The Low Entropy Conjecture: The challenges of Modern Irish nominal declension

LSA Workshop on Challenges of Complex Morphology to Morphological Theory

July 27, 2010

Rob Malouf
San Diego State University

Farrell Ackerman
University of California at San Diego

The Paradigm Cell Filling Problem

Speakers of languages with complex morphology and multiple inflection classes must generalize beyond direct experience, since it's implausible to imagine they will have encountered each form of every word

Paradigm Cell Filling Problem: Given exposure to an inflected wordform of a novel lexeme, what licenses reliable inferences about the other wordforms in its inflectional family? (Ackerman, Blevins, & Malouf 2009)

Some questions

1. How are wordforms organized into patterns within a morphological system?
2. How can one identify implicative relations between these units?
3. How might the implicative organization of a system contribute to licensing inferences that solve the paradigm cell filling problem?
4. How does this organization, and the surface inferences it licenses, contribute to the robustness and learnability of complex morphological systems?

The basic background

Inflectional morphology can exhibit spectacular complexity in:

i. syntagmatic, morphophonemic, suprasegmental structure of individual words;

ii. the size of inventories for morphosyntactic distinctions formall expressed by words;

iii. paradigmatic patterns that (classes of) words participate in.

This is the **External Complexity** or **E-complexity** of a morphological system

Our guiding intuition

Morphological systems **must** be simple in ways that allow them to be learned and used by native speakers, irrespective of how complex words and paradigms may appear according to external measures.

Speakers must generalize beyond their direct experience:

> Morphological systems must permit speakers to make accurate guesses about unknown forms of lexemes based on only a few known forms.

This is the **Internal Simplicity** or **I-simplicity** of a system

Our hypothesis: I-simplicity

I-simplicity is measurable and quantifiable

Principle of Low Paradigm Entropy: Paradigms tend to have low expected conditional entropy, where Paradigm entropy is the average of conditional entropies among all pairs of words.

Gradation in first declension nouns in Saami (Bartens 1989:511)

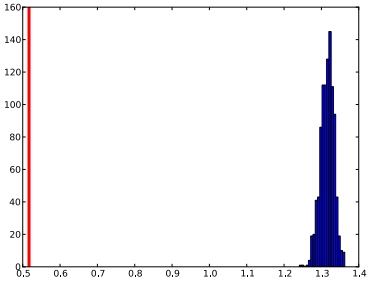
	'Weakening'		'Strengthening'	
	Sing	Plu	Sing	Plu
Nominative	**bihttá**	bihtát	baste	**basttet**
Gen/Acc	bihtá	bihtáid	**bastte**	**basttiid**
Illative	**bihttái**	bihtáide	bastii	**basttiide**
Locative	bihtás	bihtáin	**basttes**	**basttiin**
Comitative	bihtáin	bihtáiguin	**basttiin**	**basttiiguin**
Essive		**bihttán**		basten
		'piece'		'spoon'

$$H(\text{LOC.PL}|\text{NOM.SG}) = H(\text{NOM.SG, LOC.PL}) - H(\text{NOM.SG})$$
$$= 1.0 - 1.0$$
$$= 0.0$$

Results based on uniform type freqs

Language	Declensions	Cells	Realizations	Paradigm entropy	Bootstap Avg	Bootstrap *p*
Arapesh	26	2	41	0.630	0.630	1.000
Burmeso	2	12	24	0.000	0.000	1.000
Fur	19	12	80	0.517	1.316	0.001
Kwerba	4	12	26	0.428	0.523	0.001
Ngiti	10	16	68	0.484	1.019	0.001
Nuer	16	6	12	0.793	0.811	0.160
Russian	4	12	26	0.538	0.541	0.383

Fur



Upper bounds from idealized data

Bonami et. al. (2011), Sims (2011 LSA Institute) point out the challenges for testing the Low Entropy Conjecture beyond using grammar descriptions:

Realistic data bases:

1. Need for veridical representations of spoken words, rather than phonological idealizations.
2. Need for type and token frequencies
3. Need to recognize small, reliable implicational patterns that partition the whole paradigm into sub-paradigms

A passing effort: Tundra Nenets

"Given any Tundra Nenets inflected nominal word form, what are the remaining 209 forms of this lexeme for the allowable morphosyntactic feature property combinations CASE: {nom, acc, gen, dat, loc, abl, pro}, NUMBER: {singular, dual, plural}, POSSESSOR: {3 persons 3 numbers}?"

Corpus of 4,334 nominals with type and token frequencies

Nom Sg|Acc Pl; Acc Pl|Nom Sg

Nom Sg	Acc Pl	
ngano	nganu	'boat'
lyabtu	lyabtu	'harnessed deer'
ngum	nguwo	'grass'
xa	xawo	'ear'
nyum	nyubye	'name'
yi	yíbye	'wit'
myir	myirye	'ware'
wí´	wíngo	'tundra'
we´	weno	'dog'
nguda	ngudyi	'hand'
xoba	xob	'fur'
sawənye	sawənyi	'magpie'
tyírtya	tyírtya	'bird'

Modern Irish nominal inflection

Modern Irish presents a challenging test case for the Low Entropy Conjecture

Definite nominal paradigm (excluding vocative):

caibidil	'chapter'	*an chaibidil*	DEF.COM.SG
		na caibidle	DEF.GEN.SG
		an gcaibidil	DEF.PREP.SG
		na caibidlí	DEF.COM.PL
		na gcaibidlí	DEF.GEN.PL
		na caibidlí	DEF.PREP.PL

Definiteness, case, and number marked by an article, prefix, consonant mutation, stem alternation, syncope, and/or suffix

"This kind of complexity makes students, teachers, and linguists alike scratch their heads in wonder and fear." (Carnie 2008, 6)

Modern Irish nominal inflection

Carnie (2008) presents an updated and elaborated analysis of the nominal system

- Two genders
- Forty singular declensions
- Sixty-five plural types

Carnie also gives class membership and full paradigms for 1,216 nouns, exemplifying 220 gender/declension/plural class combinations

This gives us a detailed inventory of inflection classes, with information about the realization of wordforms, and from the word list we can estimate the type frequency of each

Modern Irish paradigm entropy

Paradigm entropy = 1.529 bits (!)

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		2.955	0.034	3.936	3.690	3.936	2.910
GEN.SG	0.000		0.021	2.089	2.055	2.089	1.251
PREP.SG	0.000	2.942		3.926	3.680	3.926	2.895
COM.PL	0.625	1.733	0.649		0.003	0.000	0.602
GEN.PL	0.795	2.116	0.819	0.419		0.419	0.914
PREP.PL	0.625	1.733	0.649	0.000	0.003		0.602
E[COL]	0.409	2.296	0.434	2.074	1.886	2.074	**1.529**

Modern Irish paradigm entropy

GEN.SG as a principal part

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		**2.955**	**0.034**	3.936	3.690	3.936	2.910
GEN.SG	**0.000**		**0.021**	2.089	2.055	2.089	1.251
PREP.SG	**0.000**	**2.942**		3.926	3.680	3.926	2.895
COM.PL	0.625	1.733	0.649		0.003	0.000	0.602
GEN.PL	0.795	2.116	0.819	0.419		0.419	0.914
PREP.PL	0.625	1.733	0.649	0.000	0.003		0.602
E[COL]	0.409	2.296	0.434	2.074	1.886	2.074	1.529

Modern Irish paradigm entropy

Plurals are mostly inter-predictable

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		2.955	0.034	3.936	3.690	3.936	2.910
GEN.SG	0.000		0.021	2.089	2.055	2.089	1.251
PREP.SG	0.000	2.942		3.926	3.680	3.926	2.895
COM.PL	0.625	1.733	0.649		**0.003**	**0.000**	0.602
GEN.PL	0.795	2.116	0.819	**0.419**		**0.419**	0.914
PREP.PL	0.625	1.733	0.649	**0.000**	**0.003**		0.602
E[COL]	0.409	2.296	0.434	2.074	1.886	2.074	1.529

Modern Irish paradigm entropy

Singular forms are very bad predictors of plural forms

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		2.955	0.034	**3.936**	**3.690**	**3.936**	2.910
GEN.SG	0.000		0.021	**2.089**	**2.055**	**2.089**	1.251
PREP.SG	0.000	2.942		**3.926**	**3.680**	**3.926**	2.895
COM.PL	0.625	1.733	0.649		0.003	0.000	0.602
GEN.PL	0.795	2.116	0.819	0.419		0.419	0.914
PREP.PL	0.625	1.733	0.649	0.000	0.003		0.602
E[COL]	0.409	2.296	0.434	2.074	1.886	2.074	1.529

Some reasons to doubt

Inflection class analysis based on a long philological tradition

Classes are based on the standard written language, but Modern Irish orthography is not very transparent

neamhthruamhéalach /nʲauhruavʲeːlah/

Consonant mutations are not easy to identify in spoken forms

'boat'	*an bád*	/ən baːd/	COM.SG
	an mbád	/ən maːd/	PREP.SG
'bag'	*an mála*	/ən maːlə/	COM.SG
	an mála	/ən maːlə/	PREP.SG

Classes often restricted to particular genders, phonological patterns, semantic patterns, etc.

An alternative

Declension systems like Carnie's aren't designed for answering the questions we're asking

A better alternative (for this task) is to derive the classes directly from the lexicon

Ideally, we would find all and only the patterns that obtain in the lexicon, without coming at the descriptive problem with any prior assumptions

But this is very difficult (and likely impossible in principle)

Discovering inflection classes

The first step is to convert Carnie's paradigms into a phonological transcription using TCD's Abair speech synthesis system

caibidil	'chapter'	/kabʲəd^{ʲ}əl^{ʲ}/
an chaibidil	DEF.COM.SG	/ənxabʲid^{ʲ}il^{ʲ}/
na caibidle	DEF.GEN.SG	/nəkabʲid^{ʲ}l^{ʲ}ə/
an gcaibidil	DEF.PREP.SG	/əngabʲid^{ʲ}il^{ʲ}/
na caibidlí	DEF.COM.PL	/nəkabʲid^{ʲ}l^{ʲ}iː/
na gcaibidlí	DEF.GEN.PL	/nəgabʲid^{ʲ}l^{ʲ}iː/
na caibidlí	DEF.PREP.PL	/nəkabʲid^{ʲ}l^{ʲ}iː /

Discovering inflection classes

Perform a multiple alignment using modified edit distance

	k	a	bʲ	ə	dʲ	ə	lʲ	
ən	x	a	bʲ	i	dʲ	i	lʲ	
nə	k	a	bʲ	i	dʲ	∅	lʲ	ə
ən	g	a	bʲ	i	dʲ	i	lʲ	
nə	k	a	bʲ	i	dʲ	∅	lʲ	iː
nə	g	a	bʲ	i	dʲ	∅	lʲ	iː
nə	k	a	bʲ	i	dʲ	∅	lʲ	iː

Discovering inflection classes

Remove anything that stays constant across the paradigm to find the signature of this lexeme's inflection class

	k	ə	ə	
ən	x	i	i	
nə	k	i	∅	ə
ən	g	i	i	
nə	k	i	∅	iː
nə	g	i	∅	iː
nə	k	i	∅	iː

Discovering inflection classes

Applying this discovery procedure yields 950 declensions (not very useful for pedagogy!)

These 'classes' are essentially small phonological neighborhoods, which serve as the domains for analogies

A nagging concern: many of the neighborhoods are very small (consisting of a single lexeme in the 1,200 word sample) and may not reflect any useful generalizations

Next steps

- Look at larger neighborhoods, perhaps by using a feature representation for segments
- Scale up to a larger sample of nouns, to see how neighborhoods fill in

Modern Irish paradigm entropy

Much lower paradigm entropy

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		1.003	0.808	0.976	0.104	1.011	0.780
GEN.SG	0.723		0.840	0.039	0.602	0.010	0.443
PREP.SG	0.304	0.617		0.594	0.110	0.622	0.449
COM.PL	0.770	0.113	0.892		0.603	0.123	0.500
GEN.PL	0.467	1.245	0.976	1.172		1.250	1.022
PREP.PL	0.724	0.003	0.838	0.041	0.600		0.441
E[COL]	0.598	0.596	0.871	0.565	0.404	0.603	**0.606**

Modern Irish paradigm entropy

GEN.SG is still very predictive, but it's not quite a 'principal part'

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		1.003	0.808	0.976	0.104	1.011	0.780
GEN.SG	**0.723**		**0.840**	**0.039**	**0.602**	**0.010**	**0.443**
PREP.SG	0.304	0.617		0.594	0.110	0.622	0.449
COM.PL	0.770	0.113	0.892		0.603	0.123	0.500
GEN.PL	0.467	1.245	0.976	1.172		1.250	1.022
PREP.PL	0.724	0.003	0.838	0.041	0.600		0.441
E[COL]	0.598	0.596	0.871	0.565	0.404	0.603	0.606

Conditional entropies are more even overall

Most forms are mostly predictable from most other forms, with no one form as the key

Modern Irish paradigm entropy

Singular forms are reasonably good predictors of plural forms

H(COL|ROW)	COM.SG	GEN.SG	PREP.SG	COM.PL	GEN.PL	PREP.PL	E[ROW]
COM.SG		1.003	0.808	**0.976**	**0.104**	**1.011**	0.780
GEN.SG	0.723		0.840	**0.039**	**0.602**	**0.010**	0.443
PREP.SG	0.304	0.617		**0.594**	**0.110**	**0.622**	0.449
COM.PL	0.770	0.113	0.892		0.603	0.123	0.500
GEN.PL	0.467	1.245	0.976	1.172		1.250	1.022
PREP.PL	0.724	0.003	0.838	0.041	0.600		0.441
E[COL]	0.598	0.596	0.871	0.565	0.404	0.603	0.606

Token frequency

All reported entropy calculations were weighted by **type** frequency (the number of lexemes in each class)

The **token** frequency of each inflected form may also be relevant for learning

Form	Predicting	Predicted	Frequency*
COM.SG	0.780	0.598	155,960
GEN.SG	0.443	0.596	71,614
PREP.SG	0.449	0.871	161,699
COM.PL	0.500	0.565	31,711
GEN.PL	1.022	0.404	22,660
PREP.PL	0.441	0.603	40,794

Conclusions

While detailed descriptive accounts of morphological systems can provide a useful **entry point** for analysis, they can also lead, as in Carnie's fine description of Irish nominal declension, to misleading conclusions about Paradigm Entropy (Bonami et. al. 2011).

In order to explore the validity of the Low Entropy Conjecture as well as the numerous ways that it may obtain in morphological systems, the choice of **representation** (for both forms and classes) crucial

Including accurate type (and token?) **frequency** is also important for getting an accurate picture of lexical organization

Paraphrasing Bonami et. al. (2011): this is **"tedious work"**, but it's both doable and necessary if we really want to understand morphological systems.

Sources

Farrell Ackerman et.al. 2009. "Parts and whole: Implicative patterns in inflectional paradigms." In J. P. Blevins & J. Blevins eds. *Analogy in Grammar.* Oxford University Press.

Olivier Bonami et. al. 2011. Measuring inflection complexity. Presented at Quantitative Methods in Morphology Workshop, Center for Human Development, UCSD.

Andrew Carnie. 2008. *Irish Nouns: A Reference Guide.* Oxford University Press.

Amelia Kelly. Text-to-Speech Synthesis for Irish. Presented at University of Amsterdam, February 2009. http://abair.ie

Adam Kilgarriff, Michael Rundell and Elaine Uí Dhonnchadha. 2006. "Efficient corpus development for lexicography: Building the New Corpus for Ireland." *Language Resources and Evaluation* 40(2):127–152. http://corpas.focloir.ie/ [via Michal Boleslav Měchura]

Andrea Sims. 2011. Information theory and paradigmatic morphology. Presentation at Information-theoretic Approaches to Linguistics Workshop LSA Summer Institute, Bolder, Colorado.