Paradigmatic and distributed realization

Jim Blevins (jpb39)
M11-12/LB9/12.02

Why words?

- Words are the smallest free forms in a language, i.e., the smallest units that exhibit an independent distribution, can occur as utterances, (and FWIW characterize the initial production stage of learners).
- There are well-known discrepancies between what are sometimes termed ‘grammatical’ and ‘phonological’ words (Bloomfield 1914).
- But these discrepancies cannot be overcome by considering smaller units: there is no discrepancy between the ‘grammatical’ and ‘phonological’ morpheme because there is no notion of a ‘phonological’ morpheme.
- The word-level discrepancies might also be interpreted as evidence that the diagnostic properties adopted in linguistic descriptions are in fact just secondary cues that reinforce a fundamentally statistical structure.

The cost of overabstraction

- Classical WP approaches present philological case studies that show that sub-word units exhibit less stability in form and meaning than the words from which they are abstracted:

> *The word is a more stable and solid focus of grammatical relations than the component morpheme by itself. Put another way,* ***grammatical statements are abstractions, but they are more profitably abstracted from words as wholes than from individual morphemes****. (Robins 1959: 128)*

- The same contrasts characterize implicational relations. In Estonian declensions, words, associated with a cell, are more informative than their parts in isolation.
- Entropy relations defined over words confirm the high degree of mutual information exhibited by inflectional systems (but there are no comparative studies that probe the informativeness of sub-word units).

Anti-Uniformitarianism

In many ways, and quite apart from any phonological markers, the word is a unique entity in grammar, and not just a stage in the progression 'from morpheme to utterance'. ***As a grammatical element the word is unique in its relative fixity of internal morphemic structure, its focal status in relation to syntactically relevant categories, and, in inflected words, the stability of its paradigms.*** *All of these factors make it a strong basis for grammatical description, both morphological and syntactic. The assumption of a simple ascent in order of size from single morpheme to complete sentence, ignoring or blurring the distinction of morphological structuring and syntactic structuring, achieves its apparent simplicity at the cost of neglecting or distorting patent structural features of languages. (Robins 1959: 137)*

The utility and "psychological reality" of words

For the purposes of ordinary life, the word is the smallest unit of speech. Our dictionaries list the words of a language; for all purposes except the systematic study of language, this procedure is doubtless more useful than would be a list of morphemes. (Bloomfield 1933: 178)

it is not clear that, when native speakers learn a flexional language, they do not themselves learn words as wholes (Matthews 1991: 188)

But is not the word, one may object, as much of an abstraction as the radical element? Is it not as arbitrarily lifted out of the living sentence as the minimum conceptual element out of the word? Linguistic experience, both as expressed in standardized, written form and as tested in daily usage, indicates overwhelmingly that ***there is not, as a rule, the slightest difficulty in bringing the word to consciousness as a psychological reality****. (Sapir 1921: 32f.)*

Naive measures of unit informativeness

- Can these types of traditional claims and analyses be validated by robust, general-purpose methodologies for probing the informativeness of linguistic units and boundaries?
- Specifically, is there a 'naive' method for probing the optimal unit size for a description of the regularities in a language?
- One strategy might measure entropy 'peaks' over an utterance; another could operationalize Kolmogorov complexity in terms of compressibility, using standard compression utilities (Juola 1998; Brown & Evans 2010; Moscoso del Prado Martín 2011).

An entropy-based informativeness measure?

- Despite evidence that words occur between uncertainty 'peaks' in an utterance (Pléh & Juhász 1995), challenges arise in operationalizing Shannon entropy.
- Standard entropy measures presuppose that objects to be encoded are outcomes of a known random source, and only the characteristics of that random source (its probability distribution) determine the encoding.
- Estimating the probability distribution of a random source of unit boundaries quickly leads to a number of practical challenges. One set of issues arises in evaluating the different language models (i.e., different probabilistic descriptions) of the language under consideration. These models can be based on character or word sequences but could involve any kind of grammar.
- Assessing the different formalisms that could provide the basis for these descriptions raises further complications.

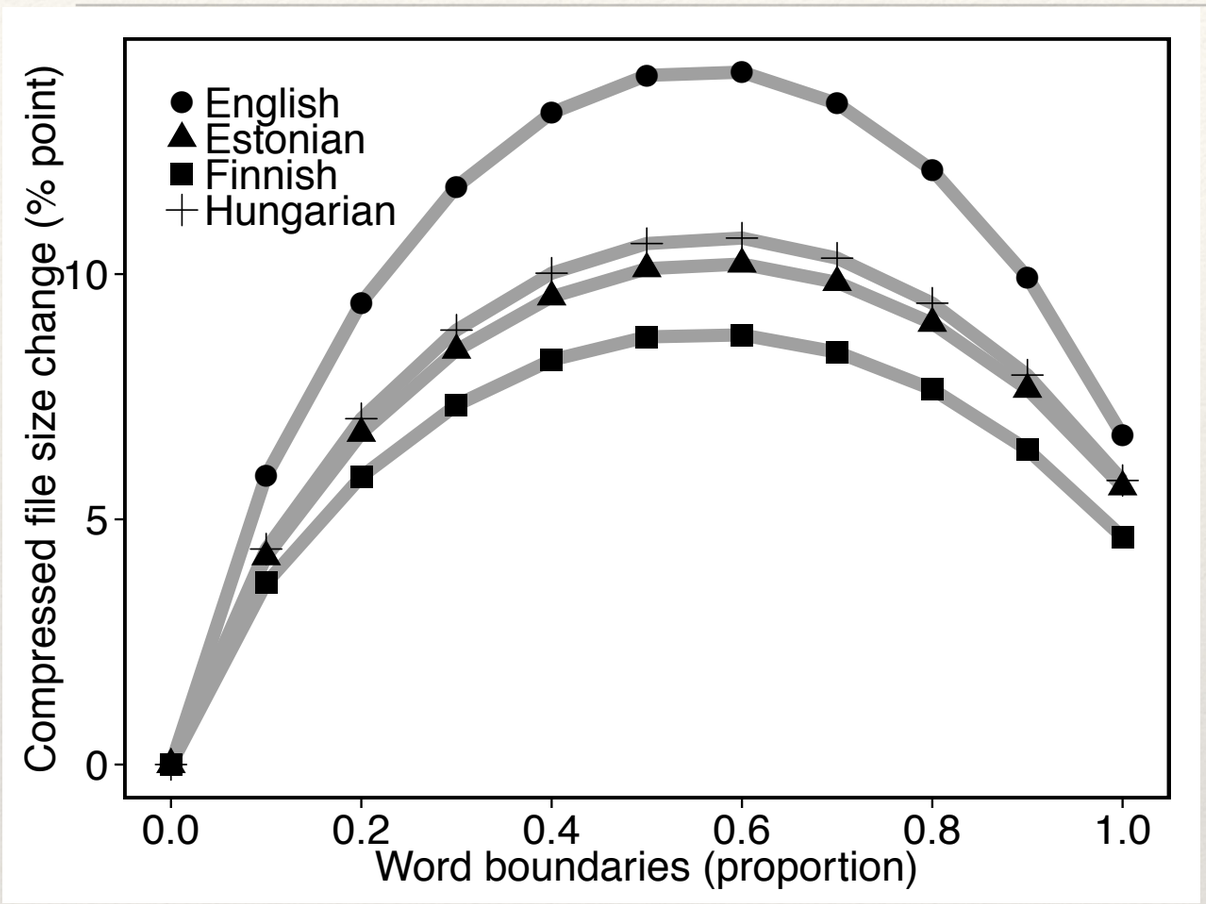
A compression-based informativeness measure

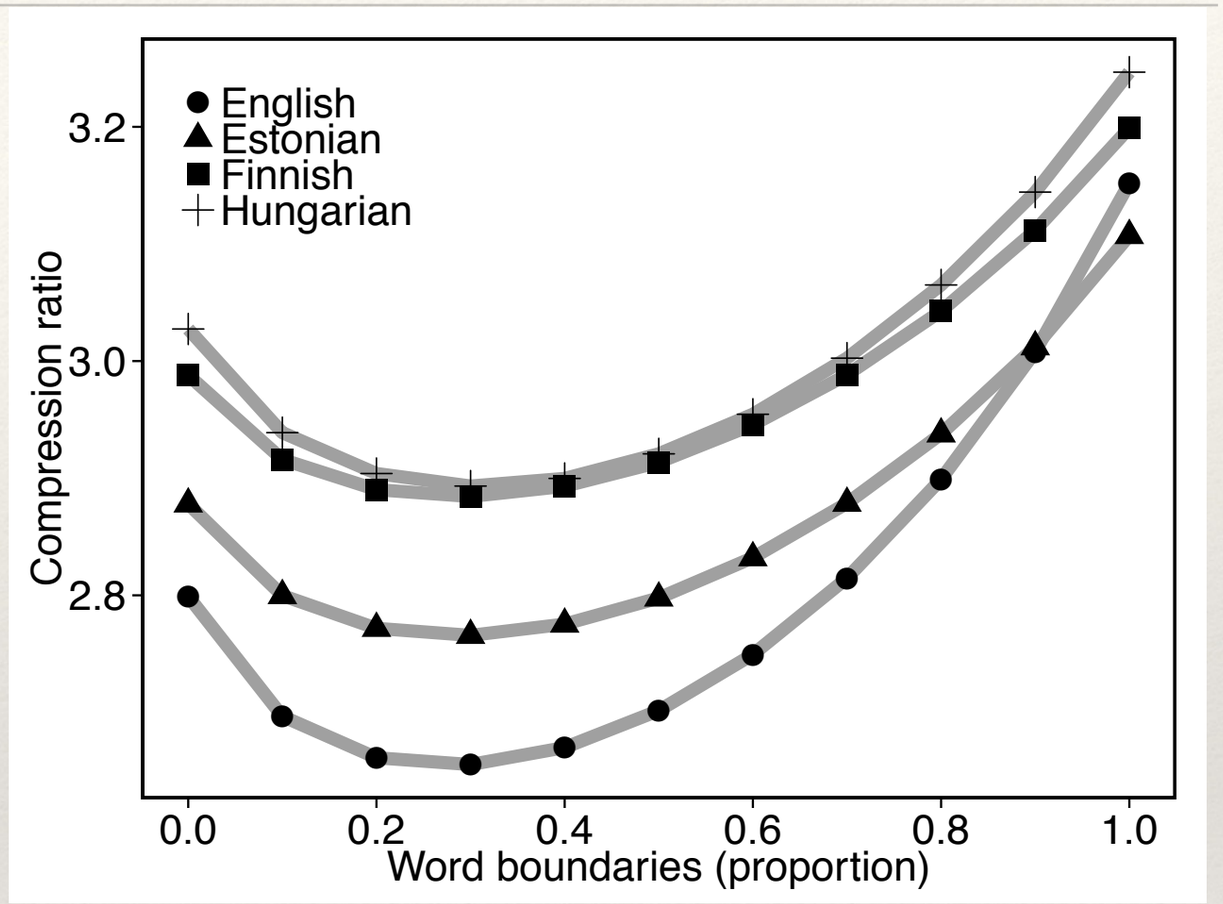
- A practical advantage of Kolmogorov complexity is that it does not depend on probabilistic assumptions about a source, and can assess the information content of an object in isolation.
- Standard formulations of Kolmogorov complexity in terms of minimal description length (MDL; Rissanen 1978) require the choice of a description language. One approach specifies a fixed language (Sagot & Walther 2013).
- A further advantage of an MDL operalization is that it can also use general-purpose (Lempel-Ziv) compression algorithms to approximate the redundancy-free ideal formalized by Kolmogorov complexity.
- It is especially useful to be able to apply a 'naive' measure directly to unannotated corpora to explore the information expressed by boundaries.

Informativeness of unit boundaries

- The status of morphological unit types can be evaluated in terms of their informativity, measured by assessing the influence that the corresponding unit boundary types have on text compressibility.
- Geertzen, Blevins & Milin (2016) examined the informativeness of sentence, word and morph boundaries in parallel English, Estonian, Finnish and Hungarian components of the Europarl corpus.

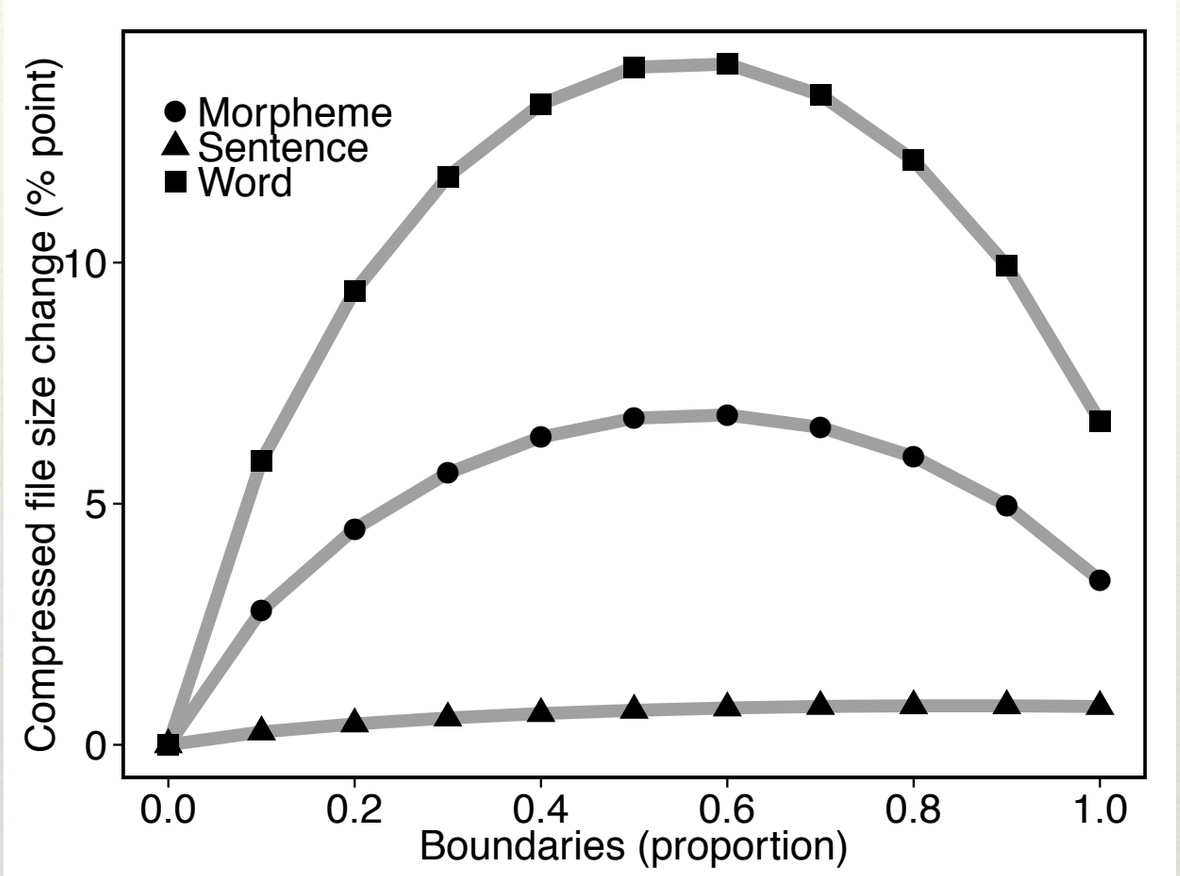
Effect of word boundaries on compression

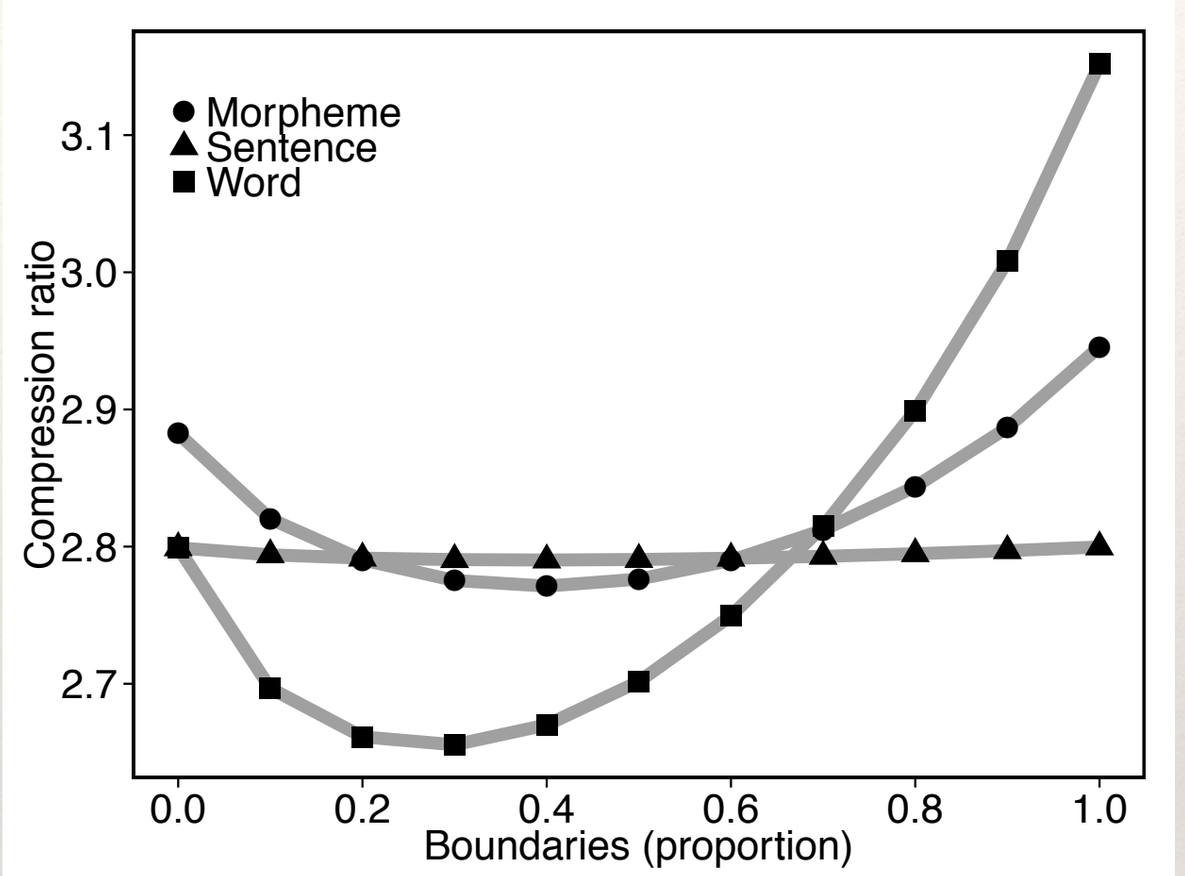




- Finnish shows the smallest increase in compression file size and English shows the greatest increase, with Estonian and Hungarian in the middle. Interestingly, the growth curve for English shows a steeper increase than those of the other languages, implying that word boundaries carry more information for English than for the other languages in the sample. The higher information load carried by word boundaries in English accords with the observation that the morphological structure of English is not only 'poorer' but also less transparently segmentable than that of the other languages.

Informativeness of boundaries in English





- It is possible to compare the informativeness of different unit types in English, given gold-standard-level morphological segmentations (which were not available for the other languages).
- In the first figure, we see that word boundaries make the greatest contribution to reduction in compressed file size, followed by morph and sentence boundaries.
- This pattern is mirrored by the second graph in which again shows that the addition of word boundaries achieves the highest compression ratio, followed by morph and sentence boundaries.

The statistical basis of morphological structure

- The results of this study support two general conclusions:

1. Word boundaries are the most informative boundary type (in English), and the demarcation of words provides the most informative description of the regular patterns of form variation in a language.
2. Unit boundaries primarily reflect a statistical structure, and the typological variability of 'diagnostic' linguistic cues reflects the fact that these cues serve a secondary reinforcing function.

Observations about predictability at word boundaries are consistent with two different kinds of assumptions about what constitutes a word: either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units (but to a lesser degree than the beginning of a word predicts its end).
(Goldwater et al. 2009: 22)

Some open questions

- How sparsely attested are the inflectional and derived forms of languages of different types in corpora?
- How are forms distributed in dense acquisition databases of morphologically complex languages?
- What is the performance of different learning models on naturalistic data, and how well do they model the course of acquisition by human learners?
- Why are words the most stable morphological units?

Why use information theory?

- Classical WP models operate with notions of 'variation' and 'structure' that cannot be reduced to static inventories of 'items' and 'combinatoric principles'.
- Information theory provides a way of modelling 'variation' in terms of frequency-weighted measures of uncertainty (surprisal/entropy/etc.).
- Conditional entropy also provides a probabilistic measure of 'informativeness' that is more flexible than logical implication, while providing a realistic measure of the difficulty of using a system on the basis of partial knowledge.
- Entropy-based measures distinguish patterns in finite systems (unlike metrics based on the Chomsky hierarchy) and correlate closely with response latencies (since they were initially applied in models of morphological processing).

What is the role of information theory?

- Information theory measures are a useful tool for identifying structure, not a substantive component of implicational models:
  - Shannon entropy (Shannon 1948) is a special case of a more general family of entropy notions (Rényi 1961), and measures like conditional entropy belong to a family of related notions, which includes normalized mutual information, information gain (from machine learning), etc.
    - Frequency is confounded with other factors, including a variety of measures involving contextual distribution and crowd knowledge.
- Any measures that can provide a description of dynamic, system-level properties (e.g., set-theoretic measures (Stump & Finkel 2013)) could do just as well, since global entropy estimations serve a mostly diagnostic purpose.

The limits of biuniqueness

One motive for the post-Bloomfieldian model consisted, that is to say, in a genuinely factual assertion about language: namely, that there is some sort of matching between minimal 'sames of 'form' (morphs) and 'meaning' (morphemes). *Qua* factual assertion this has subsequently proved false: for certain languages, such as Latin, the correspondence which was envisaged apparently does not exist … One is bound to suspect, in the light of such a conclusion, that the model is in some sense wrong. (Matthews 1972:124)

Realizational Perspectives

- Recall the feature-form patterns exhibited by the ancient Greek verb form *elelýkete* 'you have unfastened' (Matthews 1991: 173)

Features	Forms					
	e	le	lý	k	e	te
PAST	e				e	
PERFECTIVE		le	lý	k		
2PL						te
ACTIVE				k	e	te

- A model can either (i) minimize the ambiguity of an analysis or (ii) minimize the size of units, but it cannot do both at once.

Separationist morphology (Beard 1995)

- For any given word class or inflection class, it is possible to identify the morphosyntactic features that are distinctive for that class.
- For each feature, it is possible to identify admissible values.
- Multiplying out all of the possible feature-value pairs determines an abstract paradigm structure, defined independently of any item.

An interpretive view of variation I

In short, the theory of syntactic features developed earlier can incorporate the traditional paradigmatic treatment directly. The system of paradigms is simply described as a system of features, one (or perhaps some hierarchical configuration) corresponding to each of the dimensions that define the system of paradigms. Interpretive phonological rules, some quite specific, some of considerable generality, then operate on the phonological matrix of the lexical entry, giving, finally, a phonetic matrix. (Chomsky 1965: 172)

An interpretive view of variation II

The present paper is intended to supply a part of this formulation. It is restricted to inflectional problems alone: to be more precise, it deals with that subsection of the grammar (we will call it the *inflectional component*) which assigns a realization, or various alternative realizations, to each grammatical word. (Matthews 1965: 142)

Distinctive nominal features in English

- In English, the distinctive inflectional features for nouns are just:
  - PERSON, with the values{1st, 2nd, 3rd},
  - NUMBER, with the values {SING, PLU}
- All nouns are 3rd person and there is no lexical case marking.
- Hence the general 2-cell noun paradigm structure in English is defined entirely by the two contrastive number values:
  - [NUM SING]
  - [NUM PLU]

Stem entries in English

- The 'inherent' or 'intrinsic' features that are associated with all forms of an item are associated with the lexical entry of the item.
- In addition, entries specify the basic root or stem of an item:

  [PER 3RD, /baik/], [PER 3RD, /naif/],[PER 3RD, /ʃip/],
  [PER 3RD, /mæn/], …

Paradigm structure in English

- The paradigm of a particular item are then defined by combining the distinctive features from the structure with the 'inherent' features contained in the lexical entry of the item.
- Hence the example nouns have the following 2-cell paradigms:

  [PER 3RD, NUM SING, /baik/], [PER 3RD, NUM SING, /naif/],[PER 3RD, NUM SING, /ʃip/], [PER 3RD, NUM SING, /mæn/],

  [PER 3RD, NUM PLU, /baik/], [PER 3RD, NUM PLU, /naif/],[PER 3RD, NUM PLU, /ʃip/], [PER 3RD, NUM PLU, /mæn/], …

Simple rules of exponence

- The surface form associated with a paradigm cell is then defined by **realisation** or spell-out rules that interpret features in the cell.
- Realization rules only apply if the required features are present.
- If no rules apply, the cell is realized by the basic stem of the entry.
- Regular plural exponence rule in English (Matthews 1991):
  - [NUM PLU, X] → X + /z/