EUROCALL Nottingham 2011

The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

The Call Triangle: student, teacher and institution

An electronic dictionary and translation system for Murrinh-Patha

Melanie Seiss[a], Rachel Nordlinger[b]*

[a]*University of Konstanz, Germany*
[b]*University of Melbourne, Australia*

Abstract

This paper presents an electronic dictionary and translation system for the Australian language Murrinh-Patha. Its complex verbal structure makes learning Murrinh-Patha very difficult. Design learning materials or a dictionary which is easy to understand and to use also presents a challenge. This paper discusses some of the difficulties posed by the Murrinh-Patha verb system and proposes electronic resources which build on deep language processing to perform the required tasks.

Keywords: intelligent CALL; electronic dictionary; machine translation; endangered languages

1. Introduction

Murrinh-Patha is a polysynthetic language spoken in the Northern Territory of Australia (e.g. Blythe 2009). Many aspects of the grammar of Murrinh-Patha make the language difficult to learn as a second language. In fact, these aspects even make it difficult to design a dictionary which is easy for non-native speakers to understand and to use. In this paper, we present an electronic dictionary and translation system which is intended to help non-Murrinh-Patha speakers to learn and understand simple Murrinh-Patha sentences. The system uses deep language processing to overcome the challenges posed by the Murrinh-Patha verb.

2. Murrinh-Patha speakers and the language situation

Murrinh-Patha is spoken by approximately 3,000 people in and around Wadeye (Port Keats), a small community approximately 400 kilometers south of Darwin. As has been documented by Kelly et al. (2010), the language is, despite its actual small number of speakers, not considered endangered by the Murrinh-Patha speakers themselves. Murrinh-Patha is the first language of most of the speakers in the community. While older community members (45+) have reasonable knowledge of English as a second language due to their mission experiences, younger speakers have only little knowledge of English. The language of daily interaction in the community is Murrinh-Patha and parents expect that children will learn English at school. Wadeye is relatively isolated, which means that there is little incidental traffic or visitors and consequently, English is not often used in daily life.

* *E-mail addresses*: melanie.seiss@uni-konstanz.de, racheln@unimelb.edu.au

However, some English speaking people such as school teachers or nurses usually live in Wadeye for a shorter or longer period of time. The electronic applications we present are intended for people from this group who would like to learn some initial vocabulary of Murrinh-Patha beyond the simplest phrases. In a further step, however, the implementation of the Murrinh-Patha grammar could also be used to design CALL applications for Murrinh-Patha speakers to learn English.

3. Complexities of the Murrinh-Patha verb

In this section, the challenges a learner of Murrinh-Patha faces and which also make designing a Murrinh-Patha dictionary difficult are discussed. Most of the complexity of the Murrinh-Patha system lies in its complex verbal structure (Nordlinger 2010) and the phonological changes that apply when the verbal complex is constructed.

The verb may consist of up to 9 different morphemes. The lexical component of the verb, often called the lexical stem, is deeply embedded inside the verbal word. This can be seen in (1), in which the lexical stems *rta* and *bert* mainly determine the semantic meaning of the verb.

(1) a. mangan-nhi-**rta-**ngintha 'We two hugged you (sg).'
b. mangan-nhi-**bert-**ngintha 'We two grabbed you (sg).'

The first morpheme in the verbal complex, *mangan* in (1), is called the classifier stem. There are probably 38 different classifier stems which have, unlike auxiliaries, finely distinct semantic meanings. Each of the 38 classifier stems has approximately 50 different surface forms as the classifier stems inflect for tense, person and number. The examples in (2) show verbs with a different classifier stem, *ba*.

(2) a. **ba**-nhi-**ngkardu**-nu-ngintha 'We two will see you (sg).'
b. **ba**-ngintha-**warnta**-nu 'We two will split it open.'

Which classifier stem is chosen is determined by the lexical stem, i.e. lexical stems select for classifier stems. A lexical stem may combine with one or more classifier stems and the combination determines the complete lexical meaning of the verb.

Designing a paper dictionary for Murrinh-Patha verbs has two options, both of which are problematic: The first option would be to list all forms of the classifier stem with the lexical stem. However, this is impractical as this would involve over 50 entries (for all the forms of the classifier stem) of the same verb. Moreover, other material, such as markers for direct objects (*nhi* in (1) and (2a)) and subject number (*ngintha* in (2b)), can also intervene between classifier and lexical stem in the second morpheme slot.

The alternative option is to list the lexical stem and classifier stem as distinct entries. This is what is done in the dictionary of the related language Ngan'gi (Reid & McTaggart 2008). However, users have to be very advanced in their understanding of the verbal structure to be able to use such a dictionary. They have to know how to decompose a verb into its various morphemes to be able to extract the lexical stem and look it up in the dictionary.

This is made even more difficult due to the high degree of syncretism in the classifier stem forms and due to the application of phonological rules to morpheme combinations. As can be seen in (3), when the lexical stem *ngkardu* combines with the classifier stem *bam*, the nasalization *ng* is lost and the actual surface form is *bamkardu*.

(3) ba**mk**ardu ~ bam-ngkardu 'He/she saw him/her.'

Such phonological processes make it difficult to decompose the verb into its single morphemes unless one already has an advanced knowledge of the language. Additionally, there are many more complexities which make learning Murrinh-Patha difficult. For example, Murrinh-Patha distinguishes 7 different number-related categories in total. There is a distinction between singular, dual, paucal (small groups) and plural, a distinction between sibling and non-sibling in the dual and paucal categories as well as a distinction between female and male in the non-sibling category. This system is further complicated by the fact that these categories are encoded in different parts of the verb, i.e. the categories are determined by a combination of the inflection on the classifier stem and separate morphemes which appear later in the verbal word.

As has been mentioned above, the Murrinh-Patha verb may consist of up to 9 different morphemes. This in itself is already quite complex. However, the system is even more complex and difficult to learn

because dependencies between the different morphemes exist. For example, as can be seen in (4), the dual subject number marker *ngintha* usually attaches between the classifier and lexical stem as in (4a). However, if a direct object marker is present, e.g. *nhi* in (4b), the subject marker can only be realized after the lexical stem.

(4) a. Bam-**ngintha**-ngkardu 'We 2 saw it.'
b. Bam-**nhi**-ngkardu-**ngintha** 'We 2 saw you.'

These complexities make learning Murrinh-Patha and using paper dictionaries to learn Murrinh-Patha very difficult. In the following section we present some electronic resources we built which are intended to make learning Murrinh-Patha easier.

4. Building electronic resources for Murrinh-Patha

The system we present does not assume an advanced understanding of Murrinh-Patha. It can be used as a simple look-up system as well as a tool which brings the learner closer to understanding the complexities of the language. The system comprises different parts.

Part one is a translation system which can translate simple English sentences into Murrinh-Patha. It is especially intended as a resource which helps to find the correct verb form and to study the structure of the verbal complex. As the Murrinh-Patha number system is more complex than the English one, the system asks the user to disambiguate the English input when a plural is used. As output, the user is presented with the Murrinh-Patha sentence including the phonological changes. Additionally, however, the user may obtain information about the different stems and markers used to build the verb complex.

Part two comprises an electronic dictionary. It can be used to look up Murrinh-Patha words, phrases and sentences. It offers the user English glosses and paraphrases which may be more helpful to the user than a plain sentence translation as in this case, finer-grained meaning distinctions may be preserved. Additionally, the user may generate the verb they searched for with different number and person information.

The system uses deep language processing to perform the required tasks. The user input is automatically analyzed linguistically. For the morphological analysis, an xfst morphology (Beesley and Karrtunen 2003) has been compiled, which is able to decompose the complex verb into its morphemes. This is used in the electronic dictionary. First, the user input is analyzed and the lexical stem and classifier stem are extracted. The system then looks up the combination in an internal dictionary and presents the user with the dictionary entry for the combination. Thus, the system performs the morphological analysis for the user.

The system also includes an XLE grammar (Crouch et al 2011) for Murrinh-Patha and English. These grammars carry out a syntactic analysis, e.g. they analyze the sentence with respect to subject, object etc. The grammars are used in the translation system in combination with XFR rewrite rules (Crouch et al 2011). The basic idea behind this translation system is that the user input is analyzed by the English grammar, which builds an abstract representation of the sentence. Then the English words are translated on a word by word basis into Murrinh-Patha. Finally, the Murrinh-Patha grammar generates a valid sentence from the abstract representation. This ensures that the Murrinh-Patha output is always grammatical, which is important in a learning system.

5. Conclusion

This paper presented challenges posed by the Australian language Murrinh-Patha, both to language learners and to designers of learning materials and dictionaries. These challenges can be addressed by developing applications which are able to perform linguistic analyses of the user input and generate grammatical output. The applications thus show that deep language processing can be very helpful in designing applications for computer assisted language learning.

6. References

Beesley, K. R. & Karttunen, L. (2003). Finite State Morphology. Stanford: CSLI Publications.

Blythe, J. (2009). *Doing Referring in Murriny Patha conversation*. Syndey, University of Sydney, PhD thesis.

Crouch, D., Dalrymple, M., Kaplan, R., King, T., Maxwell, J.T. & Newman, P. (2011). *XLE Documentation*. Palo Alto Research Center.

Nordlinger, R. (2010). Verbal Morphology in Murrinh-Patha: Evidence for Templates. *Morphology 20(2),* 321-341.

Kelly, B., Nordlinger, R. & Wigglesworth, G. (2010). Indigenous perspectives on the vitality of Murrinh-Patha. In Y. de Treis & R. de Busser (Eds.), *Selected Papers from the 2009 Conference of the Australian Linguistics Society.*

Reid, N & McTaggart, P. (2008). *Ngan'gi Dictionary*. Armidale, NSW: Australian Linguistics Press.