Handling cross-cutting properties in automatic inference of lexical classes: A case study of Chintang

Olga Zamaraeva, Kristen Howell, Emily M. Bender

University of Washington

February 26, 2019

Grammar Inference

- Overarching Goal: to bring the benefits of *Grammar Engineering* to descriptive and documentary linguists
- How: automate the process of defining implemented grammars
- Short-term Goal: to create a feedback loop that brings some benefits before grammars are broad coverage

Precision (implemented) Grammars

- Bidirectional: parse and generate
- Produce syntactic and semantic representations
- Prioritize precision: the proportion of *correct* parses
- Uses Include:
  - Linguistic hypothesis testing (Müller, 1999; Bender, 2008; Fokkens, 2014)
  - Comparing analyses over corpora (Bender, 2010)
  - Creating treebanked corpora (Bender et al., 2012)

Precision (implemented) Grammars

rules.tdl

```
head-comp := head-comp-phrase.
```

irules.tdl

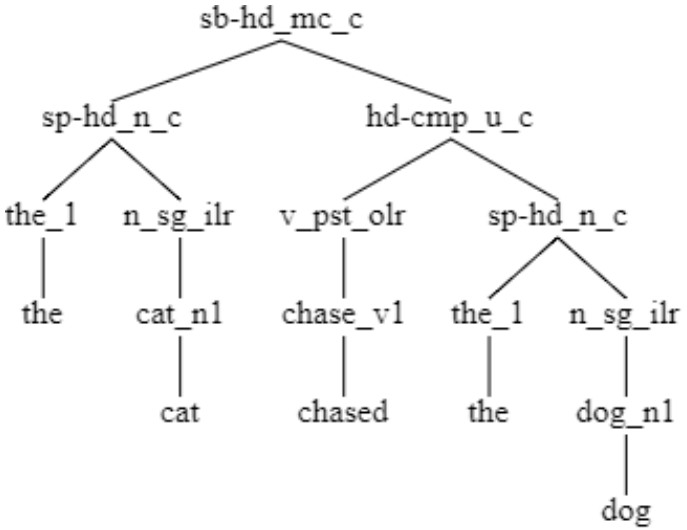
```
plural-suffix :=
%suffix (* s)
```

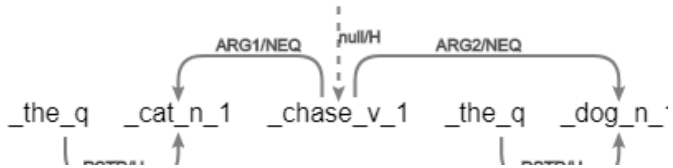
matrix.tdl

```
norm-ltop-lex-item := lex-item &
 [ SYNSEM [ LOCAL.CONT [
```

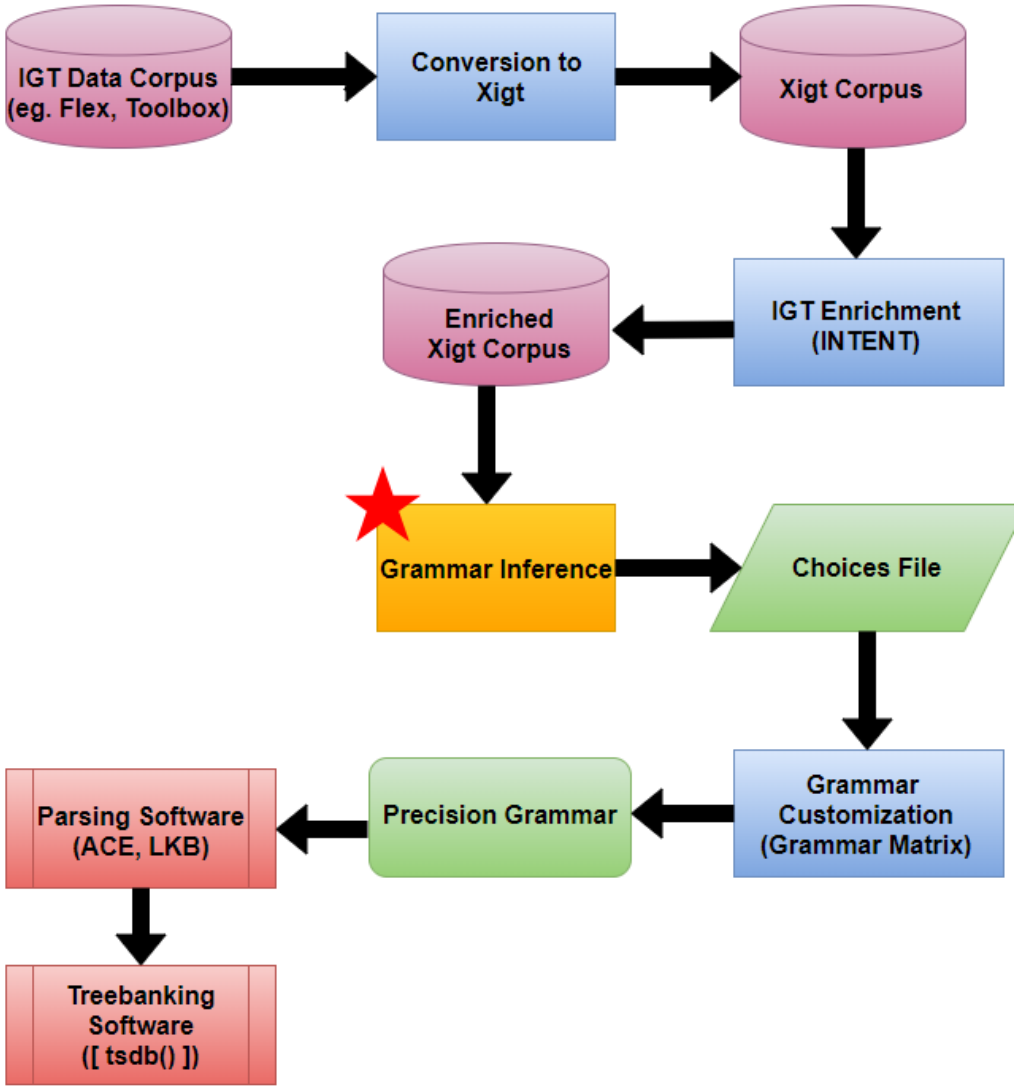
eng.tdl

```
head-spec-phrase := basic-head-
 spec-phrase & head-final &
 [ NON-HEAD-DTR.SYNSEM.OPT - ].

common-noun-lex := obl-spr-noun-
 lex &
 [ SYNSEM.LOCAL.CONT.HOOK.INDEX [
        PNG.PER 3rd ],
    INFLECTED.NUM-FLAG - ].

1sg-pronoun-noun-lex := no-spr-
 noun-lex &
 [ SYNSEM.LOCAL.CONT.HOOK.INDEX [
        PNG [ PER 1st,
              NUM sg ] ].
```





AGGREGATION Project



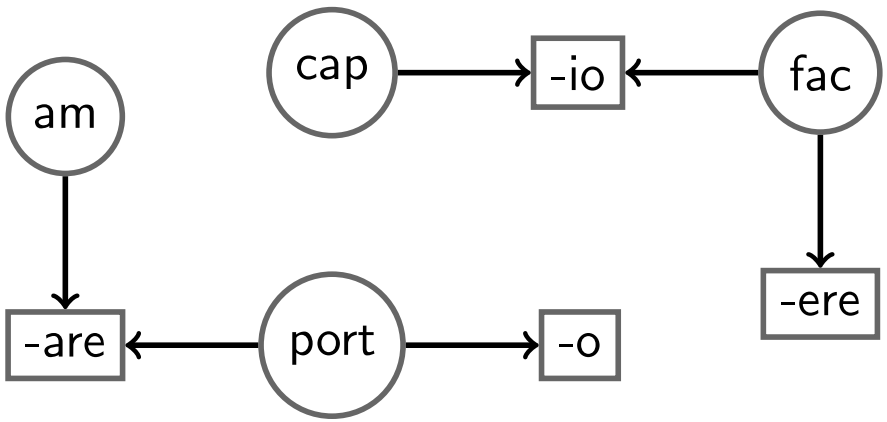
The Task at Hand: Handling Cross Cutting Categories

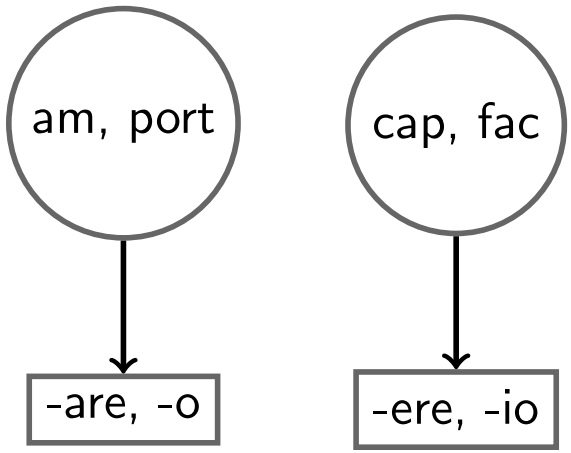
- We automatically infer a precision grammar from a corpus of Interlinnear Glossed Text (IGT)
- Previous work (Bender et al., 2014) inferred separate grammars that supported:
  - morphological inflection
  - case and argument requirements
- We integrate the two to infer **lexical classes** that inflect and have transitivity/case-frame requirements

Methodology: Morphotactic Inference (MOM) (Wax, 2014; Zamaraeva et al., 2017)

- Target the morpheme-segmented line of IGT
- Build a graph of stems and affixes
- Collapse items with overlapping edges

Latin Verbs: *am-are* ('to love'), *port-are* ('to carry'), *port-o* ('I carry'), *cap-io* ('I take'), *fac-io* ('I do'), and *fac-ere* ('to do')





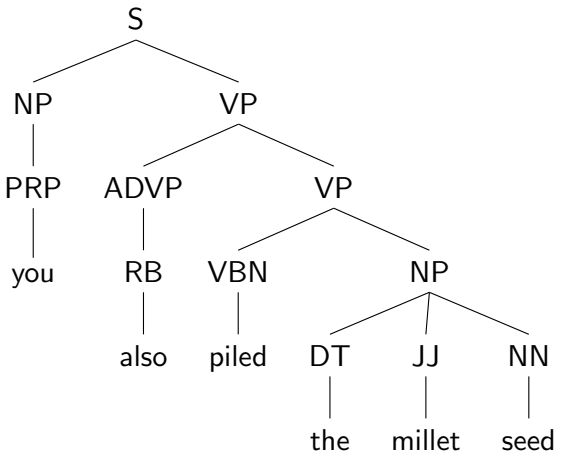
Methodology: Case and Valence Inference

- Infer case system based on the relative frequency of grams in the corpus (eg. ergative-absolutive) (Bender et al., 2014; Howell et al., 2017)
- Infer transitivity for individual verbs in the corpus

Sambok biuyaŋ abhuŋsanduthoe.

sambok	biu-yaŋ	a-bhuŋs-a-dhend-u-tha-u-e
millet	seed-ADD	2S/A-pile.up-PST-TEL-3P-TEL-3P-IND.PST
NOUN	NOUN	VERB

'You also piled up the millet seeds.' [ctn] (Bickel et al. 2013)



- Assign case frame according to transitivity
  - eg. intransitive verbs: subj: abs
    transitive verbs: subj: erg, obj: abs

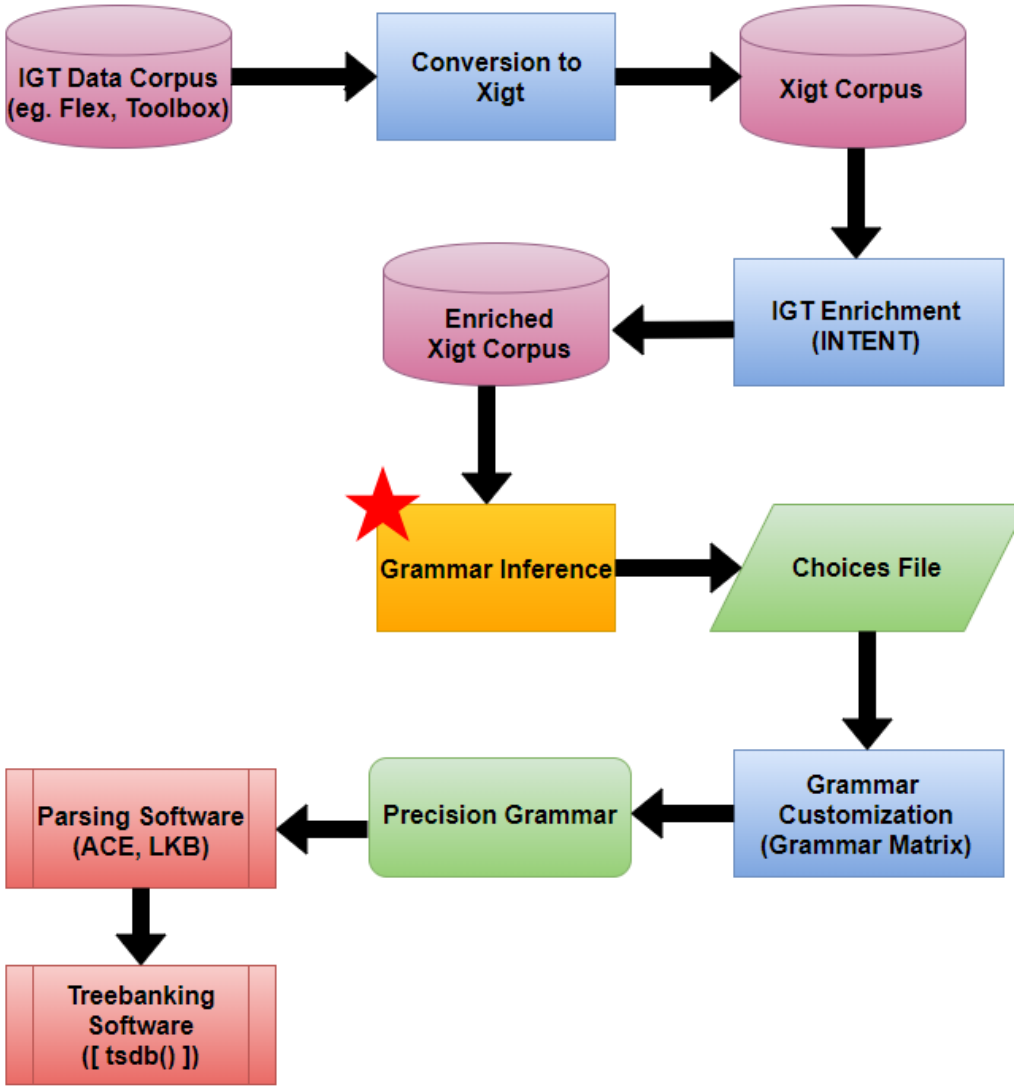
Methodology: Creating Integrated Lexical Classes

- Include case frame information in the morphotactic inference system's input
  - In order to maximize examples of morphological patterns, verbs for which case frame inference failed are inlcluded in a 'dummy' category
- The morphotactic inference system infers inflectional classes and checks to see if a verb's case frame is compatible with a lexical class before adding it

Case Study

We evaluate the benefits of this integration by replicating the case study used in the previous stage: Bender et al. 2014

AGGREGATION Pipeline



Case Study: Chintang

- Chintang (ISO639-3: ctn)
- Sino -Tibetan language of Nepal
- ∼5,000 speakers (Bickel et al., 2010; Schikowski et al., 2015)
- Relatively free word order, although V-final orders are most common (Schikowski et al., 2015)
- Ergative-absolutive case system, with exceptions (Stoll and Bickel, 2012; Schikowski et al., 2015)

Case Study: Dataset

- Chintang Language Resource Program (CLRP; https://www.uzh.ch/clrp/)
- 10,862 instances of Interlinnear Glossed Text (IGT)
- Split into train (8863 IGT), dev (1069 IGT) and test (930 IGT) sets
- Annotation is very detailed

Unisaŋa khatte mo kosiʔ moba.

u-nisa-ŋa	khatt-e	mo	kosi-iʔ	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] (Bieckel et al., 2013)

Grammars

Comparison choices files come from Bender et al. (2014)
In all grammars, all arguments can drop

Grammar	Lexicon	Morphology	Word Order	Case System
BASELINE	full form	none	default	default
ORACLE	Toolbox	hand-defined	v-final	erg-abs
FF-AUTO-GRAM	full form	none	v-final	erg-abs
MOM-DEFAULT-NONE	inferred	inferred	default	default
INTEGRATED	inferred	inferred	default	erg-abs

default word order: free
default case: none

Lexical Information in each Grammar

Grammar	# verb entries	# noun entries	# verb affixes	# noun affixes
ORACLE	899	4750	233	36
BASELINE	3005	1719	0	0
FF-AUTO-GRAM	739	1724	0	0
MOM-DEFAULT-NONE	1177	1719	262	0
INTEGRATED	911	1755	220	76

Results

Grammar	lexical coverage		parsed		correct		readings
ORACLE	116	(12.5%)	20	(2.2%)	10	(1.1%)	1.35
BASELINE	38	(0.4%)	15	(1.6%)	8	(0.9%)	27.67
FF-AUTO-GRAM	18	(1.9%)	4	(0.4%)	2	(0.2%)	5.00
MOM-DEFAULT-NONE	39	(4.2%)	16	(1.7%)	3	(0.3%)	10.81
INTEGRATED	105	(11.3%)	32	(3.4%)	15	(1.6%)	91.56

Error Analysis: INTEGRATED system

- The vast majority of sentences that failed did not succeed at lexical analysis
  - Of those the majority failed because an affix or stem was not in the training data
- The remaining 73 sentences did not succeed at syntactic analysis
  - 6 contained an NP who's case was not compatible with the verb's case frame (such as a locative NP modifier)
  - 23 did not contain a word that the grammar analysed as a verb
  - 44 contained multiple verbs

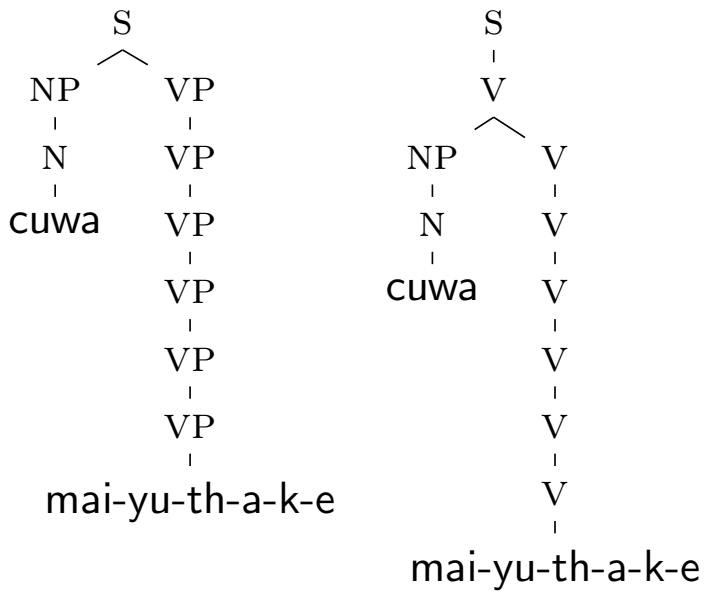
Error Analysis: INTEGRATED vs. ORACLE

	Morpheme not in Grammar	Morphemes Can't Combine	Words Can't Combine
ORACLE	46	3	6
INTEGRATED	58	0	1

Ambiguity

INTEGRATED has high ambiguity due to multiple entries for verbs and affixes

(1) cuwa mai-yu-th-a-k-e
water NEG-be.there-NEG-PST-IPFV-IND.PST
'Was there water?"



Future Work

- Add non-inflecting lexical rules to MOM
- Infer alternate case frames for verbs
- Extend this methodology to other morpho-syntactic features
- Extend syntactic inference to account for more phenomena

Conclusion

- Integrating morphological and syntactic inference improves coverage
  - With better coverage, inferred grammars can help linguists discover patterns of the combinatorics in their data
- This methodology can be extended to other morpho-syntactic features
- We are scaling up quickly! If you have an IGT corpus and want an inferred grammar, come talk to us.

Conclusion

- Integrating morphological and syntactic inference improves coverage
  - With better coverage, inferred grammars can help linguists discover patterns of the combinatorics in their data
- This methodology can be extended to other morpho-syntactic features
- We are scaling up quickly! If you have an IGT corpus and want an inferred grammar, come talk to us.

Thank you!

Acknowledgements

- This material is based upon work supported by the National Science Foundation under Grant No. BCS-1561833. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
- We would like to thank Alex Burrell for providing code from a previous project that we built on in order to infer transitivity.
- We thank previous AGGREGATION research assistants for converting the corpus into Xigt and enriching it with INTENT.

References I


Bender, Emily M (2008). "Grammar engineering for linguistic hypothesis testing". In: *Proceedings of the Texas Linguistics Society X conference: Computational linguistics for less-studied languages*. Citeseer, pp. 16–36.


— (2010). "Reweaving a grammar for Wambaya". In: *Linguistic Issues in Language Technology* 3.1.


Bender, Emily M, Joshua Crowgey, et al. (2014). "Learning Grammar Specifications from IGT: A Case Study of Chintang". In: *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Baltimore, Maryland, USA: Association for Computational Linguistics, pp. 43–53. URL: http://www.aclweb.org/anthology/W14-2206.


Bender, Emily M, Scott Drellishak, et al. (2010). "Grammar Customization". In: *Research on Language & Computation* 8, pp. 1–50. ISSN: 1570-7075.


Bender, Emily M, 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: *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*. Ed. by John Carroll, Nelleke Oostdijk, and Richard Sutcliffe. Taipei, Taiwan, pp. 8–14.

References II


Bender, Emily M. et al. (2012). "From Database to Treebank: Enhancing Hypertext Grammars with Grammar Engineering and Treebank Search". In: *Electronic Grammaticography*. Ed. by Sebastian Nordhoff and Karl-Ludwig G. Poggeman. Honolulu: University of Hawaii Press, pp. 179–206.


Bickel, Balthasar, Martin Gaenszle, et al. (2013). *Tale of a Poor Guy*. Accessed: 22 October 2018. URL: https://corpus1.mpi.nl/qfs1/media-archive/dobes_data/ChintangPuma/Chintang/Narratives/Annotations/phengniba_tale.tbt.


Bickel, Balthasar, Manoj Rai, et al. (2010). "The syntax of three-argument verbs in Chintang and Belhare (Southeastern Kiranti)". In: *Studies in ditransitive constructions: a comparative handbook*, pp. 382–408.


Fokkens, Antske Sibelle (2014). "Enhancing Empirical Research for Linguistically Motivated Precision Grammars". PhD thesis. Department of Computational Linguistics, Universität des Saarlandes.


Georgi, Ryan (2016). "From Aari to Zulu: Massively Multilingual Creation of Language Tools Using Interlinear Glossed Text". PhD thesis. University of Washington.


Goodman, Michael Wayne et al. (2015). "Xigt: Extensible Interlinear Gloss Text for Natural Language Processing". In: *Language Resources and Evaluation* 49 (2), pp. 455–485.

References III


Müller, Stefan (1999). *Deutsche Syntax deklarativ: Head-Driven Phrase Structure Grammar ffffdfffdr das Deutsche*. Linguistische Arbeiten 394. Tübingen: Max Niemeyer Verlag.


Schikowski, Robert, NP Paudyal, and Balthasar Bickel (2015). "Flexible valency in Chintang". In: *Valency Classes: a Comparative Handbook*.


Stoll, Sabine and Balthasar Bickel (2012). "How to measure frequency? Different ways of counting ergatives in Chintang (Tibeto-Burman, Nepal) and their implications". In: *Potentials of Language Documentation: Methods, Analyses, and Utilization*. Ed. by Frank Seifart et al. University of Hawai'i Press, pp. 83–89.


Wax, David (2014). "Automated Grammar Engineering for Verbal Morphology". MA thesis. University of Washington.


Zamaraeva, Olga et al. (2017). "Computational Support for Finding Word Classes: A Case Study of Abui". In: *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 130–140.