Parsing a Polysynthetic Language

Petr Homola
Codesign, s.r.o.
phomola@codesign.cz

Abstract

We present the results of a project of building a lexical-functional grammar of Aymara, an Amerindian language. There was almost no research on Aymara in computational linguistics to date. The goal of the project is two-fold: First, we want to provide a formal description of the language. Second, NLP resources (lexicon and grammar) are being developed that could be used in machine translation and other NLP tasks. The paper presents formal description of selected properties of Aymara which are uncommon in well-researched Western languages. Furthermore, we present an experimental machine translation system into Spanish and English.

1 Introduction

Aymara is an Amerindian language spoken in Bolivia, Chile and Peru by approx. two million people. It is a polysynthetic language that has many lexical and structural similarities with Quechua but the often suggested genetic relationship between these languages is still disputed.

The only research on Aymara in the field of computational linguistics we know about is the project described in (Beesley, 2006). The presented project uses Lexical-Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Bresnan, 2001) to formally describe the lexicon, morphology and syntax of Aymara in a manner suitable for natural language processing (NLP). The grammar we have implemented is capable of parsing complex sentences with embedded clauses. We have also done experiments with machine translation (MT) into Spanish and English; the results are presented in Section 4.

Aymara is a polysynthetic language with a very complicated system of polypersonal agreement (see Section 2.3 for a brief description). A rare property of words in Aymara is the so-called vowel elision (sometimes called 'subtractive morphology') which is quite hard to describe formally. We show how vowel elision can be dealt with in the lexicon.

The paper is organized as follows: Section 2 presents selected properties of Aymara, many of them absent from well-researched languages such as English, and their formal analysis in LFG. Section 3 introduces a dependency-based abstraction of f-structures which brings formal grammars closer cross-linguistically. Section 4 describes our experiment with MT from Aymara into Spanish and English. Finally, we conclude in Section 5 and give an outlook for further research.

2 Some Properties of Aymara

In this section, we focus on some properties of Aymara at the level of morphology and syntax which are mostly absent from Western languages such as English, and sketch their analysis in LFG. A detailed description of the language can be found in (Hardman et al., 2001; Adelaar and Muysken, 2007; Cerrón-Palomino and Carvajal, 2009; Briggs, 1976).

2.1 Agglutinative Morphology

Aymara has a very rich inflection. Suffixes of various categories can be chained to build up long words that would be expressed by a sentence in languages like English. For example, *alanxarusksmawa* (*ala-ni-xaru-si-ka-sma-wa*) means "I am preparing myself to go and buy it for you".

In concordance with the principle of lexical integrity (Bresnan, 2001), we deal with morphology in the lexicon. Ishikawa (1985) has suggested to use word-internal (sublexical) rules to analyze structurally complex words in agglutinative languages. We have adopted this analysis.

Proceedings of Recent Advances in Natural Language Processing, pages 562–567,
Hissar, Bulgaria, 12-14 September 2011.

2.2 Vowel elision

Aymara uses vowel elision as morphosyntacic marking, as illustrated in (1) and (2).[1]

(1) *aycha manq'ani*
meat eater
"who eats much meat"

(2) *aych manq'ani*
meat-ELI eat-$\text{FUT}_{3\to3}$
"(s)he will eat meat"

There are three types of vowel elision that interact with each other. *Object elision* marks a noun or pronoun as direct object, such as in (3) (as opposed to (4)).

(3) *khits uñji*
whom-ELI see-$\text{NFUT}_{3\to3}$
"Whom does he/she see?"

(4) *khitis uñji*
who see-$\text{NFUT}_{3\to3}$
"Who does see him/her?"

Noun compound elision occurrs in NPs. The final vowel of noun attributes gets elided if they have three or more syllables, as illustrated in (5) and (6).

(5) *aymar aru*
Aymara-ELI language
"the Aymara language"

(6) *qala uta*
stone house
"stone house"

Complement elision is applied to all words that are arguments or adjuncts of a verb except for the final word of a clause.[2]

Whereas object elision concerns the nucleus of a word (the stem with an optional possessive and/or plural suffix), noun compound and complement elisions concern the final vowel of a word (the vowel of the last suffix or the stem if there are no suffixes). Vowel elision is dealt with in the lexicon. As for noun compound elision, all nouns with more than two syllables get (↑ COMPEL) = + if the final vowel of the word nucleus is elided and (↑ COMPEL) = − if it is not. Nouns with two vowels do not define this attribute, i.e., it can be unified with both values. The corresponding rule for compound nouns is given in (7).

(7)
$$\begin{array}{lcccc} \text{N}' & \to & (\text{N}') & & \text{N} \\ & & (\uparrow \text{MOD}) = \downarrow & & \uparrow = \downarrow \\ & & (\downarrow \text{COMPEL}) = + & & \end{array}$$

2.3 Polypersonal agreement

Being a polysynthetic language, Aymara has polypersonal conjugation, i.e., the finite verb agrees with the subject and with another argument which may be the object (direct or indirect) or an oblique argument. An example is given in (8).

(8) *Uñjsma*
see-$\text{NFUT}_{1\to2}$
"I see/saw you."

The morpholexical entry for *uñjsma* is given in (9).[3] Note that the PRED value for both subject and object is optional.[4]

(9)

uñjsma	V	(↑PRED) ='uñjaña⟨(↑SUBJ)(↑OBJ)⟩'
		(↑TAM TENSE) = NON-FUT
		(↑TAM MOOD) = INDIC
		((↑SUBJ PRED) = 'PRO')
		(↑SUBJ PERS) = 1
		((↑OBJ PRED) = 'PRO')
		(↑OBJ PERS) = 2

The verb agrees with the subject and with the most animate argument which may be a patient, addressee or source, e.g., *um churäma*-$\text{FUT}_{1\to2}$ "I will give you water" (addresse), *aych aläma*-$\text{FUT}_{1\to2}$ "I will buy meat from you" (source) etc. However, there are verbal suffixes which can make the verb agree with other arguments, such as the beneficiary, e.g., *aych churarapitäta*-BEN,$\text{FUT}_{2\to1}$ "You will give him/her bread for me" (the verb agrees with the beneficiary instead of the addressee). All these agreement rules are encoded in the lexicon.

2.4 Free Word Order

At the clause level, the word order in Aymara is not restricted although SOV is preferred. There is also no evidence for a VP, thus we assume a flat phrase structure. The rules for matrix clauses are given in (10).[5]

[1] In the glosses, $\text{FUT}_{3\to3}$ means future tense. The numbers express the person of the subject and an additional argument, mostly object.

[2] Object and noun compound elision has the gloss ELI in our examples.

[3] TAM means Tense-Aspect-Mood.

[4] Both arguments can be dropped.

[5] In the functional annotation, κ is either '−' (no case) or a semantic case and GF is the corresponding grammatical function.

(10) S → $\mathcal{X}^+$

where $\mathcal{X}$ is V ↑=↓ or NP/CP (↓ CASE) = κ ⇒ (↑ GF) =↓

CP → (C) ↑=↓ , S ↑=↓

As can be seen, word order in a clause is free with the exception of an optional complementizer (see (11) and (12)) which can be placed at the beginning of the clause or at its end.

(11) *Ukat juti*
then come-NFUT$_{3\rightarrow3}$
"Then (s)he came."

(12) *Jutät ukaxa...*
come-FUT$_{2\rightarrow3}$ if
"If you will come..."

There are no discontinuous constituents and complement clauses can be embedded in the matrix sentence. Since Aymara is not discourse-configurational (see the next subsection), the word order, despite of being free, is usually unmarked (SOV) and if it is different then mostly for stylistic reasons.

2.5 Topic-Focus Articulation

We have adopted the approach proposed by King (1997). Thus we use an i(nformation)-structure to approximate topic-focus articulation (TFA).[6]

A simple example of two sentences which differ only in TFA is given in (13) (the word *qullqirï* is a verbalized noun).

(13) *Jumax qillqirïtawa*
you-SG,TOP be-a-writer-NFUT$_{2\rightarrow3}$,FOC
"You are a writer."

Jumaw qillqirïtaxa
you-SG,FOC be-a-writer-NFUT$_{2\rightarrow3}$,TOP
"It is you who is the writer."

The morpholexical entries for *jumax* and *jupaw* and corresponsing i-structures for the sentences in (13) are given in (14) and (15), respectively.

[6]The difference is that we use only two discourse functions, TOP or FOC, with the possibility for words being discourse-unspecified (the term 'discourse-neutral' is used sometimes). This is exactly how morphological marking of TFA works in Aymara.

(14) *jumax* PRON (↑PRED) = 'PRO'
(↑PERS) = 2
(↑PRED FN) ∈ (↑$_i$TOP)

$$\begin{bmatrix} \text{TOP} & \{\text{'jumax'}\} \\ \text{FOC} & \{\text{'qillqiri'}\} \end{bmatrix}$$

(15) *jumaw* PRON (↑PRED) = 'PRO'
(↑PERS) = 2
(↑PRED FN) ∈ (↑$_i$FOC)

$$\begin{bmatrix} \text{TOP} & \{\text{'qillqiri'}\} \\ \text{FOC} & \{\text{'jumax'}\} \end{bmatrix}$$

The i-structure is very important for correct translation. For example, the sentence *Chachax liwrw liyi* would be translated as "The man read(s) a book" whereas *Chachaw liwrx liyi* would be better translated as "The book is/was read by a man".[7]

3 Lexical Mapping Theory and D-Structures

Although f-structures abstract to some extent from language specific features (such as differential object marking, see (16) where the Spanish dative phrase and the Polish genitive phrase would be in accusative in German), there are still many differences even between relatively closely related languages.[8]

(16) *Ayer visité a Juan*
yesterday visit-PAST,1SG to Juan
"I visited Juan yesterday."

Nie mam samochodu
NEG have-PRES,1SG car-SG,GEN
"I don't have a car."

Wong and Hancox (1998) examine the use of a(rgument)-structures in machine translation

[7]Unlike some other languages with morphological topic and/or focus markers, such as Japanese (cf. examples from (Kroeger, 2004): *Taroo-wa*-TOP *sono hon-o*-ACC *yondeiru* "Taroo is reading that book." vs. *Sono hon-wa*-TOP *Taroo-ga*-NOM *yondeiru* "That book, Taroo is reading"), Aymara allows their co-occurrence with case suffixes without limitation.

[8]For example, the East Baltic language Latvian has only agent-less passives (i.e., in LFG, it completely lacks OBL$_{ag}$, cf. (Forssman, 2001)), whereas its closest and partially mutually intelligible relative Lithuanian has and frequently uses agents in passives.

(MT). In LFG, a-structures are another level of linguistic representation which provides the lexico-syntactic interface. The mapping between a-structures and f-structures is defined by the so-called Lexical Mapping Theory (LMT; see (Bresnan, 2001)). We will give a brief overview of LMT here.

LFG assumes that there is a prominence hierarchy of semantic roles. We use the hierarchy shown in (17) (proposed by Bresnan (2001)):

(17) agent $\succ$ beneficiary/maleficiary $\succ$ experiencer/goal $\succ$ instrument $\succ$ patient/theme $\succ$ locative

Argument grammatical functions (GF) are assigned features *objective* and *restricted* as in (18). The markedness hierarchy of GFs is given in (19).

(18)

	-r	+r
-o	SUBJ	OBL_θ
+o	OBJ	OBJ_θ

(19) $\text{SUBJ} \succ \text{OBJ}, \text{OBL}_\theta \succ \text{OBJ}_\theta$

Verbs in LFG have an a-structure that expresses their valence. The arguments of each verb are ordered according to the hierarchy in (17) and annotated with $-o, -r, +o, +r$. General LMT principles determine how the arguments are mapped onto GFs. The initial role is mapped onto SUBJ if classified with $[-o]$. Otherwise, the leftmost role classified $[-r]$ is mapped onto SUBJ. Other roles are mapped onto the lowest compatible GF according to the hierarchy in (19). There are two other constraints: Every verb must have a SUBJ and each role must be associated with a unique function, and conversely.

Bresnan (2001) argues that LMT allows for natural treatment of passives, ditransitives and other constructions which have been handled by lexical rules in earlier versions of LFG.

We use the information provided by f-structures, i-structures, c-structures and a-structures to create a dependency-based representation of parsed sentences (a tectogrammatical tree in the terminology of Sgall et al. (1986)). The main reason is that we already have a module that generates English and Spanish sentences from (tectogrammatical) syntax trees.

In the following, we will use the term d(ependency)-structure to refer to dependency trees induced by LFG structures. Table 1 gives a brief overview of which information at different levels of linguistic representation in LFG is used in d-structures.

LFG layer	**information in d-structures**
c-structure	original word order
f-structure	dependencies and coreferences
i-structure	topic-focus articulation
a-structure	valence

Table 1: Information provided by LFG layers to d-structures

The skeleton of a d-structure is provided by the f-structure. According to a generally accepted principle of deep syntax (tectogrammatics) only autosemantic (content) word are represented by nodes in d-structures. In LFG, autosemantic words are associated with projections of lexical categories, i.e., f-structures with the PRED attribute (see (Bresnan, 2001) for a detailed discussion of lexical and functional categories and the so-called 'coheads'). Thus a d-structure derived from (20) would have three nodes for the words *dog, chases* and *cat*.

(20)
$$\begin{bmatrix} \text{PRED} & \text{'chase}\langle(\uparrow\text{SUBJ})(\uparrow\text{OBJ})\rangle\text{'} \\ \text{TENSE} & \text{PRES} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'dog'} \\ \text{SPEC} & \begin{bmatrix} \text{DEF} & + \end{bmatrix} \end{bmatrix} \\ \text{OBJ} & \begin{bmatrix} \text{PRED} & \text{'cat'} \\ \text{SPEC} & \begin{bmatrix} \text{DEF} & - \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

The edges are labelled with semantic roles. This is possible due to the bi-uniqueness of the mapping between roles and GFs (see above). However, there is one exception: The initial role is assigned a special label which we call 'actor' (ACT, which is equvalent to what Bresnan (2001) marks $\hat{\theta}$ and calls 'logical subject'). This partially reflects the shifting of actants in tectogrammatics as defined by Sgall et al. (1986).[9]

So far, we have an unordered tree (f-structures are unordered by definition).[10] We define an ordering based on information structure, as proposed

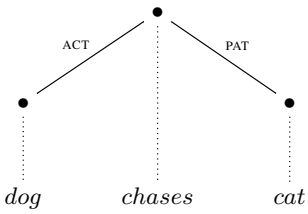
[9]The edge labels are theory specific and somewhat arbitrary. For example, Butt et al. (1999) distinguish between 'semantic' and 'non-semantic' prepositions. As a consequence, the complement in *He relies on the book* is an OBJ and therefore PAT in the corresponding d-structure although *on the book* is not a direct object in the traditional dependency grammar.

[10]Generally, the skeleton rendered by f-structures may contain a cycle, i.e., a node with more than one mother nodes.

for deep syntax by Sgall et al. (1986). Thus we use the i-structure to define a partial ordering on the nodes of the d-structure (TOP $\prec$ 'discourse-unspecified' $\prec$ FOC). The nodes in each of the three topic-focus domains are ordered according to their original ordering in the sentence (which is captured by c-structures).[11]

The resulting d-structure is given in (21).[12]

(21)



Let us briefly point out some properties of d-structures as defined above. Most of them directly correspond to properties of deep syntax (tectogrammatical) trees.

1. There is a bi-unique mapping between d-structure nodes and autosemantic (content) words. Synsemantic (auxiliary/function) words are represented as attributes of nodes. This is a direct consequence of LFG 'co-heads'.

2. 'Dropped' words (e.g., subject and/or object pronouns in so-called pro-drop languages) are re-established in d-structures as a consequence of the LFG Principle of Completeness since PRED attributes are instantiated in the lexicon if needed (cf. (Bresnan, 2001)).

3. Edge labels in d-structures reflect semantic relations rather the GFs which are more language specific.

4. The ordering of d-structure nodes is partially determined by topic-focus articulation.

This is how LFG handles coreferences, such as in the sentence *I want to go home* where the complement clause is an open complement (XCOMP) in the f-structure of 'want' and (↑ SUBJ) = (↑ XCOMP SUBJ). To obtain a well-formed tree, we reflect the path of length 1 in the f-structure as an edge and the remaining (conflicting) functional paths as co-references.

[11] In free word-order languages, NPs and PPs usually have more rigid word order than clause arguments and adjuncts, thus in an MT system, the module for syntactic synthesis of the target language would reorder the d-structure according to language specific word-order rules.

[12] The attributes associated with nodes can be obtained from corresponding f-structures (in LFG, all lingustic levels are interlinked).

However, there are several differences. For example, d-structures can be non-projective (tectogrammatical trees are projective by definition (Sgall et al., 1986)) which is a direct consequence of how long-distance dependencies are represented in f-structures. Furthermore, one word can be represented by more than one d-structure nodes (such as in languages with incorporation).

Butt et al. (1999) give a detailed description of the process of parallel grammar development. In our approach, the correspondence between original LFG structures and d-structures poses some (mostly technical) limitations on grammar writers. For example, f-structures of synsemantic words (functional categories) must be 'coheads' of their functional categories (however, this is a general requirement in modern LFG according to Bresnan (2001)). Also, GFs must conform to the strict constraints imposed by LMT.

Table 2 show how many c-structures, f-structures and d-structures are identical (two d-structures are identical if they have the same structure and edge labels) in a parallel Aymara-Spanish corpus of 1,000 sentences.

level	**identical representation**
c-structure	7.8%
f-structure	38.3%
d-structure	69.5%

Table 2: Identical c-, f- and d-structures in a parallel corpus

4 Machine Translation

In this section, we briefly present the results of an MT experiment from Aymara into Spanish and English. All modules of the system were developed in SWI Prolog (Wielemaker, 2003).

It is obvious (cf. Section 2) the there are very few structural similarities between Aymara and Spanish or English, thus a 'direct' or 'shallow' approach to MT, as proposed by Dyvik (1995), would not lead to quality translation. As has been said above, we have developed an LFG grammar for Aymara. Kaplan and Wedekind (2000) have shown that the generation of sentences out of a f-structure according to an LFG grammar yields a context-free language. However an LFG grammar developed for parsing may not be suitable for generation (due to overgeneration). That is why we use d-structures as defined in Section 3.

Evaluation results are given in Table 3.

language pair	WER
Aymara-Spanish	22.3%
Aymara-English	24.8%

Table 3: Evaluation of MT into Spanish and English

While the error rate is not low, it is acceptable given the fact that the source language is structurally very different from the target language. Most translation errors can be tracked to diverging valency of verbs in both languages.

5 Conclusions and Further Research

We have presented a formal grammar for Aymara and pointed out some interesting properties of the language and how they can be dealt with in the LFG framework.

As can be seen, the LFG framework can be easily used to develop formal grammars of polysynthetic languages such as Aymara. While the rules we have developed cover a large part of the Aymara syntax, the lexicon we have now needs to be expaned. Currently, we are focusing on refining sublexical rules.

We have chosen LFG for our grammar because it has a solid formal foundation while providing grammars that can be used directly in NLP. However, we are developing the grammar for use in MT and LFG's f-structures are still relatively language-specific. To overcome this limitation, we have developed a fully automatic procedure which induces d(ependency)-structures (deep syntax trees) that are at a higher level of abstraction. Our d-structures are not only more suitable for cross-lingual NLP tasks such as MT but they also disclose that LFG is, in its core, a dependency-based formalism.

References

Willem Adelaar and Pieter Muysken. 2007. *The Languages of the Andes*. Cambridge University Press.

Kenneth R. Beesley. 2006. Finite-state Morphological Analysis and Generation for Aymara. In *Proceedings of the Global Symposium on Promoting the Multilingual Internet*.

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

Lucy Therina Briggs. 1976. *Dialectal Variation in the Aymara Language of Bolivia and Peru*. Ph.D. thesis, University of Florida.

Miriam Butt, Tracy Holloway King, María-Eugenia Niño, and Frédérique Segond. 1999. *A Grammar Writer's Cookbook*. CSLI Publications.

R. Cerrón-Palomino and J. Carvajal Carvajal. 2009. Aimara. In M. Crevels and P. Muysken, editors, *Lenguas de Bolivia*. Plural Editores, La Paz, Bolivia.

Helge Dyvik. 1995. Exploiting Structural Similarities in Machine Translation. *Computers and Humanities*, 28:225–245.

Berthold Forssman. 2001. *Lettische Grammatik*. Verlag J.H. Roell, Dettelbach.

Martha Hardman, J. Vásquez, and J. Yapita de Dios. 2001. *Aymara. Compendio de estructura fonológica y gramatical*. Instituto de Lengua y Cultura Aymara.

Akira Ishikawa. 1985. *Complex Predicates and Lexical Operations in Japanese*. Ph.D. thesis, Stanford University.

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

Ronald M. Kaplan and Jürgen Wedekind. 2000. LFG Generation Produces Context-free Languages. In *Proceedings of COLING-2000, Saarbrücken*.

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

Paul R. Kroeger. 2004. *Analyzing Syntax*. Cambridge University Press.

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

Jan Wielemaker. 2003. An overview of the SWI-Prolog programming environment. In Fred Mesnard and Alexander Serebenik, editors, *Proceedings of the 13th International Workshop on Logic Programming Environments*, pages 1–16, Heverlee, Belgium, december. Katholieke Universiteit Leuven. CW 371.

Shun Ha Sylvia Wong and Peter Hancox. 1998. An Investigation into the Use of Argument Structure and Lexical Mapping Theory for Machine Translation. In *Proceedings of the 12th Pacific Asia Conference on Linguistics, Information and Computation*, Singapore.