From grammar induction to learning and processing

Sebastian Bank

Department of Linguistics
University of Leipzig

10th International Morphological Processing Conference
June 22–24, 2017, SISSA, Trieste

From language acquisition to grammar induction

Linguistic input $\xrightarrow[\text{acquisition}]{?}$ Lexicon, Grammar

Question: How do speakers actually do it? (*empiry*)

	SG	PL
1	spiele	spielen
2	spielst	spielt
3	spielt	spielen

$\xrightarrow[\text{induction}]{?}$

i. $\langle$`spiel`, PLAY$\rangle$
ii. $\langle$`t`, [+2]$\rangle$
iii. $\langle$`en`, [+pl]$\rangle$
o_1: i. precedes { ii., iii.}
b_1: ii. blocks iii.

Question: How can one even do it? (*theory*)

Three problems for inflectional analysis (Trommer & Bank fc.)

Present tense word forms of the German intransitive verb *spielen* 'to play'

	SG	PL
1	spiel-e	spiel-en
2	spiel-s-t	spiel-t
3	spiel-t	spiel-en

A. segmentation (stemming, subsegmentation, when to stop)
B. form-meaning pairing
   - what is the meaning of *-en*? {1PL, 3PL} sharing [+pl] (*form→meaning*)
   - what is the form for [+2]? {*spielst*, *spielt*} sharing *-t* (*meaning→form*)
C. imperfect distributions
   - given ⟨*-en*, [+pl]⟩, the unaccounted absence of *-en* in 2PL (*overinsertion*)
   - given ⟨*-t*, [+2]⟩, the unaccounted presence of *-t* in 3SG (*underinsertion*)

A, B, and C are interdependent: consider all of them together.

More problems for inflectional analysis

	SG	PL
2	spiel-st	spiel-t

PRESENT

	SG	PL
2	spiel-te-st	spiel-te-t

PAST

D. possible forms
   - do we need to consider *-s*? (*search-space*)
E. linear order
   - does the linearization need to be templatic? (*ordered partition*, *transitivity*)
F. blocking
   - how can one marker suppress another one? (*in-* vs. *extrinsic*, within *slot*)

STEM		SUFFIX I		SUFFIX II	
spiel	PLAY	-te	[+past]	-st	[+2 −pl]
				-t	[+2]

Grammar induction as iterative local optimization

Search space: combinatorial explosion from segmentations × groupings

	SG	PL
1	spiele	spielen
2	spielst	spielt
3	spielt	spielen

	MINIM		MAXIM	
candidate	ovr	und	cov	len
⟨spiel, [ ]⟩	0	0	6	5
⟨en, [−2 +pl]⟩	0	0	2	2
⟨n, [−2 +pl]⟩	0	0	2	1
⟨st, [+2 −pl]⟩	0	0	1	2
⟨t, [+2]⟩	0	1	2	1
⟨e, [+1 −pl]⟩	0	2	1	1
⟨en, [+pl]⟩	1	0	2	2

better

Incremental learning algorithm using greedy search (~*Harmonic serialism* in OT)

1. combine all (sub)strings found in the paradigm with all meanings
2. rank the form-meaning pairs by their quality, pick the best one
3. add the winner to the lexicon, erase its occurrences from the paradigm
4. as long as the paradigm is not yet empty go back to step 1

Comparing the accuracy of form-meaning pairs

CELL HAS MEANING	CELL HAS FORM: **true**	CELL HAS FORM: **false**
true	true positives	false positives
false	false negatives	true negatives

false positives = **over**insertion (= lower precision: $\frac{tp}{tp+fp}$)

false negatives = **under**insertion (= lower recall $\frac{tp}{tp+fn}$)

-n [+pl] HAS MEANING	HAS FORM: **true**	HAS FORM: **false**
true	2	1
false	0	3

-t [+2] HAS MEANING	HAS FORM: **true**	HAS FORM: **false**
true	2	0
false	1	3

*OverIns ≫ *UnderIns ≫ Coverage! ≫ Length!

	SG	PL
1	spiel-e_1	spiel-e_1-n
2	spiel-st	spiel-t_1
3	spiel-t_2	spiel-e_2-n

PRESENT

	SG	PL
1	spiel-te	spiel-te-n
2	spiel-te-st	spiel-te-t_1
3	spiel-te	spiel-te-n

PAST

Lexicon

a. ⟨spiel, [ ]⟩ b. ⟨te, [+past]⟩ c. ⟨n, [−2 +pl]⟩
d. ⟨st, [+2 −pl]⟩ e. ⟨e, [+1 −past]⟩ f. ⟨e, [+3 +pl −past]⟩
g. ⟨t, [+2 +pl]⟩ h. ⟨t, [+3 −pl −past]⟩

*OverIns ≫ Coverage! ≫ *UnderIns ≫ Length!

	SG	PL
1	spiel-e_1	spiel-e_1-n
2	spiel-s-t_1	spiel-t_1
3	spiel-t_2	spiel-e_2-n

PRESENT

	SG	PL
1	spiel-te	spiel-te-n
2	spiel-te-s-t_1	spiel-te-t_1
3	spiel-te	spiel-te-n

PAST

Lexicon

a. ⟨spiel, [ ]⟩ b. ⟨te, [+past]⟩ c. ⟨n, [−2 +pl]⟩
d. ⟨t, [+2]⟩ e. ⟨s, [+2 −pl]⟩ f. ⟨e, [+1 −past]⟩
g. ⟨e, [+3 +pl −past]⟩ h. ⟨t, [+3 −pl −past]⟩

*UnderIns ≫ *OverIns ≫ Coverage! ≫ Length!

	SG	PL
1	spiel-e	spiel-e-n
2	spiel-st	spiel-t_1
3	spiel-t_2	spiel-e-n

PRESENT

	SG	PL
1	spiel-te	spiel-te-n
2	spiel-te-st	spiel-te-t_1
3	spiel-te	spiel-te-n

PAST

Lexicon

a. $\langle$`spiel`, [ ]$\rangle$ b. $\langle$`te`, [+past]$\rangle$ c. $\langle$`n`, [−2 +pl]$\rangle$
d. $\langle$`st`, [+2 −pl]$\rangle$ e. $\langle$`e`, [−2 −past]$\rangle$ f. $\langle$`t`, [+2 +pl]$\rangle$
g. $\langle$`t`, [+3 −pl −past]$\rangle$

Blocking g. blocks e.

*UnderIns ≫ Coverage! ≫ *OverIns ≫ Length!

	SG	PL
1	spiel-e	spiel-e-n
2	spiel-st_1	spiel-t_2
3	spiel-t_4	spiel-e-n

PRESENT

	SG	PL
1	spiel-t_1-e	spiel-t_1-e-n
2	spiel-t_1-e-st_2	spiel-t_1-e-t_3
3	spiel-t_1-e	spiel-t_1-e-n

PAST

Lexicon

a. $\langle$`spiel`, [ ]$\rangle$ b. $\langle$`e`, [ ]$\rangle$ c. $\langle$`n`, [−2 +pl]$\rangle$
d. $\langle$`st`, [+2 −pl −past]$\rangle$ e. $\langle$`st`, [+2 −pl +past]$\rangle$ f. $\langle$`t`, [+past]$\rangle$
g. $\langle$`t`, [+2 +pl −past]$\rangle$ h. $\langle$`t`, [+2 +pl +past]$\rangle$ i. $\langle$`t`, [+3 −pl −past]$\rangle$

Blocking d., g., i. block b.

Automation of analysis

- try different optimization strategies by ranking or new constraints
- reproducible: apply to different data sets, compare

Predictions (*vagueness* vs. *ambiguity*)

- expected details of segmentation and syncretism vs. homophony
- for acquisition, processing, artificial grammar experiments
- e.g. depth of underspecification: Opitz et al. (2013)

Eagerness in generalization

- division between big generalizations and 'homophony residue'
- or a more balanced strategy?

Restricting considered forms using *standalone* occurrences

Swahili present, imperfective, and subjunctive verbal agreement prefixes (Seidel 1900)

Task: Split present/imperfective forms into agreement + tense

	SG	PL
1	nina-	tuna-
2	una-	mna-
3	ana-	wana-

PRESENT

	SG	PL
1	nili-	tuli-
2	uli-	mli-
3	ali-	wali-

IMPERFECT

	SG	PL
1	ni-	tu-
2	u-	m-
3	a-	wa-

SUBJUNCT

Observation: Easier once the zero-marked subjunctive is available

na- and *li-* are '*cran*-affixes': (*free* vs. *bound* for affixes)
they lack a 'free' occurrence until an adjacent 'free' form is learned

Hypothesis background: pervasiveness of zero-marking

Analysis can focus on 'free' forms (Class 1), or '*cran*-' forms (Class 2), but may never need to go through all possible 'bound' forms (Class 3)

Class 1 segmentation of German *spielen* 'to play', underlying strings

	SG	PL
1	ʃpiːl-ə	ʃpiːl-n̩
2	ʃpiːl-st	ʃpiːl-t
3	ʃpiːl-t	ʃpiːl-n̩

PRESENT

	SG	PL
1	ʃpiːl-tə	ʃpiːl-tø̸-n̩
2	ʃpiːl-tə-st	ʃpiːl-tə-t
3	ʃpiːl-tə	ʃpiːl-tø̸-n̩

PAST

-s is a '*cran*-affix' (requires class 2)

a helpful side-effect of zero-marking (or *functional motivation*?)

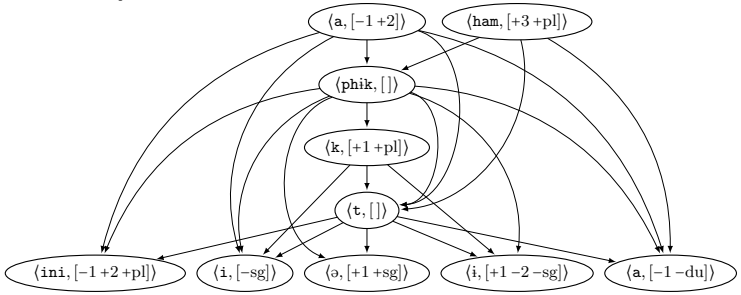
Predictions

- complexity hierarchy: Class 1 ⊂ Class 2 ⊂ Class 3
- subanalysis depth in acquisition, possible generalizations
- typological pilot study: all sample languages (verbal TAM and agreement marking on the same side of the stem) have *some* zero-marking for TAM or agreement

Templatic linearization: 'no cyclic walks'

In many descriptive grammars and theoretical frameworks, order is templatic

Lexicon, linear precedence relations (at least 5 layers ~ *slots*)



Still relatively easy to implement, but: (*transitive closure*)

- considers 'virtual' linear relations, i.e. between forms that never cooccur
- often still leaves multiple possible options for the template
- is it worth the conceptual and computational costs? (n^3)

Consistent linearization: 'don't mix prefix and suffix forms'

Nonpast tense forms of the Dumi intransitive verb *phɨkni* 'to get up' (van Driem 1993)

Consistency: Learners consider linear relations at the marker (*type*) level.

(apart from, or instead of the *token* level, modulo 'reordering')

(pairs may also be unordered, free variation)

	1EXCL	1INCL	2	3
SG	phɨk-t-ə	▪	a-phɨk-t-a	phɨk-t-a
DU	phɨk-t-i	phɨk-t-i	a-phɨk-t-i	phɨk-t-i
PL	phɨk-k-t-i	phɨk-k-t-i	a-phɨk-t-ini	ham-phɨk-t-a

Relatively easy to add to the incremental learning algorithm:

- keep track on which side of each form each learned marker is located
- don't combine forms if a learned marker would be on both sides
- allows for **cycles** (AB, BC, CA), no *transitive* reasoning

Observations

- linearization cycles in derivation (Muysken 1988; Ryan 2010)
- in inflection rather variable affix order (Bickel et al. 2007)
- often same result with basic vs. templatic consistency

Support for templates?

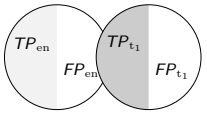
- may emerge for independent reasons (today's morphology = yesterday's syntax)
- the same surface linearization may be expressed by *n* templates
- artificial grammar experiments?

One marker's absence due to presence of another one

Blocking: Learners consider suppression relations at the marker (*type*) level.

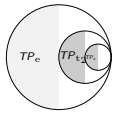
overlapping blocking

	SG	PL
1	spiel-e	spiel-**en**
2	spiel-s-**t**$_1$	spiel-**t**$_1$~~/e/~~**n**
3	spiel-t$_2$	spiel-**en**



Pāṇinian blocking

	SG	PL
1	spiel-**e**	spiel-**e**-n
2	spiel-st	spiel-t$_1$
3	spiel-**t**$_2$~~/e/~~	spiel-**e**-n



(under specific conditions = *extrinsic* vs. *intrinsic*, absense of Pāṇinian blocking = *extended exponence*)

Harder to add to the incremental algorithm than linearization:

- keep track of *possible* blockers for each unaccounted absence (*false positive*)
- ensure there is a (transitively) consistent way to select one for each
- simpler if only Pāṇinian blocking is considered (inherently *transitive*)

Some frameworks try to abandon overlapping blocking. (e.g. Stump 2001)

Pāṇinian blocking is often considered to be *automatic*.

(*Subset Principle*, *Elsewhere Principle*, *Blocking Principle*, ...)

Pāṇinian blocking can be functionally motivated (drop redundant markers).

Predictions

- details of segmentation and syncretism vs. homophony
- possibility or cost of retracting overgeneralizations (rule vs. exception)
- trade-off: subsegmentation vs. extended exponence

Merging linear precedence relations with suppression

Blocking-linearization correspondence: 'do not block a prefix by a suffix or vice versa'

Simplified: Only markers within the same slot (rule block, position class) can block each other. Slots are ordered *vertically*, e.g. Anderson (1992):

> *[...] the morphological rules of inflection to be organized in **blocks**, where the relation among rules within the same block is a **disjunctive** one. Rules within such a block are mutually exclusive, in that the first one applicable is the only one that applies. [...] The blocks themselves [...] are related by (conjunctive) sequence.*

Further pins down the template slot for each marker.
Markers that never cooccur do not need linear ordering. (and 'should not have it'?)

Hard to get right with the incremental learner:

- ensure there will be no *indirect* linear relation between blocker/blockee
- use either massive look-ahead or heuristics that might overfilter (how?)
- is an elegant grammatical structure worth the effort?

Feature coherence within a slot might further reduce slot options. (category-slot)

With Pāṇinian blocking only, some common feature per slot can emerge more-or-less automatically. (but not its uniqueness within the template)

Observed blocking-linearization correspondence and feature coherence may emerge for independent reasons.

Predictions

- possible kinds of blocking, segmentation, homophony
- (un)learnability of blocking across the stem?
- blocking domains

Goal: separate essential from accidental properties of inflectional grammar

Conclusion

- the automation of analysis allows to compare hypotheses from theoretical morphology in terms of concrete reproducible outcomes
- inducing a full grammar can be challenging, especially with many required interdependencies between different grammar domains
- more concrete questions for empirical investigation

Opitz, Andreas, Stefanie Regel, Gereon Müller & Angela D. Friederici. (2013). Neurophysiological evidence for morphological underspecification in German strong adjective inflection. *Language* 89(2). 231–264.

Ryan, K. M. (2010). Variable affix order: Grammar and learning. *Language*, 86: 758–791.

Seidel, A. (1900). *Swahili Konversationsgrammatik*. Julius Groos, Heidelberg.

Stump, G. T. (2001). *Inflectional Morphology*. Cambridge: Cambridge University Press.

Trommer, J. & Bank, S. (fc.). *Inflectional Learning as Local Optimization*. Morphology.

Anderson, S. R. (1992). *A-Morphous Morphology*. Cambridge: Cambridge University Press.

Bank, S. (in prep.). *Automatic induction for morphological grammar*. PhD Thesis, University of Leipzig.

Bank, S. & Trommer, J. (2015). Learning and the complexity of Ø-Marking. In: Matthew Baerman, Dunstan Brown & Greville G. Corbett (eds.), Understanding and measuring morphological complexity, Ofxord: Oxford University Press, 185–204.

Bickel, Balthasar, G. Banjade, M. Gaenszle, E. Lieven, N. P. Paudyal, & I. Purna Rai. (2007). Free prefix ordering in Chintang. *Language* 83 (1).

van Driem, G. (1993). *A Grammar of Dumi*. Walter de Gruyter, Berlin and New York.

Muysken, P. (1988). Affix order and interpretation: Quechua. Martin Everaert, Arnold Evers, Riny Huybregts, and Mieke Trommelen, editors, *Morphology and modularity*. Dordrecht: Foris, 259–279.