Clustering Affixes into Position Classes for Low-Resource Languages: A Case Study of Chintang

Olga Zamaraeva
University of Washington
olzama@uw.edu

1. Morphotactics: Position Classes

- The study of morpheme ordering constraints
- Complex morphotactics is underrepresented in NLP systems

Finnish Verb Morphotactics:

(1) *root + passive + tense/mood + person + particle*

2. Precision Grammars and the Grammar Matrix

- Machine-readable sets of linguistic generalizations
  English Resource Grammar (Flickinger, 2000)
- Based on HPSG theory of syntax (Pollard & Sag, 1994)
- Grammar Matrix: A grammar engineering development kit (Bender et al., 2010, 2002; Goodman & Bender, 2010; O'Hara, 2008)
- LKB System (parsing, generation; visualization) (Copestake, 2002)

```
prp-lex-rule := infl-lex-rule & non-fin-lex-rule-super &
  [ SYNSEM.LOCAL [ CONT.HOOK.INDEX.E.ASPECT prog,
                   CAT.HEAD.FORM prp ] ].
```

Figure 1: *English position class precision grammar snippet.*

An example of a morphological parse by the LKB system with non-branching S and VP rules:

S – VP – V (+ing) – V (walk)

3. Data: Low-Resource Languages and Interlinear Glossed Text (IGT)

- ~90% of the world's languages (Krauss, 1992) are low-resource
- Complex morphotactics is common
- Documenting is urgent for endangered varieties
- Chintang (ISO-639: ctn) IGT collection, shared by Bickel et al. (2013); ~9K IGT
- Gold standard exists for morphological rules (Bender et al., 2012)

unisaŋa	khatte	mo	kosi	moba
u-nisa-ŋa	khatt-e	mo	kosi-i	mo-pe

3sPOSS-younger.brother-ERG.A take-IND.PST DEM.DOWN river-LOC DEM.DOWN-LOC

'The younger brother took it to the river.' [ctn] (Bickel et al., 2013)

Figure 2: *Sample Chintang IGT*

4. Baseline: Morphotactic constraints as a DAG

- Two-Level Morphology (Koskenniemi, 1984)
- Affixes are nodes, input relationships are directed edges
- No cycles allowed in the Grammar Matrix (Can affixes cycle?..)
- Wax (2014) infers morphotactics from IGT by input overlap heuristic

1) Ivan po-kup-a-et sobak-u
Ivan.NOM PREF1-buy-IMPF-3SG.PRES dog-ACC
Ivan is buying a dog.

2) Ivan vy-kup-i-l sobak-u
Ivan.NOM PREF2-buy-PFV-3SG.M.PST dog.ACC
Ivan bought the dog out. [rus]

5. Baseline contd.

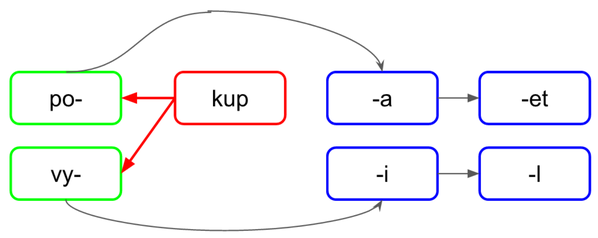

Figure 4: *Input: Affix graph. Method: Input overlap.*

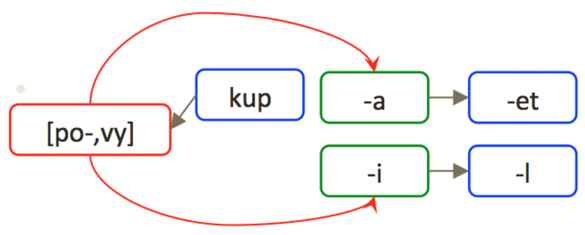

Figure 5: *Method: Input overlap.*

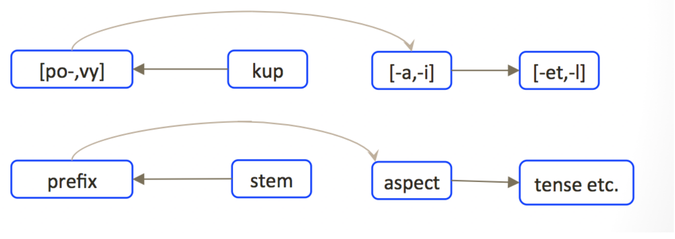

Figure 6: *Output: Position classes graph.*

6. Method: Clustering affixes

- Build the original affix DAG from the input IGT
- Use k-means to group affixes-nodes
  Choose k manually: match baseline (263), match the gold standard (54), and the literature (13)
- Reconstruct the DAG using whichever edges do not form cycles

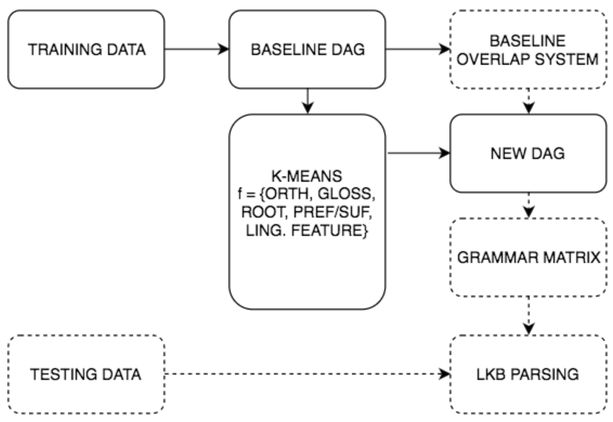

Figure 7: *General method.*

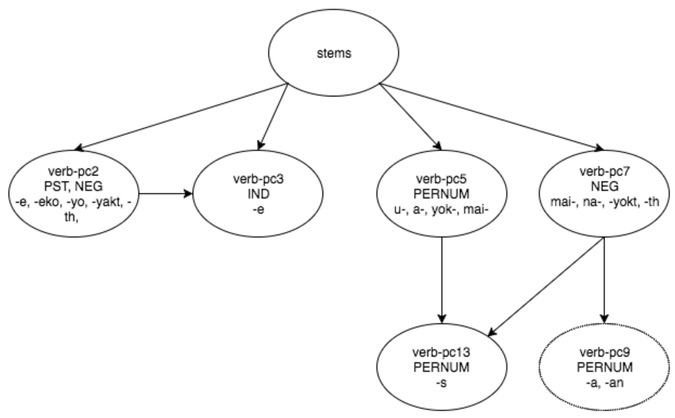

Figure 8: *Abridged Chintang final graph, k=13. To a degree, edges and clusters reflect actual Chintang morphotactics. Some edges and nodes are not shown.*

7. Results

- Evaluation by LKB morphological parsing
- Baseline: Affix input overlap (Wax, 2014)
- V-measure: based on the gold standard grammar and a manually labeled subset of data (1000 IGT)

Language	System	k/PC	% parsed
	Wax (2014)	263	**92.95**
	k-means	263	85.91
ctn	Oracle	54	76.05
	k-means	54	**92.60**
	k-means	13	85.58

homogeneity	completeness	v-measure
0.896	0.723	0.800

8. Discussion

- Both baseline and k-means strongly outperform the hand-built grammar
- Automatic procedures have high recall but lower precision than the hand-built grammar
- k-means is capable of clustering together affixes which participate in non-canonical phenomena
- Active learning seems a promising direction for future research

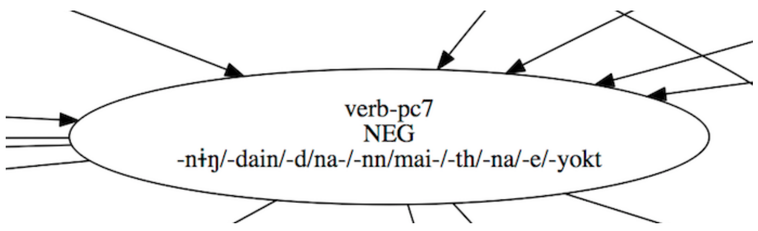

Figure 9: *Chintang circumfix mai- -yokt in one cluster.*

9. Acknowledgments

Thanks to Emily M. Bender for the poster LaTeX source.

My work on this project is partially supported by Microsoft Graduate Women Fellowship.

References

Bender, E. M., Drellishak, S., Fokkens, A., Poulson, L., & Saleem, S. (2010). Grammar customization. *Research on Language & Computation*, *8*(1), 23–72. Retrieved from http://dx.doi.org/10.1007/s11168-010-9070-1 (10.1007/s11168-010-9070-1)

Bender, E. M., Flickinger, D., & Oepen, S. (2002). The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In J. Carroll, N. Oostdijk, & R. Sutcliffe (Eds.), *Proceedings of the workshop on grammar engineering and evaluation at the 19th international conference on computational linguistics* (pp. 8–14). Taipei, Taiwan.

Bender, E. M., Schikowski, R., & Bickel, B. (2012). Deriving a lexicon for a precision grammar from language documentation resources: A case study of Chintang. In *Coling* (pp. 247–262).

Bickel, B., Gaenszle, M., Rai, N. K., Rai, V. S., Lieven, E., Stoll, S., … Rai, I. P. (2013). *Tale of a poor guy.* (Accessed online on 15-January-2013)

Copestake, A. (2002). *The LKB system.* Stanford University.

Flickinger, D. (2000). On building a more effcient grammar by exploiting types. *Natural Language Engineering*, *6*(01), 15–28.

Goodman, M., & Bender, E. M. (2010). What's in a word? refining the morphotactic infrastructure in the LinGO Grammar Matrix customization system. In *Workshop on morphology and formal grammar, paris.*

Koskenniemi, K. (1984). A general computational model for word-form recognition and production. In *Proceedings of the 10th international conference on computational linguistics* (pp. 178–181).

Krauss, M. (1992). The world's languages in crisis. *Language*, *68*(1), 4–10.

O'Hara, K. (2008). *A morphotactic infrastructure for a grammar customization system.* Unpublished doctoral dissertation, University of Washington.

Pollard, C., & Sag, I. A. (1994). *Head-Driven Phrase Structure Grammar.* Chicago, IL and Stanford, CA: The University of Chicago Press and CSLI Publications.

Wax, D. (2014). *Automated grammar engineering for verbal morphology.* Unpublished master's thesis, University of Washington.