Zipfian distributions and decompositional analysis

Jim Blevins (jpb39)
M11-12/LB9/05.02

The sparsity of the input

- A more fundamental challenge derives from the fact that speakers will **never** encounter all of the forms of even fairly frequent items.
- Instead, their 'primary linguistic input' will be sparse and biased, consisting largely of a small number of extremely frequent forms.
- A majority of open-class items will be represented by a small subset of their inflected forms, and most items by one or two forms:
  - In large corpora, upwards of 50% of the words are hapax legomena (words that occur once in the corpus) and another roughly 10% are dis legomena (words that occur twice).

Zipf's Law

The frequency of a word in a corpus is (close to) inversely proportional to its rank:

$$P_n \propto \frac{1}{n^a}$$

- where P_n is the frequency of the nth-ranked word (and a can be taken to be close to 1).
- I.e., the second most frequent word (rank 2) occurs roughly half as often as the most frequent word, the third most frequent word (rank 3) roughly a third as often, etc.

George K. Zipf
(*The Psycho-biology of Language* 1932; *The Principle of Last Effort* 1948)

Winner-take-all word distributions

- Word distributions do not in fact exactly conform to Zipf's law.
- The deviations are greatest for extremely high-frequency and low-frequency items (Zipf 1949).
- Various refinements have been suggested (e.g., Mandelbrot 1965) and other power laws proposed.
- Although these laws also have limitations (Baayen 2000), they all concern the **measurement**, not the **existence**, of distributional biases.
- There is also a range of proposals to explain **why** Zipfian biases arise.

WORD	COUNT	PERCENT	bar graph
the	53.10 B	7.14%	the
of	30.97 B	4.16%	of
and	22.63 B	3.04%	and
to	19.35 B	2.60%	to
in	16.89 B	2.27%	in
a	15.31 B	2.06%	a
is	8.38 B	1.13%	is
that	8.00 B	1.08%	that
for	6.55 B	0.88%	for
it	5.74 B	0.77%	it
as	5.70 B	0.77%	as
was	5.50 B	0.74%	was
with	5.18 B	0.70%	with
be	4.82 B	0.65%	be
by	4.70 B	0.63%	by
on	4.59 B	0.62%	on
not	4.52 B	0.61%	not
he	4.11 B	0.55%	he
i	3.88 B	0.52%	i
this	3.83 B	0.51%	this
are	3.70 B	0.50%	are
or	3.67 B	0.49%	or
his	3.61 B	0.49%	his
from	3.47 B	0.47%	from
at	3.41 B	0.46%	at
which	3.14 B	0.42%	which
but	2.79 B	0.38%	but
have	2.78 B	0.37%	have
an	2.73 B	0.37%	an
had	2.62 B	0.35%	had
they	2.46 B	0.33%	they
you	2.34 B	0.31%	you
were	2.27 B	0.31%	were
their	2.15 B	0.29%	their
one	2.15 B	0.29%	one
all	2.06 B	0.28%	all
we	2.06 B	0.28%	we
can	1.67 B	0.22%	can
her	1.63 B	0.22%	her
has	1.63 B	0.22%	has
there	1.62 B	0.22%	there
been	1.62 B	0.22%	been
if	1.56 B	0.21%	if
more	1.55 B	0.21%	more
when	1.52 B	0.20%	when
will	1.49 B	0.20%	will
would	1.47 B	0.20%	would
who	1.46 B	0.20%	who
so	1.45 B	0.19%	so
no	1.40 B	0.19%	no

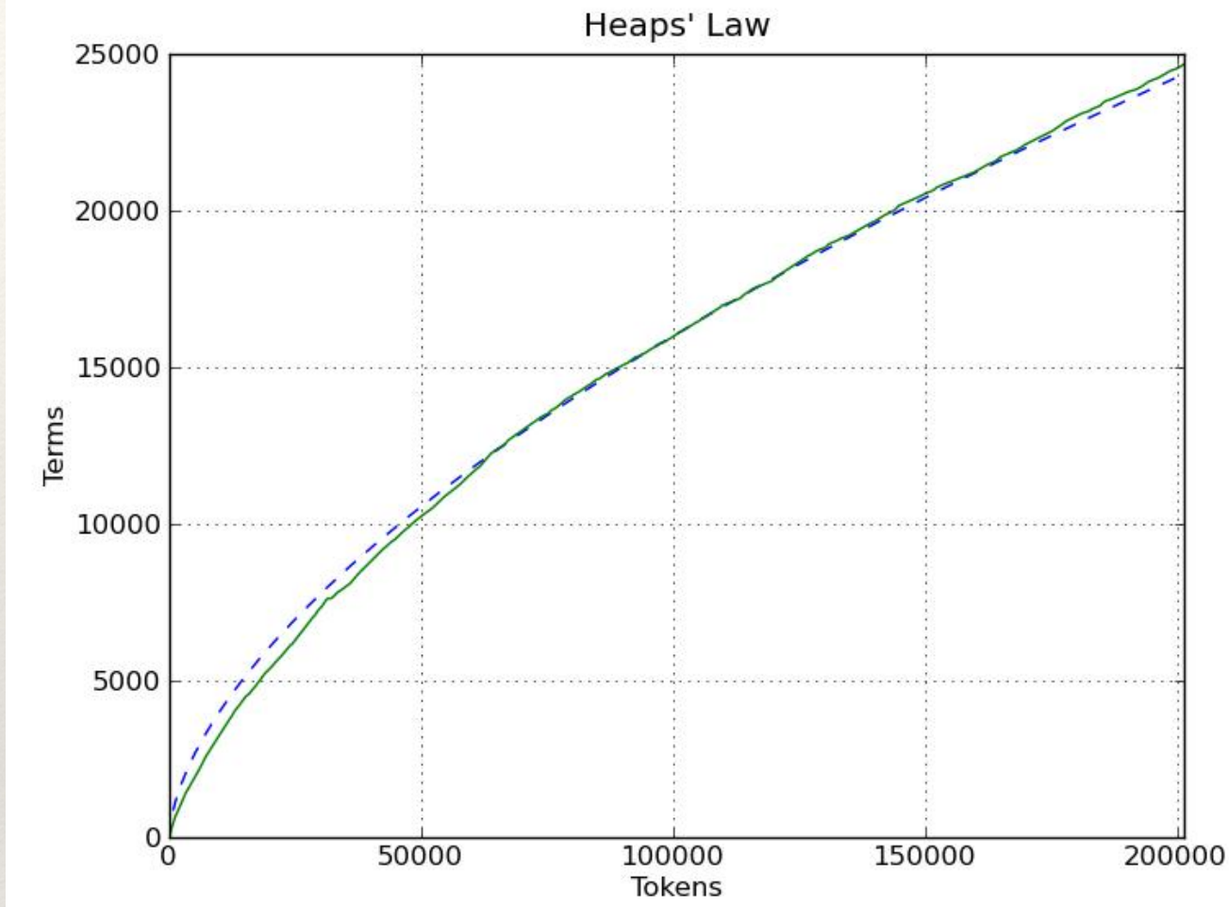
Rank and frequency of 50 most common English words in 23GB Google Books Ngrams (http://norvig.com/mayzner.html)

Herdan's/Heap's Law

The number of distinct word types is proportional to the size of a corpus:

$$V_R = Kn^{\beta}$$

- V_R is the number of word types in a corpus of size n (K and β are free parameters set empirically).
- I.e., as a corpus increases in size, there is a sharply diminishing return in coverage of the lexicon.



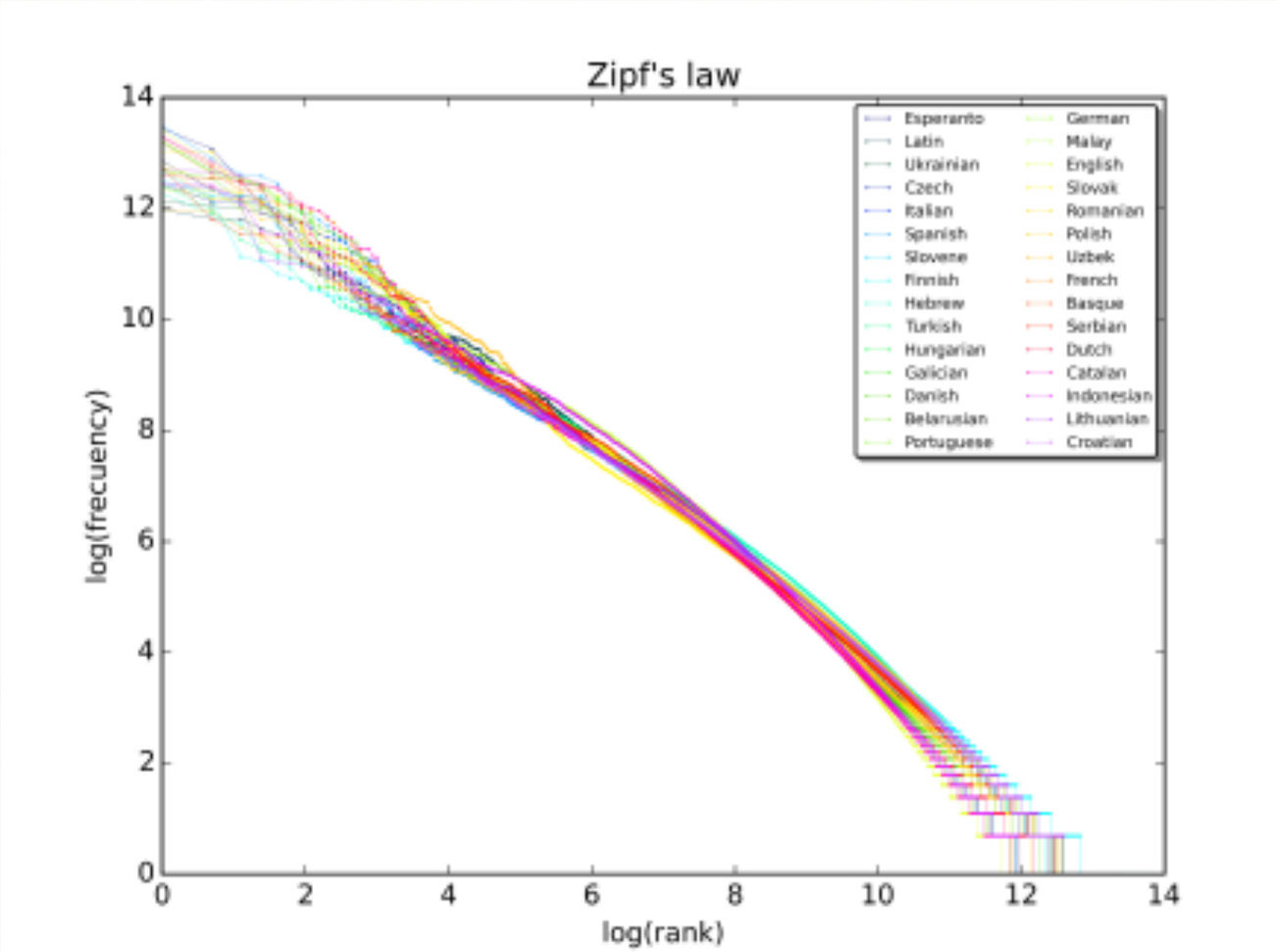
(inf.ed.ac.uk)

The law is noted independently by Herdan (1960) and Heaps (1978).

The ubiquity of Zipfian distributions

- Whatever the source of the biases described by Zipf's law, they are remarkably robust, and relevant to acquisition.

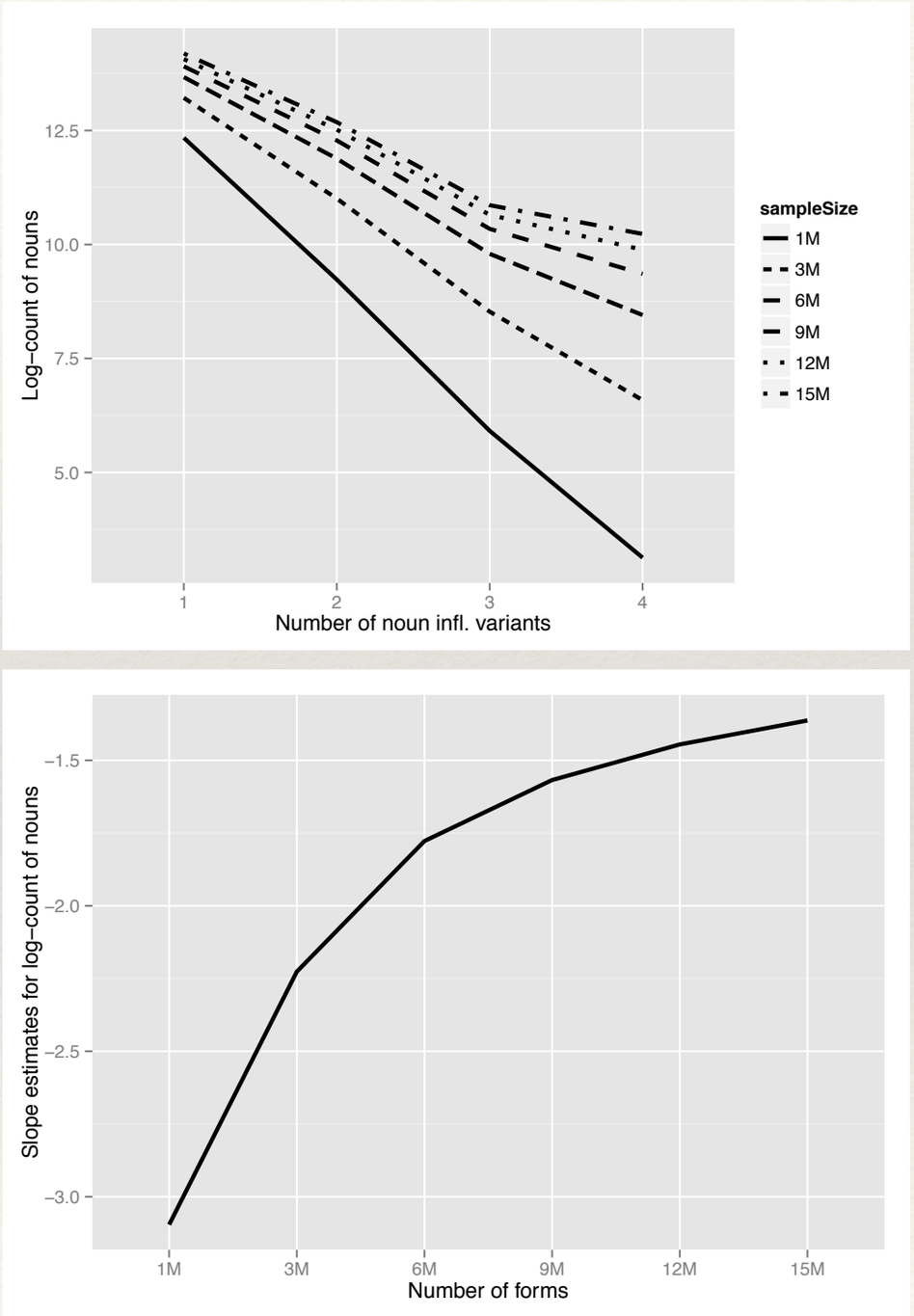
"while Zipfian distributions are ubiquitous across natural language, their consequences for learning are only beginning to be explored" (Kurumada et al. 2013: 440)



Zipf plot of the first 10M words of Wikipedias in 30 languages (wikiwand.com, 2015)

The Zipfian Paradigm Cell Filling Problem (ZPCFP)

- Although studies are preliminary, there is evidence that corpora obey Zipf's law at all sample sizes (Blevins et al. 2016).
- Hence the inflected forms that speakers encounter vastly underdetermine the systems that they come to acquire.
- Given the variation in morphological systems, speakers must acquire full systems **from structure in the input**.



You can't get there from here: Inflectional sparsity in the German SdeWaC corpus

Two solutions to the ZPCFP

- **Atomistic:** On the basis of the forms that they encounter, speakers are able to (i) identify the shape and function of **recurrent parts**, and (ii) formulate general **combinatoric principles** that allow them to reconstitute encountered forms and deduce unencountered forms.
- **Implicational:** The forms that speakers encounter are parts of networks of elements, related by patterns of mutual implication. The 'meaning' conveyed by a form includes **intramorphological** information about shape, function and distribution of related forms.

The generality of the ZPCFP challenge

- Both of these alternatives face non-trivial challenges: Zipfian biases do not 'fall out' in **any** linguistic model of 'lexical knowledge'.
- We first review the challenges that arise for an atomistic solution.
- Sparse and biased data presents a separate set of challenges for an implicational solution as noted clearly by Hockett:

> *in his analogizing … [t]he native user of the language … operates in terms of all sorts of internally stored paradigms, many of them doubtless only partial; and he may first encounter a new basic verb in any of its inflected forms. (Hockett 1967: 221)*

Ecological validity of morphological theory

- If we want a theory or model of human morphological 'knowledge', proposals must be anchored to challenges faced by learners/users.
- 'Deep learning' models are capable of learning a morphological system, but they do not (yet) provide a cognitively-viable model of human capabilities, given (i) the implausible amount of training data they require, (ii) difficulties that arise in interpreting their outputs, (iii) uncertainty about the neurological relevance of their architecture, and (iv) various discrepancies reflecting the fact that they have not been developed to mimic the stages of learning, error patterns, or any cognitively-relevant dimension of learning.
- The treatment of a morphological system as an 'abstract object' of some kind is equally implausible, and largely reflects the descriptive goal of constructing a pedagogical grammar or corpus.

Decompositional analysis

- Within the Post-Bloomfieldian tradition, systems are disassembled into static inventories, and reassembled by combinatorial rules.
- Within this tradition, a system is fully described by inventories of elements and rules that define a set of 'well-formed' expressions.

every language has its own grammar. The grammar, or grammatical system, of a language is (1) the morphemes used in the language, and (2) the arrangements in which these morphemes occur relative to each other in utterances. (Hockett 1958: 129)

Agglutinative structure of Finnish nouns

- Finnish nominals exhibit a generally suffixal structure in which the noun stem is followed by a number marker, then a case marker, then a possessive marker and a final (optional) discourse particle:

	Sing	Plu
Adessive	talolla	taloilla
Ablative	talolta	taloilta
Allative	talolle	taloila

- There are deviations from agglutination, (as discussed in L5) but the structure in standard descriptions provides a reasonable approximation of the agglutiative ideal.

Minimum meaningful parts of Finnish nouns

- Each individual part of a Finnish nominal can be assigned a function and a fixed position in the syntagmatic expansion:

Feature	λ	**Sing**	**Plu**	**Ades**	**Abla**	**Alla**
Form	talo	Ø	-i	-lla	-lta	-lle

Preconditions of an atomistic solution

- It must be possible to disassemble forms into inventories of recurrent atomic elements **with no loss of information**.
- Atomic elements must be analyzable **in isolation**.
- It must be possible to provide a full description of a morphological 'system' in terms of **static inventories** of elements, and principles that govern their arrangement.
- Disassembled parts must be **genuinely independent**, and not just provide distributed representations of larger units.

Challenges for an atomistic solution

Atomistic approaches create a number of general challenges:

1. The **Segmentation** Challenge: It is often difficult to identify a principled basis for segmenting forms into smaller units.

2. The **Interpretation** Challenge: Where it is possible to identify morphotactic units, it may be difficult to identify a principled basis for assigning them context-independent meanings.

3. The **Residue** Challenge: In cases where it is possible to identify and interpret morphotactic units, this structure may still represent a historical residue that merely provides a context in which speakers learn to discriminate a larger form.

The Segmentation Problem

In a fusional language, if one seeks to arrive at constant segments … conflicts arise in the placing of the cuts. ***One comparison of forms suggests one placement, while another comparison suggests another. Often, in fact, no constant segment can be isolated at all which corresponds to a given constant meaning.*** *Situations of this kind often permit of more than one solution according to different manners of selecting and grouping environments. (Lounsbury 1953: 172)*

… ***there is as yet no consensus on how to segment complex words****, even in extremely well-studied languages, and even less consensus on how to determine which word pieces function as the (principal) exponents of any given set of inflectional properties. (Spencer 2012: 104)*

Fusional preterites in Spanish

The order of morphemes is fixed: (derivational prefix(es)) + lexical stem + theme vowel + tense marker (sometimes including an empty morph) + person marker. Some forms, however, have fused in the course of history and a neat segmentation is not always possible. The preterit is the most difficult paradigm to analyse, since ***the theme vowel is sometimes indistinguishable****, and* ***segmenting the second and third person plural markers in the regular way****, /- is, -n/,* ***leaves an awkward residue*** *that occurs nowhere else in the system. (Green 1997: 99)*

Challenges for an atomistic solution

Atomistic approaches create a number of general challenges:

1. **The Segmentation Challenge: It is often difficult to identify a principled basis for segmenting forms into smaller units.**
2. The **Interpretation** Challenge: Where it is possible to identify morphotactic units, it may be difficult to identify a principled basis for assigning them context-independent meanings.
3. The **Residue** Challenge: In cases where it is possible to identify and interpret morphotactic units, this structure may still represent a historical residue that merely provides a context in which speakers learn to discriminate a larger form.

Declensional *-s* in Latin

The distribution of *-s* varies across the paradigms of the items that Bloomfield 1933 cites:

	Sg	Plu	Sg	Plu	Sg	Plu
Nom	amīcu**s**	amīcī	manu**s**	manū**s**	faciē**s**	faciē**s**
Gen	amīcī	amīcōrum	manū**s**	manuum	faciēī	faciērum
Dat	amīcō	amīcī**s**	manuī	manibu**s**	faciēī	faciēbu**s**
Acc	amīcum	amīcō**s**	manum	manū**s**	faciem	faciē**s**
Abl	amīcō	amīcī**s**	manū	manibu**s**	faciē	faciēbu**s**
Voc	amīce	amīcī	manu**s**	manū**s**	faciē**s**	faciē**s**
	‘friend’ (II)		‘hand’ (IV)		‘face’ (V)	

Plural *-t* in Georgian

Future indicative paradigm of K'VLA ‘kill’ (Tschenkeli 1958: §31)

	1Sg	1Plu	2Sg	2Plu	3
1Sg	—	—	mogk’lav	mogk’lav**t**	movk’lav
1Plu	—	—	mogk’lav**t**	mogk’lav**t**	movk’lav**t**
2Sg	momk’lav	mogk’lav	—	—	mok’lav
2Plu	momk’lav**t**	mogk’lav**t**	—	—	mok’lav**t**
3Sg	momk’lavs	mogvk’lavs	mogk’lavs	mogk’lav**t**	mok’lavs
3Plu	momk’laven	mogvk’laven	mogk’laven	mogk’laven	mok’laven

The ‘supine stem’ in Latin

Conjugational forms based on the ‘supine stem’ (Matthews 1972, 1991)

	Supine	Past Pass Prt	Fut Act Prt	
AMŌ	**amāt**um	**amāt**us	**amāt**ūrus	‘to love’
MONEŌ	**monit**um	**monit**us	**monit**ūrus	‘to advise’
TEGŌ	**tēct**um	**tēct**us	**tēct**ūrus	‘to cover’
CAPIŌ	**capt**um	**capt**us	**capt**ūrus	‘to take’
AUDIŌ	**audīt**um	**audīt**us	**audīt**ūrus	‘to hear’

Non-natural distributions in Cushitic

Syncretisms in Afar (Hayward 1980: 130) and Somali/Dhasaanac (Tosca 2007: 266)

	Afar Non-Perf Neg S	Somali Gen Past	Dhasaanac Pos Perfect
1Sg	**mageḍa**	**furay**	**furi**
2Sg	**mageḍḍa**	**furtay**	fuɗɗi
3Sg.M	**mageḍa**	**furay**	**furi**
3Sg.F	**mageḍḍa**	**furtay**	fuɗɗi
1Plu	magenḍa	furnay	
1Plu.Incl			**furi**
1Plu.Excl			fuɗɗi
2Plu	mageḍḍan	furteen	fuɗɗi
3Plu	mageḍan	fureen	**furi**

Challenges for an atomistic solution

Atomistic approaches create a number of general challenges:

1. **The Segmentation Challenge: It is often difficult to identify a principled basis for segmenting forms into smaller units.**

2. **The Interpretation Challenge: Where it is possible to identify morphotactic units, it may be difficult to identify a principled basis for assigning them context-independent meanings.**

3. The **Residue** Challenge: In cases where it is possible to identify and interpret morphotactic units, this structure may still represent a historical residue that merely provides a context in which speakers learn to discriminate a larger form.

Contextual ‘tuning’

- Even in cases where it is possible to segment forms and interpret segments, it may still not be possible to define a morphological system in terms of the disassembly and reassembly of forms, because the minimal parts obtained by segmentation are tuned to contexts.

Opportunistic variation

- From the perspective of a Dutch speaker, singular forms like *rat* 'rat' and *geit* 'goat' **do not recur** in the corresponding plurals *ratten* 'rats' and *geiten* 'goats' but instead have a distinctive prosodic profile that speakers are sensitive to.
- A form like *geiten* can be segmented into a stem *geit* and suffix *-en*, with the stem assigned lexical properties and the suffix assigned grammatical properties.
- However, it is an error to identify the plural stem *geit* with the singular form *geit* or to associate plurality solely with the suffix *-en:* the split between *geit* and *-en* does not correlate with a division in meaning between 'caprine' and 'plurality'.
- Instead, the plural stem is tuned to its morphological environment. From a discriminative learning perspective, this suggests that the function of the affix is not characterizable just in terms of the grammatical meaning that it conveys but also involves the 'stem-tuning' context it provides for learners.

Challenges for an atomistic solution

Atomistic approaches create a number of general challenges:

1. **The Segmentation Challenge: It is often difficult to identify a principled basis for segmenting forms into smaller units.**

2. **The Interpretation Challenge: Where it is possible to identify morphotactic units, it may be difficult to identify a principled basis for assigning them context-independent meanings.**

3. **The Residue Challenge: In cases where it is possible to identify and interpret morphotactic units, this structure may still represent a historical residue that merely provides a context in which speakers learn to discriminate a larger form.**

What about lossless decomposition?

- We want to guard against alarmist conclusions from isolated examples, and avoid a situation in which an outlier tail wags a statistically predominant dog.
- Regular plurals in English appear to be segmentable into stems and suffixes, and the suffixes do not show the variation that motivates inflection classes.
- The properties of the plural seem to be amenable to an analysis as the sum of the lexical properties of the stem and the number properties of the suffix /z/.
- The surface form of the suffix is conditioned by the final segment of the stem. For forms like *pans* and *foxes*, it appears possible to isolate the stems *pan* and *fox* and the suffix /z/, and reconstitute the original forms from these parts.

Is decomposition *ever* lossless?

- Even in this case, it is unclear that decomposition is lossless.
- Disassembling *pans* and *foxes* into stems and suffixes loses the information that the noun stems occur with the regular marker.
- Neither the grammatical properties of the nouns nor the form of the noun stems predict the fact that they occur with the regular suffix rather than following the patterns exhibited by *men* or *oxen*.
- 'Morphologically conditioned allomorphy' is invoked to govern the selection of a plural strategy, where this is not predictable from the properties or form of the stems *pan* and *man*, *fox* and *ox*.

Decomposition and frequency

- Given that few nouns follow residual plural patterns in English, a speaker can learn them as exceptions and assume that other nouns follow the regular pattern.
- However, this distinction brings in considerations of frequency, and frequency is another type of information that is sacrificed by a decompositional analysis.
- In particular, the frequency of a plural form is not generally recoverable from the frequency of the stem and the frequency of the suffix.
- Frequency information could be recovered from disassembled parts under certain conditions, notably if parts were not the ‘same’ element in all of their occurrences.
- Yet under these conditions, the link between recurrence and redundancy would be broken altogether, calling into question the assumptions and motivations underlying the strategy of disassembling and interpreting of recurrent units.

From systems to items

- The intuition behind atomistic decomposition is that recurrent structure is redundant and that the goal of morphological analysis is to eliminate redundancy by reducing systems to inventories of minimal units and combinatorial rules.
- Yet there is little evidence that morphological systems are genuinely organized in ways that facilitate disassembly and reassembly of words.
- It may be possible to disassemble a word form into minimal units.
- However there are often no effective criteria for selecting 'correct' segmentations into stable roots, stems and inflections. Moreover, in all but the simplest systems it is not possible to reconstitute word forms from inventories of minimal units, because disassembly has lost information about the distribution of units.
- Hence, atomistic accounts become preoccupied with the analysis of individual forms and never get back to analyzing the properties of the morphological system.

Deviant irregularity

- Atomistic treatments of stem alternations in Estonian (or ablaut in English) reveal a fundamental analytic bias.
- The goal of these analyses is to assimilate seemingly ‘irregular’ formations to more regular affixal patterns.
- This goal reflects the view that regular formations are ‘normative’, and that irregular forms are deviations from the uniform patterns that a system strives to maintain.