Realistic data and paradigms: the paradigm cell finding problem

28.02.2020

Introduction

The authors make some emphasis on some points:

- Abstractive vs constructive morphology
- Comparison with physics(?)

> *bstractive approaches by defining their respective focus: The notion of a DERIVATION, which builds larger units from smaller elements, is central to constructive approaches. In abstractive approaches, PREDICTABILITY is the key relation. There is no requirement that one form should underlie another; a derivational relation is just a limiting case in which one of two forms in a predictability relation is a proper part of the other.*

But I don't understand how predictability isn't constructive

Abstraction:

Their definition:

> *is the process of inferring abstract systems from concrete linguistic data. In morphology, it abstracts word relations (forms, meanings, paradigmatic structure) from concrete word values (form, meaning, context).*

Their claim:

- word forms are the basic units of morphological data
- paradigms are basic units of morphological knowledge

Realistic morphological data

> *The only available information besides words would be the lexemic index required to regroup inflected forms belonging to the same paradigm and the syntactic context where the words are used.*
>
> *Basic units should be inflected word-forms. Neither morphemes, nor 'abstract' exponents should be considered as directly available to speakers. Blevins (2006)*

But why a lexemic index? There is no such thing in the raw data. What about:

- phonemes
- words (segmentation)
- semantic parsing
- syntactic contexts

Realistic morphological data

> *In a realistic perspective, data should come from empirical evidence, a wordform should be considered in context rather than in a preanalyzed lexicon with morphosyntactic and morphosemantic information not directly accessible to speakers.*

But this is not exactly what they do.

Data quantity and data quality

Data quantity

Approaches like the one by Stump and Finkel assume there is complete information about the lexicon:

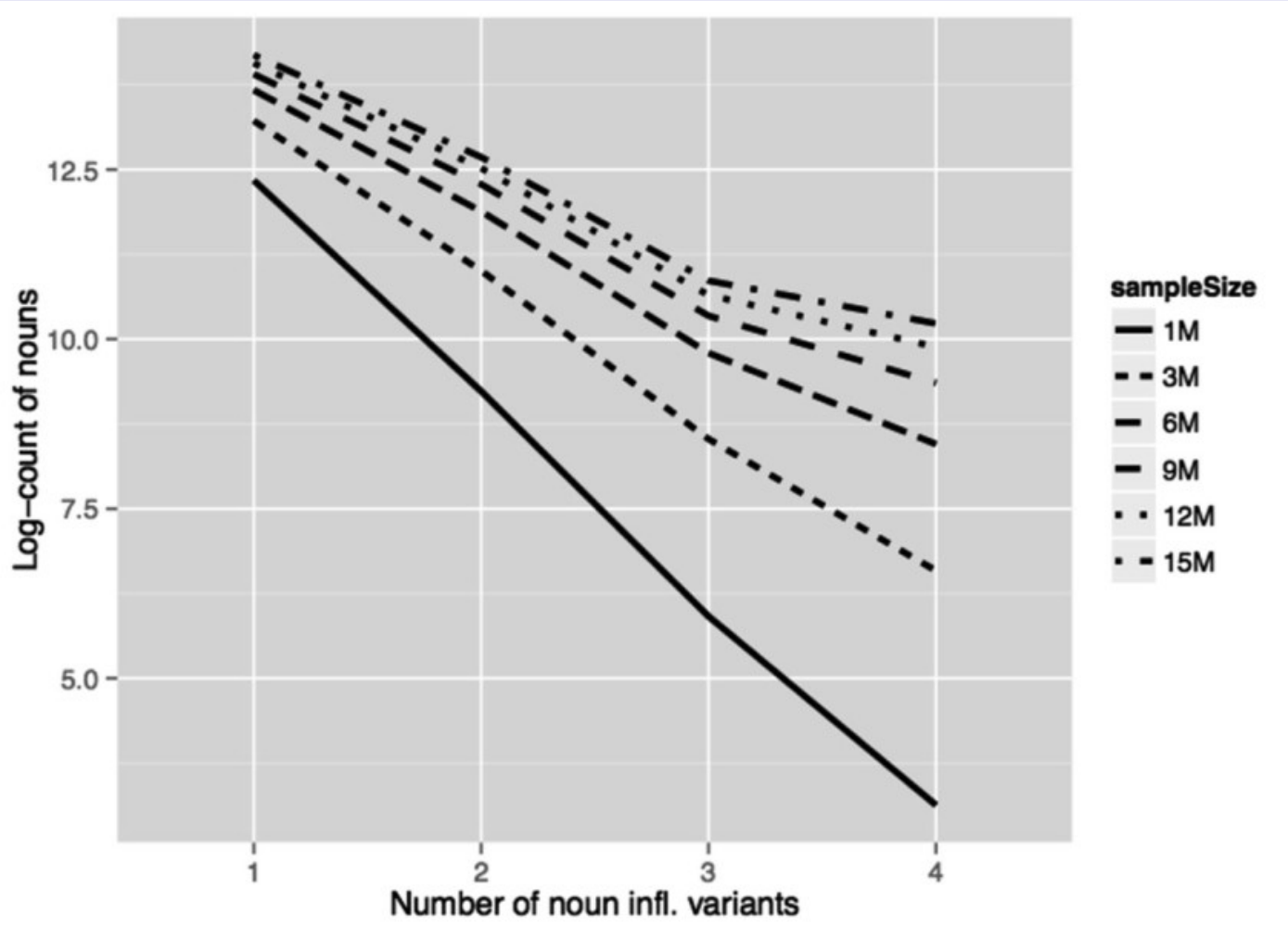
> *in the analysis of French verbs proposed by Stump and Finkel (2008), ÊTRE ('be') is meant to be the most predictable verb because only one dynamic principal part êtes ('you are') is required for it to be completely predicted. This runs contrary to the general intuition that ÊTRE is the most complex verb in French. But it is in complete accordance with the fact that the knowledge base used to derive their dynamic principal-part system possesses complete information about inflection.*

But this requires that speakers:

- have encountered all the forms of all known lexemes*
- have an exemplary paradigm for every inflectional class (maybe, eg: Mauritian)

Data quantity

However, things are difficult with larger paradigms:



Data quantity

Regarding Bonami and Beniamine:

> *while joint predictiveness is used to answer the PCFP starting from a sparse dataset to estimate the difficulty of predicting the complete one, the predictiveness calculations themselves rely on the complete dataset.*

Data quality

- Similar to infl. classes, infl. features are abstractions
- Infl. features emerge from discriminating contexts

(8) Marie ne croit pas que Paul [wœg].
Marie NEG=believe NEG COMP Paul wug.**IND/SBJV**.PRS.3SG
'Marie does not believe that Paul wugs'

(9) a. Marie ne croit pas que Paul [dɔʁ].
Marie NEG=believe NEG COMP Paul sleep.**IND**.PRS.3SG
'Marie does not believe that Paul sleeps' (*realis*)

b. Marie ne croit pas que Paul [dɔʁm].
Marie NEG=believe NEG COMP Paul wug.**SBJV**.PRS.3SG
'Marie does not believe that Paul sleeps' (*irrealis*)

Realistic paradigm derivation

They redefine the PCFP as:

> *Given exposure to a sample of inflected forms with enough lexemes, what licenses reliable inferences about the paradigm shape of these lexemes?*

Morphosyntactic paradigms

Table 6 Latin indicative conjugation morphosyntactic paradigm

INDICATIVE					
PRESENT	IMPERFECTIVE	FUTURE	PERFECT	PLU-PERFECT	FUTURE PAST
PRS.1SG	PST.IPFV.1SG	FUT.1SG	PRF.1SG	PPRF.1SG	FUT.PST.1SG
PRS.2SG	PST.IPFV.2SG	FUT.2SG	PRF.2SG	PPRF.2SG	FUT.PST.2SG
PRS.3SG	PST.IPFV.3SG	FUT.3SG	PRF.3SG	PPRF.3SG	FUT.PST.3SG
PRS.1PL	PST.IPFV.1PL	FUT.1PL	PRF.1PL	PPRF.1PL	FUT.PST.1PL
PRS.2PL	PST.IPFV.2PL	FUT.2PL	PRF.2PL	PPRF.2PL	FUT.PST.2PL
PRS.3PL	PST.IPFV.3PL	FUT.3PL	PRF.3PL	PPRF.3PL	FUT.PST.3PL

Table 7 *amare* tabular paradigm

	INDICATIVE					
	PRESENT	IMPERFECTIVE	FUTURE	PERFECT	PLU-PERFECT	FUTURE PAST
1SG	amo	amabam	amabo	amavi	amaveram	amavero
2SG	amas	amabas	amabis	amavisti	amaveras	amaveris
3SG	amat	amabat	amabit	amavit	amaverat	amaverit
1PL	amamus	amabamus	amabimus	amavimus	amaveramus	amaverimus
2PL	amatis	amabatis	amabitis	amavistis	amaveratis	amaveritis
3PL	amant	amabant	amabunt	amaverunt	amaverant	amaverint

Optimal morphomic paradigm

Table 9 English verbs: The optimal morphomic paradigm

Cell	Syncretic Feature Sets	BE	SING	TALK	SET
$Sync_1$	IND.PRS.1.SG	*am*	*sing*	*talk*	*set*
$Sync_2$	IND.PRS.2.SG, IND.PRS.1.PL, IND.PRS.2.PL, IND.PRS.3.PL	*are*	*sing*	*talk*	*set*
$Sync_3$	IND.PRS.3.SG	*is*	*sings*	*talks*	*sets*
$Sync_4$	IND.PST.1.SG, IND.PST.3.SG	*was*	*sang*	*talked*	*set*
$Sync_5$	IND.PST.2.SG, IND.PST.1.PL, IND.PST.2.PL, IND.PST.3.PL, SUBJ.PST.1.SG, SUBJ.PST.2.SG, SUBJ.PST.3.SG, SUBJ.PST.1.PL, IND.PST.2.PL, SUBJ.PST.3.PL	*were*	*sang*	*talked*	*set*
$Sync_6$	SUBJ.PRS.1.SG, SUBJ.PRS.2.SG, SUBJ.PRS.3.SG, SUBJ.PRS.1.PL, IND.PST.2.PL, SUBJ.PRS.3.PL, INF	*be*	*sing*	*talk*	*set*
$Sync_7$	PRS.PCTP	*being*	*singing*	*talking*	*setting*
$Sync_8$	PST.PCTP	*been*	*sung*	*talked*	*set*

Syncretisms

There are three types of syncretisms:

- Neutralizations
- Systematic
- Intersective

Intersective Syncretisms

Realistic data and paradigms: the paradigm cell finding problem

Table 19 BERG and NAME morphosyntactic paradigms

BERG	SG	PL	NAME	SG	PL
NOM	*Berg*	*Berge*	NOM	*Name*	*Namen*
ACC	*Berg*	*Berge*	ACC	*Namen*	*Namen*
DAT	*Berg*	*Bergen*	DAT	*Namen*	*Namen*
GEN	*Berges*	*Berge*	GEN	*Namens*	*Namen*

Table 20 the German nouns OMP with exemplary morphomic paradigms

Cell	Syncretic Feature Sets	MANN	MUTTER	KIND	KAMERA	STUDENT
$Sync_1$	NOM.SG	*Mann*	*Mutter*	*Kind*	*Kamera*	*Student*
$Sync_2$	ACC.SG, DAT.SG	*Mann*	*Mutter*	*Kind*	*Kamera*	*Studenten*
$Sync_3$	GEN.SG	*Mannes*	*Mutter*	*Kindes*	*Kamera*	*Studenten*
$Sync_4$	NOM.PL, ACC.PL, GEN.PL	*Männer*	*Mütter*	*Kinder*	*Kameras*	*Studenten*
$Sync_5$	DAT.PL	*Männern*	*Müttern*	*Kindern*	*Kameras*	*Studenten*

Paradigm emergence

Given exposure to a sample of inflected forms with enough lexemes, what licenses reliable inferences about the optimal morphomic paradigm shape of the category?

Two stage process:

- from form sets to the morphomic paradigms of lexemes
- from intersections of the morphomic paradigms of lexemes to OMPs

> *We assume that inflectional morphology begins when words start to be clustered into lexemes through semantic similarity.*

However, this would have to be shown to be correct.

> *With discriminative learning, speakers should be able to discover relevant infor- mation for each lexeme context class and infer some appropriate features.*

Has also not been shown (as far as I know).

An emergent OMP for French conjugation

- French lexicon
- $\sim$ 6500 verbs
- $\sim$ 328000 inflected forms
- phonological description from BDLEX
- word frequencies from Lexique3

French morphosyntactic paradigm

	1sg	2sg	3sg	1pl	2pl	3pl
Ind. present	pi1S	pi2S	pi3S	pi1P	pi2P	pi3P
Ind. imperfective	ii1S	ii2S	ii3S	ii1P	ii2P	ii3P
Ind. future	fi1S	fi2S	fi3S	fi1P	fi2P	fi3P
Cond. present	pc1S	pc2S	pc3S	pc1P	pc2P	pc3P
Subj. present	ps1S	ps2S	ps3S	ps1P	ps2P	ps3P
Ind. simple past	ai1S	ai2S	ai3S	ai1P	ai2P	ai3P
Subj. imperfective	is1S	is2S	is3S	is1P	is2P	is3P
Imperative	–	pI2S	–	pI1P	pI2P	–
Non-finite forms	inf	pP	ppMS	ppMP	ppFS	ppFP

Samples

I'm confused:

> *Here, we selected an increasing number of tokens at ran- dom in the lexicon according to the same token frequencies. Our samples range from 10,000 to- kens to 20,000,000 tokens and each lexical entry contains the following information:*

- form
- lexeme
- morphosyntactic tag
- number of occurrences

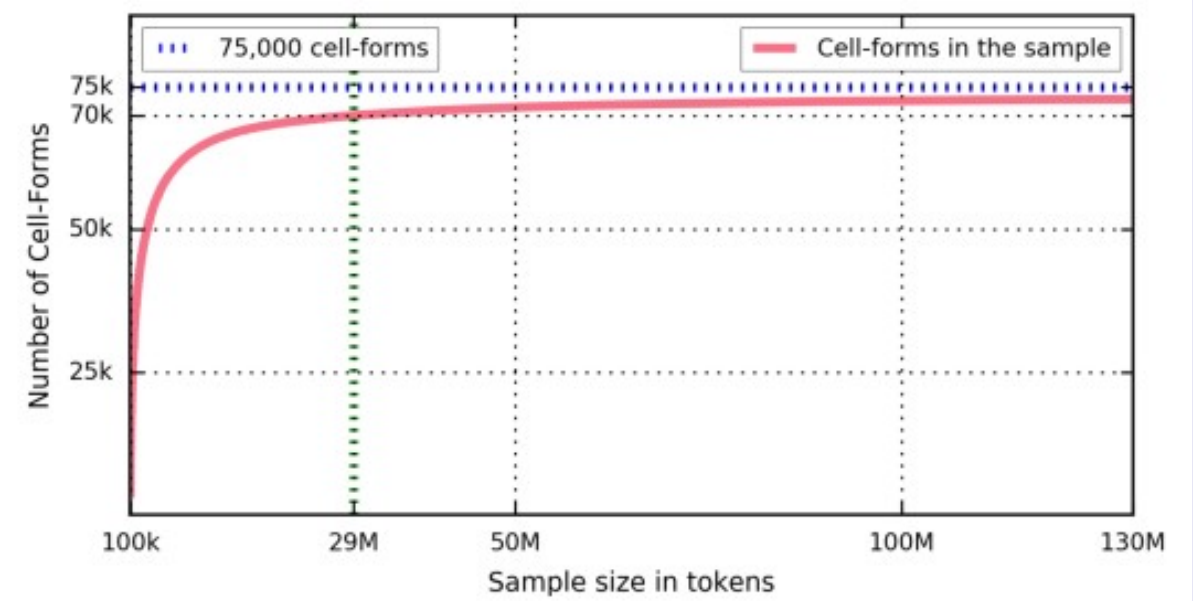
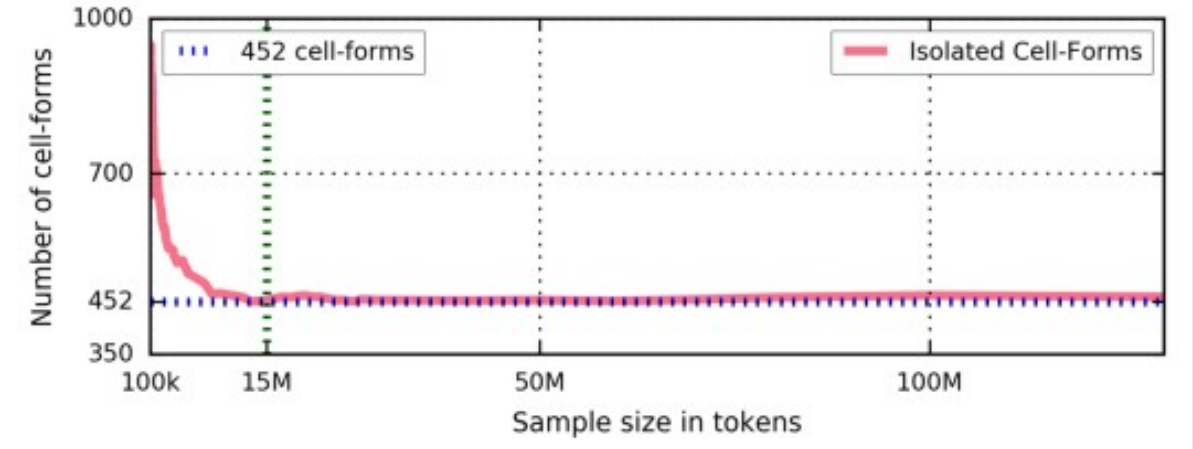


Fig. 9 Number of Cell-Forms according to Sample Size (Color figure online)



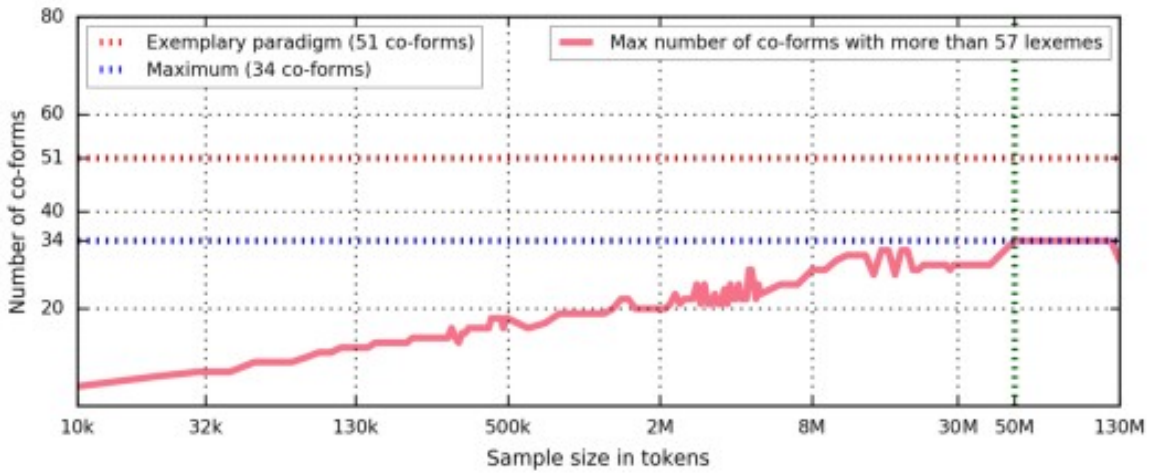


Fig. 13 Number of co-forms with at least 57 lexemes for different sample sizes (Color figure online)

The OMP processing

The French OMP

Table 29 The ideal OMP: 31 morphomic cells

	1sg	2sg	3sg	1pl	2pl	3pl
Ind. present	1	2	2	3	4	5
Ind. imperfective	6	6	6	7	8	6
Ind. future	9	10	10	11	9	11
Cond. present	9	9	9	12	13	9
Subj. present	14	14	14	15	16	14
Ind. simple past	17	18	18	19	20	21
Subj. imperfective	22	22	18	23	24	22
Imperative	–	25	–	26	27	–
	inf	pP	ppMS	ppMP	ppFS	ppFP
Non-finite forms	28	29	30	30	31	31

The algorithm

- find morphomic cells which partition the morphosyntactic paradimg
- a morphomic cell must have the same phonological content for all morphosyntactic cells inside it
- look for unifiable cells based on the morphosyntactic tags

Co-pair: couple of forms instatiating a pair of cells

> *A pair of cells C1 , C2 is unifiable, if for all the lexemes having co-pairs for (C1 , C2), the forms in C1 and C2 are identical.*

> *The proximity of the resulting OMPs to the ideal OMP depends on the availability of a sufficient number of co-pairs for every pair of cells of the morphosyntactic paradigm.*

Table 30 10,000 tokens: 4085 forms & 22 morphomic cells

	1sg	2sg	3sg	1pl	2pl	3pl
Ind. present	1	2	2	~~3/26~~	4	5
Ind. imperfective	6	6	6	7	8	6
Ind. future	9	10	10	11	9	11
Cond. present	9	9	9	~~12/31~~	~~13/18~~	9
Subj. present	14	14	14	~~15/16~~	~~16/15~~	14
Ind. simple past	~~17/18~~	18	18	~~19/31~~		21
Subj. imperfective			18			~~22/14~~
Imperative	–	25	–	~~26/3~~	27	–
	inf	pP	ppMS	ppMP	ppFS	ppFP
Non-finite forms	28	29	30	30	31	31

The shades indicate morphomes, the cells in black have no corresponding forms in the sample

Table 31 100,000 tokens: 16,484 forms & 31 morphomic cells

	1sg	2sg	3sg	1pl	2pl	3pl
Ind. present	1	2	2	3	4	5
Ind. imperfective	6	6	6	7	8	6
Ind. future	9	10	10	11	9	11
Cond. present	9	9	9	12	13	9/9
Subj. present	14	14	14	15	16	14
Ind. simple past	17	18	18	19	20	21
Subj. imperfective	22		18	23		22
Imperative	–	25	–	26	27	–
	inf	pP	ppMS	ppMP	ppFS	ppFP
Non-finite forms	28	29	30	30	31	31

Table 32 1 million tokens: 43,314 forms & 32 morphomic cells

	1sg	2sg	3sg	1pl	2pl	3pl
Ind. present	1	2	2	3	4	5
Ind. imperfective	6	6	6	7	8	6
Ind. future	9	10	10	11	9	11
Cond. present	9	9	9	12	13	9
Subj. present	14	14	14	15	16	14
Ind. simple past	17	18	18	19	20	21
Subj. imperfective	22	22	18	23	24	22
Imperative	–	25	–	26	27	–
	inf	pP	ppMS	ppMP	ppFS	ppFP
Non-finite forms	28	29	30	30	31	31

OMP

The benefits of the OMP:

- Reduces number of copairs needed
- Reduction of systematic syncretism
- Simplifies the PCFP
- Fewer comparisons needed

Conclusion

Reliable inferences can be obtained in two stages:

- Emergence of paradigm structure
- Emergence of systematic relations between cells

Additionally:

- Morphosyntactic paradigms are an arbitrary generalization of the zero syncretism situation
- The reduction of systematic syncretisms with the OMP questions the status of syncretisms and polysemy