Lushootseed semantic composition and the structure of the lexicon

Joshua Crowgey
University of Washington

Abstract: In this paper, I present an outline of my dissertation research question which regards the semanic compositionality of Lushootseed roots and stems, and the ramifications of corresponding structures in the lexicon. I touch on the relevant literature and then discuss a computational environment for modeling the problem and investigating the issue.

Keywords: UBCWPL, Implemented Grammar, Semantic Structure, Lushootseed

1 Introduction

While Salishan languages are known for wearing their derivational structure "on their sleeves" (Davis and Matthewson 2009:1098), there is some debate about the productivity of these derivational structures and their encoding with respect to the atomic elements of a lexicon in a synchronic grammar. Davis and Matthewson (2009), argue for what they call the "primacy of the root", the hypothesis that Salishan roots do contribute semantic content in the synchronic lexicon. This is contra Hess (1993)[1] who claims that using the verb-stem as the basis for lexical organization maximizes the perspicuity of the model—that treating verb roots as atomic units obscures the natural classes which fall out of the data when looking at stems. But the question of what morphological level to treat as atomic in a synchronic model of a Salishan language also has ramifications on semantic composition. I assume a neo-Davidsonian semantics in which verbs introduce event-type variables which are qualified by lexical predicates (Copestake et al. 2005; Davidson 1967; Parsons 1990) and nouns introduce individual-type variables which can serve as role-players in predicates. This basic split into lexical items which introduce events versus those which introduce individuals has ramifications on monotonic semantic compositionally because no matter which level of verbal decomposition one takes as basic (verb root or verb stem), nominalization prefixes are allowed to attach outside of inflectional markers of aspect. So the question arises whether or not to allow the basic units of verbs (be they stems or roots) to be underspecified with respect to semantic variable type (essentially making them semantically neutral with respect to semantic type), only specifying them once they have enough morphology attached that they can no longer become nominal. An alternative might be to allow semantic rewrites of variable types, potentially breaking standard assumptions about the monotonicity of semantic composition.

[1] Davis and Matthewson also cite Mattina (1996) and Willet (2003) as supporting the stems based account.

In Papers for the International Conference on Salish and Neighbouring Languages 50,
University of British Columbia Working Papers in Linguistics 40,
Natalie Weber, Erin Guntly, Zoe Lam, and Sihwei Chen (eds.), 2015.

In this short paper, I provide a little more detail about the problem to be investigated. After that, I present a methodology for exploring the issue and describe the preliminary implementation.

2 Roots and stems

Davis and Matthewson (2009) summarize arguments against the primacy of the root hypothesis as being along two lines: the first being that the semantic relation which holds between roots and derived forms is not always compositional and the second being the existence of accidental gaps in derivational paradigms along with the phenomenon of certain bound roots which do not surface without affixation. Davis and Matthewson reject these arguments by pointing out that existence of idiosyncrasy does not preclude the utility of any generality otherwise gained by describing the non-idiosyncratic forms by rule. That is, a system which accommodates exceptional forms is needed anyway and the phenomenon of linguistic exceptions to general rules is found at all levels of linguistic analysis.

While I accept Davis and Matthewson's arguments in principle, from the perspective of constructing a working synchronic grammar of Lushootseed, it's not clear what is to be gained by creating rules to systematically build up by formal composition forms which are semantically non-compositional. What's more, if Hess (1993) is right about the natural classes of verbs "falling out" of the data when operating with a stems-based approach, the structures in a stems-based grammar should offer greater explanatory value for the synchronic processes of the language than the alternative. However, I hold-off drawing any conclusion on this topic and instead turn to an related issue, one of semantic variable types and its relation to multicategoricity.

3 Semantic variables

I assume a semantic compositionality principle where the meaning of a complex sign is a function of the meaning of its parts (Szabó 2013). Furthermore, as mentioned in the introduction, I assume a neo-Davidsonian approach to semantic representation in which verbal predicates take a characteristic argument corresponding to the event the predicate describes and noun-like predicates take a characteristic argument corresponding to the individual. Minimal Recursion Semantics[2] (MRS) (Copestake et al. 2005) provides a practical encoding of this semantic representation which is amenable to implementation in machine-readable grammars. In

[2]MRS saves space and computation time by allowing an underspecification of quantifier scope in just those places where syntax allows quantifier scope ambiguity as well as a mechanism for constraining quantifier scope where syntax requires it. Algorithms for producing the set of fully scope-resolved logical forms from an underspecified MRS have been publshed. In this way, MRS provides a compact and efficient representation for sentential semantics but can be translated to logical form when required by particular downstream applications.

MRS, the variable types which fill the argument positions of the elementary predications are either e-type (events/states), x-type (instances, individuals), or h-type.[3] For example, the three elementary predications involved in an MRS representation for the English sentence 'Kim walks' are shown in (1). Note that the predications corresponding to the proper name quantifier and the name itself take the variable *x2* as their first argument. Similarly, the predication for walk takes the event-type variable *e8* as a characteristic argument.

(1) h1:proper_q_rel (x2, h3, h4)
h5:named_rel (x2, "Kim")
h7:walk_v_rel (e8, x2)

Nominalization (or deverbalization) and verbalization patterns in Lushootseed, however, pose challenges for this representation scheme. When a Lushootseed noun such as *sqʷšab* ('fog') is analyzed as derived from the verb *qʷšab* ('be foggy'), this begs the question of how to introduce the individual type variable which would typically be introduced by a noun. If we just add a new variable, we risk a proliferation of variables (of conflicting types) which correspond to a single lexical item. If we remove or delete the event-type variable from the underlying verb, we break monotonicity in semantic composition. Another alternative would be to utilize a lexical underspecification in which the characteristic variable of a given item would be underspecified for semantic type in the lexicon. In this scenario, morphological rules would hold off constraining the type until enough morphological material had been added to make a concrete determination. I note that the underspecification alternative, *prima facie*, fits nicely with the primacy of the root hypothesis discussed above. That is, if constraints on monotonicity in semantic composition suggests an analysis in which lexical items have to be underspecified as to semantic type, perhaps the root, not the stem, is the place to hang these underspecified variables.

4 An empirical question

The analytical challenges discussed above and the question of which set of analyses is to be preferred is, under certain assumptions, an empirical one. Or, it can become an empirical one given: (1) a fleshed-out, concrete theory of grammatical structures which allows implementation; (2) a testing environment including procedures for applying the competing grammatical models to linguistic data in order to evaluate the analyzed structures as well as procedures to evaluate structures of the grammatical models themselves; and (3) competing implementations corresponding to competing sets of hypotheses about linguistic structures. To this sort of quantitative analysis one can also add a qualitative evaluation of the competing grammatical models; that is, one can ask which grammar is easier to update and

[3]The latter type is only used MRS-internally for specifying the scope-tree constraints; that is, h-type variables do not map onto semantic variables in a typical logical-form language (http://moin.delph-in.net/ErgSemantics/Basics).

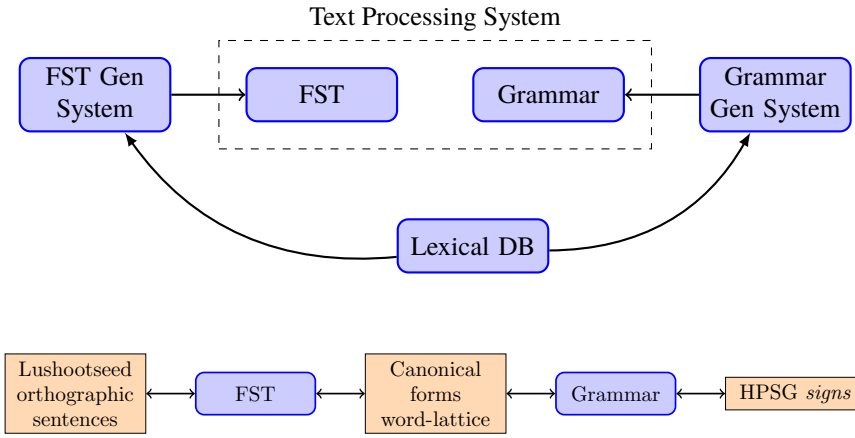

Figure 1: *Above*: the two subsystems of a text processing system for testing are generated by independent systems which share a lexical resource. *Below*: the text processing system forms a pipeline connecting orthographic sentences to semantic and syntactic structures.

maintain, which is "simpler" or more elegant (or just less baroque), which is preferable along the lines of having structures which fall into correspondence with other grammars of other languages. In sum, a mathematically precise theory of grammar along with a testing environment allows empirical analysis of linguistic questions. As things stand, (1) is provided by Head-driven Phrase Structure Grammar (Pollard and Sag 1994) as implemented in the DELPH-IN tools and environment.[4] Point (2) is likewise provided by the DELPH-IN tools. In my dissertation project, I am building out (3) in an effort to make the analytical questions discussed above into empirical ones.

Adopting the design architecture argued for in Bender and Good (2005), I am maintaining separate modules for morphophonology and morphosyntax, each of these sharing a lexical database. In order to maximize efficiency in updating and maintaining the system, the two modules are generated by metagrammar systems which access the lexical database for shared resources. The two metagrammar systems and the database fit together to generate a text-processing system. The overview of this architecture is shown schematically in Figure 1.

The electronic versions of the texts in Beck and Hess (2014, 2015) form a development and testing corpus that can be used to compare the performance of competing versions of the grammar. Furthermore, Fokkens (2011, 2014) describes new methodologies for comparing alternative analyses within grammar engineering. The crucial insight is that by using a metagrammar, one can alleviate the grammar-design problem in which early decisions are overly influential on the space of analysis available for later decisions.

[4] http://delph-in.net/

In sum, I am developing a metagrammar for Lushootseed (and supporting testing environment) intended to serve as a proving-ground for the research questions outlined above and summarized in (2).

(2) What are the ramifications of competing grammatical analyses regarding semantic variables and lexical structure? Specifically, how do different sets of analyses effect:

- the performance of the grammar (coverage, accuracy, overgeneration: when applied to a test corpus)
- the maintainability of the grammar (as a system to be extended, updated, refined)
- the "elegance" of the grammar and its utility to provide insight into the linguistic structures at play in a Lushootseed sentence

5 Conclusion

In my dissertation project, I flesh out a computational model of Lushootseed grammatical structure in order to tease apart the predictions of these competing hypotheses and to examine their ramifications on the design of a working grammar, one in which "all parts have to hang-together"[5] I suggest that the creation of such a test-environment can lead to new insights on questions of linguistic analysis like those described above.

While the initial morphophonological system was described in Crowgey (2014), work has just begun on the morphosyntactic grammar. I have written extraction procedures to backport the original morphophonological system into the lexical database, and to then allow export back to the original structures (future development will take place only in the lexical database—for maintainability and synchrony with the development of the morphosyntactic system). In future work I will report on the results of these efforts. Ideally, the empiricization of the philosophical and analytical questions described above can provide fodder for further refinement of theories on Lushootseed lexical structure and its relation to other Salishan languages.

References

Beck, D. and Hess, T. (2014). *Tellings from Our Elders: Lushootseed syəyəhub. Volume 1, Snohomish texts*. UBC Press, Vancouver.

Beck, D. and Hess, T. (2015). *Tellings from Our Elders: Lushootseed syəyəhub. Volume 2. Tales from the Skagit Valley*. UBC Press, Vancouver.

[5] Miellet quotation (Miellet 1903) translated by Emily Bender in Bender (2008).

Bender, E. M. (2008). Grammar engineersing for linguistic hypothesis testing. In Gaylord, N., Palmer, A., and Ponvert, E., editors, *Proceedings of the Texas Linguistics Society X Conference: Computational Linguistics for Less Studies Languages*, pages 16–36. Stanford CSLI Publications ONLINE.

Bender, E. M. and Good, J. (2005). Implementation for Discovery: A bipartite lexicon to support morphological and syntactic analysis. *Chicago Linguistic Society*, 41(The Panels).

Copestake, A., Flickinger, D., Pollard, C., and Sag, I. A. (2005). Minimal Recursion Semantics: An introduction. *Research on Language & Computation*, 3(4):281–332.

Crowgey, J. (2014). Constructing a Morphophonological Analyzer for Lushootseed. In Weber, N., Sadlier-Brown, E., and Guntly, E., editors, *University of British Columbia Working Papers in Linguistics*, volume 37, pages 39–63.

Davidson, D. (1967). The logical form of action sentences. *The logic of decision and action*, pages 81–95.

Davis, H. and Matthewson, L. (2009). Issues in salish syntax and semantics. *Language and Linguistics Compass*, 3(4):1097–1166.

Fokkens, A. (2011). Metagrammar engineering: Towards systematic exploration of implemented grammars. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, Portland, Oregon, USA. Association for Computational Linguistics.

Fokkens, A. (2014). *Enhancing Emipirical Research for Linguistically Motivated Precision Grammars*. PhD thesis, Universität des Saarlandes, Saarbrüken.

Hess, T. (1993). A schema for the presentation of Lushootseed verb stems. *American Indian linguistics and ethnography in honor of Lawrence C. Thompson*, pages 113–126.

Mattina, N. (1996). *Aspect and Category in Okanagan Word Formation*. PhD thesis, Simon Fraser University, Burnaby, British Columbia.

Miellet, A. (1903). *Introduction à l'étude comparative des languges indo-européennes*. Hatchet, Paris.

Parsons, T. (1990). *Events in the semantics of English: a study in subatomic semantics*. MIT Press, Cambridge, MA.

Pollard, C. and Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar*. University of Chicago Press, London.

Szabó, Z. G. (2013). Compositionality. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Fall 2013 edition.

Willet, M.-L. (2003). *A Grammatical Sketch of Nxa'amxcin (Moses-Columbia Salish)*. PhD thesis, University of Victoria, Victoria, British Columbia.