Information Theory and Paradigmatic Morphology

Andrea D. Sims
The Ohio State University
sims.120@osu.edu

Information-theoretic Approaches to Linguistics Workshop
Boulder, Colorado
July 15-16, 2011

Introduction

- What is a paradigm?
  - "... the set of all the inflected forms which an individual word assumes" (Spencer 1991:11).
  - "... a definition of the set of morphological contrasts that a given class of lexemes can make" (Spencer 2004:72).

	singular	plural
nom	mamá	mamáđes
gen	mamás	mamáđon
acc	mamá	mamáđes

	singular	plural
nom	patéras	patéres
gen	patéra	patéron
acc	patéra	patéres

Table 1: Modern Greek nouns 'mother' and 'father'

Introduction

- Paradigmatic and syntagmatic relations
  - Syntagmatic: relations holding between linearly (or hierarchically) ordered units

stem + case/number morph

  - Paradigmatic: relations holding between units (words) that are contrastive

acc = nom

gen.sg mamás → acc.sg. mamá
acc.sg mamá → gen.sg. mamás

	singular	plural	
nom	mamá	mamá	đes
gen	mamá s	mamá đon	
acc	mamá	mamá đes	

	singular	plural
nom	patér as	patér es
gen	patér a	patér on
acc	patér a	patér es

Table 1: Modern Greek nouns 'mother' and 'father'

Introduction

- Most working morphologists view a morpheme-based, purely syntagmatically-oriented model as insufficient, and instead posit the inflectional paradigm as a theoretical primitive
- Inflectional paradigms (and inflection classes) become objects of inquiry
  - **What is the internal organization of the inflectional paradigm?**
    - Syncretism and similar phenomena (e.g. Baerman et al. 2005, Zwicky 1985, Stump 2001, among many others)
    - Interpredictability or predictiveness of paradigm cells (e.g. Albright and Hayes 2002; Finkel and Stump 2007,2009; Ackerman et al. 2009)
  - **Do languages differ substantially in the complexity of their paradigms' organization?**
    - Low Entropy Conjecture (Malouf and Ackerman 2011)

Heteroclisis in Croatian

	CLASS I NEUT. (*more* 'sea')	Heteroclite (*dijete* 'child')		CLASS II (*žena* 'woman')
	SING	SING	PLUR	SING
NOM	mor-e	dijet-e	djec-a	žen-a
ACC	mor-e	dijet-e	djec-u	žen-u
GEN	mor-a	djetet-a	djec-e	žen-e
DAT-LOC	mor-u	djetet-u	djec-i	žen-i
INST	mor-om	djetet-om	djec-om	žen-om

Table 2

Inferring unobserved forms

- **Paradigm Cell Filling Problem** (PCFP, Ackerman et al. 2009) : When a speaker encounters a novel lexeme, what licenses reliable inferences about the lexeme's remaining forms?
  - Hypothesis: Speakers use knowledge of implicational relations to analogically derive unencountered word-forms.
  - Suggest that knowledge of paradigmatic implications is best captured in information-theoretic terms (see also Bonami et al. 2011, Milin et al. 2009, Moscoso del Prado Martín et al. 2004).
    - Paradigm entropy
- **Low Entropy Conjecture** (Malouf and Ackerman 2010): Cross-linguistically, paradigms tend to have low expected conditional entropy
  - Guiding intuition: "Morphological systems must permit speakers to make accurate guesses about unknown forms of lexemes based on only a few known forms."
  - Internal Simplicity vs. External Complexity

An observation and a question

- **A simple observation:** In a language with small inflectional paradigms, each word-form will be (on average) encountered more often.
- **Question**

  If low paradigm entropy is a response to the PCFP...,

  and to the extent that the PCFP is unlikely to be equally a problem in all languages...,

  should we expect to find cross-linguistic differences in the extent to which paradigms are internally organized by sets of strong implicational relations?

  - External simplicity and internal complexity??

Goals for the remainder of the talk

- Explore form-level paradigm cohesion as an empirical question through detailed exploration of a single language (Modern Greek)
- Show that a system with few paradigm cells can turn out to be complex from the perspective of the interpredictability of the cells
  - More or less the converse of the Low Entropy Conjecture
- Show that the complexity of the system has structural consequences (in the form of paradigmatic gaps)
- Speculate about potentially differential role of interpredictability of cells cross-linguistically
  - Why this data is not (necessarily) at odds with the LEC

- How predictable is any one paradigm cell based on any other?

	singular	plural
nom	mamá	mamáđes
gen	mamás	mamáđon
acc	mamá	mamáđes

- A paradigm exhibits *cohesion* to the extent that all of its forms are interconnected via a single set of implicational relations

Lack of paradigm cohesion

	CLASS I NEUT. (*more* 'sea')	Heteroclite (*dijete* 'child')		CLASS II (*žena* 'woman')
	SING	SING	PLUR	SING
NOM	mor-e	dijet-e	djec-a	žen-a
ACC	mor-e	dijet-e	djec-u	žen-u
GEN	mor-a	djetet-a	djec-e	žen-e
DAT-LOC	mor-u	djetet-u	djec-i	žen-i
INST	mor-om	djetet-om	djec-om	žen-om

Table 2: Croatian nouns

To what extent is Modern Greek like *dijete?*

Overview of Modern Greek nominal system

Singular Formatives		
NOM	GEN	ACC
X	X*s*	X
X*s*	X	X*o*
X*os*	X*u*	X*os*
X*o*	X*us*	X*a*
X*as*	X*os*	X*s*
X*ma*	X*tos*	X*ma*
X*mo*	X*matos*	X*mo*

Plural Formatives		
NOM	GEN	ACC
X*es*	X*on*	X*es*
X*is*	X*eon*	X*is*
X*i*	X*don*	X*i*
X*a*	X*iđon*	X*a*
X*đes*	X*ton*	X*đes*
X*iđes*	X*maton*	X*iđes*
X*ta*	X*anton*	X*ta*
X*mata*		X*mata*
X*antes*		X*antes*
		X*us*

Table 3: Greek nominal formatives

Modern Greek nominal stress (partial)

	singular	plural
nom	mam**á**	mam**á**đes
gen	mam**á**s	mam**á**đon
acc	mam**á**	mam**á**đes

Pattern 1: Columnar stress throughout

	singular	plural
nom	tur**í**stas	tur**í**stes
gen	tur**í**sta	turist**ó**n
acc	tur**í**sta	tur**í**stes

Pattern 2: Gen.pl. stress is final or penultimate

	singular	plural
nom	anal**ó**ɣio	analóɣia
gen	analoɣ**í**u	analoɣ**í**on
acc	anal**ó**ɣio	anal**ó**ɣia

Pattern 3: Both gen.sg. and gen.pl. stress are final or penultimate

	singular	plural
nom	**á**nθropos	**á**nθropi
gen	anθr**ó**pu	anθr**ó**pon
acc	**á**nθropo	anθr**ó**pus

Pattern 4: Acc.pl., gen.sg. and gen.pl. stress are penultimate

Table 4: Morphological stress formatives

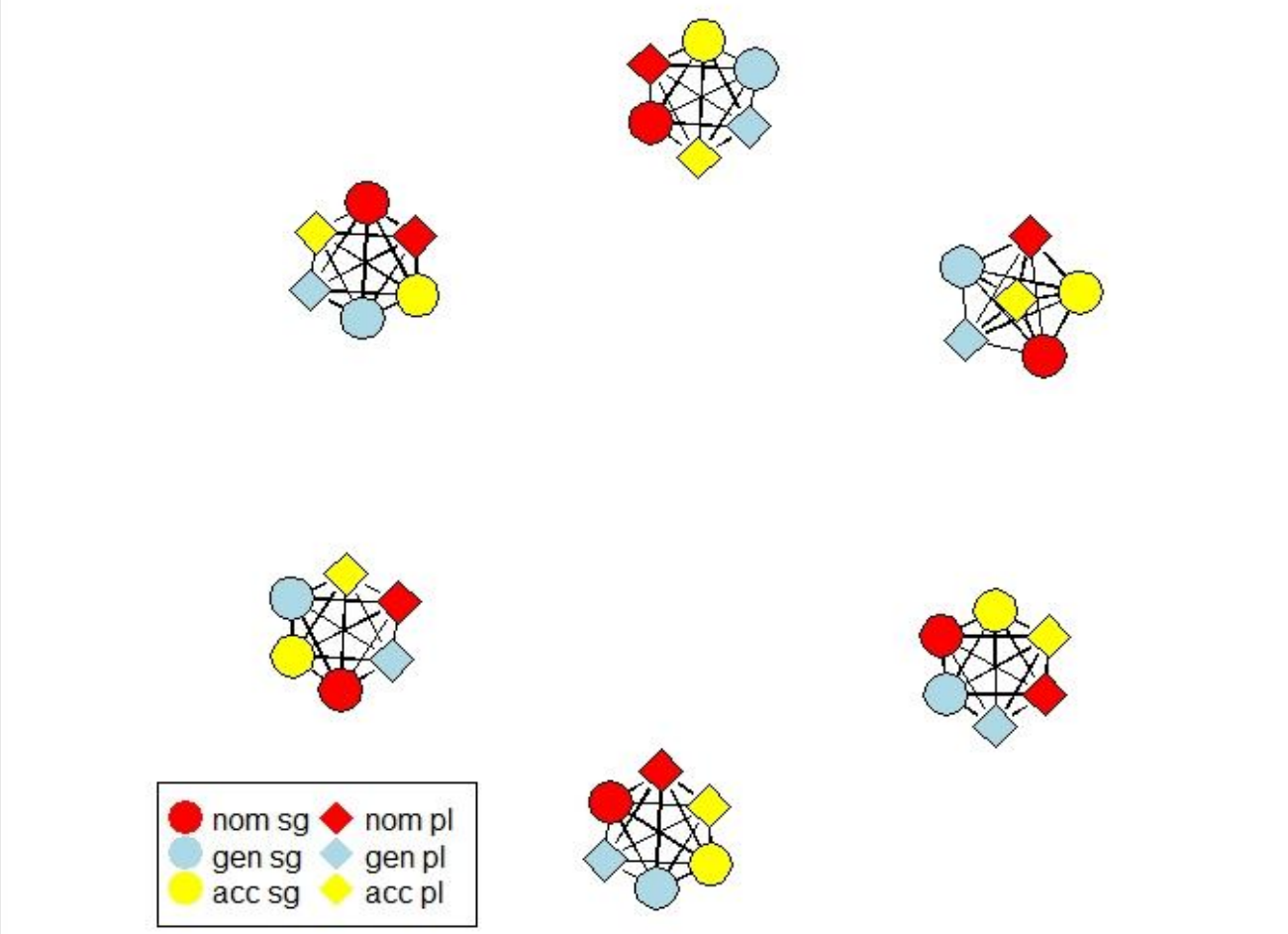
Overlapping inflection classes

	'force'	'mother'	'green-grocer'	'tourist'	'father'
nom.sg.	đínami	mamá	manávi-s	turísta-s	patéra-s
gen.sg.	đínami-s	mamá-s	manávi	turísta	patéra
acc.sg.	đínami	mamá	manávi	turísta	patéra
nom.pl.	đínami-s	mamá-đes	manávi-đes	turíst-es	patér-es
gen.pl.	đinám-eon	mamá-đon	manávi-đon	turist-ón	patér-on
acc.pl.	đínami-s	mamá-đes	manávi-đes	turíst-es	patéres

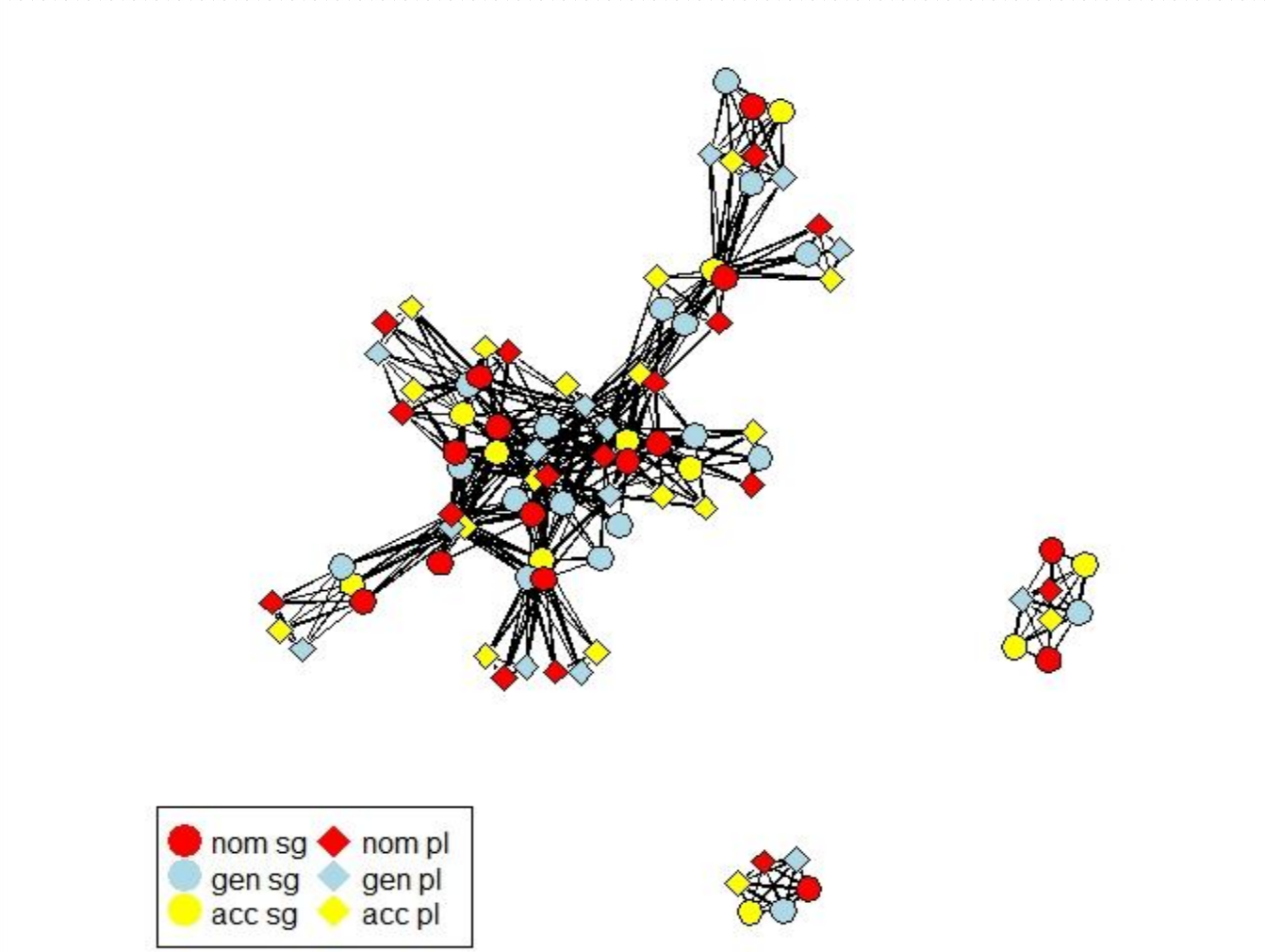
Table 5: A few Greek lexemes

Among Greek nouns, how widespread are overlapping formatives?

Hypothetical, "ideal" network of implicational relations



Actual Greek nominal network



Operationalizing interconnectedness

- X: a morphosyntactic property set (MSPS, e.g. NOM.SG)
- x: an exponent of a MSPS (e.g. -*a*)
  - Thus, x is a value of X
- p(x), where $x \in X$, is the probability of the exponent
- p(x) is probably dependent on many things:
  - type frequency
  - token frequency
  - behavior of phonologically similar items
  - ?? (See Bonami et al. 2011 for how these factors can affect paradigm entropy calculations)
- **I use type frequency to estimate probability of occurrence of x**: $p(x) = N_X / N$
  - Sample of ~27,300 distinct lexemes from the online version of the Triantafillidis Institute dictionary

Conditional entropy

$$H(Y|X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x)$$

- H(Y|X) is the entropy of MSPS Y, given that we know the value of x

- Intuitively, a measure of the extent to which the word-form in one paradigm cell can be predicted from the word-form in another cell *in general* (i.e. without regard for the particular exponent x).

Conditional entropy: H(Y|X)

		Predicted					
		nom.sg	gen.sg	acc.sg	nom.pl	gen.pl	acc.pl
Predictor	nom.sg	NA	3.40	0.18	5.77	11.84	6.55
	gen.sg	4.00	NA	2.06	6.83	9.31	6.83
	acc.sg	2.96	4.95	NA	4.74	7.55	5.11
	nom.pl	6.45	7.03	5.29	NA	5.56	0.838
	gen.pl	10.71	8.29	8.44	5.04	NA	5.04
	acc.pl	6.46	6.19	5.13	0	5.20	NA

(Average predictability of word-form in cell Y, given predicting paradigm cell X)

Paradigm entropy

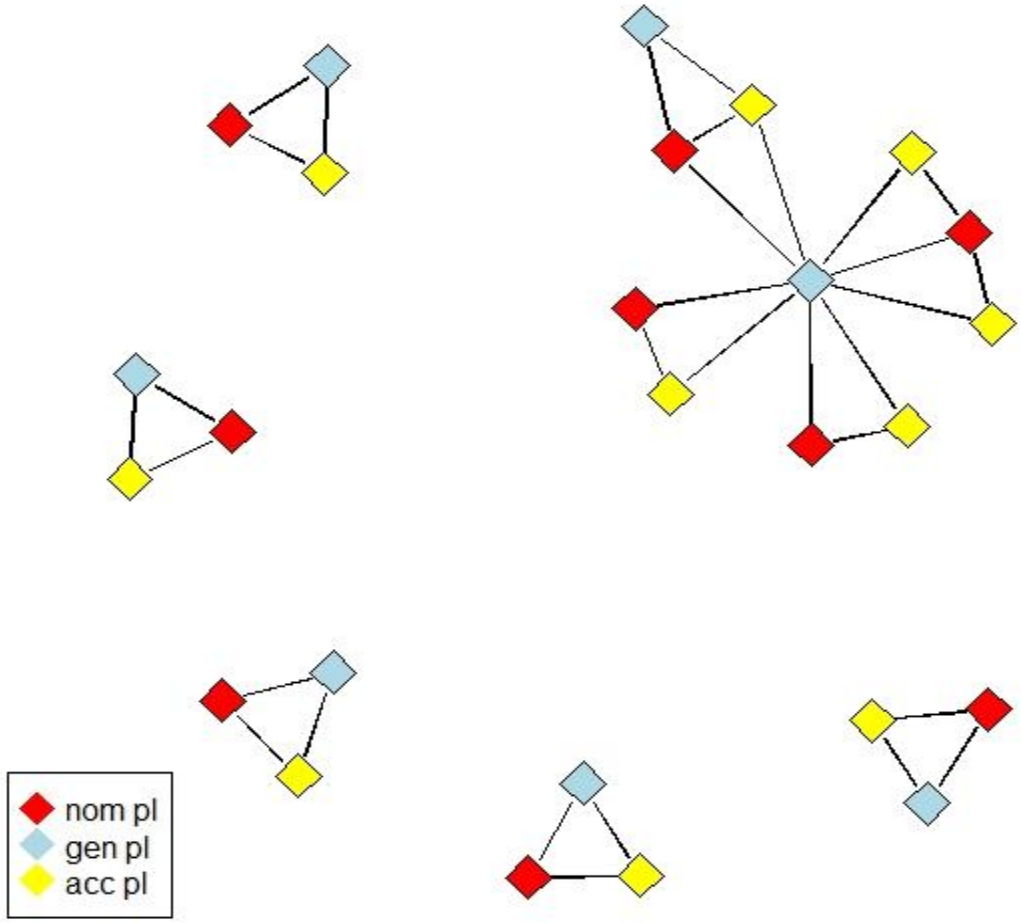
- To extend this to the entire paradigm...

$$E[H(C_1|C_2)] = \sum_{C_1, C_2 \in P} p(C_1, C_2)\, H(C_2|C_1)$$

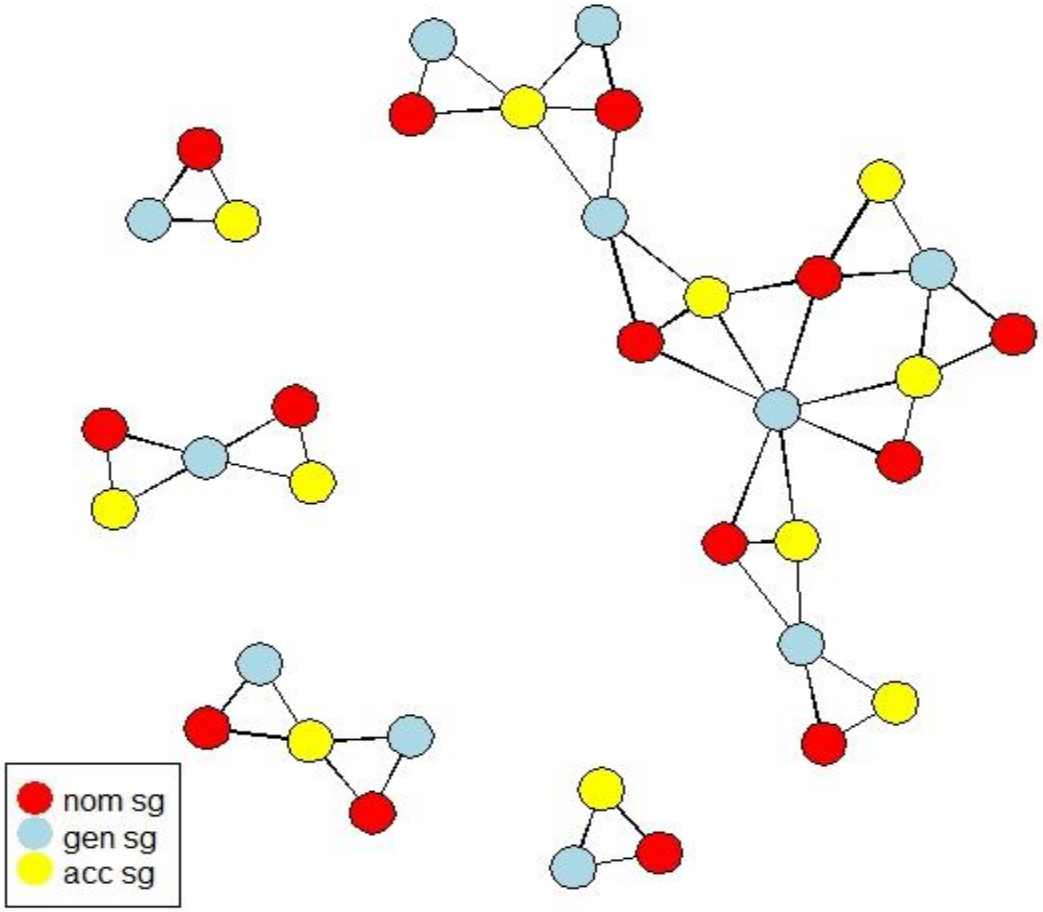
- Ackerman et al. (2009) call this *paradigm entropy.*

paradigm entropy for Modern Greek = 5.59 bits

Greek nominal network: plural forms



Greek nominal network: singular forms



Lack of paradigm cohesion

- The paradigm entropy value of 5.59 bits is quite high, but it is also somewhat misleading.
- Taken by themselves (and collapsing across morphological stress patterns), singular and plural formatives present an obviously much simpler system.
  - The "problem" arises primarily in the many-to-many mappings between singular, plural, and stress
- In other words, the Greek nominal system exhibits surprisingly little paradigm cohesion (And this despite having only six cells, and lots of syncretism!) ...
- ... but this doesn't mean that formatives randomly co-occur
  - three paradigmatic subsystems of implicational relations

Conditional entropy: H(Y|X)

		Predicted					
		nom.sg	gen.sg	acc.sg	nom.pl	gen.pl	acc.pl
Predictor	nom.sg	NA	3.40	0.18	5.77	11.84	6.55
	gen.sg	4.00	NA	2.06	6.83	9.31	6.83
	acc.sg	2.96	4.95	NA	4.74	7.55	5.11
	nom.pl	6.45	7.03	5.29	NA	5.56	0.838
	gen.pl	10.71	8.29	8.44	5.04	NA	5.04
	acc.pl	6.46	6.19	5.13	0	5.20	NA

In the sample of ~27,300 lexemes,
~1,560 are defective in the genitive plural.

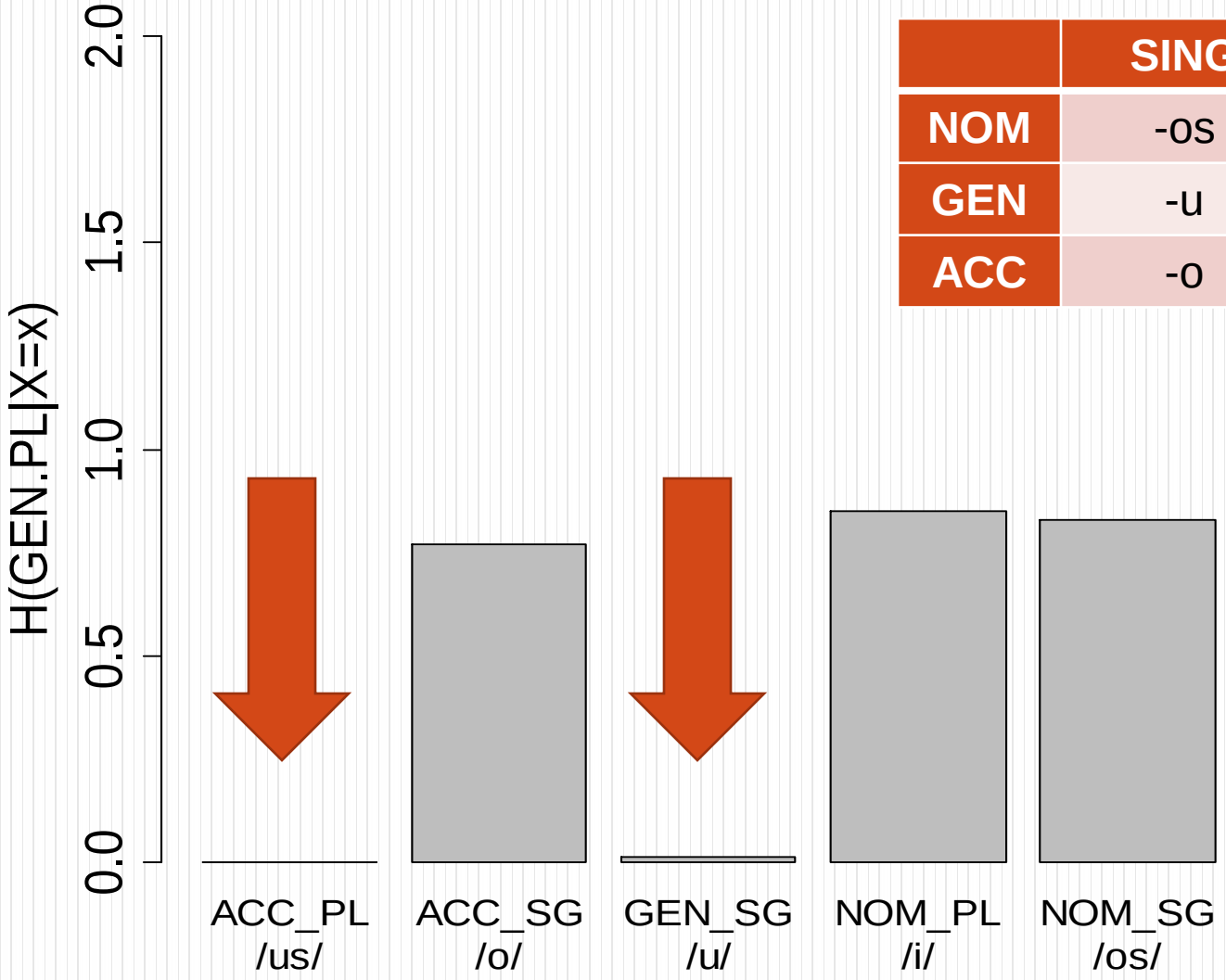
A closer look at individual classes

$$H(Y \mid X=x) = -\sum_{y \in Y} p(y \mid x) \log_2 p(y \mid x)$$

- Same as before, except now we want to know the predictability of MSPS Y *given a particular value x for X*.
  - e.g. How predictable is the genitive plural word-form, if the nominative singular exponent is /-a/?

Class O17-O20, O34-O36

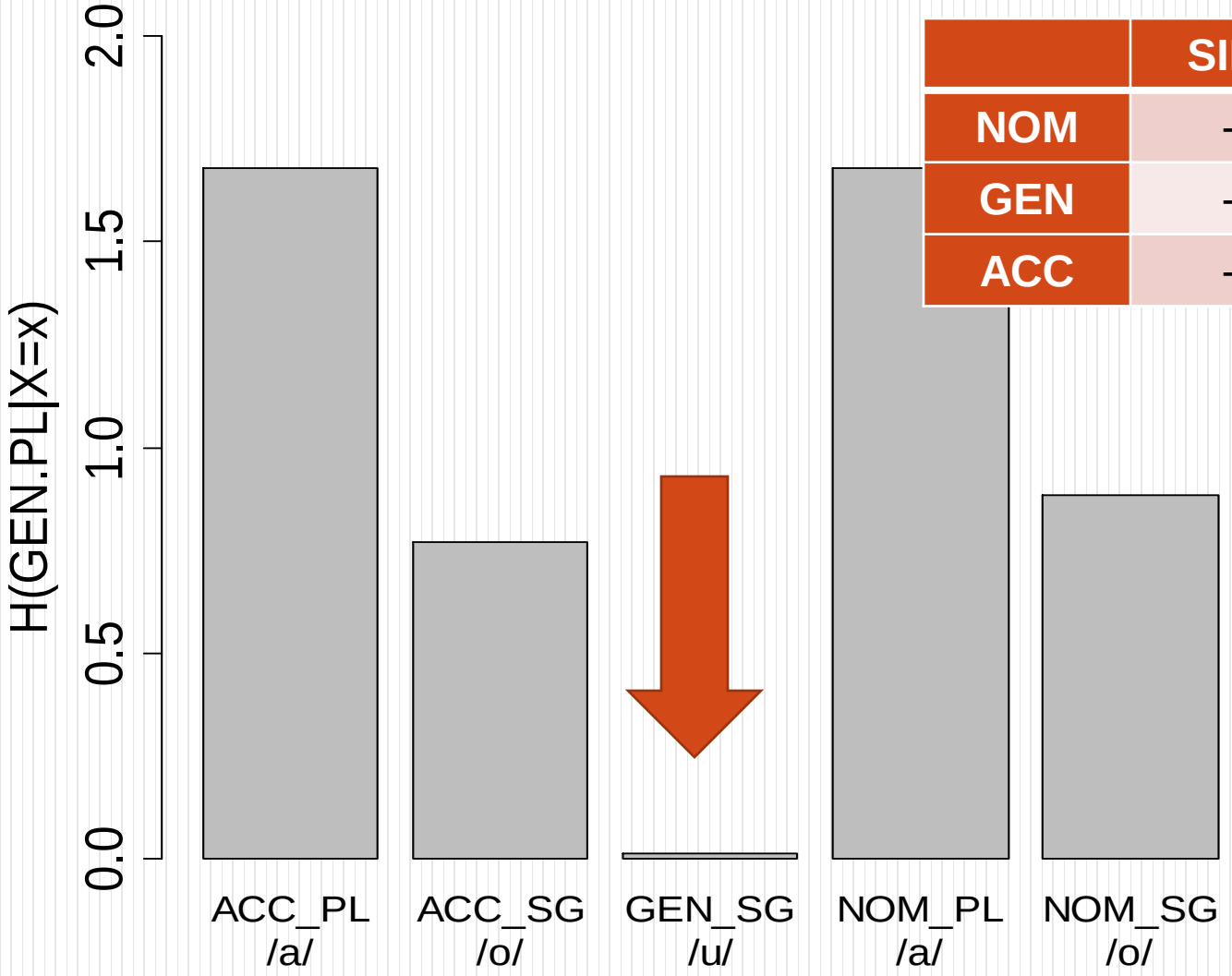
	SING	PLURAL
NOM	-os	-i
GEN	-u	??
ACC	-o	-us



X={ACC_PL,…}, x={/us/,…}

Class O39-O42

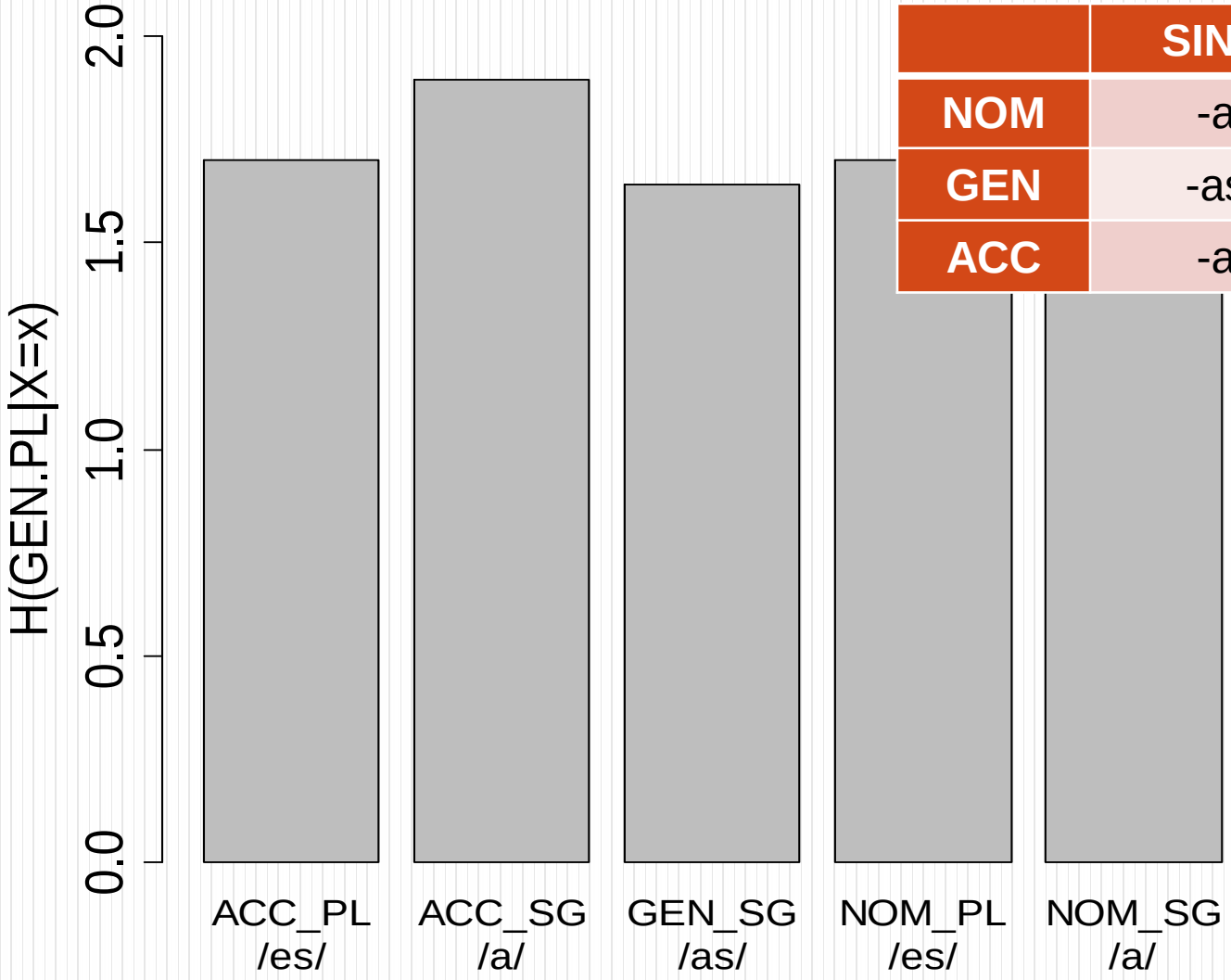
	SING	PLURAL
NOM	-o	-a
GEN	-u	??
ACC	-o	-a



X={ACC_PL,...}, x={/a/,...}

Class O24-O28, O45

	SING	PLURAL
NOM	-a	-es
GEN	-as	??
ACC	-a	-es



X={ACC_PL,...}, x={/es/,...}

Defectiveness in the genitive plural

- 17% of class O24-O28, O45 are defective in the genitive plural (N of gaps = 1380)
  - 89% of all paradigmatic gaps fall into this class, but this class constitutes only 29% of all nouns
- 0% of class O17-O20, O34-O36 are defective in the genitive plural

- In other words, defectiveness occurs exactly where the genitive plural is not well predicted from (nor is predictive of) **any** other cell in the paradigm.
  - Defectiveness as a (historical?) consequence of lack of paradigm cohesion

Conclusions

- What is a paradigm?
  - Defined by set of (morphosyntactic) contrasts, but this does not entail that the cells function as a cohesive unit *at the level of form*.
  - We need theoretical models that allow for paradigmatic relations, but the internal organization (boundaries?) of the paradigm is a separate question
- What is the internal organization of the (Greek) inflectional paradigm?
  - Greek is by one measure a quite simple inflectional system (few paradigm cells). But from the perspective of implicational relations holding between cells, it is anything but simple
    - Conceptually, similar to system-wide heteroclisis
  - External simplicity and internal complexity

Some speculation

- Do languages differ substantially in the internal organization of their paradigms (as measured by paradigm entropy, etc.)?
- Low Entropy Conjecture and Greek
  - To the extent that low paradigm entropy is a response to the Paradigm Cell Filling Problem, and the PCFP is not equally a problem in all languages, it shouldn't be surprising that high entropy is (sometimes) found in languages with few paradigm cells.
    - Each cell has greater potential "independence"
  - At the same time, defectiveness in the genitive plural indicates that Greek is nonetheless sensitive to paradigm cohesion.
    - Implicational relations in the paradigm still matter.
- Paradigm entropy is useful for a macroscopic view, but so much of the interesting stuff is in the details!
- A more systematic exploration must be left for future research.

Thank you!

Thanks to Jeff Parker for his comments on
a preliminary version of this presentation.