A Machine Translation Toolchain for Polysynthetic Languages

Petr Homola
Codesign, s.r.o.
Palackého 541
252 29 Dobřichovice
phomola@codesign.cz

Abstract

We present a set of free tools for building rule-based machine translation systems for polysynthetic languages. As there are no large corpora for most of the "small" languages, it is often impossible to use statistical methods. There are some free MT tools but very little work has been done on polysynthetic languages. The aim of this project is to provide computational tools for morphological and syntactic processing for such languages.

1 Introduction

The paper describes a set of tools for natural processing of polysynthetic languages. There are quite a few definitions of polysynthesis. Baker (1996), for example, defines a 'polysynthesis parameter' within the Chomskyan framework. However his definition is quite strict and excludes many languages that are traditionally considered polysynthetic (such as Greenlandic). We use Mattissen's (2006) definition which is closer to the understanding of polysynthesis of most researchers in the field. According to her, a language is polysynthetic if it contains "complex, polymorphemic verb forms which allow, within one word unit, for components in the form of non-root bound morphemes with quite 'lexical' meaning or optionally for the concatenation of lexical roots".

Due to typological differences from Western languages, polysynthetic languages are quite a challenge for many theories of formal grammar. Our implementation is based on Lexical Functional Grammar (Kaplan and Bresnan, 1982; Bresnan, 2001). The system consists of a morphological analyzer, rule-based parser, transfer module and morphological generator. As an example of a polysynthetic word, consider the Aymara sentence *qullqinipachänwa* which corresponds to a complete English sentence:

(1) *qullqi-ni-pacha-:n-wa*
money-POSS-EVID-PAST$_{3\to 3}$-FOC
"Apparently s/he had a lot of money."

The paper is organized as follows: Section 2 describes how we analyze polysynthetic languages morphologically and syntactically. Section 3 gives an overview of the transfer phase. Finally we offer some conclusions in Section 4.

2 Morphological and Syntactic Analysis

2.1 Lexicon

Some polysynthetic languages, such as Aymara, have no closed morphological tagset since a stem can be nominalized and/or verbalized several times by adding various derivational suffixes recursively without any theoretical limit. The output of the morphological analyzer is a set of f(eature)-structures which contain morphosyntactic and lexico-semantic information. For example, the f-structure for the Aymara word form *uñjsma* "I see/saw you" is defined by the following morpholexical annotation:

$$(2)\quad \begin{array}{l} (\uparrow \text{PRED}) = \text{`see}\langle(\uparrow \text{SUBJ})(\uparrow \text{OBJ})\rangle\text{'} \\ (\uparrow \text{TENSE}) = \text{pres}|\text{simple\_past} \\ (\uparrow \text{SUBJ PERSON}) = 1 \\ ((\uparrow \text{SUBJ PRED}) = \text{`pro'}) \\ (\uparrow \text{OBJ PERSON}) = 2 \\ ((\uparrow \text{OBJ PRED}) = \text{`pro'}) \end{array}$$

The functional equations in (2) encode the verb's lemma and valency (in the PRED attribute),

© 2012 European Association for Machine Translation.

polypersonal agreement (the person of SUBJ and OBJ) and the fact that both arguments can be dropped (in which case the value of the argument's PRED attribute is 'pro').

The lexicon contains an entry for the stem and a separate entry for the suffix:[1]

```
(r v1 uñj uñja (v2)
  ((SUBJ ((ANIM 1)))) - V)
(s v2 sma (tf) ((TENSE nfut)
  (SUBJ ((PERS 1) (PRED pro ?))
   OBJ ((PERS 2) (PRED pro ?)))))
```

Valence is defined in a separate file together with lexical rules. It contains the category, lemma, a list of grammatical functions (SUBJ and OBJ, **!** means that the GF is mandatory, **?** means that it is optional) and corresponding semantic roles (ACT(OR) and PAT(IENT)).

```
(V uñja (! SUBJ ACT) (! OBJ PAT))
```

2.2 Syntax

Many polysynthetic languages are nonconfigurational. Hale (1983) was the first to define and describe nonconfigurationality and its impact to syntax. The general rule which describes the structure of matrix sentences is lexocentric (see (Bresnan, 2001) for more examples):

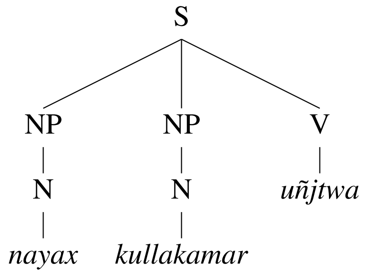
(3) $$S \rightarrow \mathcal{C}^{+}$$

$$\text{where } \mathcal{C} \text{ is } \underset{\uparrow=\downarrow}{\mathrm{V}} \text{ or } \underset{(\uparrow \mathrm{GF})=\downarrow}{\mathrm{NP} \mid \mathrm{PP}}$$

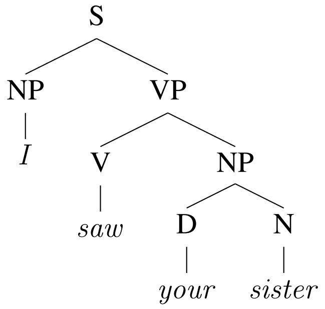
As an example, compare the c(onstituent)-structure (5) of the Aymara sentence given in (4) with the c-structure (6) of its English translation. The corresponding f-structures (the English one is given in (7)) are structurally identical and differ only in the values of the PRED attributes.

(4) *Naya-x kullaka-ma-r*
I-TOP sister-POSS2-ALL
uñj-t-wa
see-SIMPLE_PAST1-FOC
"I saw your sister."

(5)



(6)



(7) $$\begin{bmatrix} \mathrm{PRED} & \text{`see}\langle(\uparrow\mathrm{SUBJ})(\uparrow\mathrm{OBJ})\rangle\text{'} \\ \mathrm{TENSE} & \textsc{past} \\ \mathrm{SUBJ} & \begin{bmatrix}\mathrm{PRED} & \text{`I'} \\ \mathrm{PERSON} & 1\end{bmatrix} \\ \mathrm{OBJ} & \begin{bmatrix}\mathrm{PRED} & \text{`sister'} \\ \mathrm{CASE} & \textsc{all} \\ \mathrm{POSS} & \begin{bmatrix}\mathrm{PRED} & \text{`pro'} \\ \mathrm{PERSON} & 2\end{bmatrix}\end{bmatrix}\end{bmatrix}$$

As can be seen, the c-structure of the Aymara sentence is flat since the language has no VP. As has been pointed out by Kruijff (2000), phrase structures represent the process of syntactic derivation whereas f-structures (which roughly correspond to dependency trees in depedency-based grammars) are the result of this derivation. Hale (1983) argues that in this kind of languages phrase structures do not encode syntactic relations but only word order.

However, most polysynthetic languages are discourse-configurational and if there are no articles nor other markers which would express information structure, constituency has to be used to analyze topic-focus articulation. So while the lexocentric rule given in (3) is approriate for the analysis of languages such as Aymara and Quechua because they have topic and focus markers (the suffixes *-x* and *-wa* in (4)), languages like Abkhaz or Guaraní express information structure mainly by word order. We use a set of X′-rules that are very similar to what Meurer (2007) uses for Georgian. The core of the context-free grammar is given in (8).

[1] Stem entries are denoted by **r** and contain the form as it occurs in the word (*uñj*), lemma (*uñja*), start and end state in the corresponding finite state automaton (**v1** and **v2** respectively), attribute-value pairs for the f-structure and category for c-structures (**V** for verbs etc.). Suffix entries are denoted by **s** and contain the states of the automaton (**v2** and **tf**), the form of the suffix (*-sma*) and attribute-value pairs for the f-structure associated with the suffix.

(8)
S → XP$^+$
I$'$ → I (S)
IP → (XP) I$'$
IP → XP IP

The subtree headed by S belongs to the focus, the verb and the specifier of I$'$ may belong to the topic or to the focus and all XPs adjoined to IP are part of the topic. Independent i(nformation)-structures introduced by King (1997) are used to capture topic-focus articulation.

2.3 Valency

Valency is very important for the correct assignment of grammatical functions. As an example, let us have a look at Abkhaz, an ergative language with transitive and intransitive bivalent verbs that have different morphosyntactic alignment. Compare, for example, the order of personal affixes in (9) and (10).

(9) *У-з-б-оит*
OBJ2SG,MASC-SUBJ1SG-see-PRES
"I see you."

(10) *С-у-с-уеит*
SUBJ1SG-IOBJ2SG,MASC-hit-PRES
"I hit you."

As can be seen, some Abkhaz bivalent intransitive verbs such as *асъıара* "to hit" are translated as transitive verbs in English. This situation is somewhat similar to oblique objects in Turkish (Çetinoğlu and Butt, 2008). We use a valency lexicon of verbs that contains both grammatical functions and semantic roles. The roles are used in the transfer phase.

3 Transfer

The transfer module can be used for experiments with direct (word-to-word), shallow (NPs and PPs) and deep syntactic transfer. The output of the transfer module can be used to calculate WER (word error rate) if reference translations are available.

3.1 Structural Transfer

In the LFG framework, the transfer module usually operates on f-structures (Kaplan et al., 1989). However, f-structures are too language-specific (cf. the f-structure of (10) with the f-structure of its English translation which differ in grammatical functions). The inventory and use of grammatical functions is language-specific (there are, for example, languages without secondary objects, with double subjects etc.) which suggests that one should abstract from them and use a more general concept instead. In LFG, a-structures with thematic roles seem to be more suitable for bilingual transfer.

We use deep syntax trees (henceforth DSTs) in the transfer phase. DSTs can be obtained automatically from interlinked c-structures, f-structures, i-structures and a-structures using the following algorithm:

1. F-structures can be interpreted as dependency trees with autosemantic words (i.e., f-structures with the PRED attribute) as nodes and grammatical functions as edge labels.[2]

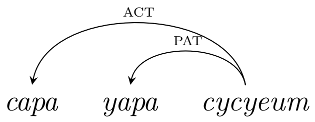
2. Annotate the edges of the DST with thematic roles (using the grammatical functions from the f-structure and lexical mapping).

3. Order the nodes using information structure (see (Sgall et al., 1986) for discussion).

Let us use a variant of (10) as an example:

(11) *Сара уара*
I you-2SG,MASC
с-у-с-уеит
SUBJ1SG-IOBJ2SG,MASC-hit-PRES
"I hit you."

Using the algorithm sketched above, the f-structure of (11) yields the DST in (12):

(12)


It is obvious that the tree in (12) is identical in Abkhaz and English (except for the PRED values) whereas the f-structures are different (PAT corresponds to OBL$_\theta$ in Abkhaz and to OBJ on English).

Table 1 summarizes what various LFG layers contribute to DSTs.

[2] Generally, we get a directed acyclic graph (DAG). However, edges resulting from structure sharing can be interpreted as coreferences and ignored in DSTs. Formally, we get DSTs from DAGs induced from f-structures as minimum spanning trees. We use Prim's algorithm where the weight of edges is their distance from the root node.

LFG layer	information in DSTs
c-structure	original word order
f-structure	dependencies and coreferences
i-structure	topic-focus articulation
a-structure	thematic roles

Table 1: Information provided by LFG layers to DSTs

3.2 Lexical Transfer

Having converted f-structures to DSTs, the transfer is mostly lexical, i.e., the PRED values associated with nodes are translated to the target language. Because the translation of many words depends on the context, word sense disambiguation (WSD) is needed. Nonetheless, this is a very complicated problems itself and as a semantic and pragmatic task it is independent of the syntactic framework of LFG or any other rule-based parser.

A very simple and comparatively viable solution is the use of a statistical ranker that selects the most probable translation according to a language model. Thus in our experiments we nondeterministically generate all possible translations and select the best sentence using a trigram based target language model.

A simple bilingual entry for the pair Aymara-English looks as follows:

```
(l V ((PRED uñja)) ((PRED see)) ())
```

It contains the category (to distinguish between identical word forms with different POS tags, such as *book* in English), a skeletal f-structure for the source language and an f-structure for the target language (most entries contain only the PRED attribute).

4 Conclusions

We have presented a set of tools developed for natural language processing of polysynthetic languages. Examples given in this paper demonstrate several typological features of polysynthetic languages which do not occur in well-researched Western languages and show how we analyze them in the LFG framework.

To test the tools we have developed an MT system from Aymara to Quechua. The WER (word error rate) measured on narrative texts is around 10%. An ongoing experiment with translation from Aymara to English indicates a WER around 30% but final results are not available yet.

Linguistic resources used in the modules are defined in separate files, there are files for the morphological lexicon, parser rules, valence lexicon (valence frames and lexical rules) and transfer (structural and lexical rules). The code is strictly separated from data. All tools are implemented in portable C++ (using the new C++11 standard and STL) and were tested on Mac OS X (clang/LLVM) and MS Windows (Visual Studio 2010).

References

Baker, Mark C. 1996. *The Polysynthesis Parameter*. Oxford University Press.

Bresnan, Joan. 2001. *Lexical-Functional Syntax*. Blackwell Textbooks in Linguistics, New York.

Çetinoğlu, Özlem and Miriam Butt. 2008. Turkish Non-canonical Objects. In Butt, Miriam and Tracy Holloway King, editors, *Proceedings of the LFG Conference*.

Hale, Kenneth L. 1983. Warlpiri and the grammar of non-configurational languages. *Natural Language & Linguistic Theory*, 1:5–47.

Kaplan, Ronald M. and Joan Bresnan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In Bresnan, Joan, editor, *Mental Representation of Grammatical Relations*. MIT Press, Cambridge.

Kaplan, Ronald M., Klaus Netter, Jürgen Wedekind, and Annie Zaenen. 1989. Translation By Structural Correspondences. In *Proceedings of 4th EACL*, pages 272–281.

King, Tracy Holloway. 1997. Focus Domains and Information-Structure. In Butt, Miriam and Tracy Holloway King, editors, *Proceedings of the LFG Conference*.

Kruijff, Geert-Kan. 2000. A Dependency-based Grammar. Technical report, Charles University, Prague, Czech Republic.

Mattissen, Joanna. 2006. On the Ontology and Diachrony of Polysynthesis. In Wunderlich, Dieter, editor, *Advances in the theory of the lexicon*, pages 287–354. Walter de Gruyter, Berlin.

Meurer, Paul. 2007. A computational grammar for Georgian. In *Proceedings of the 7th International Tbilisi Conference on Logic, Language, and Computation*.

Sgall, Petr, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. D. Reider Publishing Company.