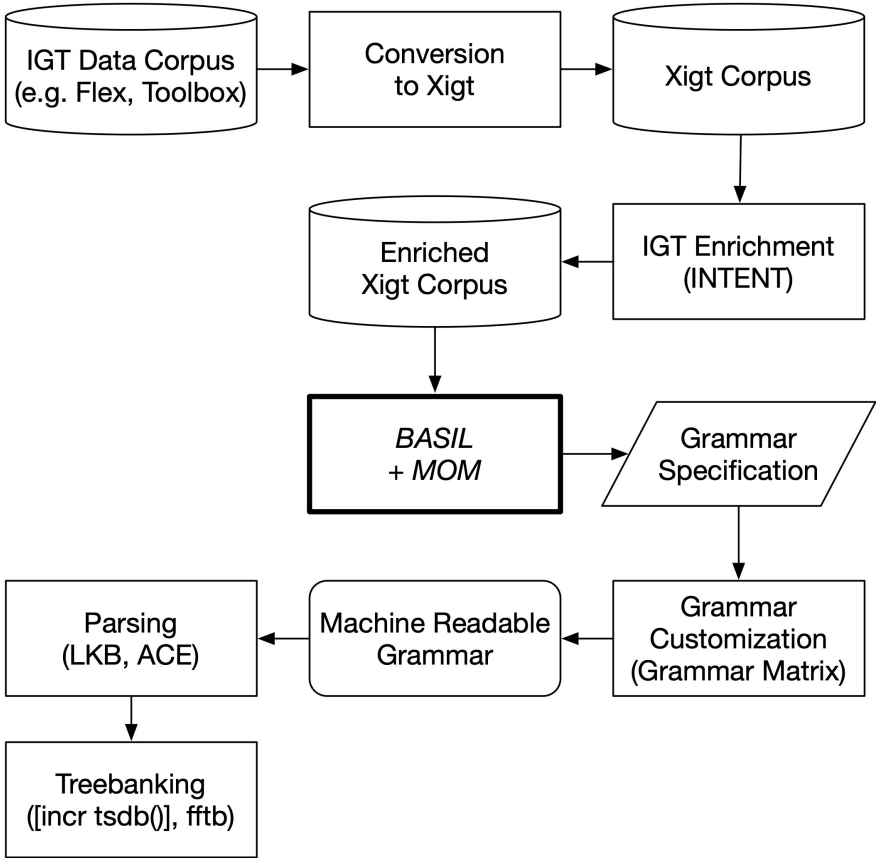
Tracing and Reducing Lexical Ambiguity in Automatically Inferred Grammars

Liz Conrad

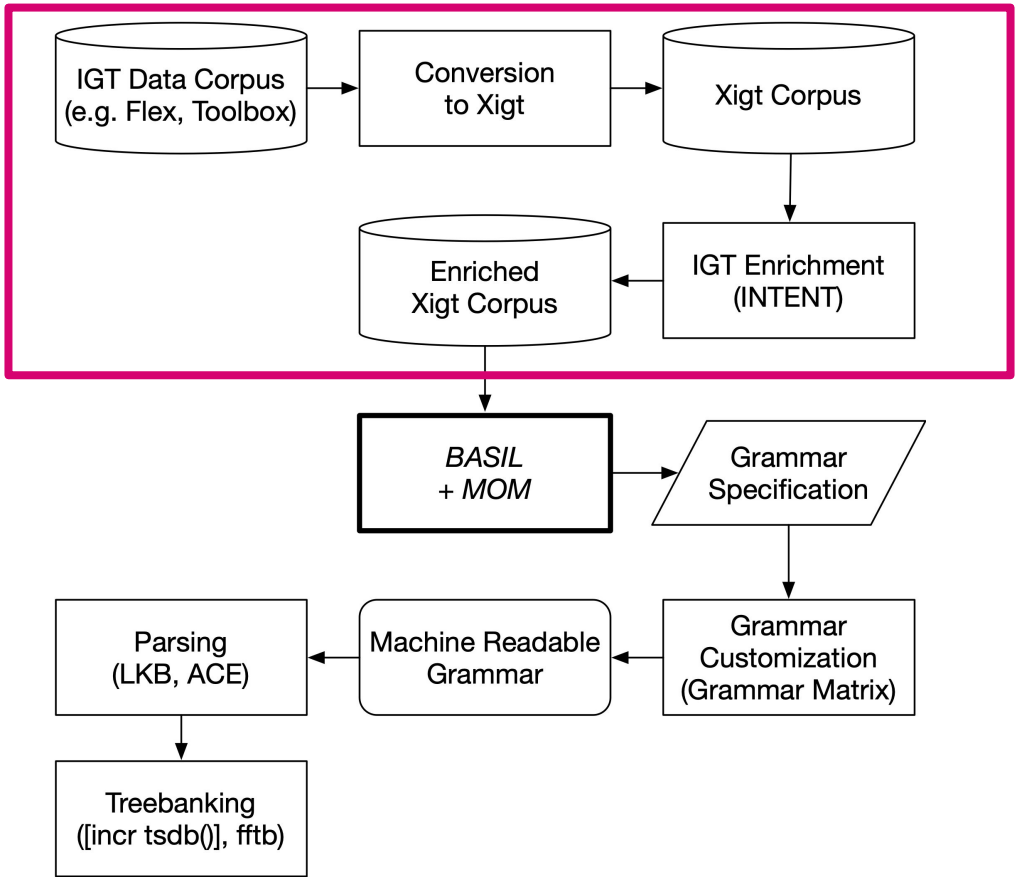
DELPH-IN 2021

The AGGREGATION Project



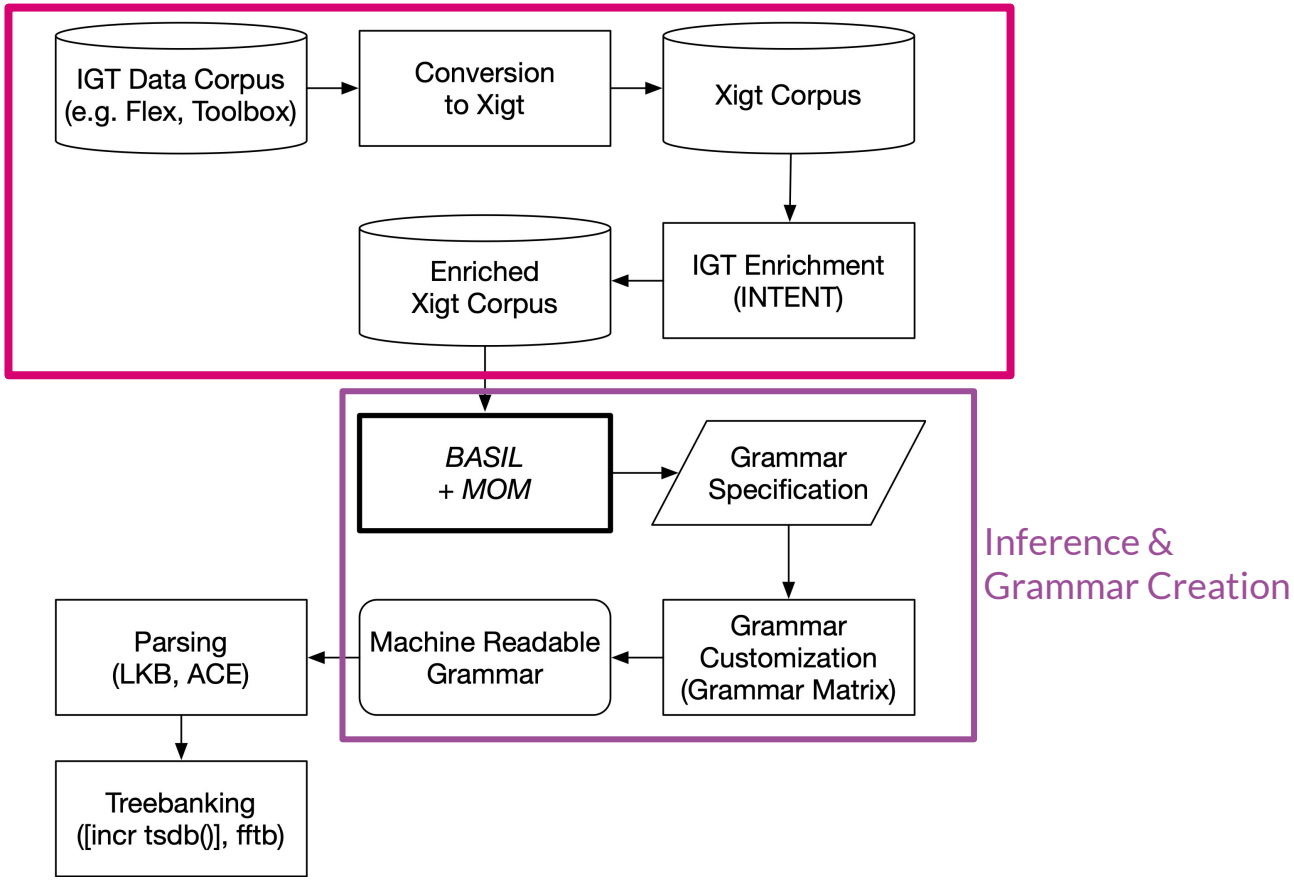
The AGGREGATION Project (Howell 2020, p17)

Data Formatting



The AGGREGATION Project (Howell 2020, p17)

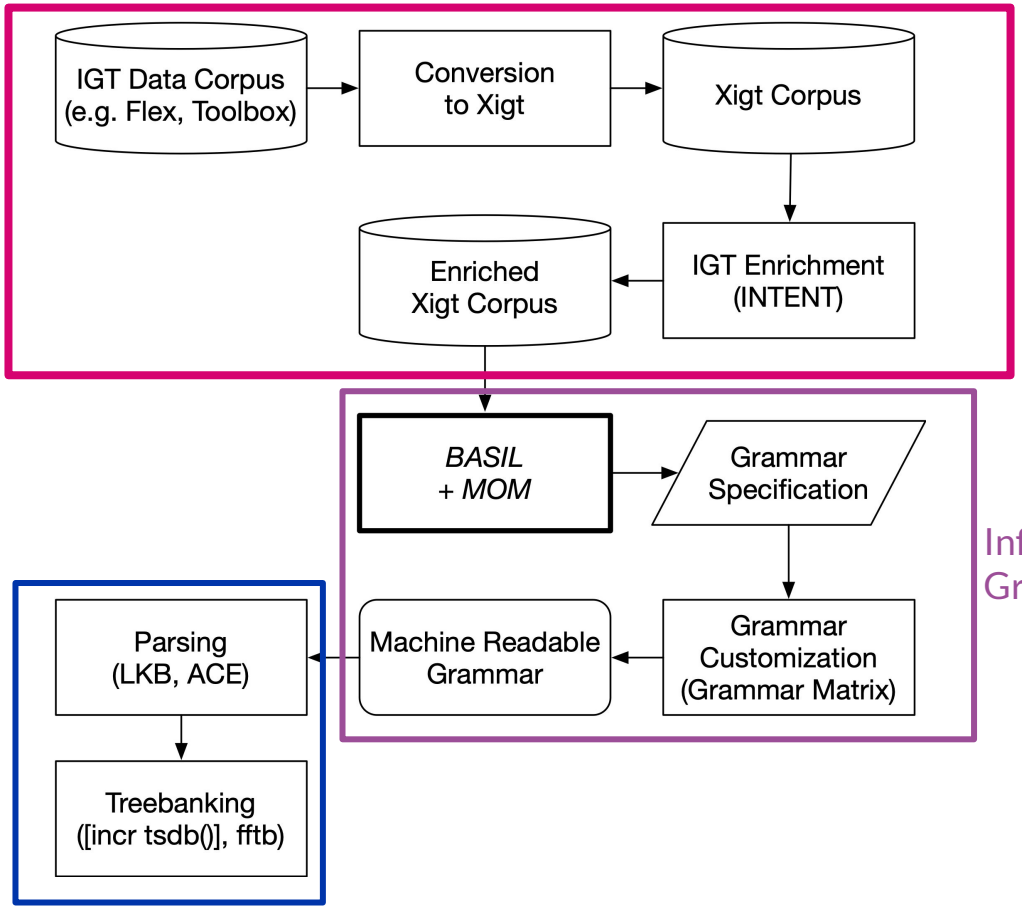
Data Formatting



Inference &
Grammar Creation

The AGGREGATION Project (Howell 2020, p17)

Data Formatting



Inference &
Grammar Creation

Grammar Usage

The AGGREGATION Project (Howell 2020, p17)

Evaluation Metrics

Evaluation Metrics

1. **Ambiguity** (avg. # readings per sentence)
2. **Coverage** (# sentences parsed)
3. **Persistence of treebanked readings**
    a. How many of the sentences that were treebanked and had a valid reading are still parsing with a reading isomorphic to the treebanked reading?
4. **“Good loss” of coverage**
    a. How many of the sentences that were treebanked and had *no* valid reading are no longer parsing?

Data

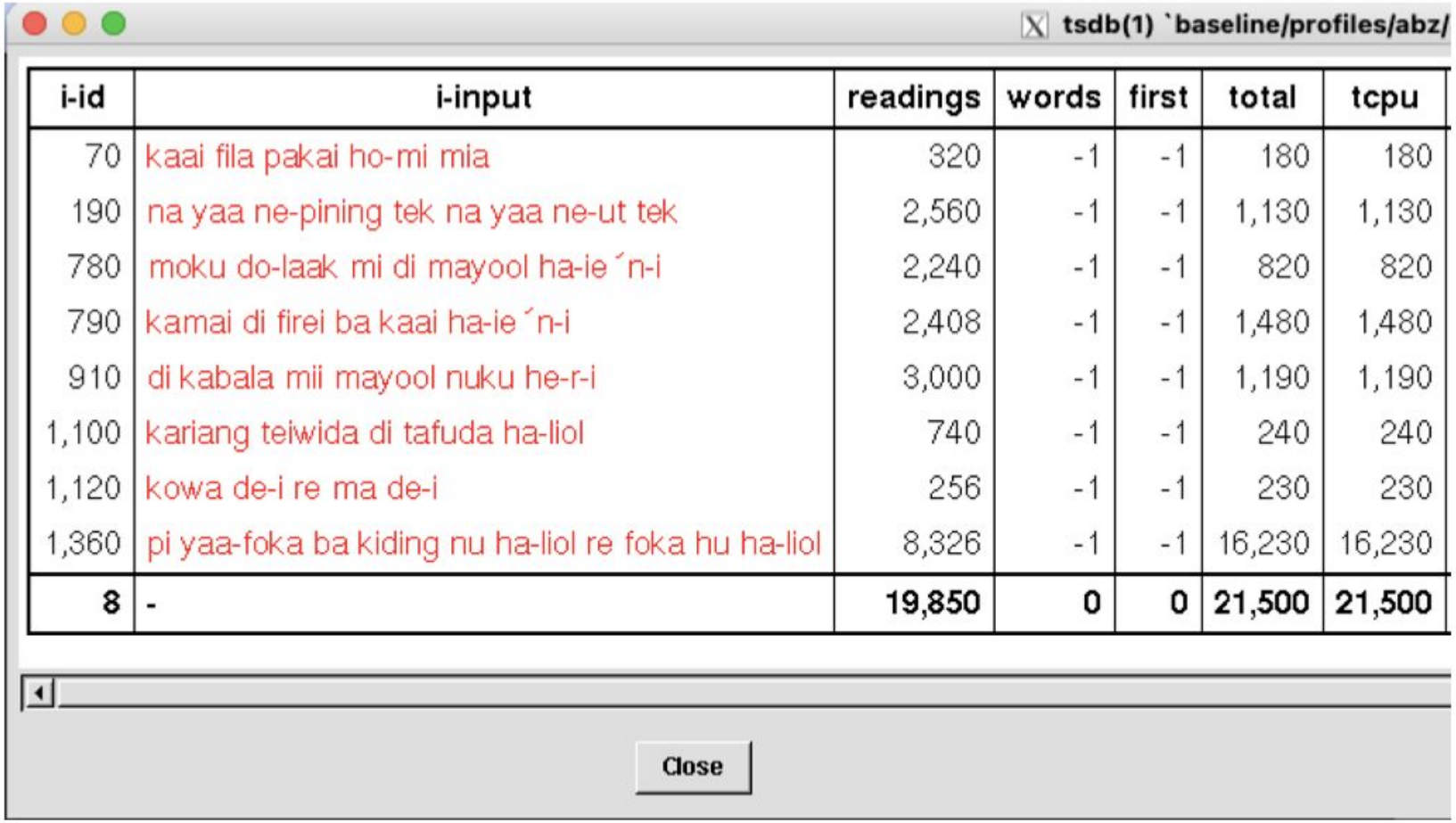
DEVELOPMENT LANGUAGES

1. Abui [abz]
2. South Efate [erk]
3. Hiaki [yaq]

TEST LANGUAGES

1. Tsova-Tush [bbl]
2. Meitei [mni]
3. Wakhi [wbl]

Finding Ambiguity Culprits



i-id	i-input	readings	words	first	total	tcpu
70	kaai fila pakai ho-mi mia	320	-1	-1	180	180
190	na yaa ne-pining tek na yaa ne-ut tek	2,560	-1	-1	1,130	1,130
780	moku do-laak mi di mayool ha-ie´n-i	2,240	-1	-1	820	820
790	kamai di firei ba kaai ha-ie´n-i	2,408	-1	-1	1,480	1,480
910	di kabala mii mayool nuku he-r-i	3,000	-1	-1	1,190	1,190
1,100	kariang teiwida di tafuda ha-liol	740	-1	-1	240	240
1,120	kowa de-i re ma de-i	256	-1	-1	230	230
1,360	pi yaa-foka ba kiding nu ha-liol re foka hu ha-liol	8,326	-1	-1	16,230	16,230
8	-	**19,850**	**0**	**0**	**21,500**	**21,500**

(Kratochvíl 2019)

“di kabala mii mayool nuku he-r-i” [abz] **… 3000 readings**

?	?	VP1-TOP-COORD	di kabala mii || mayool nuku he-r-i
?	?	DECL-HEAD-OPT-SUBJ	di kabala mii mayool nuku he-r-i
?	?	SUBJ-HEAD	di kabala || mii mayool nuku he-r-i
?	?	VP1-BOTTOM-COORD	kabala mii mayool nuku he-r-i
?	?	VP1-TOP-COORD	kabala mii || mayool nuku he-r-i
?	?	VP1-TOP-COORD	mii || mayool nuku he-r-i
?	?	VP1-TOP-COORD	di || kabala mii
?	?	COMP-HEAD	di kabala || mii
?	?	NP4-TOP-COORD	di || kabala
?	?	VP1-BOTTOM-COORD	kabala mii
?	?	COMP-HEAD	kabala || mii
?	?	BASIC-HEAD-OPT-COMP	di
?	?	BARE-NP	di
?	?	verb50-verb-lex	di
?	?	noun175-noun-lex	di
?	?	NOUN-PC14-SYNTH-LOC-LEX	kabala
?	?	NOUN-PC14-SYNTH-IN-LEX	kabala
?	?	NOUN-PC14-SYNTH-ON-LEX	kabala
?	?	NP4-BOTTOM-COORD	kabala
?	?	noun107-noun-lex	kabala
?	?	noun85-noun-lex	kabala
?	?	BASIC-HEAD-OPT-COMP	mii
?	?	verb132-verb-lex	mii
?	?	verb50-verb-lex	mii
?	?	verb17-verb-lex	mii
?	?	verb9-verb-lex	mii
?	?	NOUN-PC14-SYNTH-ON-LEX	mayool
?	?	NOUN-PC14-SYNTH-IN-LEX	mayool
?	?	NOUN-PC14-SYNTH-LOC-LEX	mayool
?	?	noun85-noun-lex	mayool
?	?	noun107-noun-lex	mayool
?	?	VERB-PC6_LRT1-PREFIX3	he-r-i
?	?	VERB-PC9_LRT2-PREFIX	he-r-i
?	?	VERB-PC18_LRT1-PREFIX	he-r-i
?	?	verb17-verb-lex	he-r-i
?	?	verb32-verb-lex	he-r-i
?	?	verb50-verb-lex	he-r-i

?	?	VERB-PC11_LRT1-PREFIX	ha-took
?	?	NOUN-PC14-SYNTH-LOC-LEX	ha-took
?	?	NOUN-PC14-SYNTH-IN-LEX	ha-took
?	?	NOUN-PC12_LRT1-PREFIX1	ha-took
?	?	NOUN-PC14-SYNTH-ON-LEX	ha-took
?	?	BARE-NP	ha-took
?	?	verb3-verb-lex	ha-took
?	?	noun85-noun-lex	ha-took
?	?	noun49-noun-lex	ha-took

From the Abui corpus [abz]

Ambiguity Culprits

1. SPURIOUS CASE RULES (the -SYNTH- rules)
2. POS mis-tagging

Reducing Ambiguity

Spurious Adpositional Case Rules

- A result of the case library (Drellishak, 2009)
  - Nouns in languages that mark case morphologically via an affix have a mandatory lexical rule applied to them, marking them for the appropriate CASE value
  - In a language with adpositional case marking, there are case-marking adpositions in the lexicon
  - But for a mixed-marking language… 🤔

Mixed-Marking (pretend) Example

- Say in a theoretical language with mixed-marking, the nominative is marked with an affix, but an adposition is used to mark the accusative
- For the nominative, a mandatory lexical rule must be applied to the noun
- For the accusative, the lexical entry of the accusative case-marking adposition must take a *non-case-marked NP* as its complement
    - But due to the mandatory nominative lexical rule, no noun lexical entry can serve as the head of a bare non-case-marked NP

Solution? Synthetic rules.

- The solution to this is to “synthesize an additional, non-spelling changing lexical rule that applies to nouns and marks the appropriate value of case,” (Drellishak, 2009)
- These rules sit in the same position class alongside the mandatory case-marking lexical rules, and mark a lexical entry as having the CASE value that a case-marking adposition would require

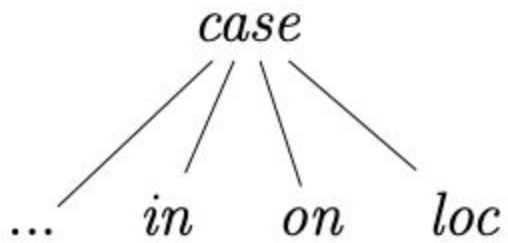
Mixed-Marking (pretend) Example

- Say in a theoretical language with mixed-marking, the nominative is marked with an affix, but an adposition is used to mark the accusative
- For the nominative, a mandatory lexical rule must be applied to the noun
- For the accusative, the lexical entry of the accusative case-marking adposition must take a *non-case-marked NP* as its complement
    - But due to the mandatory nominative lexical rule, no noun lexical entry can serve as the head of a bare non-case-marked NP
- For mixed-marking, we need…
    - non-inflecting rule that marks the noun as having CASE accusative, the result being that nouns not marked as nominative can unify with the accusative adposition

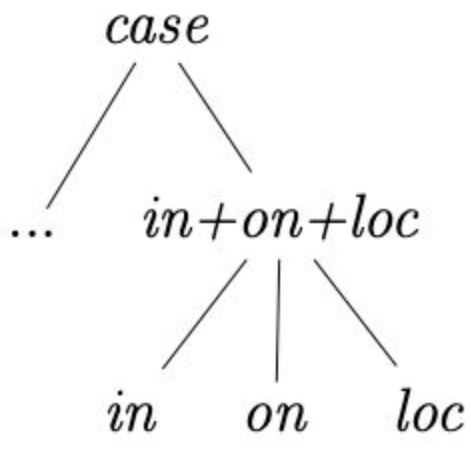
?	?	VERB-PC11_LRT1-PREFIX	ha-took
?	?	NOUN-PC14-SYNTH-LOC-LEX	ha-took
?	?	NOUN-PC14-SYNTH-IN-LEX	ha-took
?	?	NOUN-PC12_LRT1-PREFIX1	ha-took
?	?	NOUN-PC14-SYNTH-ON-LEX	ha-took
?	?	BARE-NP	ha-took
?	?	verb3-verb-lex	ha-took
?	?	noun85-noun-lex	ha-took
?	?	noun49-noun-lex	ha-took

From the Abui corpus [abz]

(4)



(5)



POS Mis-tagging

- Words being labeled as both verbs and nouns

nee	aa=	tu'ure	hiakinooka-po	nee-u	nooka-wa-m-ta
1.SG.NOM	3.SG.ACC	like	speak.in.Hiaki-LOC	1.SG.ACC-DIR	speak-PASS-PL-ACC

'I like speaking Hiaki' [yaq] (Harley, 2019)

nee	aa=	tu'ure	hiakinooka-po	nee-u	nooka-wa-m-ta
1.SG.NOM	3.SG.ACC	like	speak.in.Hiaki-LOC	1.SG.ACC-DIR	speak-PASS-PL-ACC

'I like speaking Hiaki' [yaq] (Harley, 2019)

```
<tier id="x" type="syntax" alignment="w">
    <item id="x1" alignment="w1">WkPro</item>
    <item id="x2" alignment="w2">Pro</item>
    <item id="x3" alignment="w3">V</item>
    <item id="x4" alignment="w4">PP</item>
    <item id="x5" alignment="w5">PP</item>
    <item id="x6" alignment="w6">N</item>
  </tier>
```

Solution? Mixed-morpheme check.

- The solution I implemented was a mixed-morpheme check, which looks at:
  - (1) the POS tag of the full token
  - (2) the glosses of each affix attached to the root
- If there is a mismatch such that there appears both “nouny” and “verby” information, MOM will not make a lexical entry based on this token
- The motivation behind this being if there is a mixture of “nouny” and “verby” affixes, there’s less reason to be confident about what POS the root is

Results

Abui Results

ABUI [abz]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	919.07	344.67	141 .77
Coverage by count	657/1568	657/1568	529/1568
Coverage %	41.90%	41.90%	33.73%
TB sentences with valid analysis	12	12	10
TB sentences with only invalid analyses	52	52	41
'Good loss'	–	0	11

South Efate Results

SOUTH EFATE [erk]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	9284.82	9309.95	9085.89
Coverage by count	139/1875	139/1875	146/1875
Coverage %	7.41%	7.41%	7.78%
TB sentences with valid analysis	2	2	2
TB sentences with only invalid analyses	114	114	119
'Good loss'	–	0	4

Hiaki Results

HIAKI [yaq]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	130.75	130.75	43.42
Coverage by count	231/2 235	231/2235	203/2235
Coverage %	10.34%	10.34%	9.08%
TB sentences with valid analysis	4	4	3
TB sentences with only invalid analyses	18	18	15
'Good loss'	–	0	3

Tsova-Tush Results

TSOVA-TUSH [bbl]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	509.37	512.76	416.64
Coverage by count	355/1601	356/1568	255/1601
Coverage %	22.17%	22.24%	15.93%
TB sentences with valid analysis	5	5	3
TB sentences with only invalid analyses	38	38	28
'Good loss'	–	0	10

Meitei Results

MEITEI [mni]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	1261.78	1261.78	607.06
Coverage by count	50/955	50/955	33/955
Coverage %	5.24%	5.24%	3.45%
TB sentences with valid analysis	10	10	6
TB sentences with only invalid analyses	40	40	22
'Good loss'	–	0	18

Wakhi Results

WAKHI [wbl]	Baseline	Synth Consolidation	POS Tagging
Average readings per sentence	8.26	8.26	7.56
Coverage by count	165/683	165/683	89/683
Coverage %	24.16%	24.16%	13.03%
TB sentences with valid analysis	43	43	22
TB sentences with only invalid analyses	93	93	44
'Good loss'	–	0	54

General Takeaways

- Abui was the only language in my set to benefit from the spurious case rule fix, but it helped dramatically with no loss in coverage
- South Efate didn't see improvement from either change, sadly :(
- The other languages all benefited from the POS tagging improvement

Discussion

Discussion Topics

- I’m interested in eliciting connections to linguistic questions. What does this make you think about as linguists?
- What would you like to see as next steps in theory, and what do you want to know about how this was built in light of your interests in grammar building/morphology?

References

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The grammar matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In John Carroll, Nelleke Oostdijk, and Richard Sutcliffe, editors, *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei, Taiwan, 2002.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language & Computation*, 8(1):23–72. ISSN1570-7075. URL `http://dx.doi.org/10.1007/s11168-010-9070-1.` 10.1007/s11168-010-9070-1

Scott Drellishak. 2009. Widespread but not universal: Improving the typological coverage of the Grammar Matrix. PhD thesis, University of Washington.

Heidi Harley. 2019. Hiaki text corpus. University of Arizona. Unpublished FieldWorks (FLEx) project. (Accessed August 2019).

Kristen Howell. 2020. *Inferring grammars from interlinear glossed text: extracting typological and lexical properties for the automatic generation of HPSG grammars*. University of Washington, Seattle.

František Kratochvíl. 2019. Abui Corpus. Electronic Database: Unpublished toolbox project (accessed March 2019). Nanyang Technological University, Singapore.

Nick Thieberger. 2006. Dictionary and texts in South Efate. *Digital collection managed by PARADISEC [Open Access]*. (Accessed March 2019).

David Wax. 2014. *Automated grammar engineering for verbal morphology*. University of Washington, Seattle.