# Computational methods for morphological theory

Rob Malouf, San Diego State University

# Plan

1. Foundational questions
2. Morphological complexity & Information theory
3. Morphological description & Deep Learning
4. Morphological explanation & Bayesian agents

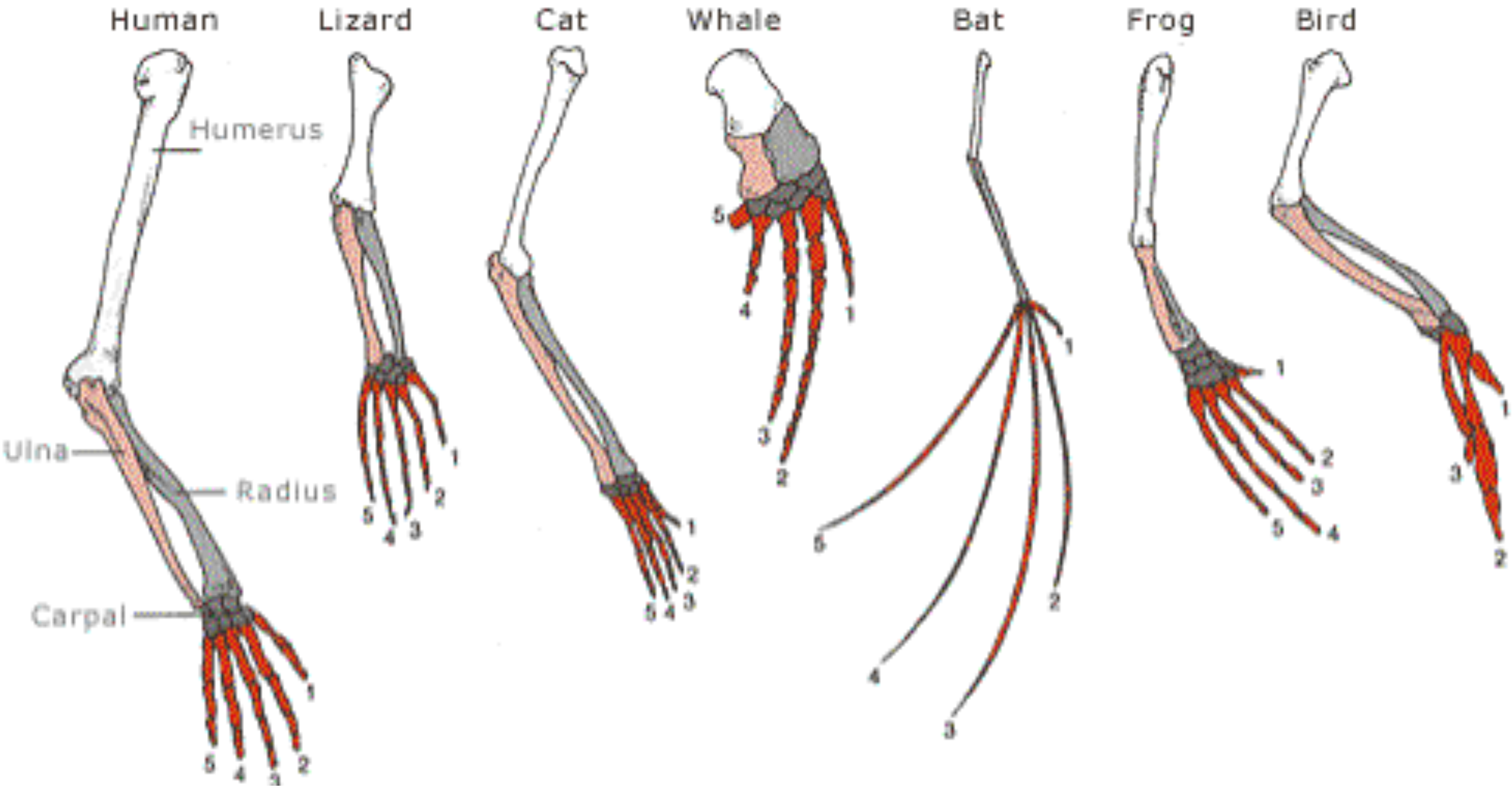
# What is Morphology?

- **Morphology** is the study of form

  μορφή (morphḗ) ‘form’ + λογία (logía) ‘explanation’

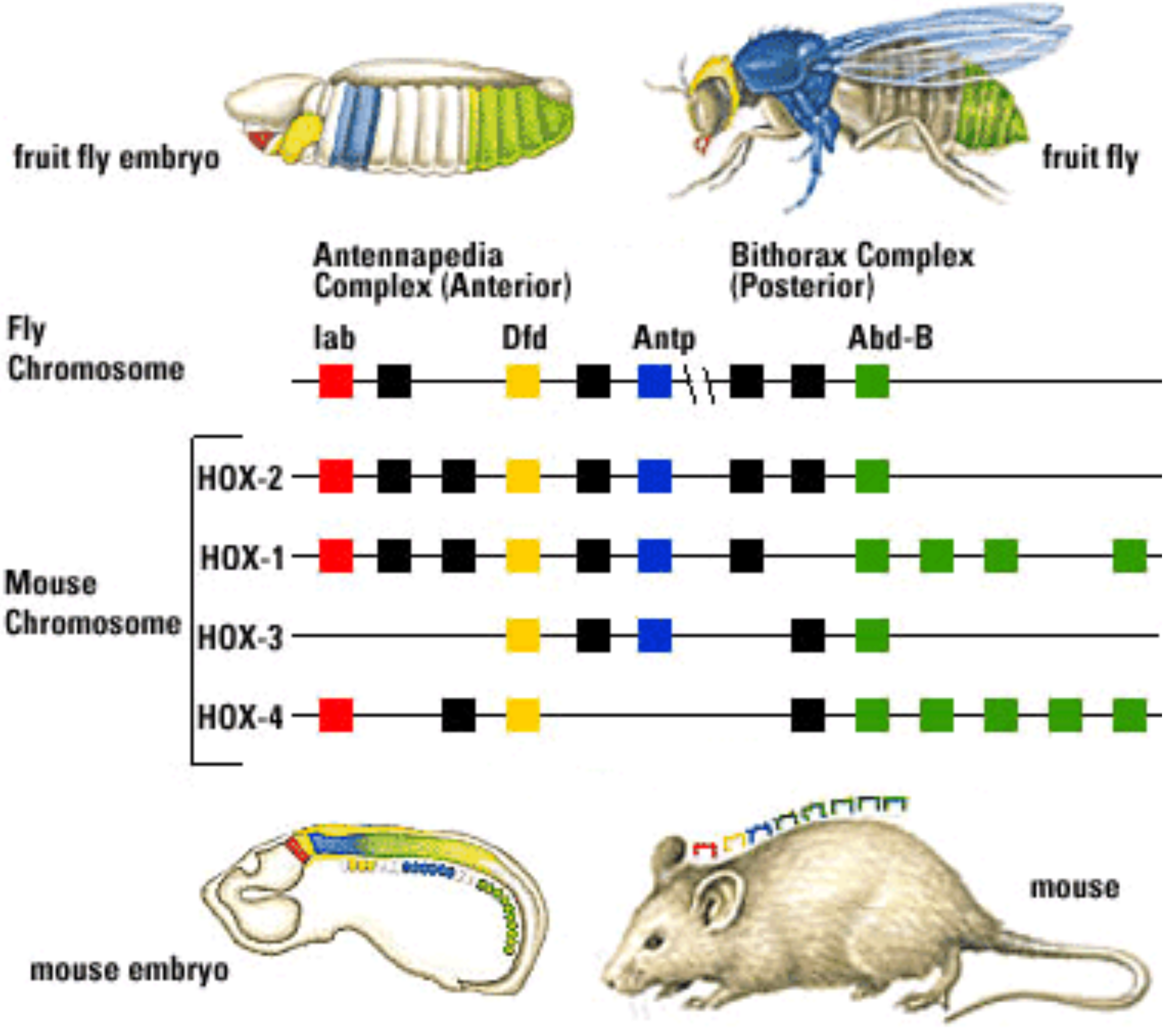
- Origins in biology: J. W. Goethe (1749–1832)
- Comparative anatomy, in contrast to **physiology** (study of function)
- In biology, form is closely associated with function, but how much?

# Comparative anatomy



# Comparative anatomy

- Cross species comparison of body design (Carroll et. al. 2005:25)



# Linguistic morphology

- The study of (biological) morphology forms a crucial foundation for evolutionary theory
- What is (linguistic) morphology good for?
- Linguistic natural history
    - Descriptive and pedagogical applications
    - Understanding historical processes
- Linguistic diversity

# Linguistic diversity

Friedrich Schlegel
1772–1829

August Schlegel
1767–1845

# Linguistic diversity

- Divided languages into **affixal**, **isolating**, and **flectional** types

  - Turkish:

    | | | | |
    |---|---|---|---|
    | *anla-* | *ma-* | *d-* | *ım* |
    | understand | NEG | PAST | 1PERS |

    ‘I did not understand’

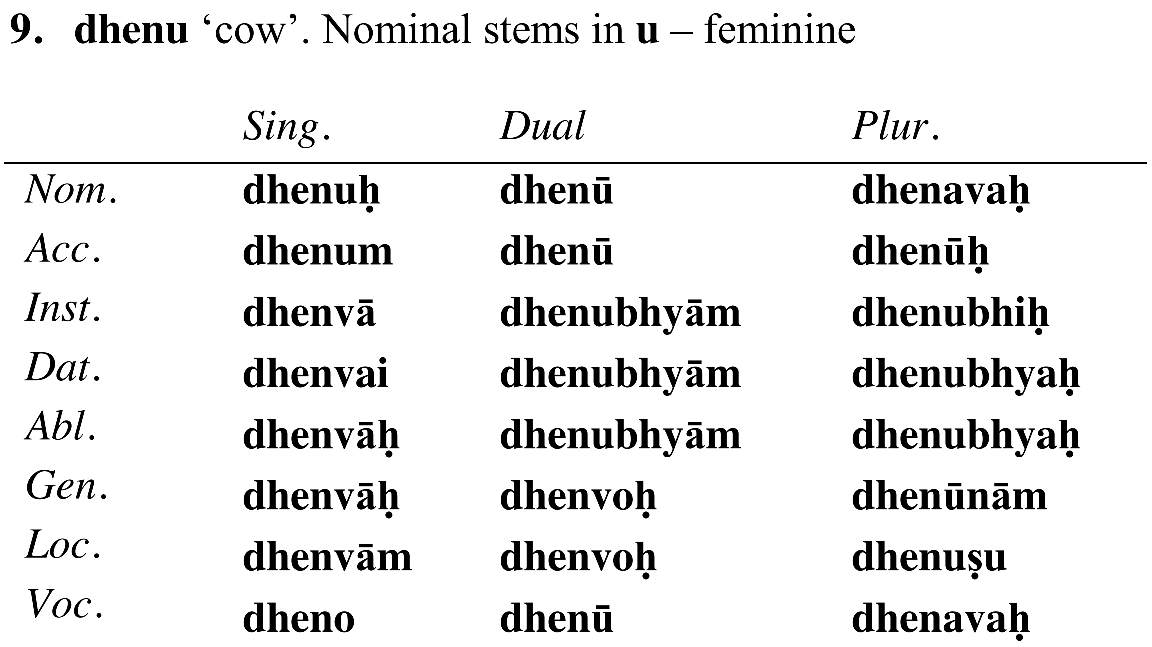
  - Classical Chinese:

    | | | | | | |
    |---|---|---|---|---|---|
    | *liù zǔ* | *wén yǐ,* | *jí* | *shí* | *fó* | *yì* |
    | six patriarch | hear finish | then | familiar | buddha | thought |

    ‘When the Sixth Patriarch had heard this he was familiar with the Buddha’s thought.’

# Linguistic diversity

- Divided languages into **affixal**, **isolating**, and **flectional** types
  - Sanskrit



**9. dhenu** ‘cow’. Nominal stems in **u** – feminine

| | *Sing.* | *Dual* | *Plur.* |
|---|---|---|---|
| *Nom.* | **dhenuḥ** | **dhenū** | **dhenavaḥ** |
| *Acc.* | **dhenum** | **dhenū** | **dhenūḥ** |
| *Inst.* | **dhenvā** | **dhenubhyām** | **dhenubhiḥ** |
| *Dat.* | **dhenvai** | **dhenubhyām** | **dhenubhyaḥ** |
| *Abl.* | **dhenvāḥ** | **dhenubhyām** | **dhenubhyaḥ** |
| *Gen.* | **dhenvāḥ** | **dhenvoḥ** | **dhenūnām** |
| *Loc.* | **dhenvām** | **dhenvoḥ** | **dhenuṣu** |
| *Voc.* | **dheno** | **dhenū** | **dhenavaḥ** |

# Linguistic diversity

- F. Schlegel (1808): affixed (Turkish-type) languages are “**a heap of atoms which every wind of chance scatters or sweeps together**”
- A. Schlegel (1818): of isolating (Chinese-type) languages, “**one might say that all their words are roots, but sterile roots which produce neither plants nor trees**”, while flectional languages “**contain a vital principle of development and growth**”
- Hierarchy of language types

  flectional > affixal > isolating

# Linguistic morphology in 1800’s

Wilhelm von Humboldt
1767–1835

August Schleicher
1821–1868

Ernst Haeckel
1834–1919

# Wilhelm von Humboldt

- Wilhelm von Humboldt (1767–1835)
- Language emerged spontaneously out of the inner creative energy of a nation in a unique way

  "**. . .every nation, quite apart from its external situation, can and must be regarded as a human individuality, which pursues an inner spiritual path of its own.**" (1836)

- A language is a combination of the inner spirit of a speaker (as a member of a nation) and the external constraints of the language as developed over history by past speakers

# Wilhelm von Humboldt

- All languages share a universal core: “**Since the natural disposition to language is universal in man . . . it follows automatically that the form of all languages must be essentially the same. . . The difference can lie only in the means, and only within the limits permitted by attainment of the goal.**” (1836)
- Connection to thought: “**The similarity of the laws of thought produces what is shared by the grammar of all languages. . . Every grammatical form may, in some way or another, be pointed out in every language . . .**” (1824)

# Wilhelm von Humboldt

- We can imagine an ideal language which most directly reflects the needs of universal thought, and "**we must be able to judge the merits and defects of existing languages by the degree to which they approximate to this one form**." (1836)

- The ideal language combines a meaning with a relation: "**The perfecting of language demands that every word be stamped as a specific part of speech, and carry within it those properties that a philosophical analysis of language perceives therein. It thus itself presupposes inflection.**"

- Ummm….

# Wilhelm von Humboldt

- Thought is universal, but not all languages allow thoughts to be articulated with the same efficiency
- In inflectional languages, word formation mirrors concept formation
- A nation with an inferior creative spirit will speak an inferior language, which further limits their intellectual development
- Sanskrit is the best, Classical Chinese is the worst
- English and French look bad, but they used to be good and so preserve an inner inflecting spirit

# August Schleicher

- August Schleicher (1821–1868)
- Significant contributions to reconstruction of Proto-Indo-European
- Introduced the ‘family tree’ as a model for language relationships
- After reading Darwin’s *Origin of Species*, suggested trees as a model for biological relationships

# August Schleicher

- Schleicher saw deep connections between human biological evolution and linguistic evolution
    - Linguistic pre-history is easier to reconstruct (at the time) than biological pre-history
    - Language = thought (monism)
    - “The formation of language is for us comparable to the evolution of the brain and the organs of speech”
    - “Animals can be ordered according to their morphological character. … To classify human beings we require . . . a higher criterion, one which is an exclusive property of man. This we find … in language.”

# August Schleicher

- Schleicher’s theory of human evolution
  - Pre-linguistic period, with no humans
  - Pre-historic period, in which proto-humans gradually developed language
  - Stages of linguistic development reflect the full self-realization of a Weltgeist
    - thesis: isolation
    - antithesis: affixation
    - synthesis: inflection

# Ernst Haeckel

- Ernst Haeckel (1834–1919)
- Darwin enthusiast, naturalist, biologist
- Built on Schleicher’s theory that linguistic evolution = biological evolution
- Best remembered for recapitulation theory (“Ontogeny recapitulates phylogeny”) and scientific racism

# Ernst Haeckel

- Human polygenesis: “We must mention here one of the most important results of the comparative study of languages, which for the Stammbaum of the species of men is of the highest significance, namely that **human languages probably had a multiple or polyphyletic origin.** … If one views the origin of the branches of language as the special and principal act of becoming human, and the species of humankind as distinguished according to their language stem, then one can say that **the different species of men arose independently of one another**.” (1868)

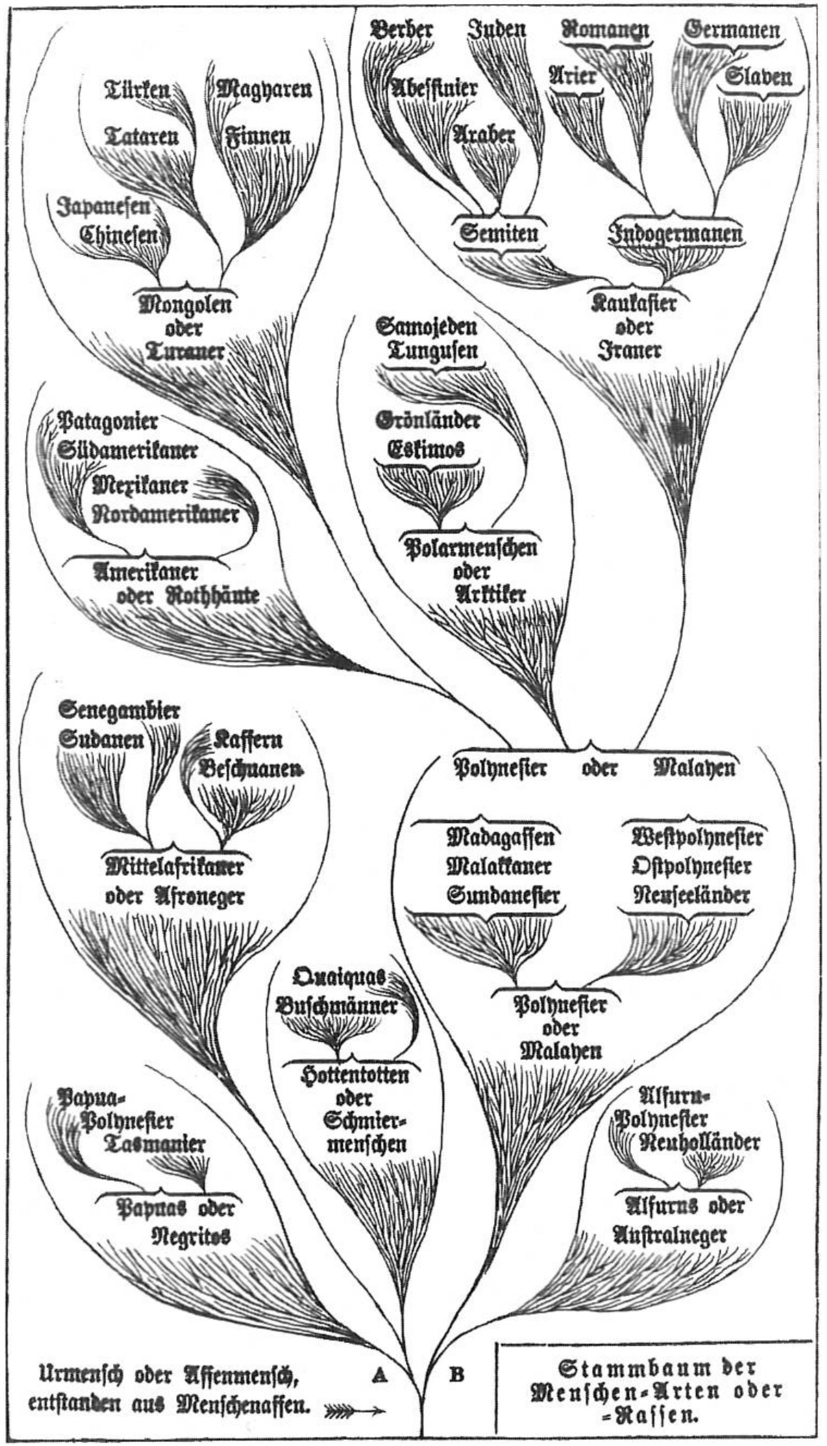


FIGURE 2.7 Haeckel's *Stammbaum* of the nine species of men. From his *Natürliche Schöpfungsgeschichte* (Berlin: Reimer, 1868).

# Linguistic morphology in 1900's

Franz Boas

1858–1942

Edward Sapir

1884–1939

# Franz Boas

- Introduction to *Handbook of American Indian Languages* (1911)
- Independence of language, culture, and ethnicity
- Uniformity of the linguistic landscape
- No relationship between physical environment and grammar
- All languages are complex and systematic
- Non-European languages are often more complex/ systematic

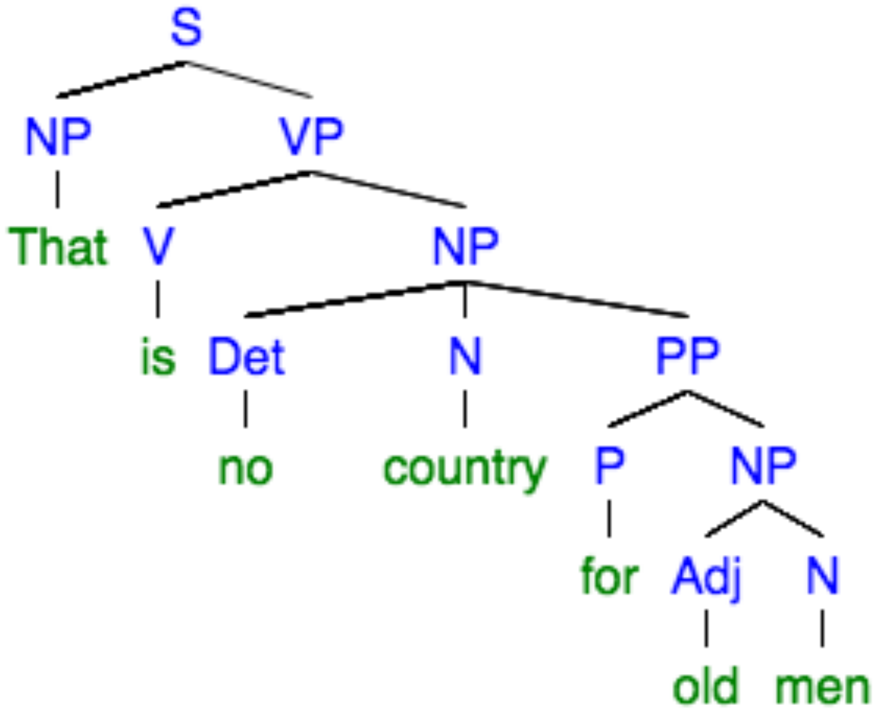
# Edward Sapir

- “… the valuation [of languages] according to whether their inflections are more or less transparent is as foolish as if one judged the merit of European armies according to the greater or lesser visibility of their trouser seams” (Mauthner 1923).

- “All attempts to connect particular types of linguistic morphology with certain correlated stages of cultural development are vain. […] When it comes to linguistic form, Plato walks with the Macedonian swineherd, Confucius with the head-hunting savage of Assam.” (Sapir 1921)

# Linguistic morphology in 1900’s

- Without (as much) racism, morphology lost a lot of its theoretical value
- Duality of patterning (Martinet, Hockett)
- Primary articulation



# Morphology

- Secondary articulation

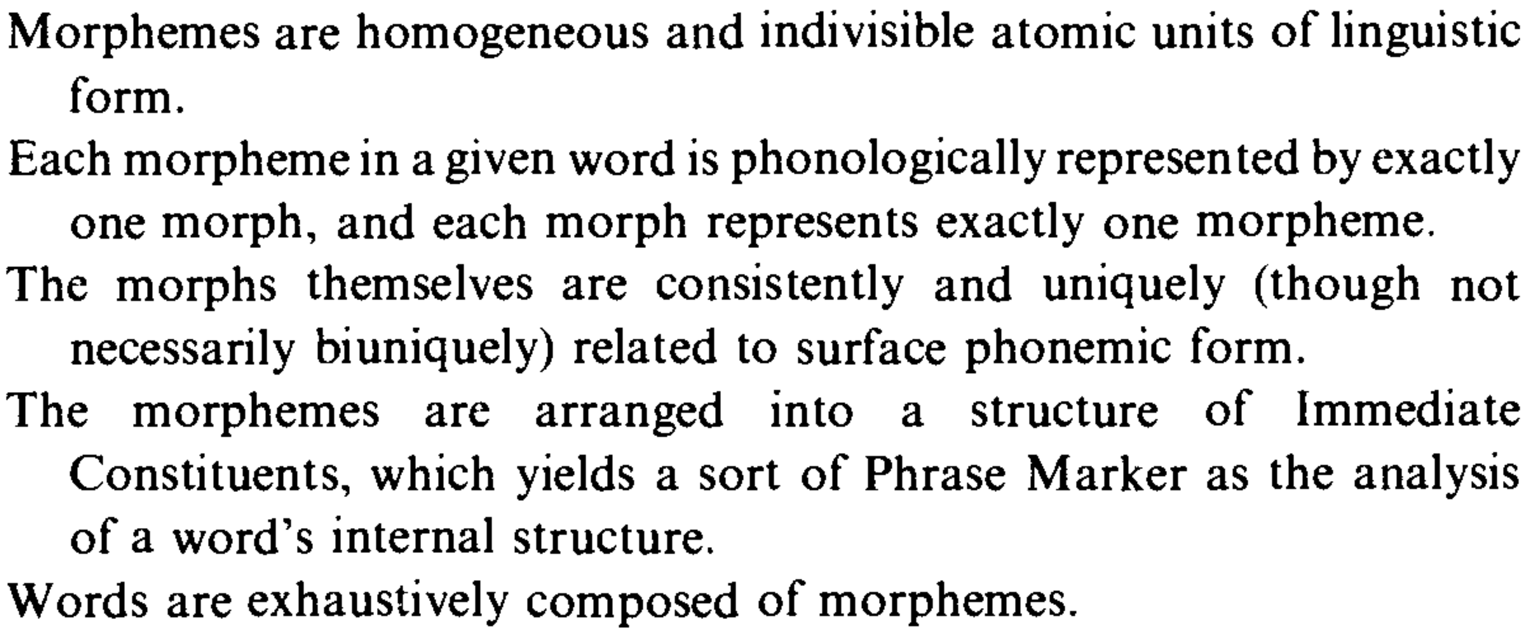
  [ðæt][ɪz][noʊ][ˈkʌntɹi][fɹ̩][oʊld][mɛn]

- Rules operate independently on the two articulations
  - *That is for country no men old
  - *zbæp, *ŋɪʃ, pæbz, ʃɪŋ

# Post-Bloomfieldians

- Item and Arrangement = morphemes + tactics

- Anderson (1992:50)



Morphemes are homogeneous and indivisible atomic units of linguistic form.

Each morpheme in a given word is phonologically represented by exactly one morph, and each morph represents exactly one morpheme.

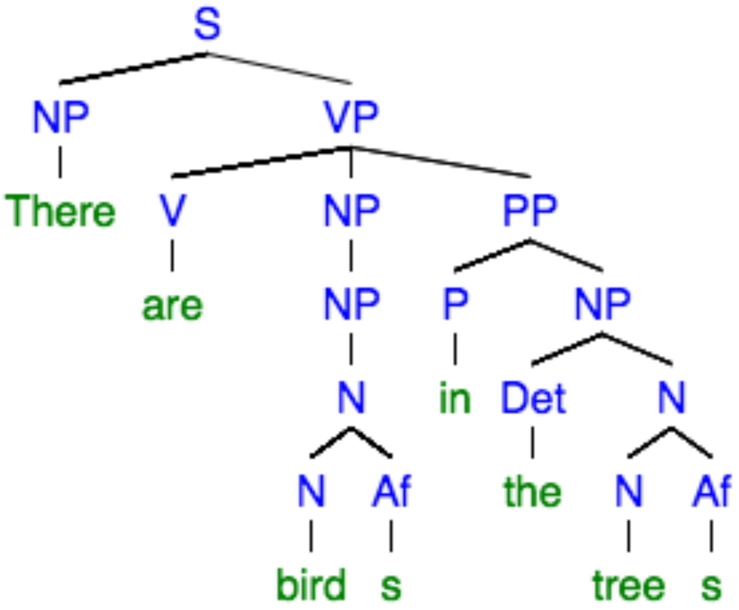
The morphs themselves are consistently and uniquely (though not necessarily biuniquely) related to surface phonemic form.

The morphemes are arranged into a structure of Immediate Constituents, which yields a sort of Phrase Marker as the analysis of a word's internal structure.

Words are exhaustively composed of morphemes.

# Post-Bloomfieldians

- One view: the smallest meaningful units in language (**morphemes**) are the basic units of the first articulation



- On this view, word structure is no different from phrase structure — it's all just grammar

# What is morphology?

- Morphology as the study of **morphemes**
    - Morphology is the study of the combination of morphemes to yield words.  (Haspelmath & Sims 2010:2)
    - Morphology is the study of morphemes and their arrangements in  forming words. (Nida 1949:1)

# What is morphology?

- At a descriptive level, word organization is (argued to be) different from phrases
- Morphology as the study of **words**
  - Morphology is the study of the systematic covariation in the form and meaning of words.  (Haspelmath & Sims 2010:3)
  - Morphology ... is simply a term for that branch of linguistics which is concerned with the ‘forms of words’ in different uses and constructions. (Matthews 1991:3)

# Word-based morphology

- General problems with segmentation into morphemes
  - Zero morphs : one meaning, no form
  - Empty morphs : no meaning, one form
  - Cumulative exponence : many meanings, one form
  - Extended exponence : one meaning, many forms

# Empty morphs

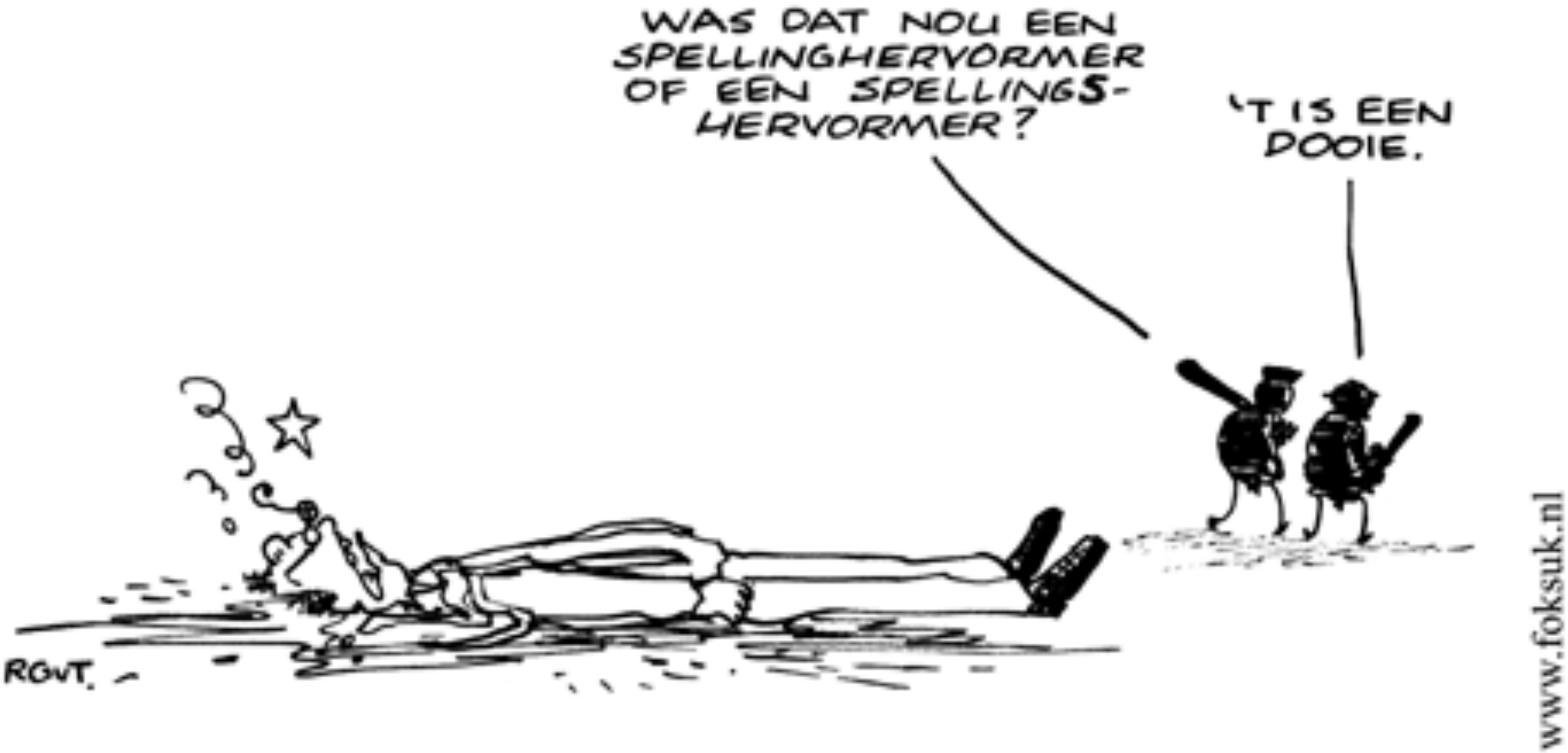
- Theme vowels

| | | | |
|---|---|---|---|
| *mán-o* | ‘hand.SG’ | *mán-o-s* | ‘hand.PL’ |
| *dí-a* | ‘day.SG‘ | *dí-a-s* | ‘day. PL’ |
| *cruc-e* | ‘crossing.SG’ | *cruc-e-s* | ‘crossing. PL’ |

- Linking elements (Booij 2005:89)

| | |
|---|---|
| *schaap* | ‘sheep’ |
| *schaap-herder* | ‘shepherd’ |
| *schaap-**s**-kop* | ‘sheep’s head’ |
| *schap-**en**-vlees* | ‘mutton’ |
| *kind* | ‘child’ |
| *kind-**er**-wagen* | ‘stroller’ |

## FOKKE & SUKKE
HOUDEN HET SIMPEL


WAS DAT NOU EEN SPELLINGHERVORMER OF EEN SPELLINGS-HERVORMER?
'T IS EEN DOOIE.
RGVT.
www.foksuk.nl

# Empty morphs

- **Cranberry morphs** distinguish words but don’t have any identifiable meaning

  *blackberry, blueberry, salmonberry, strawberry*
  *raspberry, cranberry*

  Dutch *stiefvader* ‘stepfather’

# Cumulative exponence

- Cherokee (Aronoff & Fudeman 2005:153)

| | |
|---|---|
| *ski-, skw-* | 2SG.SUBJ/1SG.OBJ |
| *stiː-* | 2DU.SUBJ/3SG.INAN.OBJ |
| *kaciːy-* | 1SG.SUBJ/3PL.AN.OBJ |
| *ciːy-* | 1SG.SUBJ/3SG.AN.OBJ |

*sə̃ːkthə̃ kaciːneːlə̃ːʔi*
apple 1SG.SUBJ/3PL.AN.OBJ-give.PERF
‘I gave them an apple.’
*ciːkoːwthiha*
1SG.SUBJ/3SG.AN.OBJ-see.PRES
‘I see him.’

# Cumulative exponence

- Adyghe (Arkadiev 2014)

| PŜAŜE ‘girl’ | SG | PL |
|---|---|---|
| ABS | *pŝaŝe-r* | *pŝaŝe-xe-r* |
| OBL | *pŝaŝe-m* | ***pŝaŝe-xe-m***<br>***pŝaŝe-me***<br>***pŝaŝe-xe-me*** |
| INS | *pŝaŝe-m-č̣’e* | *pŝaŝe-xe-m-č̣’e* |

- Agglutination, cumulation, overabundance

# Extended exponence

- Exuberant exponence (Harris 2009)

- Batsbi

-

| | | | |
|---|---|---|---|
| *oqar* | *tiši*$^{n}$ | *c’a* | ***d**-ox-**d**-iy-er* |
| they | old | house.ABS | CM-destroy-TR-IMPF |

‘They tore down the old house.’

-

| | | | |
|---|---|---|---|
| *tiši*$^{n}$ | *c’a* | *daħ* | ***d**-ex-**d**-o-**d**-anŏ* |
| old | house.ABS | PV | CM-destroy-CM.TR-PRES-CM-EV |

‘They are evidently tearing down the old house.’

-

| | | | |
|---|---|---|---|
| *tiši*$^{n}$ | *c’a* | *daħ* | ***d**-ex-**d**-o-**d**-an-iš* |
| old | house.ABS | PV | CM-destroy-CM-PRES-CM-EV-2PL.ERG |

‘Y’all are evidently tearing down the old house.’

# Extended exponence

- Exuberant exponence (Harris 2009)
- Archi

***d***-*as:á*-***r***-*ej*-***r***-*u*-*t:u*-***r***
II-of.myself-II-SUFF-II-SUFF-SUFF-II
‘my own’ [female]

***w***-*as:á*-***w***-*ej*-***w***-*u*-*t:u*-***w***
I-of.myself-I-SUFF-I-SUFF-SUFF-I
‘my own’ [male]

# Word-based morphology

- We take **words** to be minimal signs — parts of words do not have any meaning

  “In the ancient model the primary insight is not that words can be split into roots and formatives, but that they can located in paradigms.  They are not whole composed of simple parts, but are themselves the parts within a complex whole.  In that way, we discover different kinds of relation, and, perhaps, a different kind of simplicity.” (Matthews 1991:204)

# Word-based morphology

- Traditional word-based models are organized around **exemplars** or **analogies**, rather than rules
- The forms of an inflectional system are organized into paradigms
- Each paradigm contains one or more diagnostic or 'leading' forms which help guide hypotheses about what unknown members of the paradigm look like
- New items are inflected by analogy to an established paradigm

# Word-based morphology

- Word-based morphology treats word-level formations as fundamentally different from phrase-level formations
- Makes morphology theoretically interesting again!
- But lots of lingering issues . . .

# Dissent

Marantz (1997): “**The underlying suspicion behind the leading idea of Lexicalism is this: we know things about words that we don’t know about phrases and sentences**; what we know about words is like what we would want to say we know about (atomic) morphemes. This paper brings the reader the following news: Lexicalism is dead, deceased, demised, no more, passed on.... **The underlying suspicion was wrong and the leading idea didn’t work out.** This failure is not generally known because no one listens to morphologists. Everyone who has worked on the issues of domains—what are the domains for “lexical phonological rules,” what are the domains of “special meanings,” what are the domains of apparently special structure/meaning correspondences—knows that **these domains don’t coincide in the “word,” and in fact don’t correlate (exactly) with each other**. But the people that work on word-sized domains are morphologists, and when morphologists talk, linguists nap.”

# Dissent

- Bruening, Benjamin. 2018. “The lexicalist hypothesis: Both wrong and superfluous.” *Language* 94(1): 1–42.
- Haspelmath, Martin. 2011. “The indeterminacy of word segmentation and the nature of morphology and syntax.” *Folia Linguistica* 45(1): 31–80.

> The general distinction between morphology and syntax is widely taken for granted, but it crucially depends on a cross-linguistically valid concept of ‘(morphosyntactic) word’. I show that there are no good criteria for defining such a concept. Thus, I conclude that **we do not currently have a good basis for dividing the domain of morphosyntax into morphology and syntax**, and that linguists should be very careful with general claims that make crucial reference to a cross-linguistic ‘word’ notion.

# Dissent

- Here’s a secret: all morphological theories are formally equivalent
- That is, an analysis of some language in model *X* can always be converted to an analysis in model *Y*
- From Matthews (1972):

  “In dealing with problems which can be solved in more ways than one, the solutions themselves are of less interest than the reasons for making one choice rather than another.” (Newman 1967:192)

# Memory

- Memory must play a big role in morphology (simple words, suppletion) in a way that it doesn’t in syntax
- Since speakers can create and understand forms they’ve never heard before, morphology can’t be just memory
- What’s the balance of labor between retrieval and computation? Are these different for words and phrases?
- What is the relationship between memory and productivity?

# Memory

- Model of associative memory (Collins & Loftus 1975)

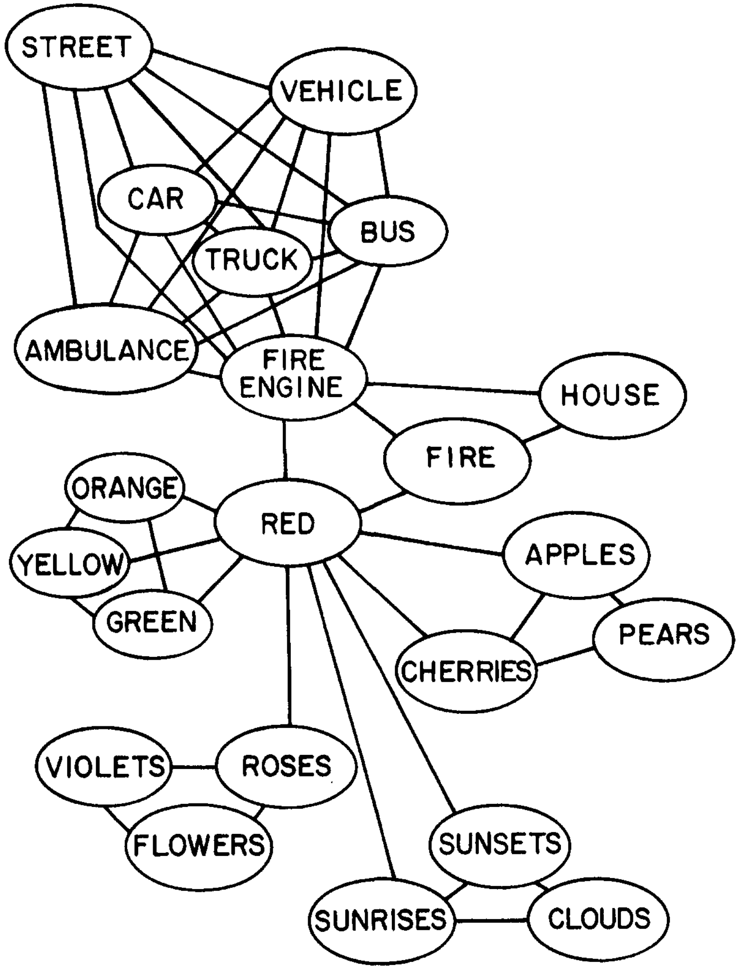


FIGURE 1. A schematic representation of concept relatedness in a stereotypical fragment of human memory (where a shorter line represents greater relatedness).

# Memory

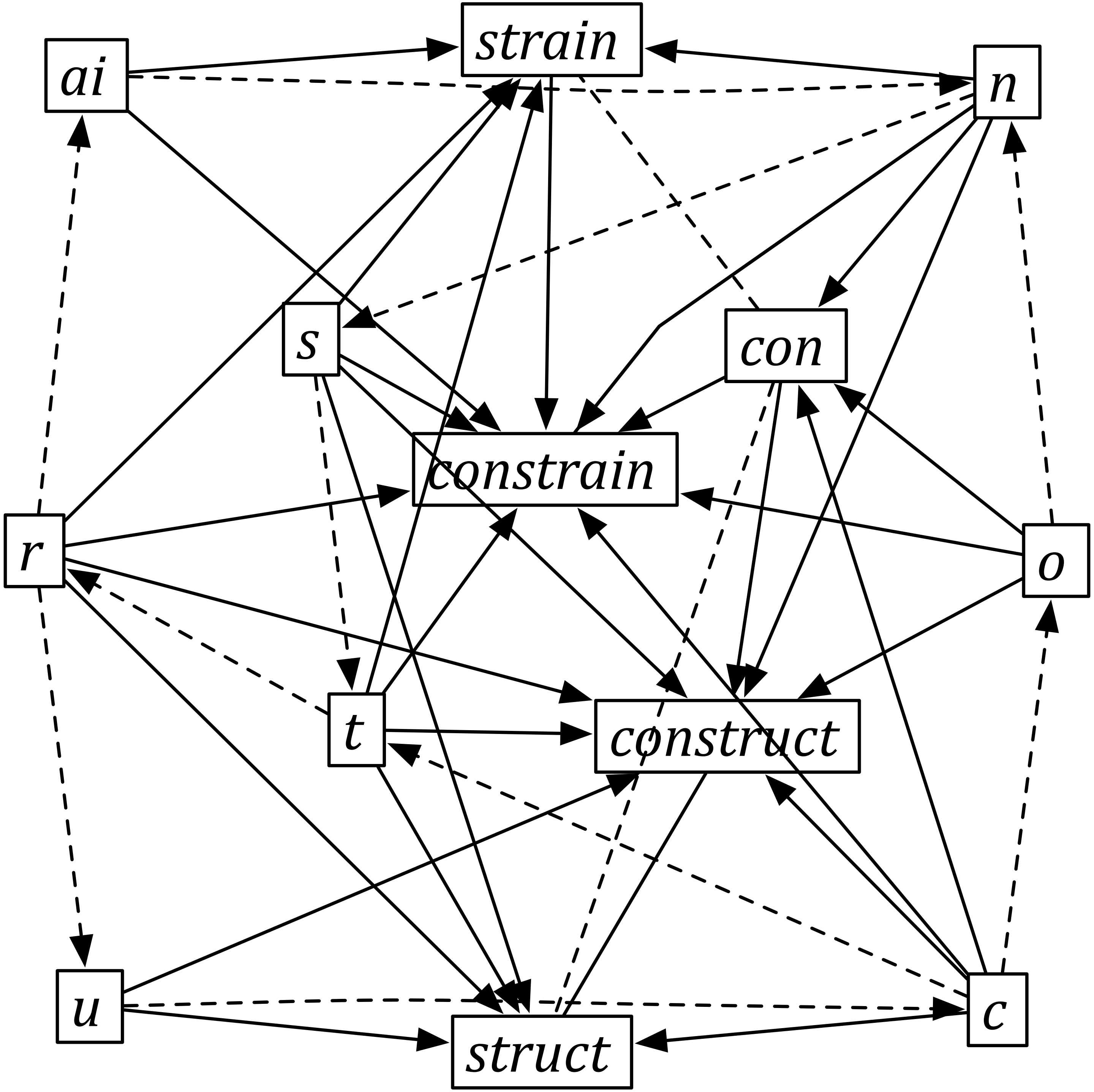
- Model of associative memory (Collins & Loftus 1975)
  - Semantic memory consists of linked concepts
  - Retrieval time is proportional to a concept’s level of **activation**
  - When a node is activated, the activation spreads through the network to related concepts

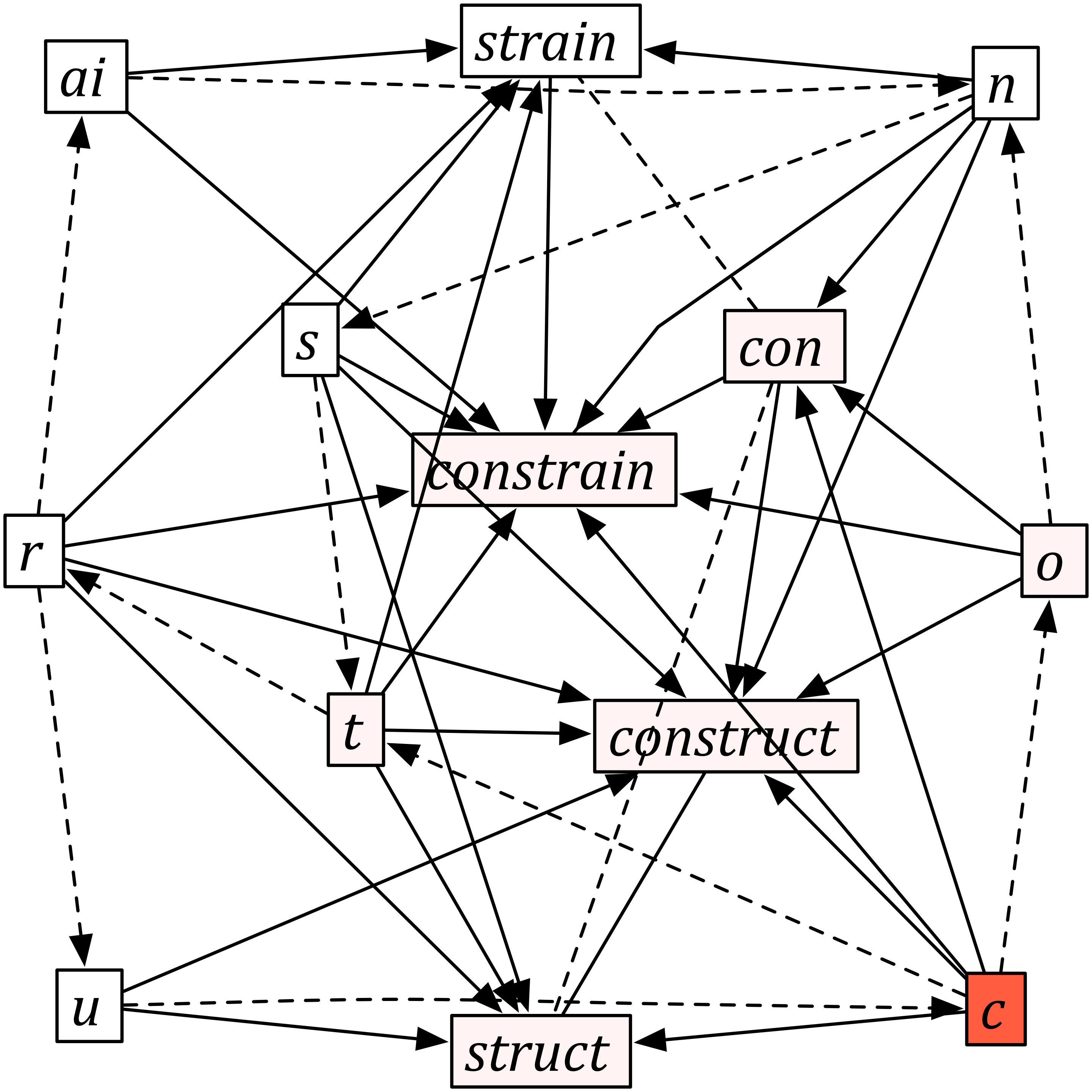
# Dual mechanism

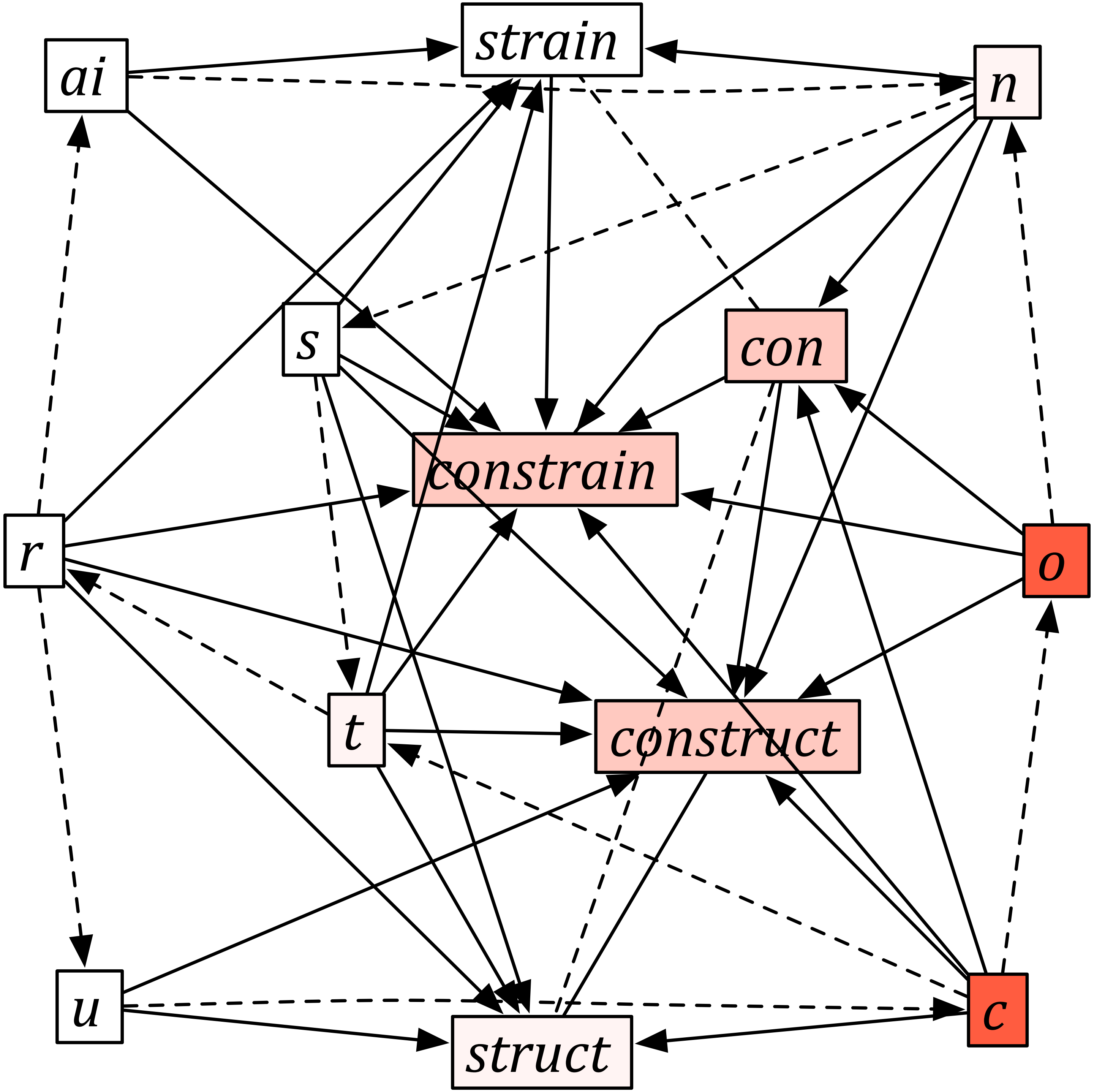
- **Dual Mechanism Theory** (Marcus, Clahsen, Pinker, et al.)
  - Irregular inflection processed by associative lexicon (*sing* ~ *sang*)
  - Regular inflection processed by computational rules (*walk* + *ed*)
- Mostly based on English verbs, but also plurals in German
  - Strong word frequency effects for irregular words, but not (?) for regular words
  - Strong root priming for regular words, not (?) for irregular words

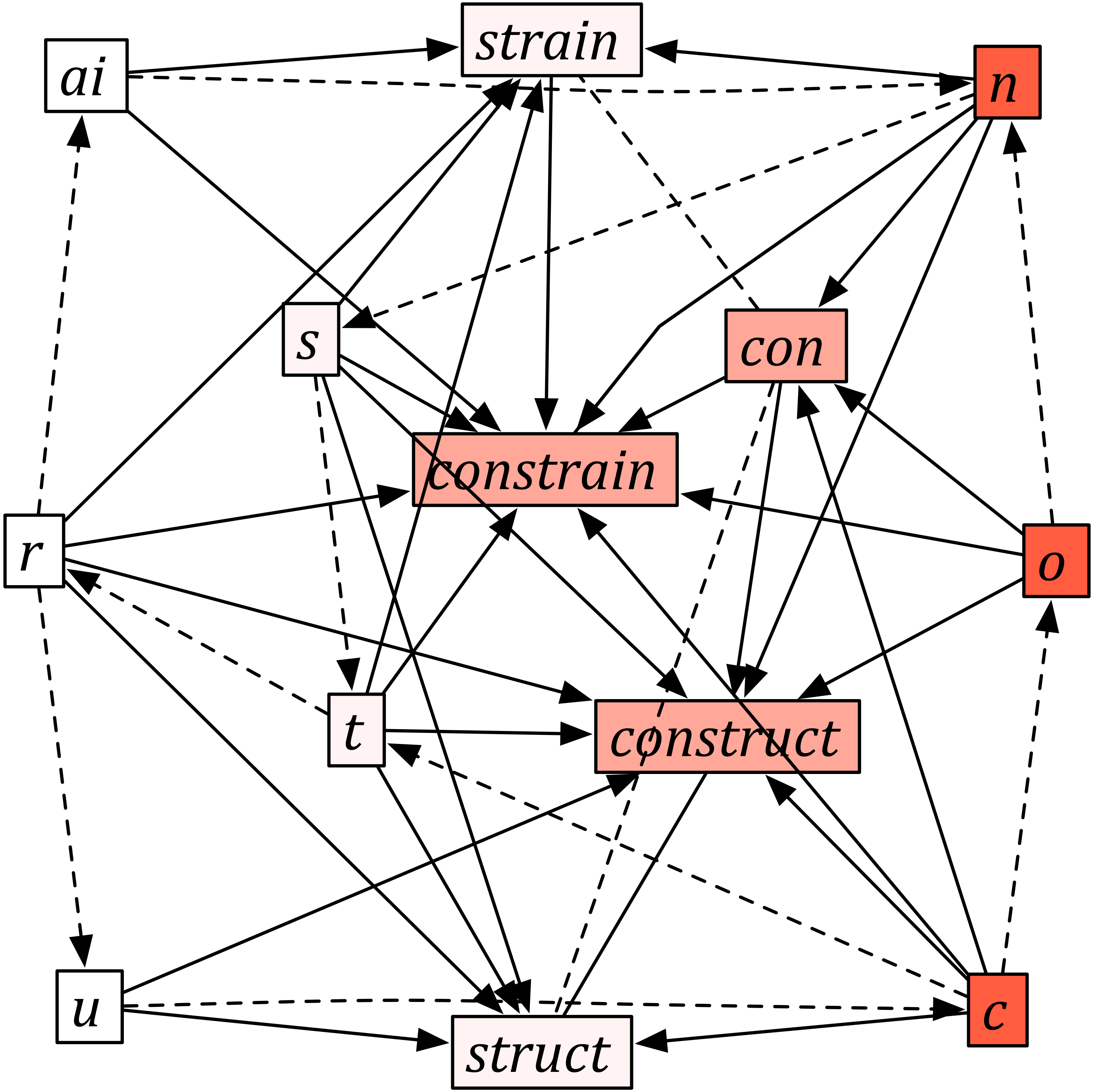
# Dual mechanism

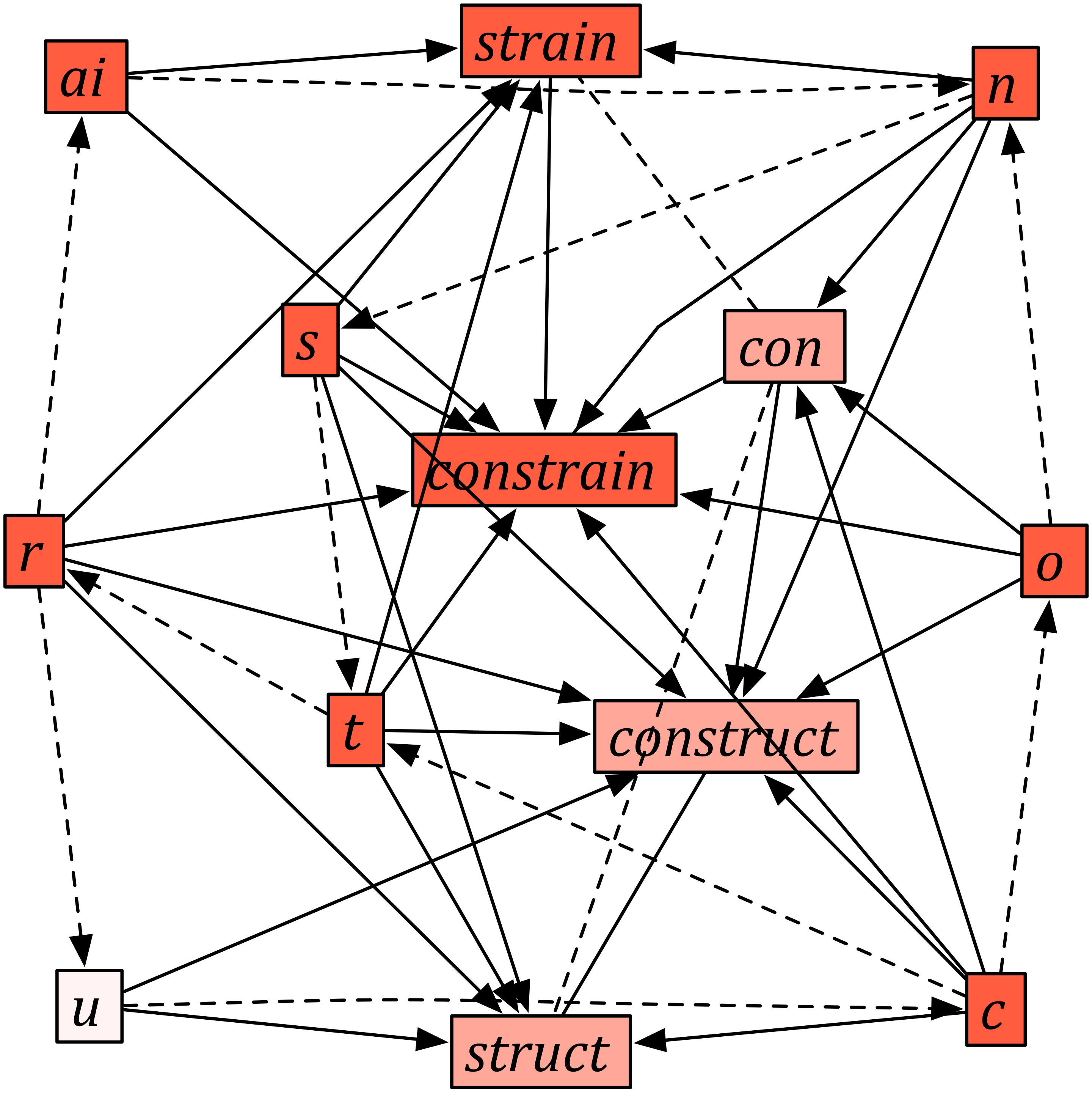
- An alternative version of the DMT proposes that both systems work in parallel (Baayen and Schreuder)
- The lexicon is a list of all (known) full words *and* all morphemes
- Words are segmented by looking up both the whole word and all substrings in the lexicon
- MATCHECK: spreading activation model of lexical lookup during visual reading
- Initial activation levels proportional to frequency

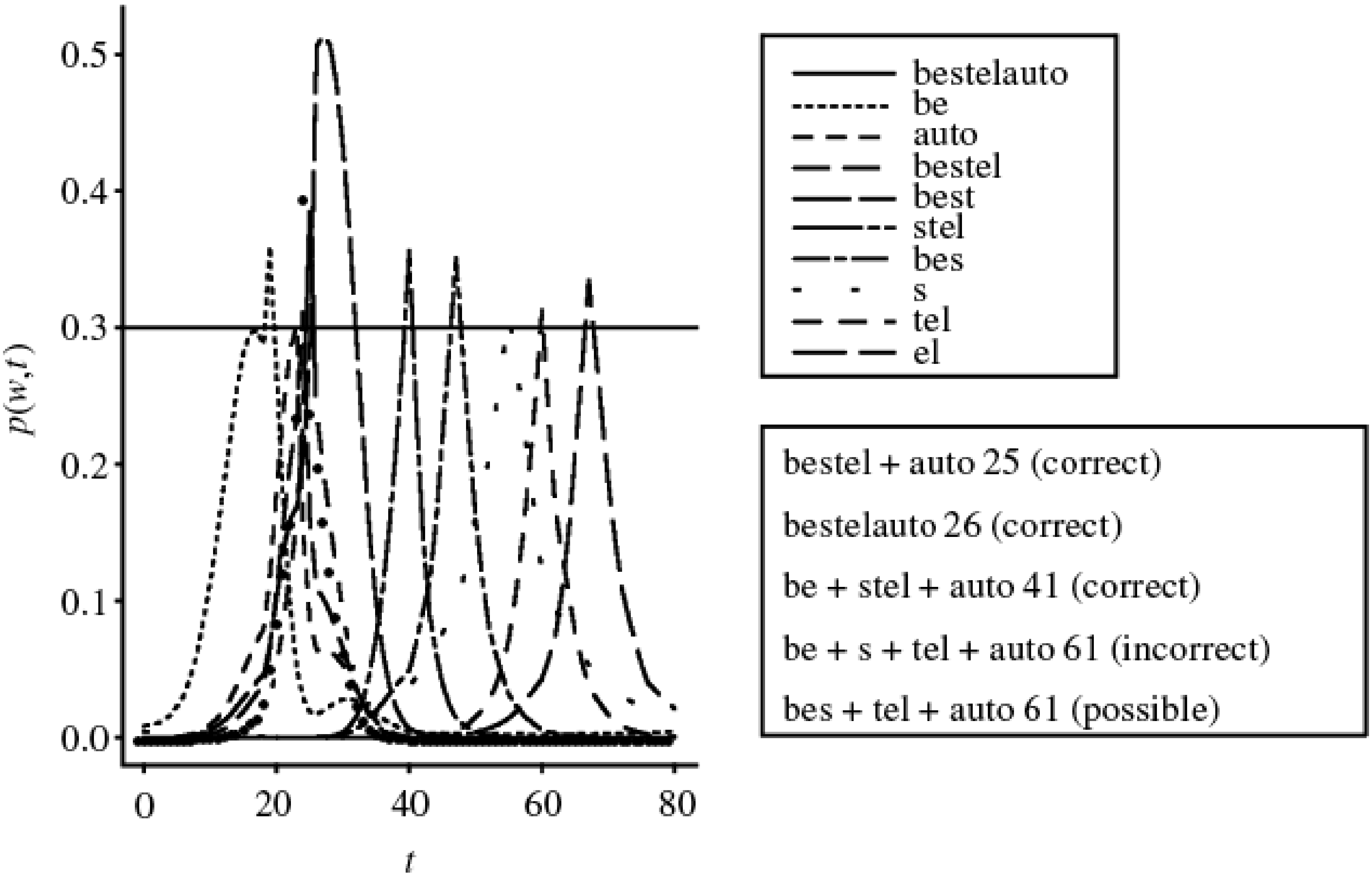

strain
ai
n
s
con
constrain
r
o
t
construct
u
c
struct


strain
ai
n
s
con
constrain
r
o
t
construct
u
c
struct


strain
ai
n
s
con
constrain
r
o
t
construct
u
c
struct


strain
ai
n
s
con
constrain
r
o
t
construct
u
struct
c


strain
ai
n
s
con
constrain
r
o
t
construct
u
c
struct



bestel + auto 25 (correct)

bestelauto 26 (correct)

be + stel + auto 41 (correct)

be + s + tel + auto 61 (incorrect)

bes + tel + auto 61 (possible)

# Productivity

- Parallel dual mechanism model implies a competition among lexical items
  - If a complex word is more frequent than its base, it will retrieved from the lexicon as a whole word
  - If a complex word is less frequent than its base, it will retrieved as separate morphs
- A hypothesis: the more often a suffix is retrieved, the more likely it is to be productive
- Hay and Baayen (2002, 2003) compared the predictions of Matcheck to various measures of productivity

# Productivity

- For any affix, the **parsing line** separates words which are retrieved whole from words which are parsed
  - words which are likely to be retrieved whole

    *government*, *pavement*

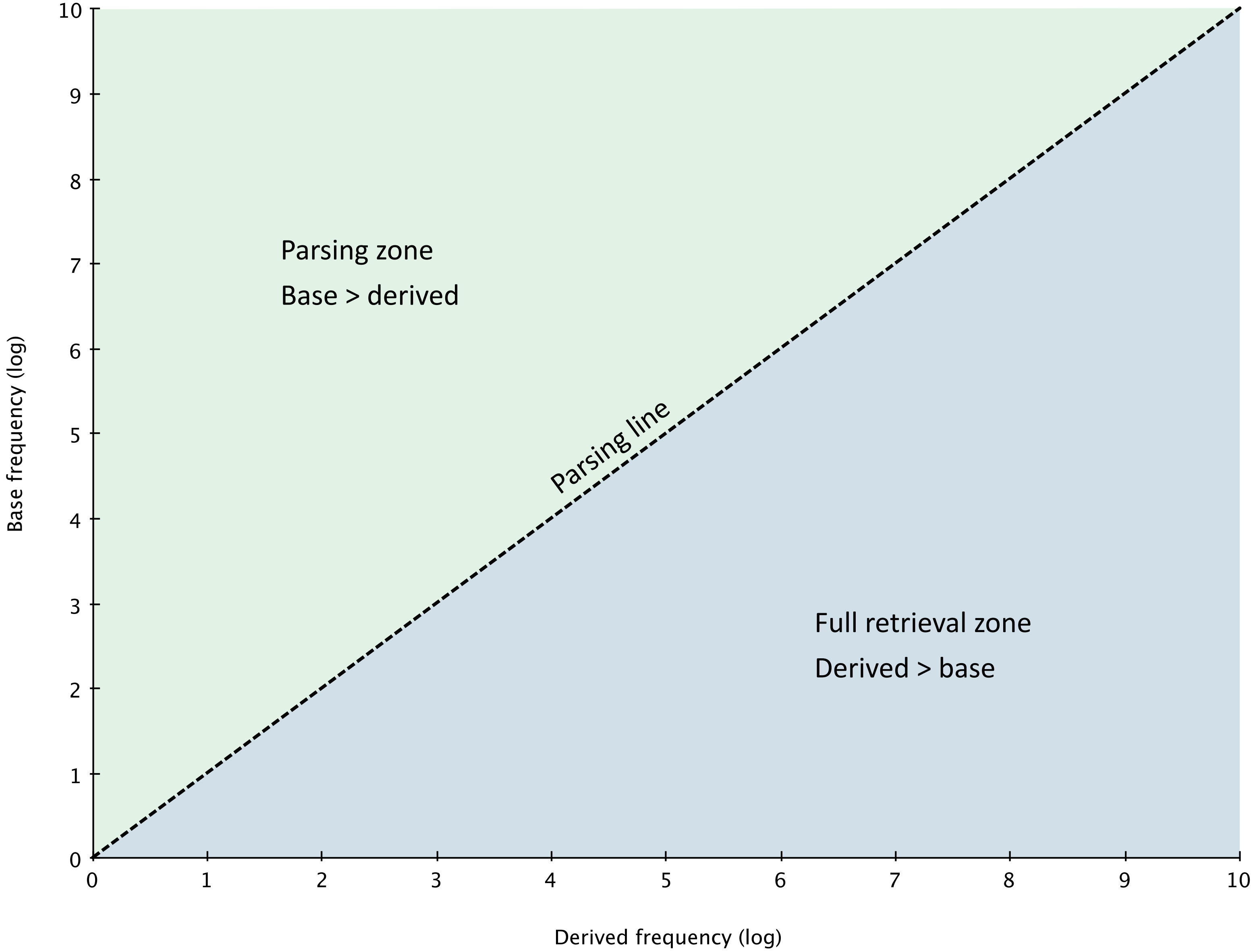
  - words which are likely to be parsed

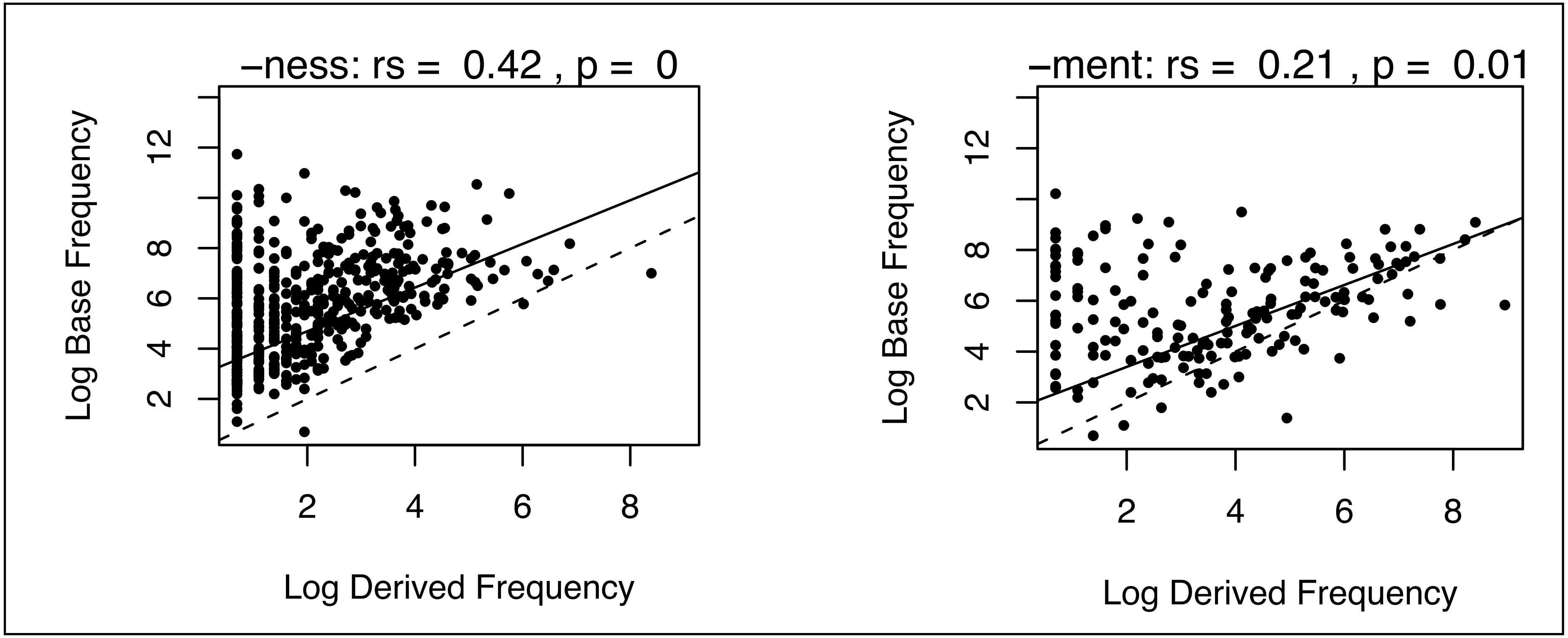
    *arrestment*, *dazzlement*

  - words which are on the line

    *argument*, *assessment*

# Productivity





−ness: rs =  0.42 , p =  0

−ment: rs =  0.21 , p =  0.01

# Productivity

- The **parsing ratio** of an affix is the proportion of words with that affix that are above the parsing line (and so are parsed rather than retrieved whole)
- Both the **type** parsing ratio and the **token** parsing ratio are relevant to productivity
- Parsing ratios correlate closely with Baayen’s category conditioned degree of productivity *P* (i.e., the chance that a particular affix token is a hapax)
- The total number of forms above the parsing line correlates well with an affix’s type frequency and with the number of hapaxes

# Productivity

- The number of forms with an affix that get parsed is a good measure of that affix's activation

- High activation affixes are more likely to be salient and productive

- Words with high activation affixes are less likely to be irregular (semantically or otherwise)

# Competition

- If both singular and plural forms are stored, then they should compete, slowing down recognition
- Kemps et al. (2005) on Dutch *boek* [buk], *boeken* [bukə]
- Experiment 1
    - Number recognition task with real plurals, real singulars, fake singulars (truncated plurals)
    - Real singulars were significantly longer and lower in pitch than fake singulars
    - RTs were slower for fake singulars than for real singulars, and the delay varied with pitch and length

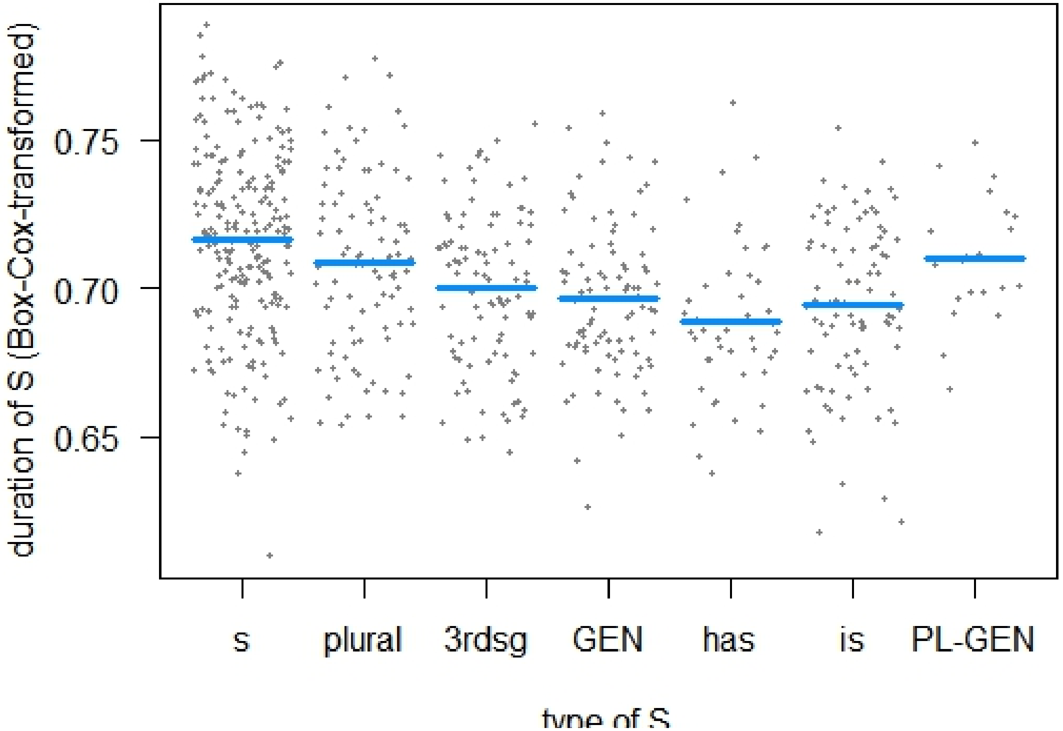
# Competition

- Experiment 2
  - Number recognition task with real singulars, spliced plurals (with plural stems), and spliced plurals (with singular stems)
  - RTs were slower for mismatched plurals than for non-mismatched plurals
- Other experiments artificially manipulated intonation and length, and again conflicting cues slowed reaction times

# Sub-phonemic variation

- Plag, et al. (2014) look at final *-s* and *-z* in the Buckeye Corpus

- Absolute and relative duration, controlling for voicing, phonetic environment, etc.



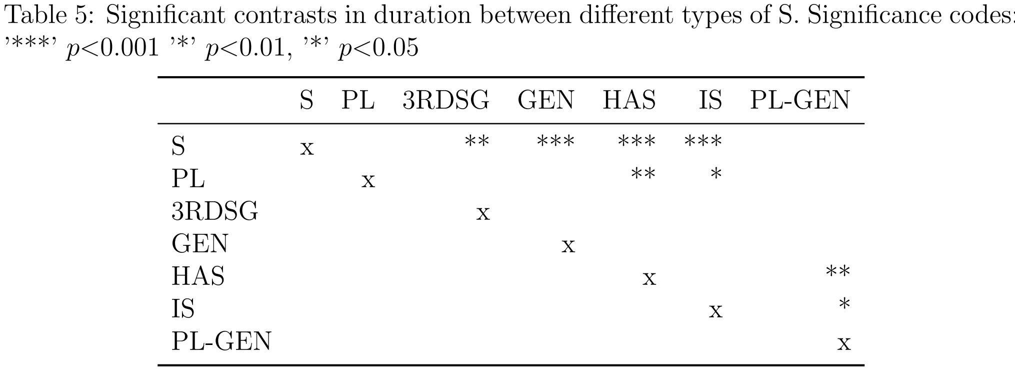
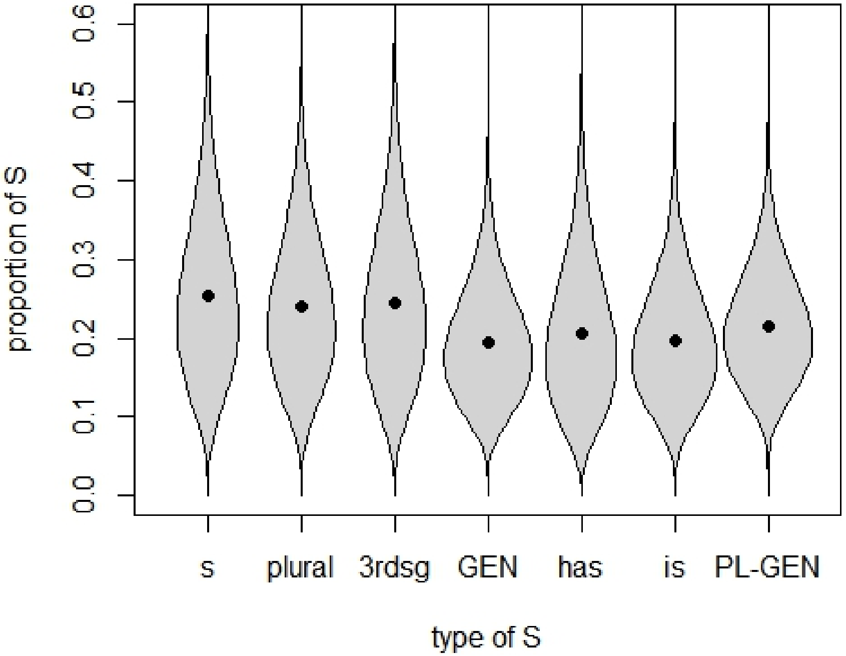


Table 5: Significant contrasts in duration between different types of S. Significance codes
’***’ $p$<0.001 ’*’ $p$<0.01, ’*’ $p$<0.05

| | S | PL | 3RDSG | GEN | HAS | IS | PL-GEN |
|---|---|---|---|---|---|---|---|
| S | x | | ** | *** | *** | *** | |
| PL | | x | | | ** | * | |
| 3RDSG | | | x | | | | |
| GEN | | | | x | | | |
| HAS | | | | | x | | ** |
| IS | | | | | | x | * |
| PL-GEN | | | | | | | x |

# Sub-phonemic variation

- Plag, et al. (2014)



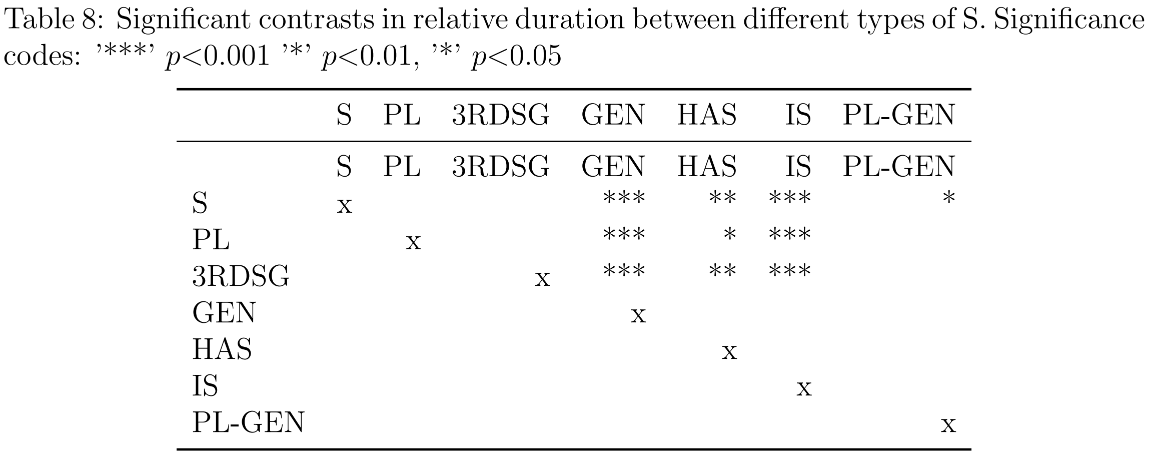


Table 8: Significant contrasts in relative duration between different types of S. Significance codes: ’***’ $p$<0.001 ’*’ $p$<0.01, ’*’ $p$<0.05

| | S | PL | 3RDSG | GEN | HAS | IS | PL-GEN |
|---|---|---|---|---|---|---|---|
| | S | PL | 3RDSG | GEN | HAS | IS | PL-GEN |
| S | x | | | *** | ** | *** | * |
| PL | | x | | *** | * | *** | |
| 3RDSG | | | x | *** | ** | *** | |
| GEN | | | | x | | | |
| HAS | | | | | x | | |
| IS | | | | | | x | |
| PL-GEN | | | | | | | x |

- Morphemic -s/-z is phonetically different from non-morphemic -s/-z

# Paradigm uniformity

- Seyfarth, et al. (2017): “Paradigm uniformity is a pressure for invariance among the phonological forms of an inflectional or derivational paradigm”

- In USian English, unstressed /tə/ usually becomes [ɾə]

*capitalistic* [ˌkæpɪ**ɾə**ˈlɪstɪk]    *capital* /ˈkæpɪ**tl̩**/

- But!

*militaristic* [ˌmɪlɪ**tʰə**ˈɹɪstɪk]    *military* /ˈmɪlɪˌ**tɛ**ɹi/

# Paradigm uniformity

- Inflected forms vs. simple homophones
- *-s/-z*

  *If we **freeze** it, it should be fine.*
  *If he **frees** it, it won’t survive.*

- *-t/-d*

  *Yeah, they made a **pact** for their trip.*
  *Yeah, they had it **packed** for their trip.*

# Paradigm uniformity

- A predictions:
  - The duration of the nucleus is longer in open syllable *free* [fɹi]
  - Paradigm uniformity leads us to expect nucleus in *frees* to be longer than in *freeze*

# Paradigmatic uniformity

- Seyfarth, et al. (2017)

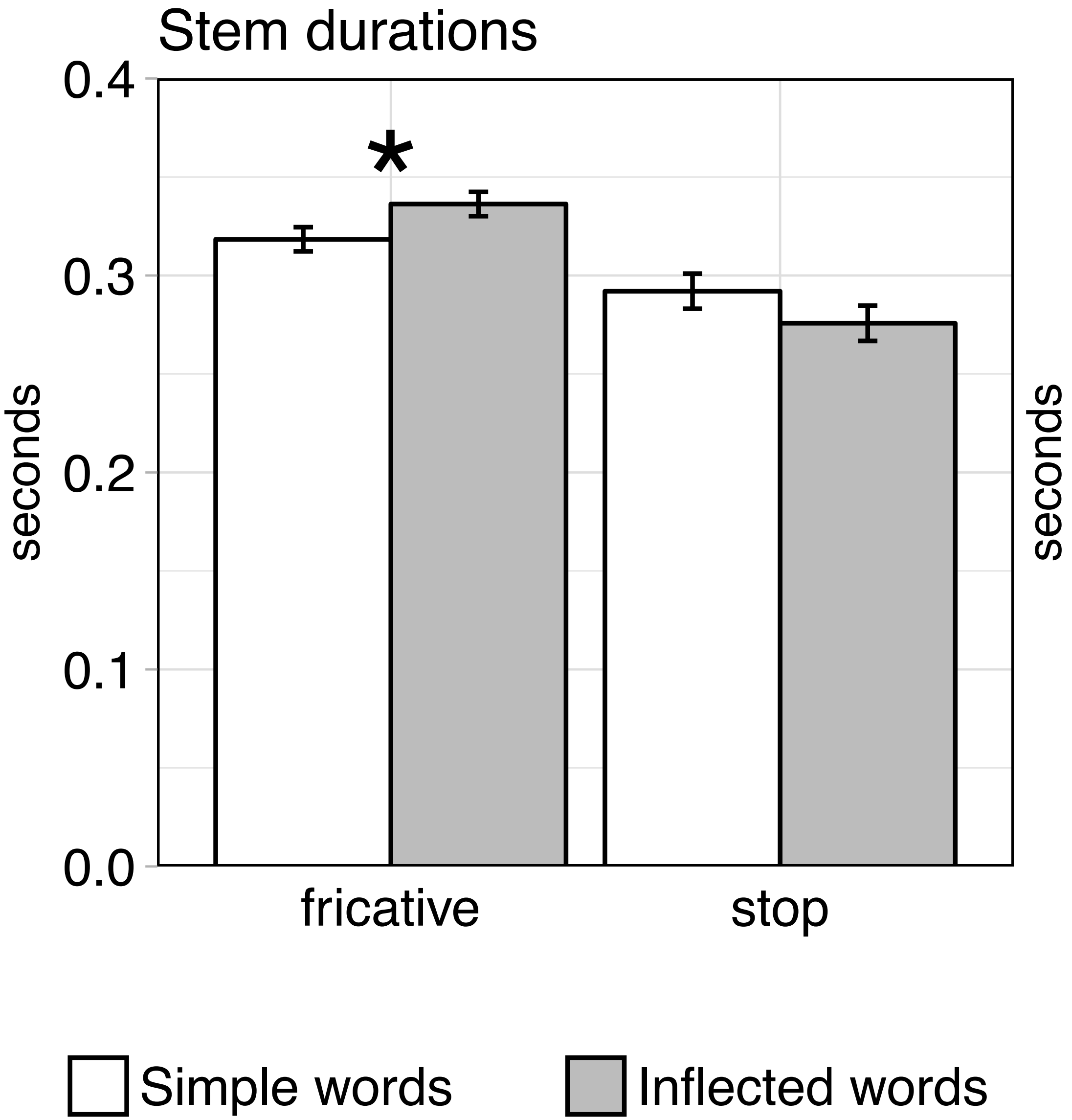
  *Two housemates are wrapping up a surprise birthday party that they put on for a friend.*  
  B: It looks like most people are leaving now. I guess I'm going to start cleaning up a little bit.  
  A: There's so much cake leftover. I don't want it to go bad.  
  B: If we freeze it, it should be fine.

# Paradigm uniformity

- Seyfarth, et al. (2017)

  *Two rural neighbors are talking about a friend, Rich, who is an avid hiker and animal-lover.*  
  B: Rich decided to take care of the injured hawk that he found yesterday.  
  A: They don’t do well in captivity. Wouldn’t it be better to let it go?  
  B: If he frees it, it won’t survive.

# Paradigm uniformity



# Linguistic morphology in the 2020’s

- Can morphology survive if we can’t first define what a word is?
- Well, can biology survive if we can’t first define what life is?
- Descriptive and pedagogical applications (dictionaries, orthographies)
- Certain kinds of questions are natural to ask when we think about words
    - grammaticalization
    - entrenchment
    - paradigmatic organization