The Development of Spoken Language Resources in Oceania

J Bruce Millar

Computer Sciences Laboratory,
Research School of Information Sciences and Engineering,
Institute of Advanced Studies,
Australian National University,
{bruce.millar@anu.edu.au}

ABSTRACT
This paper examines the issues that surround the task of creating spoken language resources in the region of Oceania. The geographical extent and fragmented landmass, the linguistic complexity and the economic diversity are presented as highly significant influential features. The parallel interests of linguistic analysts and language technologists are acknowledged and the scope for synergistic collaboration is promoted. Current initiatives are described, prospective initiatives are discussed, and some potentially helpful enabling areas of study are proposed. The benefits of such development to the region is seen in scientific, economic and cultural terms.

1. INTRODUCTION
Motivations for developing spoken language data corpora in the Oceania region range from the desire to implement appropriate speech technology, to the documentation of nearly extinct languages. Each of these two diverse domains have their own agenda for language data collection but each can provide some synergism to the agenda of the other. This paper will review three major factors relating to corpora development: (1) the overall geographical and economic structure of the region, (2) the linguistic structure of the region, and (3) the current status of, and future prospects for spoken language data resources.

2. THE OCEANIA REGION
The region of Oceania comprises a land area of some 8.5 million square kilometres comprising a large number of islands scattered over about a third of the 165 million square kilometres of the Pacific ocean. It comprises 24 distinct territorial administrations many covering multiple small islands. It is home to 30 million people who have mostly high levels of literacy and who speak some 1500 native languages. It contains a wide range of levels of economic development with a GDP per capita range from 21000 to 800 US dollars. Its territories range from the highly developed, most populous, and largest island of Australia based mostly on European and latterly Asian emigrants, to its multitude of island states whose greatest virtue in the eyes of the rest of the world is their lack of development. This widespread and diverse corner of the globe presents some interesting challenges for the developer of languages resources.

The region of Oceania comprises the four sub-regions of Australasia, Melanesia, Micronesia and Polynesia. These sub-regions have broad geographical, cultural, economic and linguistic differences that influence the way in which the development of spoken language resources can proceed. It is a part of the world characterised by distance and isolation and immense diversity, especially in terms of its economics and its linguistics. The region is dominated in terms of area and population by Australia with the next two large territories, Papua New Guinea and New Zealand having less than one quarter of the population and much less area even though they are giants in both respects compared to the remainder of the region (see table 1).

3. LINGUISTIC STRUCTURE
The region is strongly influenced linguistically by its history of migration and colonisation. This has resulted in a legacy of official languages made up of 81% English, 17% English-based pidgin, 1.5% French, and 0.6% English/French. The official language will normally be the language of education (although some vernacular-based education is used at elementary level in some places), and of formal administration. The spoken use of the official language will be accent by the phonologies of the native languages of the speakers.

At the level of native language, the pattern is much more complex arising from a history of tribal isolation and subsequent migration activity. Tryon (1995) states that "Of the 5000-6000 languages spoken in the world today, some 2000, or nearly 40% are spoken on the islands in the Pacific and Indian Oceans and in Australia". These languages are divided by linguists into three distinct families: Austronesian, Papuan and Australian. Australian languages comprising some 234 living languages (Grimes, 1996), are spoken by some 44,000 Australians of Aboriginal origin (Australian Bureau of Statistics, 1997), Papuan languages comprising some 741 individual languages (Wurm, 1982), are spoken throughout Papua New Guinea, the second most populous country in Oceania, and in eastern parts of the Indonesian archipelago. The Austronesian languages comprise some 1200 languages but only about 500 of these lie in the region of Oceania (Tryon, 1995). Some 220 of these languages are spoken in Papua New Guinea in addition to the order of 500 Papuan languages.

It is clear that at the native language level, Oceania has one of the lowest population per language ratios in the world. This fact makes the development of language resources linguistically complex and economically difficult. While there are certain anomalies, the Melanesian region extending off the north-eastern coast of Australia northerly to Papua New Guinea and easterly to Fiji, is characterised linguistically by many languages per island group (Pawley, 1995). In this region the
number of speakers of a language can be extremely small (of the order of 1000). In contrast, the regions of Micronesia, a northerly extension of Melanesia, and Polynesia, an easterly and southerly extension of Melanesia, are characterised by a single language per island group. Thus the indigenous languages of Polynesia and Micronesia are characterised by both a larger territorial range and population size than the indigenous languages of Melanesia (Pawley, 1995).

There is however an intermediate linguistic level which represents the language of trade or other activities of forced contact between disparate language speakers. Such languages can span much larger populations than “native” languages, and can rise in status to “official” languages, as with Tok Pisin in Papua New Guinea. Some of these languages have developed from being second languages (L2) used mainly for trading purposes (hence labelled Pidgin languages) to being first languages (L1) for a small number of their speakers (hence labelled Creole languages). Within the region there are six contact languages having speaker populations which are significantly larger than the majority of native languages.

A critical feature that is not always a part of linguistic analyses of the region is that of the use to which each language is put in the community. While reasonable assumptions may be made, it will be important for efficient development of language resources to assess, in this highly multilingual environment, what forms of language are used for what purposes, and in which languages the literacy of the populations are expressed.

The 1990 PNG Census asked questions about education, language and literacy. These data, now available on the web, provide “self-reported” evidence of the levels of literacy in English, two contact languages (Tok Pisin and Motu), and in any local language. It is likely that these data will be supplemented by similar material from the next census in the Solomon Islands but in Vanuatu there are no plans to include such questions (Early, 2000).

<table>
<thead>
<tr>
<th>COUNTRY</th>
<th>Area (sq. km)</th>
<th>Pop’n (000s)</th>
<th>Languages Spoken</th>
<th>GDP</th>
<th>Economy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitcairn Islands</td>
<td>47</td>
<td>0</td>
<td>English + Pitcairn + Tahitian</td>
<td>n/a</td>
<td>Subsistence+Fishing</td>
</tr>
<tr>
<td>Tokelau</td>
<td>10</td>
<td>1</td>
<td>English + Tokelauan</td>
<td>1000</td>
<td>Copra+Crafts</td>
</tr>
<tr>
<td>Niue</td>
<td>260</td>
<td>2</td>
<td>English + Niuean</td>
<td>1200</td>
<td>Subsistence</td>
</tr>
<tr>
<td>Nauru</td>
<td>21</td>
<td>11</td>
<td>English + Nauruan</td>
<td>10000</td>
<td>Phosphate(exhausted)</td>
</tr>
<tr>
<td>Tuvalu</td>
<td>26</td>
<td>11</td>
<td>English + Tuvaluan</td>
<td>800</td>
<td>Subsistence</td>
</tr>
<tr>
<td>Wallis and Futuna</td>
<td>274</td>
<td>15</td>
<td>French + Wallisian</td>
<td>2000</td>
<td>Subsistence</td>
</tr>
<tr>
<td>Palau</td>
<td>458</td>
<td>18</td>
<td>English + Palaun + 4 others</td>
<td>8800</td>
<td>Subsistence+Fishing</td>
</tr>
<tr>
<td>Cook Islands</td>
<td>240</td>
<td>20</td>
<td>English + 4 other languages</td>
<td>4000</td>
<td>Supported by NZ</td>
</tr>
<tr>
<td>American Samoa</td>
<td>199</td>
<td>64</td>
<td>English + Samoan</td>
<td>2600</td>
<td>Fishing</td>
</tr>
<tr>
<td>Marshall Islands</td>
<td>181</td>
<td>65</td>
<td>English + Japanese + 2 others</td>
<td>1450</td>
<td>Copra+Crafts</td>
</tr>
<tr>
<td>Northern Marianas</td>
<td>477</td>
<td>69</td>
<td>English + Chuuk + Carolinian</td>
<td>9300</td>
<td>Tourist+Textiles</td>
</tr>
<tr>
<td>Kiribati</td>
<td>717</td>
<td>85</td>
<td>English + Gilbertese</td>
<td>800</td>
<td>Copra+Fish</td>
</tr>
<tr>
<td>Tonga</td>
<td>748</td>
<td>109</td>
<td>English + Tongan + 2 others</td>
<td>2100</td>
<td>Agric+Tourism</td>
</tr>
<tr>
<td>Micronesia</td>
<td>702</td>
<td>131</td>
<td>English + 4 other languages</td>
<td>1760</td>
<td>Farming &amp; Fishing</td>
</tr>
<tr>
<td>Guam</td>
<td>541</td>
<td>152</td>
<td>English + Chuuk + Japanese</td>
<td>19,000</td>
<td>Military+Tourism</td>
</tr>
<tr>
<td>Vanuatu</td>
<td>14,760</td>
<td>189</td>
<td>English + French + 109 others</td>
<td>1300</td>
<td>Fishing,Finance,Tourism</td>
</tr>
<tr>
<td>New Caledonia</td>
<td>19,060</td>
<td>197</td>
<td>French + 38 other languages</td>
<td>11400</td>
<td>Nickel+Tourism</td>
</tr>
<tr>
<td>French Polynesia</td>
<td>4,167</td>
<td>242</td>
<td>French + Tahitian</td>
<td>10,800</td>
<td>Military+Tourism</td>
</tr>
<tr>
<td>Western Samoa</td>
<td>2,860</td>
<td>209</td>
<td>English + Samoan</td>
<td>2100</td>
<td>Agriculture</td>
</tr>
<tr>
<td>Solomon Islands</td>
<td>28,450</td>
<td>455</td>
<td>English + Fiji+ 120 other languages</td>
<td>2600</td>
<td>Fishing+Forestry</td>
</tr>
<tr>
<td>Fiji</td>
<td>18,270</td>
<td>812</td>
<td>English + Fijian + Hindi + 7 others</td>
<td>6700</td>
<td>Sugar+Tourism</td>
</tr>
<tr>
<td>New Zealand</td>
<td>282,680</td>
<td>3662</td>
<td>English + Maori</td>
<td>1700</td>
<td>Near Western</td>
</tr>
<tr>
<td>Papua New Guinea</td>
<td>462,840</td>
<td>4705</td>
<td>English + Tok Pisin + 714 others</td>
<td>2400</td>
<td>Agri+Minerals</td>
</tr>
<tr>
<td>Australia</td>
<td>7,686,830</td>
<td>18783</td>
<td>English + 234 other languages</td>
<td>21,200</td>
<td>Western</td>
</tr>
</tbody>
</table>

Table 1. Areas, populations, economy (CIA, 1999) and languages (Grimes, 1996) of the countries of the Oceania region

3.1 Spoken Language Data Resources

Spoken language resources in the region, in the form of audio, video, and text materials, have two distinct sources that have developed in isolation from each other. Field linguists have been collecting data in the region for many decades. Their public archiving of these data has typically been limited to text in print media. Speech technologists have tended to focus only on areas with strong economies and a high degree of linguistic uniformity. Their public archiving has spanned the material domains but with limited linguistic analysis. A collaborative phase that will bring synergistic benefit to both source communities is clearly needed.

3.1.1 The linguistic analysis perspective

A very large informal base of data exists in the private collections of many linguists who have performed intensive analyses in the region for many years. These data are the focus of initiatives involving three Australian universities to establish effective archiving of such data in digital form so that it is
Academics at the Australian National University in the Research School of Pacific and Asian Studies (RSPAS, 2000) have had a long involvement in the documentation of the linguistics of the region (e.g. Wurm, 1972; 1982; Tryon, 1995). They and their students and colleagues in other Australian universities provide a significant linguistic analysis resource.

In New Zealand, a corpus comprising 15 different styles of speech (monologue and dialogue) with some balance across speakers and across gender has been created (Holmes et al., 1998). Transcripts of the material are made available to researchers and the audio data is available for research at the University of Victoria in Wellington (Vine, 2000).

3.1.2 The language technology perspective

The limiting factor in the development of speech recognition systems in the region is the need to acquire adequate data resources of local language patterns. Initial interest in the development of these resources arose in academia (e.g. Millar et al, 1990, 1994, 1997). While this development has been slow owing to labour intensive collection and annotation techniques, there has arisen more recently a rapid expansion of speech data resources for telephone speech using the specific opportunities of commercial enterprises. One technique has been to implement a rather simple speech recognition task using data models from a distant English speaking community, where the perplexity of each step is limited in order to allow sufficient accuracy to be obtained (Forsyth, 2000). If each recognised spoken token is also verified by the user, then any use of this system can generate reliable data from which to develop local speech data resources. As these resources are expanded they can be used to build more accurate phone models which can support adequate accuracy with more difficult tasks. In this way the telephone speech applications can grow without the very heavy overheads of initial data model development. An alternative approach is to conduct an explicit data collection exercise and then integrate the data acquired with a distant model. Both kinds of process do, of course, produce data resources that are jealously guarded by their commercial developers as they have been gained by lengthy processes. Commercial enterprises do not wish to give competitors the opportunity to build sophisticated applications without the time delay inherent in their development.

A recent exception to the above is a data corpus of telephone speech of Australian English collected by CallBase Databases Ltd using the designs of the European SpeechDat-II consortium (CallBase, 2000). Speakers were solicited by conducting a mailing campaign covering most population centres in Australia. The published corpus includes the data from 1000 respondents who represent well the geographic spread of the population. The varieties of Australian English spoken by those born in Australia tend to be socio-economically determined and may not be sampled evenly simply by means of even population sampling. Some 20% of the respondents were first generation migrants from at least 50 different countries. This proportion matches well the population average of 23% overseas-born citizens but shows an over sampling of those of Chinese, South African, and American origin and an undersampling of those of British and Greek origin. The material includes a mix of phonetically rich sentence and isolated word material, a range of spoken number formats, some local proper names, date, time, and money phrases, and a spelling sequence.

This corpus is the only known non-proprietary source of telephone speech data for Australian English and as such represents a useful resource for researchers or small companies who lack the capacity to develop their own. It is being made available commercially through the European Language Resources Association. It may be used as a base on which to build larger corpora using the bootstrapping techniques mentioned above or to adapt existing phone models from other varieties of English.

4. SOME REGIONAL INITIATIVES

It is clear that speech technology can enter any English speaking population at a level of low perplexity recognition, and can be developed for higher perplexity recognition for applications supported by the 18 million population and western economy of Australia. This development from low perplexity recognition is yet to be demonstrated in other parts of the region. As English has high incidence as a official language in the region, the repetition of this process for smaller populations speaking an accented form of English can be foreseen from a linguistic stance. The existence of phonological similarity between some of the native languages of the region should facilitate this.

The penetration of English-based recognition of limited perplexity will clearly depend on the degree of accenting relative to the externally developed phone models. Performance deficits of USA English based systems when applied to Australian speakers have been examined but not to my knowledge in a way that is explicitly related to perplexity. Studies that examine the “data increment” that is required to “seed” effective transformation of recognition performance, for a defined task perplexity, from one accented form of English to another would indeed shed light on the likelihood of success for this approach.

Another useful initiative that could be examined is an analysis of the likelihood of telephone-based information access based on a contact language such as “Tok Pisin” which is accepted as the official language of 17% of the population of the region. It seems clear that Tok Pisin has the semantic capacity to readily handle spoken information queries. There may be large phonological variance in the realisation of Tok Pisin, which is within the bounds of cross-speaker intelligibility but would strain the viability of current speech recognition techniques. As a “contact” language it seems likely that the phonological mappings from its native languages may be less contorted than for an external official language such as English. It is significant for this approach that Tok Pisin has been systematised into a grammar and dictionary so the necessary information on which to build components in addition to phoneme models required for speech recognition systems are available (Mihalic, 1971).
5. CONCLUSIONS

Many interesting challenges, opportunities, and stimuli for the development of spoken language resources in the region of Oceania have been presented. As language, either spoken or written, is the major means of human information transfer and as information is becoming the dominant commodity in the developed world, then there is a strong impetus to assess these challenges, opportunities, and stimuli so that the people of Oceania can benefit. The development of digital language resources that encompass but extend beyond the strong economies of the region by building on years of painstaking linguistic analysis can provide a culturally sensitive bridge between the region and the highly-developed world that in some ways appears to be leaving it behind.

The tangible benefits that can start to flow in the region from the establishment of language resources are many. Enhanced confidence in the utility and relevance of “contact” or “local” languages is fundamental. It will grow through the introduction of word-processing with spell-checkers and appropriate hyphenation. It will proceed with simple then more complex use of speech technology for interrogating local and distant data repositories by telephone access to local automated call centres.

The region presents a unique challenge in its linguistic, communications, and cultural composition. There are causes of pessimism when we look at the distances and the cost of telecommunications but there are also causes of optimism when we consider the past evidence of goodwill and cooperation that have developed between the major economies and those in need of assistance in the region. Appropriate partnerships which link across scientific disciplines and across economics are likely to be the enablers of development of robust language resources in Oceania.

6. ACKNOWLEDGMENTS

An assessment of the kind presented in this paper is of necessity highly reliant on multi-disciplinary input. In preparing this paper I am highly indebted to the valuable insights of many colleagues who have all freely shared their expert insights with me. I specifically acknowledge the assistance of following people: Andrew Pawley, Malcom Ross, Darrel Tryon and Gerard Ward (Research School of Pacific and Asian Studies at ANU), Robert Early (University of the South Pacific, Vanuatu), Paul Warren and Bernadette Vine (University of Victoria), Ross Caughley and Robert Litteral (SIL International), Andrew Simpson (CallBase Databases Ltd), Mark Forsyth (Voicenet Australia). All responsibility for any errors introduced into this paper rests with the author.

7. REFERENCES

5. Early, R. Personal Communications on 12 April and 24 July 2000