Computational Modeling of Polysynthetic Languages

Judith L. Klavans, Ph.D.
US Army Research Laboratory
2800 Powder Mill Road
Adelphi, Maryland 20783
Judith.l.klavans.civ@mail.mil
Judith.klavans@gmail.com

Abstract

Given advances in computational linguistic analysis of complex languages using Machine Learning as well as standard Finite State Transducers, coupled with recent efforts in language revitalization, the time was right to organize a first workshop to bring together experts in language technology and linguists on the one hand with language practitioners and revitalization experts on the other. This one-day meeting provides a promising forum to discuss new research on polysynthetic languages in combination with the needs of linguistic communities where such languages are written and spoken. Finally, this overview article summarizes the papers to be presented, along with goals and purpose.

Motivation

Polysynthetic languages are characterized by words that are composed of multiple morphemes, often to the extent that one long word can express the meaning contained in a multi-word sentence in language like English. To illustrate, consider the following example from Inuktitut, one of the official languages of the Territory of Nunavut in Canada. The morpheme -*tusaa*- (shown in boldface below) is the root, and all the other morphemes are synthetically combined with it in one unit.[1]

(1) **tusaa**-tsia-runna-nngit-tu-alu-u-junga
hear-well-be.able-NEG-DOER-very-BE-PART.1.S
'I can't hear very well.'

Kabardian (Circassian), from the Northwest Caucasus, also shows this phenomenon, with the root -*še*- shown in boldface below:

(2) wə-q'ə-d-ej-z-ɣe-**še**-ž'e-f-a-te-q'əm
2SG.OBJ-DIR-LOC-3SG.OBJ-1SG.SUBJ-CAUS-**lead**-COMPL-POTENTIAL-PAST-PRF-NEG
'I would not let you bring him right back here.'

Polysynthetic languages are spoken all over the globe and are richly represented among Native North and South American families. Many polysynthetic languages are among the world's most endangered languages,[2] with fragmented dialects and communities struggling to preserve their linguistic heritage. In particular, polysynthetic languages can be found in the US Southwest (Southern Tiwa, Kiowa Tanoan family), Canada, Mexico (Nahuatl, Uto-Aztecan family), and Central Chile (Mapudungun,

[1] Abbreviations follow the Leipzig Glossing Rules; additional glosses are spelled out in full.
[2] In fact, the majority of the languages spoken in the world today are endangered and disappearing fast (See Bird, 2009). Estimates are that, of the approximately 7000 languages in the world today, at least one disappears every day (https://www.ethnologue.com).

Proceedings of Workshop on Polysynthetic Languages, pages 1–11
Santa Fe, New Mexico, USA, August 20-26, 2018.

Araucanian), as well as in Australia (Nunggubuyu, Macro-Gunwinyguan family), Northeastern Siberia (Chukchi and Koryak, both from the Chukotko-Kamchatkan family), and India (Sora, Munda family), as shown in the map below (Figure 1).

This workshop addresses the needs for documentation, archiving, creation of corpora and teaching materials that are specific to polysynthetic languages. Documentation and corpus-building challenges arise for many languages, but the complex morphological makeup of polysynthetic languages makes consistent documentation particularly difficult. This workshop is the first ever meeting where researchers and practitioners working on polysynthetic languages discuss common problems and difficulties, and it is intended as the capstone to establishing possible collaborations and ongoing partnerships of the relevant issues.

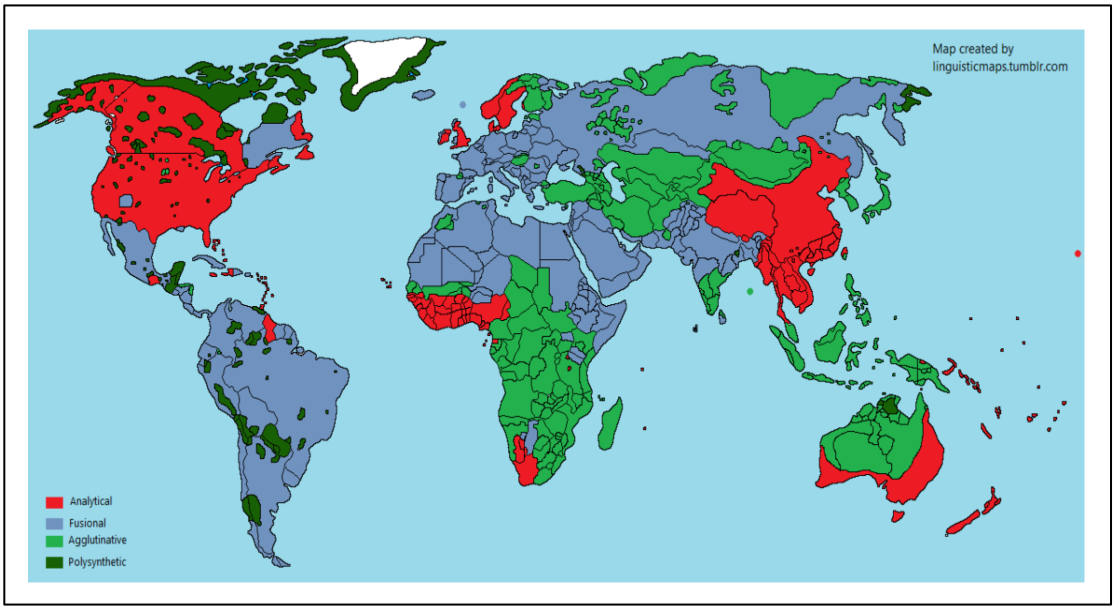


Figure 1: Polysynthetic Languages[3]

Defining Polysynthesis: An Ongoing Linguistic Challenge

Although there are many definitions of polysynthesis, there is often confusion on what constitutes the exact criteria and phenomena (Mithun 2017). Even authoritative sources categorize languages in conflicting ways.4 Typically, polysynthetic languages demonstrate holophrasis, i.e. the ability of an entire sentence to be expressed in what is considered by native speakers to be just one word (Bird 2009). In

[3] http://linguisticmaps.tumblr.com/post/120857875008/513-morphological-typology-tonal-languages

[4] For example, the article in the *Oxford Research Encyclopedia of Linguistics* on “Polysynthesis: A Diachronic and Typological Perspective” by Michael Fortescue (Fortescue, 2016), a well-known expert on polysynthesis, lists Aymara as possibly polysynthetic, whereas others designate it as agglutinative (http://www.native-languages.org).

linguistic typology, the opposite of polysynthesis is isolation. Polysynthesis technically (etymologically) refers to how many morphemes there are per word. Using that criterion, the typological cline can be represented as follows:

(3) isolating/analytic languages > synthetic languages > polysynthetic languages

Adding another dimension of morphological categorization, languages can be distinguished by the degree of clarity of morpheme boundaries. If we apply this criterion, languages can be categorized according to the following typological cline:

(4) agglutinating > mildly fusional > fusional

Thus, a language might be characterized overall as polysynthetic and agglutinating, that is, generally a high number of morphemes per word, with clear boundaries between morphemes and thus easily segmentable. Another language might be characterized as polysynthetic and fusional, so again, many morphemes per word, but so many phonological and other processes have occurred that segmenting morphemes becomes less trivial.

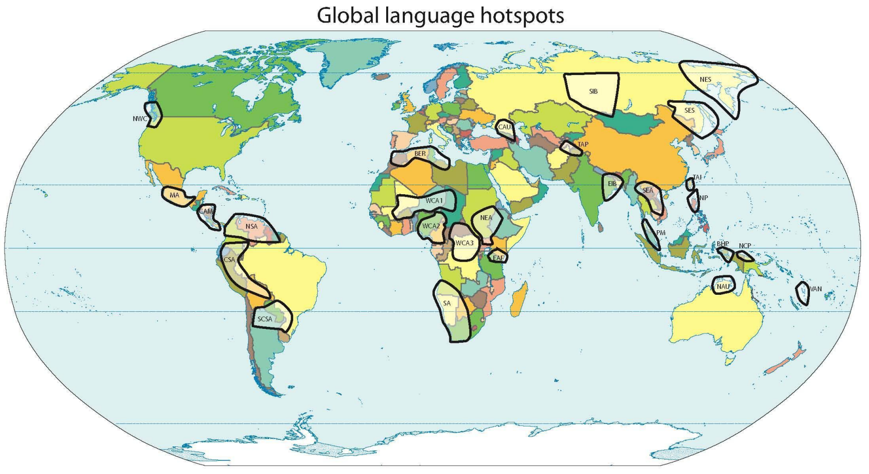
So far we have discussed the morphological aspects of polysynthesis. Polysynthesis also has a number of syntactic ramifications, richly explored in the work of Baker (Baker 1997; 2002). He proposes a cluster of correlated syntactic properties associated with polysynthesis. Here we will mention just two of these properties: rich agreement (with the subject, direct object, indirect object, and applied objects if present) and omission of free-standing arguments (pro-drop).

Polysynthetic languages are of interest for both theoretical and practical reasons. On the theoretical side, these languages offer a potentially unique window into human cognition and language capabilities as well as into language acquisition (Mithun 1989; Greenberg 1960; Comrie 1981; Fortescue et al. 2017). They also pose unique challenges for traditional computational systems (Byrd et al. 1986). Even in allegedly cross-linguistic or typological analyses of specific phenomena, e.g. in forming a theory of clitics and cliticization (Klavans 1995), finding the full range of language types on which to test hypotheses proves difficult. Often, the data is simply not available so claims cannot be either refuted or supported fully.

On the applied side, many morphologically complex languages are crucial to purposes in domains ranging from health care,[5] search and rescue, to the maintenance of cultural history. Add to this the interest in low-resource languages (from Inuktitut and Yup'ik in the North and East of Canada with over 35,000 speakers, and all the way to Northwest Caucasian), which is important for linguistic, cultural and governmental reasons. Many of the data collections in these languages, when annotated and aligned well, can serve as input to systems to automatically create correspondences, and these in turn can be useful to teachers in creating resources for their learners (Adams, Neubig, Cohn, & Bird 2015). These languages are generally not of immediate commercial value, and yet the research community needs to cope with fundamental issues of language complexity. Consequently, research on these language could have unanticipated benefits on many levels.

[5] For example, the USAID has funded a program in the mountains of Ecuador to provide maternal care in Quechua-dominant areas to reduce maternal and infant mortality rates, taking into account local cultural and language needs (https://www.usaidassist.org). Quechua is highly agglutinative, not polysynthetic; it is spoken by millions of speakers and has few corpora with limited annotation.

Finally, many of these understudied languages occur in areas that are key for health concerns (e.g. the AIDS epidemic) and international security. Consider the map in Figure 2, which shows languages identified as Language Hotspots, i.e. low resource and/or endangered. For example, many languages in the Siberian peninsula (which is of strategic political importance) are endangered and polysynthetic. Comparing the two maps in Figures 1 and 2 shows these languages are more widespread than is commonly believed. Understanding theoretical mechanisms underlying the range of language types contributes to teaching, learning, maintaining and data-mining across both speech and text in these languages and beyond.



Map by Robbie Hart and K. David Harrison, 2006
© Living Tongues Institute for Endangered Languages

Figure 2: Language Hotspots[6]

Corpus Collection and Annotation

The more language data that is gathered and accurately analyzed, the more deep cross-linguistic analyses can be conducted which in turn will contribute to a range of fields including linguistic theory, language teaching and lexicography. For example, in examining cross-linguistic analyses of headedness, Polinsky (Polinsky 2012) gathered as much data as possible to examine the question of whether the noun-verb ratio differs across headedness types. She collected as much numerical data as she could identify across a sample of languages. However, she notes that:

"[T]he seemingly simple question of counting nouns and verbs is a quite difficult one; even obtaining data about the overall number of nouns and verbs proves to be an immense challenge. The ultimate consequence is that linguists lack reasonable tools to compare languages with respect to their lexical category size. Co-operation between theoreticians and lexicographers is of critical importance: just as comparative syntax received a big boost from the micro-comparative work on closely related languages (Romance; Germanic;

[6] https://www.swarthmore.edu/SocSci/langhotspots/resources/Hotspots%20Aug%202006%20copy.jpg

Semitic), so micro-comparative WordNet building may lead to important breakthroughs that will benefit the field as a whole." (Polinsky, 2012*, p. 351*)

One of the underlying causes of this difficulty is that there are many languages for which a clear lexical division between nouns and verbs has been challenged; these languages are characterized by a large class of roots that are used either nominally or verbally, and many of these languages typically have polysynthetic features (cf. Lois & Vapnarsky 2006 for Amerindian, Aranovich 2013 for Austronesian, Testelets et al. 2009 for Adyghe, Davis & Matthewson 2009, Watanabe 2017 for Salish). Without a clear definition of what counts as a verb and what as a noun, there is no reliable way to compute significant correlations. Thus, a deeper understanding of polysynthetic phenomena may well contribute to a more nuanced understanding of cross-language comparisons and generalizations and enable researchers to pose meaningful and answerable questions about comparative features across languages.

One of the goals of the workshop is to identify and build new resources, with annotation that is effective for a range of efforts, as outlined in Levow et al. (2017). We will ensure that all materials resulting from this workshop are listed in the LDC catalog with adequate metadata giving descriptions, pointers, terms and conditions and other facts necessary for use. What we have found is that there are corpora in many different places by different types of community actors, and often they are difficult to locate and obtain. Building models and theoretical descriptions can be challenging without adequate data, and this is a gap we plan to address along with the many others involved in this endeavor.

While collections of annotated corpora (spoken and written) for major isolating, agglutinative and inflectional languages exist (https://www.ldc.upenn.edu), there are significant additional complexities involved when it comes to polysynthetic languages, including:

- tokenization - what are the boundaries for units of meaning? How are morphology and syntax delimited?
- lemmatization - where is the root? which morphemes are affixes? which are clitics?
- part-of-speech tagging
- glossing and translation into other languages

Linguistic data in these languages, be it text or audio, is scarce. This has created challenges for language analysis as well as for revitalization efforts. Only recently have researchers started collecting well-designed corpora for polysynthetic languages, e.g. for Circassian (Arkhangelskiy & Lander 2016) or Arapaho (Kazeminejad et al. 2017).

Towards a shared task

Concomitant with the collection and cataloging of corpora, as part of the workshop, we aim to formulate a shared task, that meets the goals outlined in Levow, et al. (2017), namely, to "align the interests of the speech and language processing communities with those of endangered language documentation communities." Levow et al. 2017 propose an initial set of possible shared tasks based on the design principles of realism, typological diversity, accessibility of the shared task, accessibility of the resulting software, extensibility and nuanced evaluation. In addition to coordinating with the NSF-funded EL-STEC project, we have consulted with the SIGMORPHON organizers,[7] and Morpho Challenge project. We have also collaborated with organizers of the Documenting Endangered Languages Workshops (notably Jeff Good of the University at Buffalo). We have also coordinated with the NSF-funded CoLang program (Institute on Collaborative Language Research) at the University of Florida (http://colang.lin.ufl.edu/). Given the challenges of compiling a shared task, we have planned sessions during the workshop for participants to engage together in the creation of a shared task. In this way, we will involve community activists in the task formulation, which will lead to a higher chance of actually meeting local language needs.

Related Projects and Conferences

In recent years, there has been a surge of major research on many of these languages. For example, the first Endangered Languages (ELs) Workshop held in conjunction with ACL was held in 2014 and the second in 2017.[8] The National Science Foundation and the National Endowment for the Humanities jointly fund a program for research on ELs.[9] The US government through IARPA and DARPA both have programs for translation, including for low resource languages.[10] The IARPA BABEL project focused on keyword search over speech for a variety of typologically different languages, including some with polysynthetic features.

To reiterate, an interdisciplinary workshop specifically on the challenges of dealing with polysynthesis in computational linguistics has not been held before. The languages involved in Morph-Challenge (http://morpho.aalto.fi/events/morphochallenge/) did not include polysynthetic languages, nor did SIGMORPHON (http://ryancotterell.github.io/sigmorphon2016/). Given recent advances in computational morphology, a workshop that addresses the full range of morpho-syntactic features of language, extending to and including polysynthesis, is timely.

As indicated above, this workshop brings together researchers from multidisciplinary fields to address ongoing challenges and to compare outputs of various recent approaches, resulting in a lively venue for discussion and argument. The specific goals of the proposed workshop include:

1. To bring together experts in linguistic theory and computational linguistics with those working on preserving and reviving indigenous languages.
2. To discuss the potential of technologies (e.g., text-to-speech systems, segmentation of speech files by speaker, audio-indexing, morphological analysis) to assist in language revitalization.
3. To construct and annotate data sets in these languages for use by the relevant linguistic communities; these datasets can be used for research and practical applications.

[7] https://sites.google.com/view/conll-sigmorphon2017/home?authuser=0;
http://www.aclweb.org/old_anthology/W/W16/W16-20.pdf#page=22.

[8] http://www.acsu.buffalo.edu/~jcgood/ComputEL.html; http://altlab.artsrn.ualberta.ca/computel-2/.

[9] https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=12816; https://www.neh.gov/grants/manage/general-information-neh-nsf-documenting-endangered-languages-fellowships.

[10] MATERIAL, https://www.iarpa.gov/index.php/research-programs/material and LORELEI, http://www.darpa.mil/program/low-resource-languages-for-emergent-incidents, respectively.

4. To explore a deeper understanding of polysynthesis as a linguistic phenomenon.

Discussion of Workshop Papers

Eight papers were accepted to the Workshop. The languages and technologies discussed are wide-ranging and reflected the intended nature of the meeting as inclusive and exploratory. Languages include:

- Hinóno'eitíít - Arapaho (in English:), one of the highly-endangered Plains Algonquian languages (unknown numbers, ranging from 500 to 2500 speakers)
- Nahuatl, Wixarika and Yorem Nokki - from the Uto-Aztecan language family (estimated 1.5 million speakers)
- Kwak'wala - spoken by the Kwakwaka'wakw people (which means "those who speak Kwak'wala") and highly-endangered, belonging to the Wakashan language family (estimated 250 speakers).
- Kanyen'kéha (Ohsweken dialect) - language of the Iroquoian family commonly known as Mohawk, spoken in parts of Canada (Ontario and Quebec) and the United States (New York state) with about 3500 native speakers.
- Inuktitut - one of the principal Inuit languages used in parts of Newfoundland and Labrador, Quebec, the Northwest Territories and Nunavit, recognized as an official language in the Province of Nunavut with about 40,000 speakers.
- Chuckchi - a Chukotko–Kamchatkan language spoken in the easternmost extremity of Siberia, mainly in Chukotka Autonomous Okrug, rapidly decreasing in speakers with only about 500 native speakers left, down from nearly 8000 15 years ago.

We accepted one paper on an agglutinative language, with projected hypotheses on how the techniques might apply to some of the challenges of polysynthesis, namely;

- Lezgi (лезги), a statutory language of provincial identity in Dagestan Autonomous Republic west of the Caspian sea coast in the central Caucasus and a member of the Nakh-Daghestanian languages (approx. 600,000 speakers).

Our justification for including this paper is that we believe the authors may be able to test their techniques on other languages, so this paper will serve as a baseline for future research.

The technologies range from research on Finite State Transducers (FSTs), Statistical and Rule-Based Machine Translation (SMTs), Conditional Random Fields (CRF) and CRF with Support Vector Machines (CRF-SVM), Neural Machine Translation (seq2seq) and Segmental Recurrent Neural Nets (SRNNs). Applications include morphological analysis, glossing, verb conjugation and generation, machine translation.

Although each article in the Workshop represents a specific and original contribution, either in method or in application of method to a given polysynthetic language or language group, as a whole, this collection of papers contributes to the literature that addresses the interdependence between linguistic theory, language revitalization, education and computational contributions. These relationships are reflected in the choice of invited speaker and in the panel.

Invited speaker

We are honored to have had the invited talk from Brian Maracle (Owennatekha, Turtle Clan, Mohawk), founder and teacher at the Onkwawenna Kentyohkwa Mohawk immersion school and head of the Mohawk-language school on the Six Nations Reserve near Brantford, Ontario. Maracle has been a language activist for nearly 25 years and has developed and published materials, as well as teaching adults and young people. He left a lucrative career to return to the reservation of his youth. His book Back on the Rez: Finding the Way Home (Penguin 1993) documents his path back and struggles to understand meetings held in the Kanyen'kehaka (Mohawk) language. These experiences led to his groundbreaking work in language revitalization.[11] Brian's dynamic and deep commitment to language documentation, teaching, and policy have had an impact on many people from linguists to anthropologists to teachers to elders to children and even to politicians.

Invited Panel – How Can We Work Together?

One of the goals of the Workshop is to create dialog between those language professionals who collect and annotate language data for polysynthetic languages, those who are committed to linguistic analysis, those who develop and apply computational methods to these languages, and those who are dedicated to revitalization through policy, education and community activism. Too often, these communities do not interact enough to benefit each other, so there is a lost opportunity cost all around. This lost opportunity, especially in the case of endangered languages, is one that cannot be recuperated. Thus, it is urgent to work towards the goal of leveraging each other's efforts.

Towards this end, we have organized a panel the purpose of which is to address and debate some of the controversial issues that arise in the process of establishing better communication. Some of these issues include such provocative and often divisive points such as:

- I am a teacher and none of your so-called useful tools are of any use to me. Why can't you come to my classroom and see what we really need?
- I am a computer scientist and I want to find out what is the best method to use to figure out how to morphologically analyze and label your really long words? How much text can you annotate for me so that I can train my systems?
- I am a speech recognition expert and I really need more data of spoken language transcribed into an accurate phonetic representation? Why can't you just ask people to make some recordings for me and then turn that into text?
- I am a revitalization expert and I want to establish new policy for my town so we can get a new school started. If you're a linguist, what can you tell me about other programs and how it might help in enforcing cultural identity and competence so I can convince people that we need funding?

[11] http://www.thecanadianencyclopedia.ca/en/article/brian-maracle/.

- I am developing curricula for a set of new classes for my endangered language. What kind of experience do you have in making dictionaries so my students can look up words they don't know?
- I am the child in a family where my parents and grandparents only speak their local language and not the dominant language of the government. I want to make sure that all important government documents are translated into my local language so the many elders like mine are empowered and so that I can pass this language onto my children. You are a computational linguist so how can you help? In fact, do you even care about this?

Future work will include follow up on documentation, corpus collection, revitalization, annotation, tools for analysis and methods to contribute both to the wide range of fields this research draws upon and impacts.

Acknowledgements:

Of the many workshops this author has organized and of the many program chairships she has held, she has never experienced 100% fulfillment of reviewing commitments from the program committee. The reviewers were thorough and detailed, which was deeply appreciated by the authors. More than one author thanked the organizing committee for the exhaustive in-depth anonymous reviews from the different perspectives of the fields represented. I believe this is due to the fact that each of the reviewers is committed to language maintenance, to linguistic accuracy and to the potential for computational approaches to make a contribution to teaching and learning of these rich and challenging languages.

We are especially grateful to the University of Maryland, especially Omer Preminger and Maria Polinsky for discussions on focus and purpose for the workshop; to the National Science Foundation for funding a Research Experience for Undergraduates (REU) proposal under an award to Maria Polinsky at the University of Maryland entitled "Cleaning, Organizing, and Uniting Linguistic Databases (the CLOUD Project)" [BCS 1619857] to support three undergraduates to attend the conference and to perform research on papers, authors and language revitalization efforts relevant to their undergraduate training; to Aaron Broadwell, who has headed the 2018 CoLang program and who was helpful in identifying promising undergraduates for the REU; to the COLING organizers, program committee and COLING Workshop organizers. The Army Research Laboratory in Adelphi, Maryland has supported three of the conference organizers and the National Research Council of Canada has supported two of the conference organizers in this endeavor.

Bibliography

Adams, O., Neubig,, G., Cohn, T., & Bird, S. (2015). Inducing bilingual lexicons from small quantities of sentence-aligned phonemic transcriptions. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2015).* Da Nang, Vietnam.

Aranovich, R. (2013). Transitivity and polysynthesis in Fijian. Language 89: 465-500.

Arkhangelskiy, T. A., & Lander, Y. A. (2016). Developing a polysynthetic language corpus: problems and solutions. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2016"* , June 104, 2016.

Baker, M. C. (1996). *The polysynthesis parameter.* New York: Oxford University Press.

Baker, M.C. (2002). *Atoms of language*. New York: Basic Books.

Bird, S. (2009). Natural language processing and linguistic fieldwork . *Computational Linguistics, 35* (3), 469-474.

Byrd, R. J., Klavans, J. L., Aronoff, M., & Anshen, F. (1986). Computer methods for morphological analysis. *Proceedings of the 24th annual meeting on Association for Computational Linguistics* (pp. 120-127). Stroudsberg, PA: Association for Computational Linguistics.

Comrie, B. (1981). *Language Universals and Linguistic Typology.* Oxford: Blackwell.

Davis, H., & Mattewson, L. (2009). Issues in Salish syntax and semantics. *Language and Linguistics Compass 3,* 1097-1166.

Fortescue, M. (2016). Polysynthesis: A Diachronic and Typological Perspective . In M. Aronoff (ed.) *Oxford Encyclopedia of Linguistics*. Oxford, Oxford, England: Oxford University Press.

Fortescue, M. (1994). Polysynthetic morphology . (R. E. al., Ed.) *The encyclopedia of language and linguistic, 5*, 2600–2602.

Fortescue, M., Mithun, M., & Evans, N. (Eds.). (2017). *The Oxford Handbook of Polysynthesis.* Oxford: Oxford University Press.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language . *International Journal of Linguistics,, 26*, 178–194.

Kazeminejad, G., Cowell , A., & Hulden , M. (2017). Creating lexical resources for polysynthetic languages—the case of Arapaho. *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages* (pp. 10-18). Honolulu: Association for Computational Linguistics.

Klavans, J. L. (1995). *On Clitics and Cliticization: The Interaction of Morphology, Phonology, and Syntax.* New York: Garland .

Levow, G.-A., Bender, E., Littell, P., Howell, K., Chelliah, S., Crowgey, J., et al. (2017). STREAMLInED Challenges: Aligning Research Interests with Shared Tasks. *Proceedings of ComputEL-2: 2nd Workshop on Computational Methods for Endangered Languages, .*

Lois, X., & Vapnarsky, V. (2006.). Root indeterminacy and polyvalence in Yukatecan Mayan languages. In X. Lois, & V. Vapnarsky (Eds.). L*exical categories and root clauses in Amerindian languages* (pp. 69-115). Bern: Peter Lang.

Mithun, M. (1989). The acquisition of polysynthesis. *Journal of Child Language, 16*, 285–312.

Mithun, M. (2017). Argument marking in the polysynthetic verb and its implications. In M. Fortescue, M. Mithun, & N. Evans (Eds.), *The Oxford Handbook of Polysynthesis* (pp. 30-58). Oxford, UK: Oxford University Press.

Polinsky, M. (2012). Headedness, again. *UCLA Working Papers in Linguistics, Theories of Everything. 17*, pp. 348-359. Los Angeles: UCLA.

Sadock, J. (1986.). Some Notes on Noun Incorporation. *Language , 62,*, 19–31.

Testelets Ya. (ed.). (2009). *Aspekty polisintetizma: Očerki po grammatike adygejskogo jazyka [Aspects of polysynthesis: Essays on Adyghe grammar],* (pp. 17-120). Moscow: Russian University for the Humanities.

Watanabe, H. (2017). The polysynthetic nature of Salish. In Fortescue, M., Mithun, M., & Evans, N. (Eds.). (2017). *The Oxford Handbook of Polysynthesis* (pp. 623-642). Oxford: Oxford University Press.