Jay Kaiser and Sarah Shenk
Kübler LING-L614
4 May 2016

Parsing Polysynthesis: Theoretical and Practical Methodologies

Introduction

Polysynthetic languages are those where individual words of the language consist of many different morphemes, most of which are inflectional in nature; this can go so far as that an individual word can represent an entire sentence. Example (1) provides four sentences taken from polysynthetic languages to be used for example. Sentences (1a) and (1b) come from the Siouan language Lakhota (Van Valin 2007), and sentences (1c) and (1d) come from the Iroquoian language Mohawk (Baker 1996).

(1)

a. mathó ki hená wičhá-wa-kte
bear the those 3plO-1sgS-kill
'I kill the bears.'

b. wičhá-wa-kte
3plO-1sgS-kill
'I kill them.'

c. wa'-k-hnÍnu-' ne ka-nákt-a'.
FACT-1sS-buy-PUNC NE NsS-bed-NSF
'I bought the/a bed.'

d. wa'-ke-nákt-a-hnínu-'.
FACT-1sS-bed-∅-buy-PUNC
'I bought the/a bed.'

As seen between (1a) and (1b), and paralleled through (1c) and (1d), sentences in many polysynthetic languages can be written through multiple words as could be found in more analytic languages, but the potential for merging verbal roots with bounded noun forms is ever-present and widely utilized. These types of sentence forms appear very exotic to many people, so extensive data is necessary for better comprehension of this form of morphosyntax in language.

Unfortunately, most examples of polysynthetic languages are of native North American descent, and as such they lack adequate research and documentation (Bybee 1997). Small populations of speakers, a lack of annotated text, and a comparatively few number of researchers devoting time to the research of these languages creates a kind of desert of data about polysynthetic languages in general. An example of this can be seen in the polysynthetic Siouan language Lakhota, a primary anchor of this research.

This is not made any easier by the fact that the most well-known syntactic theories, those of Chomskyan grammar, are ill-equipped to deal with phenomena that arise in polysynthetic languages. The non-configurational properties of these languages lead to traits in their syntax like free word order and discontinuous argument expressions, as well as the fact that in some polysynthetic languages the omission of noun phrases posits no error for native speakers, potentially violates the idea of the theta-criterion, which states that each argument of a sentence bears only one theta-role and vice versa, where a theta-role is some thematic label that semantically describes the argument (Chomsky 1981). Because these languages often have the ability omit arguments whose theta-roles must be filled, theories of syntax where the theta-criterion plays a prominent role, like X-bar theory and government and binding theory, cannot accurately and fully describe these phenomena (Baker 1996).

Therefore, an alternative grammatical formalism is required in order to alleviate these presented difficulties. Throughout the literature regarding this issue, two prominent formalisms have been presented as potential alternatives: Role and Reference Grammar and Lexical-Functional Grammar. Each handles the nature of polysynthetic languages differently, and both are presented here in full description centered around our understanding of each. The goal of this paper was to answer the following question: how are polysynthetic languages dealt with in syntactic parsing, and what are the implications of this?

Role and Reference Grammar

Role and Reference Grammar (RRG) is a syntactic formalism first described by Foley and Van Valin (1984) that is centered around two primary concepts: the development of a grammar formalism based not on English or other Eurocentric

languages, but instead on radically differing languages like Lakhota or Tagalog, for example; and the ability to capture and explain syntax, semantics, and pragmatics and their interaction (Van Valin 2007).

Together, these two ideas have led to the formation of a vastly different attempt at explaining language than traditional systems. RRG is a monostratal theory that posits that the surface form of a sentence alone provides the necessary syntactic representation to allow for complete parsing. Moreover, RRG rejects traditional Chomskyan clause structure representations, instead centering around a 'layered structure of the clause,' or LSC, whose primary components are, in descending encompassing order, the sentence, the clause, the core, and the nucleus. According to Van Valin, all sentences arise through some form of coordination, subordination, or cosubordination between the syntactic units of the LSC of a sentence, and these can potentially account for any and all complications of language. Figure 1 below presents the LSC of the head-marking language Lakhota and of the dependent-marking language English (Van Valin 2007).

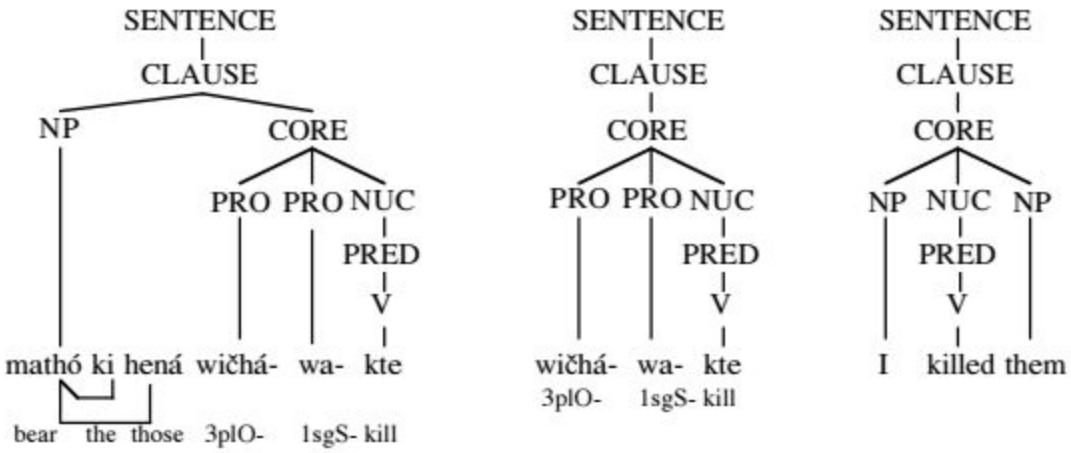

Figure 1: The 'layered structure of the clause' of Lakhota and English (Van Valin 2007)

In Lakhota and many other polysynthetic languages, head-marking morphemes are merged with verbal roots to allow for the creation of entire one-word sentences, as seen in Figure 1 and example sentences (1). RRG analyzes the head-marking nature of Lakhota in a manner distinct from many other grammar formalisms In RRG a sentence is viewed as complete and whole in its surface form, with no null phonological elements

like traces or null pronominals permitted in a sentence's parse, in contrast with certain frameworks of Government and Binding Theory (GB) where some form of trace is obligatorily used to allow for such constructions, as can be seen in Figure 2 below. Therefore, the formalism derives a different way to represent head-marking. Bound pronominal markers are represented as core arguments, while noun phrases that represent the explicit focus of the marker are assigned a clause position, representing a single discontinuous argument consisting of both the bound affix and itself (Van Valin 1999). This can be contrasted with certain frameworks of GB, for example, where this element of the sentence is instead marked as a dislocated sentential adjunct, but the bound markers are treated as the true verbal arguments (Jelinek 1984).

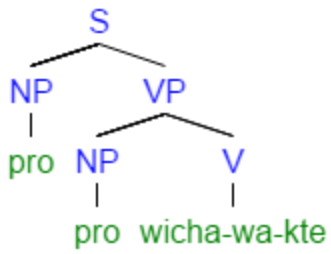

Figure 2: GB representation of the sentence (1b)

As seen in the first tree of Figure 1, the bound pronominal markers are too marked as core arguments of the sentence; however, RRG instead links the noun phrase complement *mathó ki hená* and its bound pronominal affix *wičhá* to the same argument position in the sentence semantically, in an "extra-core slot". This sentential position differs from that of traditional clausal structures in that it lacks a specific discourse function and can appear in various placements in the sentence, depending primarily on the verb placement and head-placement in the language being analyzed (Van Valin 2013). Through this, "extra-core slot" arguments appear outside of the periphery of the sentence, as RRG denotes to adjunctive sentence elements, and the bound morpheme and its noun phrase argument are regarded as a single argument, though discontinuous in nature (Van Valin 1999).

Lexical-Functional Grammar

Lexical-Functional Grammar (LFG) is a syntactic formalism first described by Bresnan (1982), which, in assuming a single level of syntactic structure, utilizes a system of lexical processes that mark associations between arguments using primitive grammatical functions. By lacking multi-stratal syntactic representations of a sentence, complex workarounds to explain difficult language phenomena like transformation or movement are not necessary in parsing even non-projective sentences. A syntactic constituent structure, the *c-structure*, is initiated with a functional structure, or *f-structure*, that combines information between the c-structure and a lexicon (Neidle 1994).

LFG is a popular framework used by many authors to analyze polysynthetic languages (Baker et al. 2010; Homola 2010, 2011, 2012; Van Valin 2013), in particular due to the flexibility of the graphical c-structure and the expressive power of the attribute-value-matrical f-structure. The following examples of analysis from the polysynthetic language Wubuy illustrate how some phenomena unique to polysynthetic languages can be interpreted using LFG.

One way that possession is expressed in Wubuy is through incorporation of the possessed entity through affixation onto the verbal element. This occurs most frequently with terms for anatomical body parts for example, and in Figure 4 we see a sentence in Wubuy, along with its English gloss and translation, that demonstrates this phenomenon.

nga-wu-yarrga-nagiina yii-ngarrugalij-(inyung)*
1SG-NEUT-flipper-cook.PR FEM.OBL-dugong-GEN
'I'm cooking the dugong's (FEM) flipper (NEUT).'

Figure 4: An example of incorporation in a Wubuy sentence (Baker et al. 2010)

Here, the object being possessed, the *flipper*, is attached to the verb as a nominal suffix, and what could be interpreted as the traditional subject, *I*, is expressed only through morphemic agreement on the verb. In order to preserve the transitive valency of the verb *cook*, a subject and object is necessary; however, there also must be a place in the sentence's f-structure to separately express both the *flipper* and the *dugong*. The verb here also indicates *flipper* as the direct object, through the agreement

of the neuter-marking morpheme in the verbal construction with the neuter noun class of *flipper*.

The proposed f-structure for the sentence found in Figure 4 is shown below in Figure 5. The distinct possession quandary can be dealt with by analyzing the lexical entry of the verb as the predicate of the sentence, requiring a subject field and an object field; then by allowing this object to subcategorize as a set with multiple units, the object can be described to contain more than one entity (here both the *dugong* and it's *flipper*), which aids not only with sentence structures like those presented here, but also in occurrences of coordination. The predicate value of the incorporated noun *flipper* indicates that it is a possession and that it requires a possessor, as marked in its objective predicate label.

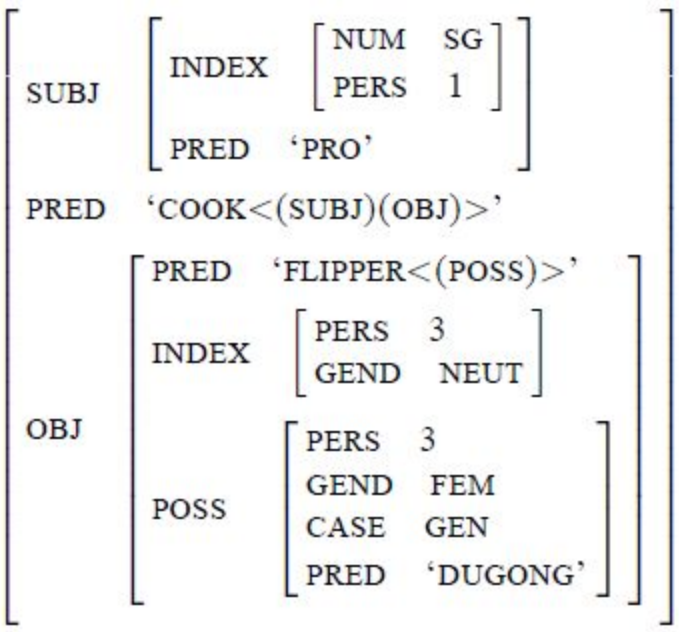

Figure 5: f-structure for the Wubuy sentence in Figure 4 (Baker et al. 2010)

This is one of many ways that LFG is well-equipped to parse unusual syntactic phenomena through the use of feature-value matrices.

Parsing and Machine Translation of Polysynthetic Languages

Parsing polysynthetic languages creates unique challenges, both because of limited pre-existing research and very few syntactic frameworks that provide satisfactory

analyses. Particular research has been done to create a series of parsing and translating tools for Aymara, a polysynthetic language from South America (Homola 2010, 2011, 2012). Like the languages previously discussed, Aymara is morphologically rich and uses affixation to create large verbal constructions that can be used to express entire clauses, as Example 2 illustrates (Homola 2010).

(2)

Alanxarusksmawa.

Ala-ni-xaru-si-ka-sma-wa.

"I am preparing myself to go and buy it for you."

Homola created a framework that uses an expands on LFG, using not only c-structure and f-structure, but an information i-structure, and an argument a-structure to capture information about the valency of verbs and the theta roles of their arguments. Because of the free word order that exists in Aymara, Homola also makes use of Lexical Mapping Theory (LMT) to assign subject and object to the arguments, as well as their theta roles. LMT uses a hierarchy of the most common theta roles and maps them onto the grammatical functions of the arguments based on the positive or negative values of two features, objective and descriptive (Homola 2010). Finally, all of these structures provide the information needed to create the dependency d-structure, which details both syntactic structure and semantic roles and can be used to map meanings between languages, as seen in Figure 6. The use of multiple structures enables LFG to express both syntactic and semantic meaning, creating the potential for robust and thorough parsing tools.

LFG layer	information reflected in d-structures
c-structure	original word order
f-structure	dependencies and coreferences between phrases
i-structure	topic-focus articulation
a-structure	valence (semantic roles and their mapping to GFs)

Figure 6: The layers of Homola's LFG formalism (Homola 2010)

After building the layered information structures needed to analyze Aymara, Homola began working on a translation system from English and Spanish to Aymara. By

creating a deeper structure that shows valency and semantic roles without restricting word order or morphological representation, Homola hoped to ease translation between a language like Aymara and Spanish, which have very little syntactic similarity. Using this approach, Homola was able to translate from Aymara to Spanish with a word-error rate of 22.3% and from Aymara to English with a word error rate of 24.8% (Homola 2011). The efficacy of LFG in the parsing of polysynthetic structures in Aymara that do not occur in traditionally studied languages can be seen through the initial high accuracy of these translations, and with further work and optimization of these methods, analysis of other under-resourced polysynthetic languages can be equally done well without statistical methods.

Related Work

The Polysynthesis Parameter (Baker 1996) works within a Chomskyan framework to provide an all-encompassing study of the Mohawk language (another polysynthetic Native American language); ultimately, the author presents theoretical depictions of polysynthetic frameworks in his search for evidence of a universal "macroparameter" of polysynthesis.

Van Valin (1977) is a comprehensive syntactic analysis of the polysynthetic Lakhota language, using RRG as mentioned above. In this manner, the author provides a detailed exposition of the theory as a means to describe the structures of the language.

Van Valin (2007) is a general overview of RRG grammar, a framework that appears frequently in the literature revolving around polysynthetic parsing. The author's own creation and developed to lay foundationally distinct from other more Eurocentric grammar formalisms, RRG is given explicit care and detail in an attempt to understand polysyntheticity.

Van Valin (1999) analyzes head-marking languages like the polysynthetic Lakhota language and explains the flaws and limitations of a traditional x-bar analysis of them. It then compares how LFG and RRG deal with analysis of this attribute of language. These two alternative syntactic formalisms provide a clearer means of describing head-marking aspects of polysynthetic languages than other formalisms,

though the paper naturally provides a clear bias towards role-and-reference grammar over lexical-functional grammar.

Van Valin (2013) analyses head-marking languages like Lakhota, though under the assumption that, as presented by Nichols (1986), there exists a typological contrast between head-marking and dependent-marking languages. In order to capture this contrast, the author again utilizes RRG, as it is more easily able to present a description of polysynthetic sentences than would traditional Chomskyan formalisms. Information throughout this paper provides an alternative perspective in head-marking language processing, and a large number of examples and comparisons provide substantial evidence centered around polysynthetic parsing.

Part two of Theoretical Perspectives on Native American Languages (Gerdts and Michelson 1989) provides six different examples of complex morpho-syntactic structures found in Native American languages, languages of which many are polysynthetic in nature. In particular, the passages written by Leslie Saxon and J. Peter Denny regarding inflection and cliticization in Dogrib and polysynthesis in Algonquian and Eskimo respectively could provide further information regarding polysynthetic aspects of languages and how they are resolved syntactically. One downside of this book is that these two specific sections lack alternative grammatical formalisms when describing the phenomena; instead, theoretical descriptions that merely describe possible explanations are used. However, this does not mean that this source does not provide important reference in the evaluation of polysynthetic structures.

Baker et al. (2010) provides detailed analyses of various phenomena in Wubuy, a polysynthetic language with free word order found in Australia. This serves as an excellent example of how to use LFG to address polysynthetic structures that don't exist in strictly-ordered languages that have fairly monomorphemic words.

A final selection of work is based in the recent research of Petr Homola, who utilizes an LFG parser and machine translator for the under-resourced polysynthetic language Aymara, found in South America. Homola (2010) builds a grammar for Aymara using the lexical-functional grammar framework, taking especial interest in various constructions of the language and how they can be accounted for. The grammar has

been devised to allow for easy branching of this research into fields of natural language processing.

Homer (2011) elaborates on using LFG for NLP by incorporating an argument a-structure that describes the valency of verbs and an information i-structure which describes topic-focus articulation. The author uses Prolog to create a dependency parser based on this four-structured (f-structure, c-structure, i-structure, and a-structure) grammar paradigm.

Homola (2012) describes a set of tools that can be used in the machine translation of polysynthetic languages. This is a four step tool chain, with a morphological analyzer of the original polysynthetic language, a rule based parser, a transfer module which holds the structures of equivalent words in the target language, and a morphological generator that builds the translation in the target language. Ultimately, the translator received a word-error rate of only 22.3% from Aymara to Spanish and 24.8% from Aymara to English, promising results for such a new development in polysynthetic parsing.

Conclusion

Polysynthetic languages prove challenging to describe in many grammatical formalisms, due to these languages' complex affixation systems. Certain frameworks like role and reference grammar and lexical-functional grammar, however, provide the flexibility required to create adequate parses of these languages.

In RRG a distinction between the word and the morpheme omits the need of trace structures or other phonologically null phenomena in explaining the head-marking nature of polysynthetic languages; moreover, the "layered structure of the clause" provides a means to parse the complex internal structure of the sentence, through a formalism far unlike that of either Chomskyan frameworks.

LFG, through layered structures of analysis, seems to be a useful and popular framework with which to evaluate polysynthetic languages. Because it is widely known and well-suited to both theoretical syntactic work and natural language processing, LFG may prove to be the most useful approach to future analysis of polysynthetic phenomena. While the work done by Homola is not an exhaustive solution to natural

language processing, it provides a promising starting point for others interested in parsing polysynthetic languages.

Although this is but a small delving into the world of polysynthetic parsing, these two formalisms provide a strong basis for future work. It is important to step back from the theoretical syntax or the computational processing and recognize that many polysynthetic languages exist in small, aging communities. The insights that these languages may provide us about how the human mind both processes its surroundings and creates syntactic structures will be lost if there is not more research in this field. Finding grammar formalisms that most easily express these unique languages will enable essential future research. Creating parsers that can accurately interpret polysynthetic languages makes automatic translation of these exotic languages a possibility. The time and effort invested in the study of these languages in turn translates to time and effort invested in the survival and preservation of linguistic knowledge that will otherwise soon be lost.

References

Baker, B., Horrack, K., Nordlinger, R., & Sadler, L. (2010). PUTTING IT ALL TOGETHER: AGREEMENT, INCORPORATION, COORDINATION AND EXTERNAL POSSESSION IN WUBUY (AUSTRALIA). In Proceedings of LFG '10 Conference. Retrieved April 12, 2016, from http://web.stanford.edu/group/cslipublications/cslipublications/LFG/15/papers/lfg10bakeretal.pdf

Baker, M. C. (1996). The Polysynthesis Parameter. New York, NY: Oxford University Press.

Bresnan, J. (1982). *The Mental representation of grammatical relations*. Cambridge, MA: MIT Press.

Bybee, J. L., Haiman, J., & Thompson, S. A. (1997). Semantic Aspects of Morphological Typology. In J. L. Bybee (Ed.), *Essays on Language Function and Language Type: Dedicated to T. Givon*. Retrieved from https://www.unm.edu/~jbybee/downloads/Bybee1997SemAsp.pdf

Chomsky, N. (1981). *Lectures on Government and Binding*. Dordrecht, The Netherlands: Foris Publications.

Foley, W. A., & Van Valin, R. D. (1984). *Functional Syntax and Universal Grammar*. Cambridge, England: Cambridge University Press.

Gerdts, D. B., & Michelson, K. (1989). Theoretical Perspectives on Native American Languages. Albany, NY: State University of New York Press.

Homola, P. (2010). Building a formal grammar for a polysynthetic language. In Proceedings of the 15th and 16th international conference on Formal Grammar (pp. 228-242). Copenhagen, Denmark. Retrieved April 12, 2016, from http://download.springer.com/static/pdf/916/chp%3A10.1007%2F978-3-642-32024-8_15.pdf?originUrl=http://link.springer.com/chapter/10.1007/978-3-642-32024-8_15&token2=exp=1460436582~acl=/static/pdf/916/chp%253A10.1007%252F978-3-642-32024-8_15.pdf?originUrl=http%3A%2F%2Flink.springer.com%2Fchapter%2F10.1007%2F978-3-642-32024-8_15*~hmac=a2ce373c3c20b22db0cb030b678c5dcbf3790adbf3bd8c5bddfb1317670dc794

Homola, P. (2011). Parsing a Polysynthetic Language. In Proceedings of Recent Advances in Natural Language Processing (pp. 562-567). Hissar, Bulgaria. Retrieved April 12, 2016, from http://aclweb.org/anthology/R11-1079

Homola, P. (2012). A Machine Translation Toolchain for Polysynthetic Languages. In Proceedings of the 16th EAMT Conference. Retrieved April 12, 2016, from http://hltshare.fbk.eu/EAMT2012/html/Papers/11.pdf

Jelinek, E. (1984). Empty Categories, Case, and Configurationality. *Natural Language & Linguistic Theory, 2*(1), june., 1984, 39-76. Retrieved from http://www.jstor.org/stable/4047560?seq=1#page_scan_tab_contents

Neidle, C. (1994). *Lexical Functional Grammar* (Rep.). Retrieved from http://www.bu.edu/asllrp/neidle-lfg.pdf

Van Valin, R. D., Jr. (1977). Aspects of Lakhota Syntax (Unpublished doctoral dissertation). University of California, Berkeley. Retrieved April 12, 2016, from http://linguistics.berkeley.edu/~survey/documents/dissertations/vanvalin-1977.pdf

Van Valin, R. D., Jr. (1999). Linguistic Diversity and Theoretical Assumptions (Rep.). Retrieved April 11, 2016, from State University of New York at Buffalo website: http://linguistics.buffalo.edu/people/faculty/vanvalin/rrg/vanvalin_papers/linguisticdivrs.pdf

Van Valin, R. D., Jr. (2007). A Summary of Role and Reference Grammar (Rep.). Retrieved April 11, 2016, from State University of New York at Buffalo website: http://www.acsu.buffalo.edu/~vanvalin/rrg/RRGsummary.pdf

Van Valin, R. D., Jr. (2013). Head-marking languages and linguistic theory (Rep.). Retrieved April 11, 2016, from State University of New York at Buffalo website: http://linguistics.buffalo.edu/people/faculty/vanvalin/rrg/VanValinHead-mrkg_lgs.pdf