

# **Information content and implicational structure in Czech nominal morphology**

James Kirby  
Department of Linguistics  
University of Chicago

12th International Morphology Meeting  
Budapest, Hungary  
27 May 2006

## Goals of the talk

- (1) To suggest that we can do without underlying representations (URs), because adopting them - and the concomitant need for concatenative operations - leads to serious representational problems.
- (2) To argue that the underlying motivation for URs - to minimize redundancy in the lexicon - may be fruitfully captured by other types of representations that avoid the problems mentioned in (1).

# Why URs?

A grammar is an analytic attempt to characterize the independent (non-redundant) information contained in a system.

Both intuitively and technically speaking, unorganized data contains no information. Grammars organize data into information by minimizing redundancy (compression).

Generative linguistics captures the distinction between predictable (redundant) and unpredictable (contrastive) information via the notion of **underlying representation** - a level of grammar at which only contrastive information is represented (Chomsky & Halle 1968). Predictable information is packaged in the form of rules (or constraint rankings, etc).

## Symbol-counting as redundancy minimization

$$\left. \begin{array}{l} \textit{jump} \\ \textit{jumps} \\ \textit{jumped} \\ \textit{jumping} \end{array} \right\} 22 \text{ bits} \qquad \textit{jump} \left\{ \begin{array}{l} \textit{s} \\ \textit{ed} \\ \textit{ing} \end{array} \right\} 10 \text{ bits}$$

The analytic tradition going back to at least Halle (1962) views redundancy minimization as a mandate to reduce symbol count. This has driven the development of theories which strive to minimize the size of the lexicon in terms of the number of symbols it contains - thus, lexica containing fewer symbols are, *ceteris paribus*, preferable to lexica containing many symbols.

## However

This program leads to various problems, not the least of which is highly abstract URs. More seriously, deriving surface forms in e.g. inflectional paradigms via concatenative operations disrupts the natural implicational structure of paradigms (Blevins 2006).

A word-based, surface-oriented approach sidesteps these issues, but (presumably) at the cost of introducing rampant redundancy into the grammar.

### **“Why Such a Theory is Likely to be Frightening to Work On:**

- It will most likely involve some redundancy.
- Hence the intuitive guides of elegance and economy that have hitherto helped us with analysis might not help here.” (Hayes 1998:14)

## Nem!

I argue that Hayes' fears are unfounded, because his notion of redundancy is grounded in tacit acceptance of the **symbol-counting** evaluation metric.

By applying an **independent information** metric (Jackendoff 1975, Bochner 1993) instead of the symbol-counting metric (Chomsky & Halle 1968), word-based schema for encoding representations may be regarded as redundancy-free as concatenative schema, but empirically and representationally preferable.

Case study: Czech nominal inflection.

## Czech nominals

Czech nominals may be broadly grouped into three genders (masculine, feminine, neuter), with an animacy distinction maintained in the masculine.

However, gender is largely unpredictable from formal properties of the stems.

	CLASS 12 'mayor'	CLASS 21 'bridge'	CLASS 28 'post office'	CLASS 39 'bone'
NOM	starosta	most	pošta	kost
GEN	starosti	mostu	pošti	kost'i
DAT	starostovi	mostu	pošt'e	kost'i
ACC	starostu	most	poštu	kost

Table 1: Partial Czech nominal singular paradigms.

## Czech nominals

**Memorization** of form-meaning correspondences fails to organize the data...

$$\left\{ \begin{array}{l} \text{starosta} \\ \text{starosti} \\ \text{starostovi} \\ \text{starostu} \end{array} \right\} \left\{ \begin{array}{l} \text{most} \\ \text{mostu} \\ \text{mostu} \\ \text{most} \end{array} \right\} \left\{ \begin{array}{l} \text{pošta} \\ \text{pošti} \\ \text{pošt'e} \\ \text{poštu} \end{array} \right\} \left\{ \begin{array}{l} \text{kost} \\ \text{kost'i} \\ \text{kost'i} \\ \text{kost} \end{array} \right\} 90 \text{ bits}$$

But **compression** (via symbol minimization) creates information out of noise:

$$\left\{ \begin{array}{l} \text{starost-} \\ \text{most-} \\ \text{pošt-} \\ \text{kost-} \end{array} \right\} \left\{ \begin{array}{ll} -a & -i \\ -ovi & -u \\ -'e & -'i \\ -\emptyset & \end{array} \right\} 28 \text{ bits}$$

## This is starting to get messy...

**However:** note that there is no one-to-one mapping between form and function, e.g. is *-u* CLASS 21 GEN/DAT or CLASS 28 ACC? Avoiding mismatches will require a bookkeeping mechanism, e.g. class diacritics:

$$\left( \begin{array}{l} \text{starost-}_{12} \\ \text{most-}_{21} \\ \text{pošt-}_{28} \\ \text{kost-}_{39} \end{array} \right) \left\{ \begin{array}{ll} -a \rightarrow \text{NOM } \{12,28\} & -i \rightarrow \text{GEN } \{12,28\} \\ -i \rightarrow \text{DAT } 39 & -'e \rightarrow \text{DAT } 28 \\ -ovi \rightarrow \text{DAT } 12 & -u \rightarrow \text{DAT } 21 \\ -u \rightarrow \text{ACC } \{12,28\} & -'i \rightarrow \{\text{GEN, DAT}\} 39 \\ -\emptyset \rightarrow \text{NOM } \{21,39\} & -\emptyset \rightarrow \text{ACC } \{21,39\} \end{array} \right\}$$

...not terribly insightful. Also not immediately clear how to calculate the symbol count...

## Stem class diacritics

As argued in e.g. Bochner (1993), Kirby (2005), Blevins (2006), stem class diacritics are little more than bookkeeping devices - and the number starts to grow very quickly as the complexity of the inflectional system grows - Fronek (1996) analyzes 50+ inflection classes; then consider Hungarian, Estonian...

A major disadvantage to this type of approach is that, through formal separation of stem and affix, all stems now look formally the same - but since formal cues cannot be used to match stems and affixes, we have disrupted the natural implicational structure present in the Czech paradigms:

NOM -a + DAT -ovi → GEN -i + ACC -u

NOM -a → GEN -i + ACC -u

GEN -u → NOM, ACC -∅ + DAT -u

...

## Another problem

Nothing in this representation would signal that some patterns are more common than others, but this is clearly the case empirically: 90% of hard masculine inanimates take *-u* in LOC SG. But 1% take just *-’e* (robustly) and the remaining 9% alternate (Janda & Townsend 2002, ČNK).

## Frequency

	CLASS 17 'rest'	CLASS 21 'castle'	CLASS 21B 'ceiling'
NOM SG	klid	hrad	strop
GEN SG	klidu	hradu	stropu
DAT SG	klidu	hradu	stropu
ACC SG	klid	hrad	strop
LOC SG	<b>o klidu</b>	<b>o hradu/'e</b>	<b>o strop'e</b>
INST SG	klidem	hradem	stropem
DIST	0.9	0.09	0.01

Table 2: Partial CLASS 17, CLASS 21, and CLASS 21B nominal paradigms.

## Frequency

If we represent this as before, we have

$$\left\{ \begin{array}{l} \text{klid-}_{17} \\ \text{hrad-}_{21} \\ \text{strop-}_{21b} \end{array} \right\} \left\{ \begin{array}{l} -\emptyset \rightarrow \text{NOM, ACC } \{17,21,21b\} \\ -u \rightarrow \text{GEN, DAT } \{17,21,21b\} \\ -em \rightarrow \text{INST } \{17,21,21b\} \\ -u \rightarrow \text{LOC } 17 \\ -u/'e \rightarrow \text{LOC } 21 \\ -'e \rightarrow \text{LOC } 21b \end{array} \right\}$$

But nothing about this representation suggests one class is more or less productive than another.

## The word-based alternative

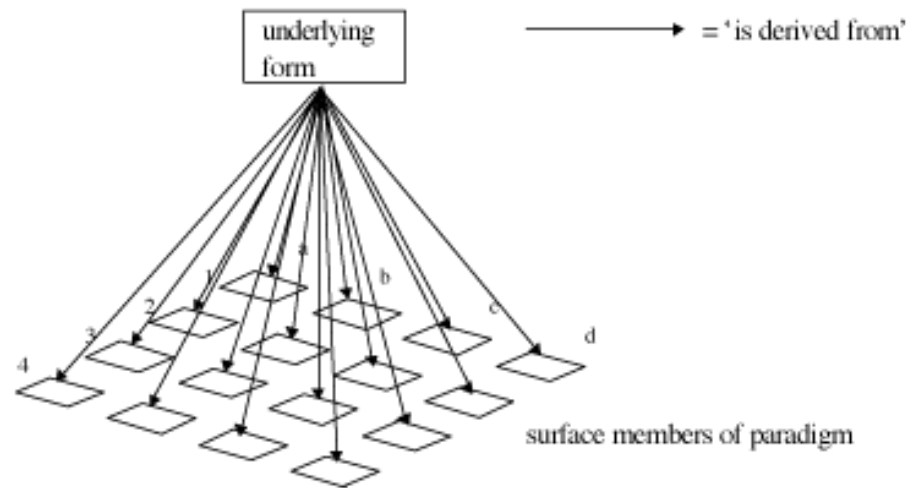
What if the lexicon were conceived of as a repository of full forms (Jackendoff (1975), Daelemans et. al. (2004), Baayen (2003), Bybee (1985, 1991), Skousen (1989, 1992), Hay & Baayen (2005), Blevins (2003, 2006) etc.)?

On such a conception, the previously mentioned problems (explosion of diacritics, absence of frequency information) are not problems at all:

- diacritics are unnecessary (no mismatch problem);
- frequency information exists as distributional lexical statistics.

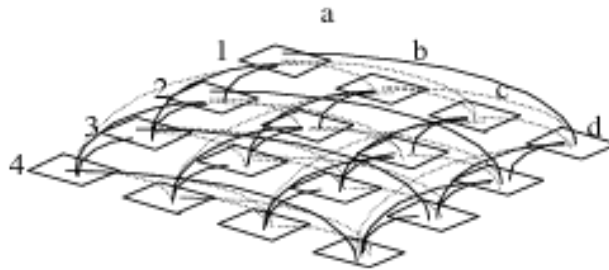
# One-to-many

On this type of approach, instead of relations from a **single underlying form** to **multiple surface forms** via rules...



# Many-to-many

...we have relations from **many** surface forms to **many other** surface forms:



(Images from Hayes 1998)

## Independent information

**But:** doesn't a many-to-many mapping violate the redundancy mandate (cf. Albright (2002))? Only if we're counting **symbols** as a means of measuring the redundancy encoded in the lexicon.

The actual **number** of the pairwise relations isn't what we want to count. There are tons of them, by anyone's model. What we want to measure is the **independent information** the grammar contains...that information which is not in some sense predictable.

Bochner (1993) following Jackendoff (1975) discusses an alternative means of measuring independent information content, which Bochner calls the **independent information** metric

## Redundancy minimization in a word-based model

Let's look at the Czech hard masculine cases again...

	CLASS 17 'rest'	CLASS 21 'castle'	CLASS 21B 'ceiling'
NOM SG	klid	hrad	strop
GEN SG	klidu	hradu	stropu
DAT SG	klidu	hradu	stropu
ACC SG	klid	hrad	strop
LOC SG	<b>o klidu</b>	<b>o hradu/'e</b>	<b>o strop'e</b>
INST SG	klidem	hradem	stropem

Table 3: Partial CLASS 17, CLASS 21, and CLASS 21B nominal paradigms.

## What do we really want to capture?

- That there exist three (six, twenty-four...) patterns;
- That these patterns do not all occur with equal frequency;
- That the cost of representing a given inflectional paradigm should be based on the amount of independent information it contributes - a number which is itself a function of the frequency with which the pattern in question is instantiated.

## Redundancy minimization in a word-based model

Think of the paradigm itself - a many-to-many mapping - as being the learned relation - a *set* (of mappings).

Whatever it 'costs' to learn this set, we only need to pay for it once, in the sense that it should only contribute to the complexity of the grammar once.

This contribution - a measure of the **independent information** contributed by the relation - becomes a function of the entropy of the distribution of inflection classes calculated over the entire lexicon.

## Redundancy minimization in a word-based model

“Lexical knowledge is two-fold, consisting of knowledge of words and knowledge of relations between words” (Koenig 1999)

Thus: the **independent information** contained in all the surface forms which belong to a given inflection class consists of...

(a) the cost of representing the idiosyncratic information of the relation between individual phonological strings and morphosyntactic/semantic information (constructivist parlance: ‘stems’), plus

(b) the cost of representing the idiosyncratic information of the sets of relations between phonological strings and morphosyntactic/semantic information (constructivist parlance: ‘affixes’/‘paradigms’), adjusted relative to the degree these patterns are represented in the lexicon.

## Redundancy minimization in a word-based model

So for the set of e.g. CLASS 17 forms like *klid*, *mír*, etc., the independent information is the sum of the following three terms:

(1) the cost  $n\mathcal{K}$  of the  $n$  various idiosyncratic lexical representations ('words'/'stems') which instantiate the pattern:

$$\begin{bmatrix} \text{klid} \\ \text{N} \\ \text{REST} \end{bmatrix}, \begin{bmatrix} \text{mír} \\ \text{N} \\ \text{PEACE} \end{bmatrix}, \begin{bmatrix} \text{dar} \\ \text{N} \\ \text{GIFT} \end{bmatrix}, \begin{bmatrix} \text{vodovod} \\ \text{N} \\ \text{WATER MAIN} \end{bmatrix} \dots$$

( $\mathcal{K}$  more or less the same for all lexical items, since all forms encode the same type of information...)

## Redundancy minimization in a word-based model

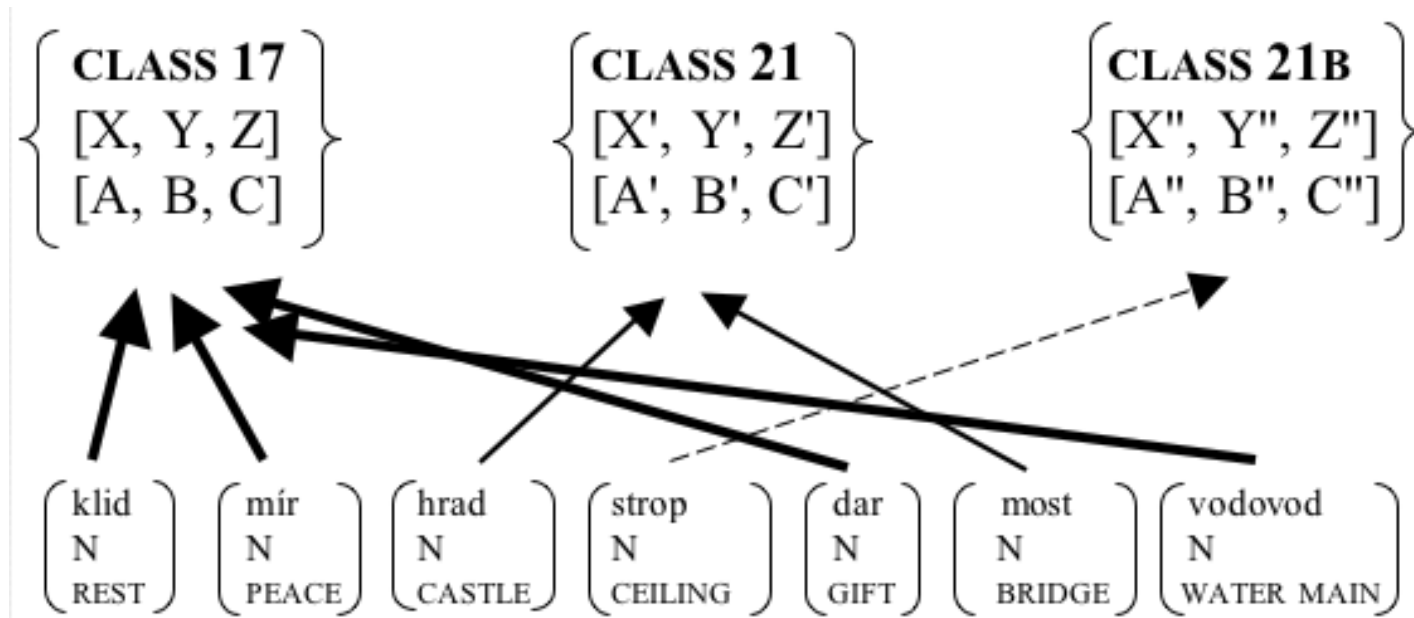
(2) the cost  $\mathcal{R}$  of knowing the CLASS 17 pattern ('relations between words'):

$$\left( \begin{array}{l} \left[ \begin{array}{l} X \\ N, Nsg \\ Z \end{array} \right], \left[ \begin{array}{l} Xu \\ N, Gsg \\ Z \end{array} \right], \left[ \begin{array}{l} Xu \\ N, Dsg \\ Z \end{array} \right], \left[ \begin{array}{l} X \\ N, Asg \\ Z \end{array} \right], \\ \left[ \begin{array}{l} Xe \\ N, Vsg \\ Z \end{array} \right], \left[ \begin{array}{l} Xu \\ N, Lsg \\ Z \end{array} \right], \left[ \begin{array}{l} Xem \\ N, Isg \\ Z \end{array} \right] \end{array} \right)$$

( $\mathcal{R}$  more or less the same for all lexical items, since virtually all forms inflect for all cases in both numbers...)

## Redundancy minimization in a word-based model

(3) The cost  $\mathcal{M}$  of knowing that the strings *klid*, *mír*, etc. and their associated lexical semantics can fill in for the variables X and Z in the above pattern:



## Redundancy minimization in a word-based model

It is this last calculation that is crucial, because we don't want  $\mathcal{M}(\text{CLASS } 17)$  to cost what  $\mathcal{M}(\text{CLASS } 21\text{B})$  costs.

We can measure  $\mathcal{M}$  (in bits) as  $-\log_2 p(x)$ , where  $p(x)$  is the frequency of occurrence of the pattern  $x$  (Bochner 1993).

Crucially, while we might be able to calculate the costs  $\mathcal{R}$  and  $\mathcal{K}$  of a lexical relation or lexical item in isolation, the cost  $\mathcal{M}$  of associating items with patterns can **only** be calculated with respect to the **entire lexicon**, because we need the size of the entire lexicon to calculate productivity of the various patterns...

**Upshot:** the **less regular a pattern, the more independent information the examples of that pattern represent.**

## Redundancy minimization in a word-based model

Distribution of hard masculine inanimate inflection classes:

	CLASS 17	CLASS 21	CLASS 21B
LOG SG	-u	-u/'e	-'e
% lexicon	90%	9%	1%

...and the independent information content in bits:

	CLASS 17	CLASS 21	CLASS 21B
LOG SG	-u	-u/'e	-'e
$\mathcal{M}$	0.02	3.47	6.64

## So what does this buy us...

- obviates need for class diacritics
- obviates need for underlying representations
- representation provides information about relative frequency
- basis for a theory of how speakers assign forms to classes (Kirby 2005)

Admitting multidirectional lexical correspondences into the grammar is really not a problem, because it's not redundant in the 'bad' sense. This is because the crucial quantity affecting the independent information calculation is not the *number of relations in a set* - this is essentially constant for all paradigms - but the *productivity* of these set-objects in the lexicon.

## Conclusion

Czech represents an instance (like Saami, Estonian, Nenets....) where attempting to decompose inflectional paradigms into ‘minimal meaningful parts’ obscures the implicational relations inherent in the system.

Word-based models, or models of morphological representation that contain rich sets of mappings, serve to highlight rather than disrupt these relations.

Furthermore, a word-based model allows for a representation that directly encodes information about pattern frequency - a marked improvement on the compositional representation - while observing the general principle of redundancy minimization.

## Conclusion

Have we allayed Hayes' fears? **Yes.** Why?

**Q: Does a word-based, surface-oriented model increase redundancy?**

A: No: such models of grammar also seek to minimize redundancy, but in a way that takes into account distributional information - and which must in turn be evaluated against the relevant metric.

**Q: Must we abandon or do without the intuitive guides of elegance and economy that have hitherto helped us with analysis?**

A: No: a word-based model is arguably much more elegant and economic than a diacritic-based one. We need not abandon notions of elegance or economy - just rethink how our models should capture them (Goldsmith 2001a,b).

# Acknowledgments

Farrell Ackermann

Eric Bakovic

Jim Blevins

Christian Hilchey

Jason Riggle

Sharon Rose

Andrew Sihler

Alan Yu

This work utilized the *Český Národní Korpus* (Czech National Corpus)

## Selected references

- ALBRIGHT, A. 2002. *The identification of bases in morphological paradigms*. Los Angeles: ULCA dissertation.
- BAAYEN, R.H. 2003. Probabilistic approaches to morphology. In *Probabilistic Linguistics*, ed. by R. Bod, J. Hay, and S. Jannedy. Cambridge: MIT Press.
- BLEVINS, J. P. 2003. Stems and paradigms. *Language* 79.4.
- BLEVINS, J. P. 2006. Word-based morphology. *Journal of Linguistics* 42.3.
- BOCHNER, H. 1993. *Simplicity in generative morphology*. Berlin: Mouton de Gruyter.
- BYBEE, J. 1985. *Morphology: a study of the relation between meaning and form*. Amsterdam: John Benjamins.
- BYBEE, J. 1991. Natural morphology: the organization of paradigms and language acquisition. In *Second language acquisition and linguistic theory*, ed. by C. Ferguson and T. Huebner. Amsterdam: John Benjamins.
- CHOMSKY, N. & M. HALLE. 1968. *The Sound Pattern of English*. New York: Harper and Row.
- DAELEMANS, E, J. ZAVREL, K. VAN DER SLOOT, & A. VAN DEN BOSCH. 2004. TiMBL:

Tilburg Memory-Based Learner - version 5.1 Reference Guide. ILK Technical Report Series 04-02.

FRONEK, J. 1998. *Anglicko-český a česko-anglický slovník*. Praha: Leda.

GOLDSMITH, J, 2001a. "Probability for linguists." University of Chicago ms.

GOLDSMITH, JOHN. 2001b. Unsupervised learning of the morphology of a natural language. *Computational Linguistics* 27.153-198.

HALLE, M. 1962. Phonology in generative grammar. *Word* 18.

HAY, J. & R. HARALD BAAYEN. 2005. Shifting paradigms: gradient structure in morphology. *Trends in Cognitive Sciences* 9(7).

HAYES, B. 1998. "On the Richness of Paradigms, and the Insufficiency of Underlying Representations in Accounting for them." Stanford talk handout.

JACKENDOFF, R. 1975. Morphological and semantic regularities in the lexicon. *Language* 51.639-671.

JANDA, L. & C. TOWNSEND. 2002. *Czech*. SEELRC online grammar.

KIRBY, J. 2005. *Implicational structure in Czech nominal morphology*. UCSD MA thesis.

KOENIG, J-P. 1999. *Lexical Relations*. Stanford: CSLI.

SKOUSEN, R. 1989. *Analogical modeling of language*. Dordrecht: Kluwer.

SKOUSEN, R. 1992. *Analogy and structure*. Dordrecht: Kluwer.