are at the corpus level: both the primary corpus level of recorded written documents and speech recordings and the next level of abstraction, transcriptions and annotations — such as part-of-speech (POS) tagging. The next level involves a jump from a corpus of *tokens* to the first order lexicon, consisting of tables and lists of *types*: word frequency lists, pronunciation lexicons, concordances, interlinear glossing structures. The second order lexicon requires the integration of all this information into a highly redundant flat tabular lexicon in which every entry contains a full specification of lexical information — which is practically never fully constructed, because of the extreme redundancy. The third order lexicon has the familiar shapes of semasiological (e.g. alphabetically organized) and onomasiological (e.g. synonym based) dictionaries. The fourth order lexicon is mainly of theoretical linguistic and lexicological interest: it involves a maximum of generalization over as many entries in the lexicon as

possible, and thereby a minimum of idiosyncratic information in each lexical entry. The third and fourth order lexicons have different structural levels: the *megastructure* (metadata about the creation process and the product), the *macrostructure* (overall organization, for example as semasiological or onomasiological dictionary), *microstructure* (which organizes the information for each lexical entry in terms of data categories) and the *mesostructure* (links to the generalizations about grammar, pronunciation etc., and to word fields in the dictionary as well as to corpus examples).

At the first order lexicon level, lists of ‘prelexical’ units of many different kinds are compiled. For linguists, the most useful kinds of lists are word lists, in particular word frequency lists, with a type/token frequency ratio to indicate to what extent a corpus is saturated (i.e. where decreasing numbers of new units occur). In lexicon creation, information about contextual word properties is required, and a popular tool for aiding this procedure is the concordance, i.e. a list of words accompanied by the contexts in which they occur, usually either lines of text, or sentences. A concordance is already a prototype dictionary, though missing other kinds of information such as pronunciation, part of speech, definition. For speech technology, other kinds of unit such as diphones and digraphs are needed (for decoding speech) as well as words and their contexts (for language models).

Examples of prelexical data at the first order lexicon level for the Eastern Turkic language Uyghur are shown in Figure 10 (cf. Gibbon 2012).



**Figure 10:** Collage of pre-lexical corpus input interface (top left), data flow (top right) and automatically created outputs (KWIC concordance, bottom left; frequency list with IPA transcription, bottom right).

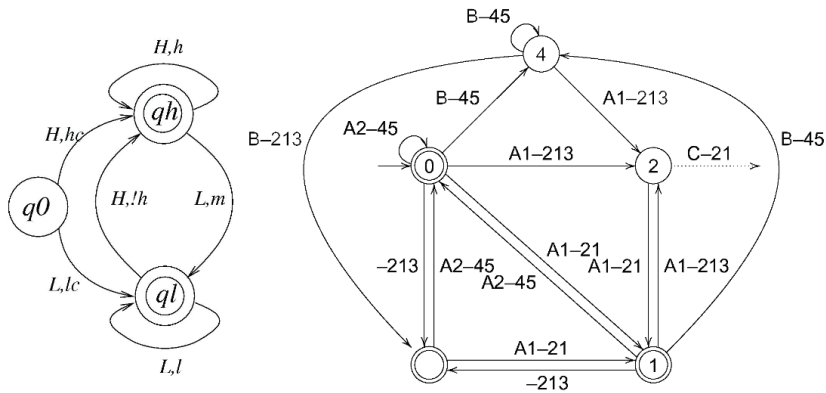
The first order lexicon extracts show a user interface for corpus input, the procedure for processing this input, and two examples of the output obtained: a KeyWord In Context (KWIC) concordance to aid lexicological and grammatical analysis, and a frequency list with IPA transliteration to aid the development of pronunciation lexicons and speech technology systems. All of these prelexical data structures are generated automatically from the corpus.

### 3.5 Formal issues: Finite Machines as resources for tone modelling and prediction

The modelling of the phonetic properties of tonal resources is only part of the story. *Human Language Resources* are not only data and not only lexicons, but also grammars and prosodic models. In order to be able to use these grammars and prosodic models in a technological context, for example in reading software for the blind and in speech synthesis software for the vocally impaired, precise models are required.

Fortunately, we know by now that the mathematics underlying these models is relatively simple: in the Chomsky Hierarchy of Formal Grammars, from the most complex (Type 0) transformational operations through context-sensitive (Type 1) grammars for cross-serial dependencies, and context-free (Type 2) grammars for classic tree constructions, the simplest type (Type 3), the regular or linear grammars, are adequate for phonological, prosodic, and most morphological patternings. These grammars can be operationally modelled by Finite Machines, i.e. Finite State Automata and Finite State Transducers. Without going into detail, these are models of finite devices, i.e. devices which only need a finite memory for their operation. They are exceptionally easy to compute; nevertheless, by iteration (repetition) they can still describe infinite numbers of patterns.

For Niger-Congo languages I discovered in the mid-1980s that the phonetic interpretation of tone sequences of terraced tone languages is straightforwardly modelled by means of Finite State Transducers (Figure 11, left).



**Figure 11:** Left: general Finite Machine for phonetic tone realisation in two-tone Niger-Congo languages (Gibbon 2007). Right: Jansche's Finite Machine for lexical tone sandhi in Tianjin Mandarin (1998).

Starting a sequence at 'q0', a high tone moves to the high state, 'qh', where further high tones circle around until a low tone appears, when the low tone moves to the low state, 'ql', and so on. Similarly, if the sequence starts with a low tone it moves to the low state 'ql', if more low tones follow they cycle around until a high tone appears, and moves to 'qh'. On each type of transition a different kind of phonetic operation takes place: from 'ql' to 'qh' downstepping (assimilation of a high tone to a preceding low tone) occurs; from 'qh' to 'ql', low tone assimilation to the preceding high tone occurs. Languages differ in the details, but the general model for 2-tone languages remains the same.

Starting from this idea, Jansche developed a Finite Machine for describing the tone sandhi of the Tianjin variety of Mandarin (Figure 11, right). Without going into the details, it is obvious the typology of the two types of tone system is very different. The key to understanding the typological difference lies in noticing that the output of the Niger-Congo system is phonetically defined, while in the Mandarin system it is phonologically defined.



Nevertheless, the same formal system of Finite State Transducers is adequate for both systems: for describing both similarities and differences, and permits the creation of *HLR/LD* for computational purposes in speech technology, in particular in building prosodic models for speech synthesizers. Indeed, Finite Machine models are ubiquitous in *HLR/LD* for speech technology. The *Hidden Markov Models (HMMs)* in speech technology are statistically enhanced Finite Machines.

## 4. Technical resource creation procedures

### 4.1 A general resource for text and speech: the lexicon

An example of a specific lexicon from a linguistic point of view was given as a case study in the previous section. Lexicon resource processing consists of the following main steps, which require appropriate software tools (note that the procedures listed after tokenization are not necessarily conducted in the order given):

1. **Tokenization.** Individual word tokens are identified, including abbreviations, numbers, prices, dates, punctuation, identification of complex layout objects such as tables.
2. **POS (part of speech) tagging.** Each token is provided with a label (or set of labels) constituting a hypothesis about its part of speech; the European EAGLES (Expert Advisory Groups for Language Engineering Systems) developed a standard POS tagset for European languages, which has been extended and applied to other languages (these sets are in flux; consult the internet for up-to-date details).
3. **Word token and word type list creation.** A list of (possibly inflected) word types is extracted from the set of tokens, often also in conjunction

with the word token frequencies.

4. **Lemmatization.** A list of lemmas is created from the list of word types, involving stemming in the simplest case, and morphological analysis in the general case.
5. **Concordancing.** A context dictionary consisting of a list of items (types, lemmas, tags, etc.) and the contexts in which they occur in the texts. The best known kind of concordance is the KWIC (KeyWord In Context), a simple list of words and their left and right context strings.
6. **Word sketching.** Extraction of a maximum of (grammatical and other kinds of similarity) information about lemmas based on their distribution in the texts.
7. **Dictionary database compilation.** Semi-automatic (moderated) entry of information into the lexical database.
8. **Manual editing** of lexicon articles (definitions, etc.).
9. **Production.** Selection, organization and formatting of lexical information for the intended dictionary megastructure.

These procedures apply, with suitable modifications, to the compilation of other types of dictionary, including dictionaries for use in multilingual, speech-based and multimodal communication systems. The steps 1-7 above pertain to three of the stages shown in Figure 9: secondary corpus processing, and first and second order lexicons. Steps 8 and 9 pertain to third order lexicons. The fourth order lexicon is a product of research and not discussed further here.

## 4.2 Resource creation procedures for text-to-speech synthesis

A *Text-To-Speech (TTS)* synthesis system requires resources for developing the following subcomponents which resemble the *HLR/LD* components outlined

in the case studies in §3, with additional smaller components and more detailed steps:

1. **Text parser.** The text parser is a special case of the kind of parser which is used in text processing in general, enhanced with phonetization and prosodic modelling information, and will not be discussed further here. The text is pre-processed in order to extract implicit information:
  1. The spelling and ultimately the pronunciation of special text components such as abbreviations and numbers must be extracted.
  2. A pronunciation lexicon, usually with additional pronunciation rules, is required.
  3. A parser is needed for disambiguating the structure by picking the correct word readings from the lexicon and delimiting the phrasing of sentences.
  4. A grapheme-to-phoneme (phonetization) component is used to derive a transcription of the speech sounds for input to the speech processing component.
  5. A prosody module is needed for deriving intonation, accentuation and timing patterns for input to the speech processing component.
2. **Signal processing component:** conversion from an interface with parsed and phonetized text with added prosodic information into a synthetic speech signal. For the signal processing component there are several different speech synthesis paradigms, including the following main types, for which paradigm specific resources are required:
  1. **Pre-recorded ‘canned’ speech.** Canned speech is typically used in straightforward information service environments such as satellite navigation systems for vehicles, and for railway station announce-

ments. Systems such as these use a restricted set of utterance templates which permit substitution of station names and times, but also permit a combinatorially large set of new utterances to be synthesized. Canned speech is in principle very comprehensible and very natural, provided that the template units are carefully designed and produced, with close attention paid to the correct prosody (intonation and accentuation), and to appropriate transitions between canned speech units.

2. **Concatenative speech synthesis.** Small units, such as phonemes, diphones, demi-syllables and sometimes larger units, are concatenated to form words and sentences. There are three main approaches, each of which requires different kinds of resource:
  1. **Diphone synthesis** is one of the first kinds of concatenative speech synthesis, and is still used. In diphone synthesis, pre-recorded speech samples containing all the diphones in the sound system of the language are used, which are concatenated in order to reproduce the patterns of the input syllable and word sequences. A diphone is essentially a pair of phonemes (speech sounds; see below).
  2. **Unit selection synthesis**, a popular variety of speech synthesis, and in general more natural than diphone synthesis, is based on selecting continuous units from a large recorded corpus. The corpus is designed to contain all the phonemes, generally all the diphones, and perhaps all the triphones (sequences of three phonemes). Units are concatenated after calculating the best possible fit (cost, weight).

3. **Hidden Markov Model (HMM) synthesis**, a recent development based on stochastic modelling of unit sequences, trained on a suitable corpus.
3. **Formant speech synthesis**. Formant synthesis is one of the earliest kinds of speech synthesis, and is based on the spectral structure of speech sounds. An acoustic signal is reconstructed from empirical information about vowels, consonants, and the pitch, intensity and duration patterning of the intended synthetic speech signal. In principle, this approach is the most flexible and parametrizable in terms of linguistic and phonetic properties, but is more difficult to use in practical systems than concatenative techniques.

Each of these components and procedures requires different *HLR/LD* steps. The text analysis component shares many features with lexicography, and the initial stages of the signal processing component shares many features, such as transcription and annotation, with linguistic speech documentation.

## 5. Conclusion: the institutional perspective

It should have become very clear that *HLR/LD* creation can only be a transdisciplinary, collaborative, international effort. The traditional division of linguistic disciplines according to languages and language families is of course well justified, in permitting intensive study of history and typology of these various language domains. But in the field of *HLR/LD* creation procedures, all languages are equal and the same fundamental formal, computational methods apply to them all. The task is huge, and empirically justified progress can only come from cooperation. However, in the prioritization of languages for *HLR/LD* creation all languages are not equal: for practical reasons, the human language

technologies set different priorities from linguistics, even though the problems and methods are very closely related.

The need for transdisciplinary and transnational cooperation becomes very clear if the criteria listed earlier are considered. High quality resources and documentation must be *independent, reusable, sustainable, searchable, interoperable, standardized, and trustworthy*. There are many institutions world-wide concerned with compliance with these quality standards in production of language resources and documentation of the kinds discussed in the present contribution. It is only possible to provide a few relevant categories of such institutions here, and to mention a few initiatives:

1. **Language and speech resource agencies:** in many regions, the two most well known being perhaps the *Linguistic Data Consortium*, in Philadelphia, and the *European Language Resources Association (ELRA)* with its operational wing, the *Evaluations and Language Resources Distribution Agency (ELDA)*, in Paris.
2. **Academic institutions:**
  1. Research institutions such as *Academia Sinica*, *CASS Beijing*, the *School of African and Oriental Studies*, London, the *Institute for Language and Internet Technology* at Eastern Michigan University, and Institutes of Technology in many countries.
  2. Empirical linguistic and human language technology teaching institutions world-wide, too numerous to mention.
3. **Academic networks:**
  1. Conferences, in particular the *Language Resources and Evaluation Conference (LREC)*, which takes place every two years.

2. Coordination initiatives in the speech technology field, in particular the *International Coordinating Committee for Speech Databases and Assessment (COCOSDA)*, of which I am currently Convenor, and, very prominently *Oriental COCOSDA*, which originated as the Asian wing of *COCOSDA*, but is now an autonomous and very successful networking organization with its own conference series.
4. **Funding agencies:** from government agencies in many countries through larger and smaller non-for-profit organizations to the World Bank — again too many to mention here.

It is a fitting conclusion for me, in my role as Convenor of *COCOSDA*, to the present contribution to recall that the achievements in the area of Asian *Human Language Resources* were internationally recognized by the award of the prestigious *Antonio Zampolli Prize* to *Oriental COCOSDA* at the recent *LREC 2012* meeting in Istanbul, 23-25 May 2012. We owe them our congratulations. The integrative international activities of *Oriental COCOSDA* have a vitality and productivity which will produce highly significant results in theory and in application in the great variety of languages of Asia, and will continue feeding into empirical work in speech and text sciences and technologies throughout the world.

### Selected further reference

The following list is highly selective, and partly oriented towards linguistics, partly towards technology, with a slight slant towards the languages of Eastern Asia.

With a rapidly developing field like this there is currently no one publication which can cover all aspects of *Human Language Resources* and *Language Documentation*, but there are many online tutorials, guidelines and international standards. Consequently there is no substitute for Internet search, particularly for institutions such as the *International Standards Organization (ISO)*, the *World-Wide Web Consortium (W3C)*, the *Text Encoding Initiative*.

Gibbon et al. (1997) deal with basic techniques for creating resources for speech technology, including evaluation methods; Gibbon et al. (2000) extend the scope to multimodal systems. The SAMPA alphabet, which is frequently used in speech transcription in speech technology, is detailed in these publications. Itahashi et al. (2010) and Tseng (2009) pay particular attention to the requirements of Eastern Asian languages and to spontaneous speech, respectively. Koehn (2010) on statistical machine translation and McEnery & Hardie (2012) discuss specific aspects of corpus processing for linguistics and technology, and Atkins et al. (2008) and Van Eynde et al. (2000) provide background to lexical resources. Witt et al. (2009) describe the formal document description languages required for archiving and dissemination of *HLR/LD*, and Mehler et al. (2012) provide an overview of all aspects of speech and text technologies, including development, resources and evaluation.



- Aidaoguangci (nickname). 2010. Putonghua shengyun peihe biao [A chart for the combination of Mandarin initials and finals]. Forum *Iask*. <http://ishare.iask.sina.com.cn/> (Last consulted 2012-09-16.)
- Atkins, Beryl T. Sue, and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford & New York: Oxford University Press.
- Gibbon, Dafydd. 2007. Formal is natural: toward an ecological phonology. *Proceedings of ICPhS XVI*, 83-88. Saarbrücken: Universität Bielefeld.
- Gibbon, Dafydd. 2012. ULex: new data models and a mobile environment for corpus enrichment. *Proceedings of the Language Resources and Evaluation Conference 2012*, 3392-3398. Paris: ELRA/ELDA.
- Gibbon, Dafydd, Roger Moore, and Richard Winski. (eds.) 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin & New York: Mouton de Gruyter.
- Gibbon, Dafydd, Inge Mertins, and Roger Moore. (eds.) 2000. *Handbook of Multimodal and Spoken Dialogue Systems Resources, Terminology and Product Evaluation*. Dordrecht & Boston: Kluwer.
- Gibbon, Dafydd, Pramod Pandey, D. Mary Kim Haokip, and Jolanta Bachan. 2009. Prosodic issues in synthesising Thadou, a Tibeto-burman tone language. *Proceedings of InterSpeech 2009*, 500-503.
- Gibbon, Dafydd, Moses Ekpenyong, and Eno-Abasi Urua. 2010. Medefaidrin: resources documenting the birth and death language life-cycle. *Proceedings of the Language Resources and Evaluation Conference 2010*, 2702-2708. Paris: ELRA/ELDA.
- He, Hu. 2004. *Xinbian Putonghua Xunlian yu Ceshi Jiaocheng* [Textbook for Mandarin Training and Testing]. Lanzhou: Lanzhou University Press.
- Itahashi, Shuichi, and Chiu-yu Tseng. (eds.) 2010. *Computer Processing of Asian Spoken Languages*. Los Angeles: Consideration Books.
- Jansche, Martin. 1998. A two-level take on Tianjin tone. *Proceedings of the 10<sup>th</sup> European Summer School in Logic, Language and Information, Student Session*, 162-174. Saarbrücken: Universität des Saarlandes.

- Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge & New York: Cambridge University Press.
- McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge & New York: Cambridge University Press.
- Mehler, Alexander, Laurent Romary, and Dafydd Gibbon. (eds.) 2012. *Handbook of Technical Communication*. Berlin: De Gruyter Mouton.
- Tseng, Shu-Chuan. (ed.) 2009. *Linguistic Patterns in Spontaneous Speech*. Taipei: Institute of Linguistics, Academia Sinica.
- Van Eynde, Frank, and Dafydd Gibbon. 2000. *Lexicon Development for Speech and Language Processing*. Dordrecht & Boston: Kluwer.
- Witt, Andreas, and Dieter Metzger. 2009. *Linguistic Modeling of Information and Markup Languages: Contributions to Language Technology*. Dordrecht & London: Springer.

## 自然語言資源之於學術研究與技術研發應用的重要性

Dafydd Gibbon

畢勒費爾德大學／口語語音資料庫協調暨標準化國際委員會

語言科學與科技研究很重要的關鍵課題是如何為學術研究與發展提供適當而且可機讀的研究基礎，包含高品質的語料庫資源、語料分析的方法、詞彙庫與語法的建立、與軟體工具的開發。典藏與記錄語言學在過去十五年來已經發展成為應用語言學門的新次領域，特別是有關瀕危語言的語言記錄與其維護。對於自然語言科技來說，自然語言資源有文字、語音、與手勢等型態。為網路訊息散播與擷取、機器翻譯、與自動語音辨識與合成等相關領域的應用提供必要的基礎。本文詳述自然語言資源發展的緣起與動機。並且強調以可理解的語言符號架構為基本設計的資料與語料庫分析必須以計算模型呈現才能提供實證的、具體的語言學研究面向。本文實際以音節語法、聲調分析、與詞彙學等分析方式討論文獻手稿的記錄典藏方法與連讀變調計算模型的研究型態。

**關鍵詞：**典藏記錄語言學，自然語言科技，語料庫語言學，詞彙學，語音學