Paradigm Learning and Subanalysis Complexity

Sebastian Bank & Jochen Trommer

University of Leipzig

CLS 48
April 19-21, 2012

minimal subanalysis

present

	sg	pl
1	glaub-**e**	glaub-**en**
2	glaub-**st**	glaub-**t**
3	glaub-**t**	glaub-**en**

past

	sg	pl
1	glaub-**te**	glaub-**ten**
2	glaub-**test**	glaub-**tet**
3	glaub-**te**	glaub-**ten**

maximal subanalysis (Müller 2005: 10)

present

	sg	pl
1	glaub-**e**	glaub-**(e)n**
2	glaub-**s-t**	glaub-**t**
3	glaub-**t**	glaub-**(e)n**

past

	sg	pl
1	glaub-**te**	glaub-**te-n**
2	glaub-**te-s-t**	glaub-**te-t**
3	glaub-**te**	glaub-**te-n**

(German verbal agreement)

Observation

Subanalysis is facilitated/influenced by markers that occur as **the only prefix or suffix in a paradigm cell** ('free' occurrence).

present

	sg	pl
1	Σ-**e**	Σ-**en**
2	Σ-**est**	Σ-**et**
3	Σ-**et**	Σ-**en**

past

	sg	pl
1	Σ-t-**e**	Σ-t-**en**
2	Σ-t-**est**	Σ-t-**et**
3	Σ-t-**e**	Σ-t-**en**

Idea

Subanalyzing learners exploit this fact to **reduce search space** of possible segmentation points. They can do so to different degree.

- Learners with different reliance on the 'free' occurrence of affix substrings should yield dramatically different subanalyses
- Each type of search-space-reduction yields a class of data complexity the learner can (and can't) successfully subanalyze
- Zero-Exponence (e.g. 3rd person), Imperatives, etc. might help to keep the observed complexities low

Hypothesis

No language has a subanalysis complexity that demands the learner to explore the full range of possible segmentations.

Approach:

- compile subanalyzing learning algorithms and evaluate them on constructed and real data sets
- perform a complexity study on a typologically diverse language sample

Outline

Learning and zero-exponence

1. Subanalysis complexity
2. Learning algorithms
3. Typological study

The subanalysis problem

Assumption

Learners isolate the stem of an inflected word form (stemming) yielding **prefix and suffix strings** that may be internally simplex or decomposable

prs	sg	pl
1	e	en
2	est	et
3	et	en

pst	sg	pl
1	te	ten
2	test	tet
3	te	ten

- full range: test → test, t+est,te+st, tes+t, t+e+st, t+es+t, te+s+t, or t+e+s+t = $2^{len(string)-1}$ possible cell string segmentations
- without subanalysis: learn portmanteau meanings for the distributions of {e, est, et, en, te, test, ten, tet}

Subanalysis complexity depends on the occurrences of potential subaffixes as independent **affix string**.

Class 0

Affix strings are potential forms (no subaffixes)

Class 1

Every subaffix S of an affix string AS also occurs as an affix string

Class 2

For every binary subanalysis of an affix string AS into $S_1 + S_2$ either S_1 or S_2 occur as an affix string

Class 3

No restriction on the occurrences of subaffixes

Class 0 ⊆ Class 1 ⊂ Class 2 ⊂ Class 3

German: Class 0

prs	sg	pl
1	Ø	n
2	st	t
3	t	n

pst	sg	pl
1	te	ten
2	test	tet
3	te	ten

Class 0 learner {st, t, n, te, test, ten, tet} → 1 analysis (16 partitions)
Class 1 learner add {} → 2^4 = 16 subanalyses (30450 partitions here)

German: Class 1

prs	sg	pl
1	Ø	n
2	st	t
3	t	n

pst	sg	pl
1	te	te-n
2	te-st	te-t
3	te	te-n

German: Class 2

prs	sg	pl
1	Ø-e	Ø-en
2	Ø-est	Ø-et
3	Ø-et	Ø-en

pst	sg	pl
1	t-e	t-en
2	t-est	t-et
3	t-e	t-en

Class 0/1 learner {e, est, et, en, te, test, ten, tet} → 1 analysis
Class 2 learner add {st, t, n} → $2^7 * 3^4 = 10368$ possible subanalyses
Class 3 learner add {es, s} → $2^{17} = 131072$ possible segmentations (12x)

German′: Class 3

prs	sg	pl
1	l-e	l-en
2	l-est	l-et
3	l-et	l-en

pst	sg	pl
1	t-e	t-en
2	t-est	t-et
3	t-e	t-en

prs	sg	pl
1	e	e|n (2)
2	e|st (2)	e|t (2)
3	e|t (2)	e|n (2)

pst	sg	pl
1	t|e (2)	t|e|n (3)
2	t|e|st (3)	t|e|t (3)
3	t|e (2)	t|e|n (3)

prs	sg	pl
1	e	e|n (2)
2	e|s|t (2^2)	e|t (2)
3	e|t (2)	e|n (2)

pst	sg	pl
1	t|e (2)	t|e|n (2^2)
2	t|e|s|t (2^3)	t|e|t (2^2)
3	t|e (2)	t|e|n (2^2)

Marker accuracy measurements

Paradigm: {(a,[-x,-y]), (b,[+x,-y]), (a,[-x,+y]), (ab,[+x,+y])}

	[-x]	[+x]
[-y]	a	b
[+y]	a	ab

Marker with	examples	false positives	false negatives
perfect precision and recall (↔)	(b,[+x])	no	no
perfect recall (→)	(a,[])	yes (1)	no
perfect precision (←)	(a,[-x]),	no	yes (1)
	(a,[+y])	no	yes (1)
neither	(b,[-y]),	yes (1)	yes (1)
	(a,[-y])	yes (1)	yes (2)

Incremental perfect precision learning

Input: the set P of (form, meaning) pairs of all affix strings

1. build the set M of potential markers without false positives
2. choose the optimal marker $O \in M$ with (a, b, c):
   - a maximal number of true positives (including subaffixes)
   - b minimal number of false negatives (excluding subaffixes)
   - c maximal number of segments
3. while any $C \in P$ has the same form as O
   - 3.1 remove the form of O from all meaning-matching $C \in P$
   - 3.2 add O to (initially empty) lexicon L
   - 3.3 recompute the next optimal marker O with identical form using step **1** and **2**
4. if all $C \in P$ have empty form output L else goto **1**

Restrict step **1** to class 0, class 1 or class 2 segmentation (checked with P)

German verbal agreement

prs	sg	pl
1	Ø	n
2	st	t
3	t	n

pst	sg	pl
1	te	ten
2	test	tet
3	te	ten

Class 1 restricted leaner

1. *te*, [+past]
2. *n*, [-2 +pl]
3. *st*, [+2 +sg]
4. t_1, [+2 +pl]
5. t_2, [+3 +sg -past]

prs	sg	pl
1	Ø	n
2	st	t_1
3	t_2	n

pst	sg	pl
1	te	te-n
2	te-st	te-t_1
3	te	te-n

Class 2 restricted leaner

1. *te*, [+past]
2. *n*, [-2 +pl]
3. t_1, [+2]
4. t_2, [+3 +sg -past]
5. *s*, [+2 +sg]

prs	sg	pl
1	Ø	n
2	s-t_1	t_1
3	t_2	n

pst	sg	pl
1	te	te-n
2	te-s-t_1	te-t_1
3	te	te-n

Estonian verbal agreement with class 0/1 restriction

prs	sg	pl
1	n	me
2	d	te
3	b	vad

imp	sg	pl
1	sin	sime
2	sid	site
3	sib	sid

1. *site*, [+2 +pl +past]
2. *sime*, [+1 +pl +past]
3. *sib*, [+3 +sg +past]
4. *sin*, [+1 +sg +past]
5. *vad*, [+3 +pl -past]
6. *te*, [+2 +pl -past]
7. *me*, [+1 +pl -past]
8. *b*, [+3 +sg -past]
9. *d*, [+2 +sg -past]
10. *n*, [+1 +sg -past]
11. *sid*, [+3 +pl +past]
12. *sid*, [+2 +sg +past]

Estonian verbal agreement with class 2 restriction I

prs	sg	pl
1	n	me
2	d_1	te
3	b	va-d_2

imp	sg	pl
1	si-n	si-me
2	si-d_1	si-te
3	si-b	si-d_2

1. *te*, [+2 +pl]
2. *si*, [+past]
3. *me*, [+1 +pl]
4. *b*, [+3 +sg]
5. *n*, [+1 +sg]
6. d_1, [+2 +sg]
7. d_2, [+3 +pl]
8. *va*, [+3 +pl -past]

Estonian verbal agreement with class 2 restriction II

(Ehala 2009: 42)

prs	sg	pl
1	n	me
2	d	te
3	b	vad

imp	sg	pl
1	s-i-n	s-i-me
2	s-i-d	s-i-te
3	s	s-id

1. *s*, [+past]
2. *te*, [+2 +pl]
3. *i*, [-3 +past]
4. *me*, [+1 +pl]
5. *d*, [+2 +sg]
6. *n*, [+1 +sg]
7. *vad*, [+3 +pl -past]
8. *id*, [+3 +pl +past]
9. *b*, [+3 +sg -past]

A Typological Pilot Study: Methodology

- Inflectional verbal paradigms of 20 areally and genetically diverse languages on the basis of Ruhlen's (1987) phyla and macroareas
- Only languages which have (at least some) subject agreement and TAM inflection on the same side of the stem are considered
- Disregarding portmanteau expression of subject agreement + TAM, non-finite verb forms, and non-segmental exponence
- **Prediction:**
  Either Agreement or TAM (or both) have at least one Ø-realization

A Typological Pilot Study: Language Sample

Language	Phylum	Macroarea	Ø-Agr	Ø-TAM	Source
Udmurt	Uralic	Eurasia	+	+	Csucs1998
Armenian	Indo-European	Eurasia	+	+	Schmitt1981
Nahuatl	Uto-Aztecan	N.America	+	+	Andrews1975
Kobon	Trans-New Guinea	Austr./N.Gui.	+	+	Davies1989
Mapudungun	Araucanian	S.America	+	+	Zuniga2000
Azerbaidjanian	S.Turkic	Eurasia	+	+	Schoenig1998
Turkana	Nilotic	Africa	+	+	Dimmendaal1983
Berber	Afroasiatic	Africa	+	+	Kossmann2007
Choktaw	Muskogean	N.America	+	+	Broadwell2006
Remo	Munda	Eurasia	+	+	AndersonHarrison2008
Kalkatungu	PamaNyungan	Austr./N.Gui.	+	+	Blake1979
Moghol	Mongolian	Eurasia	+	–	Weiers2011
Belhare	Kiranti	Se.Asia/Oc.	+	–	Bickel2003
Kannada	S.Dravidian	Eurasia	+	–	Steever1998
Somali	Cushitic	Africa	+	–	ElSolamiMewis1987
Inuktitut	Eskimo-Aleut	Eurasia	+	–	Mallon1991
Swahili	Bantu	Africa	–	+	Seidel1900
Pawnee	Caddoan	N.America	–	+	Parks1976
Manambu	Sepik	Austr./N.Gui.	–	+	Aikhenvald2008
Lenakel	CE.M-Polynes.	Se.Asia/Oc.	–	–	Lynch1978

A Typological Pilot Study: Results

- More than half of the languages (11/20) have some Ø-marking for Subject Agreement **and** TAM
- Virtually all languages (19/20) have some Ø-marking for **either** Subject Agreement **or** TAM

Lenakel

"In each case, the categories of person, tense, and number are obligatory, except that ... tense may be omitted in certain ... circumstances ... Certain tense prefixes may be omitted under certain conditions. The markers ak- and ~~im-~~ may be omitted in verbs with third person subjects when the context makes the time of action quite clear." (Lynch 1987:42,52)

Bound Affixes in Archi

(Kibrik 1998: 471; Kibrik 1991: 256)

Partial paradigms of *alnš* 'apple', *dab* 'awl', and *qlin* 'bridge'

	aInš		dab		qIin	
	SG	PL	SG	PL	SG	PL
NOM	*alnš*	*alnš-um*	*dab*	*dab-mul*	*qlin*	*qionn-or*
ERG	*alnš-li*	*alnš-um-čaj*	*dab-li*	*dab-mul-čaj*	*qlinn-i*	*qlonn-or-čaj*
GEN	*alnš-li-n*	*alnš-um-če-n*	*dab-li-n*	*dab-mul-če-n*	*qlinn-i-n*	*qlonn-or-če-n*
DAT	*alnš-li-s*	*alnš-um-če-s*	*dab-li-s*	*dab-mul-če-s*	*qlinn-i-s*	*qlonn-or-če-s*
⋮	⋮	⋮	⋮	⋮	⋮	⋮

- **-n** and **-s** occur only after **-li** and **-če/-čaj**
- **-če/-čaj** occurs only after a plural affix

  -čaj (word-final) ≈ **-če** (non-word-final)

Why **-če/-čaj** may be non-bound after all

Partial paradigms for *hagl təra* 'river', *c'aj* 'female goat', and *χ̣'on* 'cow'

	haˤtəra		c'aj		χ̣'on	
	SG	PL	SG	PL	SG	PL
NOM	*haˤtəra*	*haˤtər-mul*	*c'aj*	*c'ohor*	*χ̣'on*	*būc'i*
ERG	*haˤtər-čaj*	*haˤtər-mul-čaj*	*c'ej-t̄aj*	*c'ohor-čaj*	*χ̣'ini*	*būc'i-li*

The bound Superlative in Persian

(Mace 2003: 52)

Adjective	**Comparative**	**Superlative**	
bozorg	bozorg-tár	bozorg-tar-ín	'big'
mofid	mofid-tár	mofid-tar-ín	'useful'
moškel	moškel-tár	moškel-tar-ín	'clear'

(similar structure in Ubykh, Sanskrit, Gothic, cf. Bobaljik 2007: 12)

Swahili

(Seidel 1900: 10,11,14,18,42)

Present

	sg	**pl**
1	ni-na-pend-a	tu-na-pend-a
2	u-na-pend-a	m-na-pend-a
3	a-na-pend-a	wa-na-pend-a

Imperfective

	sg	**pl**
1	ni-li-pend-a	tu-li-pend-a
2	u-li-pend-a	m-li-pend-a
3	a-li-pend-a	wa-li-pend-a

Imperative

	sg	**pl**
1	–	–
2	pend-a	pend-eni
3	–	–

Subjunctive

	sg	**pl**
1	ni-pend-e	tu-pend-e
2	u-pend-e	m-pend-e
3	a-pend-e	wa-pend-e

Infinitive

ku-pend-a

Conclusion

- Boundness of affixes in affix strings gives a simple measure of learning complexity for inflectional systems
- Zero exponence reduces complexity because it potentially restricts hypotheses on possible segmentations
- This corresponds closely to the typological preference for inflectional systems with at least some zero exponence for every inflectional category

Bobaljik, Jonathan David (2007). *On Comparative Suppletion*. Ms., University of Connecticut

Ehala, Martin (2009). Linguistic strategies and markedness in Estonian morphology. *Sprachtypologie und Universalienforschung*, 1/2, 29 - 48.

Kibrik, Aleksandr (1991) Organising Principles for Nominal Paradigms in Daghestan Languages: Comparative and Typological Observations. In: *Paradigms: The Economy of Inflection*, ed. by Frans Plank, Berlin: Mouton de Gruyter, pp. 255–274.

Kibrik, Aleksandr (1998) Archi (Caucasian – Daghestanian). In: *Handbook of Morphology*, ed. by Andrew Spencer and Arnold Zwicky, Oxford: Blackwell, pp. 455–476.

Mace, John (2003) *Persian grammar: For reference and revision*. Routledge.

Müller, Gereon (2005). *Subanalyse verbaler Flexionsmarker*. Ms. Universität Leipzig.

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