10/16/2019

Lexical flexibility in English: A preliminary study

Daniel W. Hieber
University of California, Santa Barbara

danielhieber.com

College of William & Mary, Department of Linguistics, Williamsburg, VA

UC SANTA BARBARA

What part of speech is *friend*?

- Noun
- Verb
- Adjective

UC SANTA BARBARA

[Take a poll]

friend as Noun

- *I got a spooky box from my best **friends**.*
- *Secrets don't make **friends**, Luke.*
- *Just think I saw an old college **friend** on TV meeting Hilary Clinton.*

UC SANTA BARBARA

Twitter data from W&M students

friend as Verb

- *What's your user? I would love to **friend** you and look at it when finished!*
- *If we don't have mutual friends we can't get **friended**.*
- *I accidentally downloaded Facebook and created a profile and **friended** a bunch of people.*

friend as Adjective

- *the guy became the national symbol of **friend** zone in just a day*
- *Facebook just put me in the damn **friend** zone with my wife*
- *can someone help me with some **friend** drama?*

UC SANTA BARBARA

What does the dictionary say?

- Dictionary.com: verb, noun
- Merriam-Webster: verb, noun
- Why not adjectives?

UC SANTA BARBARA

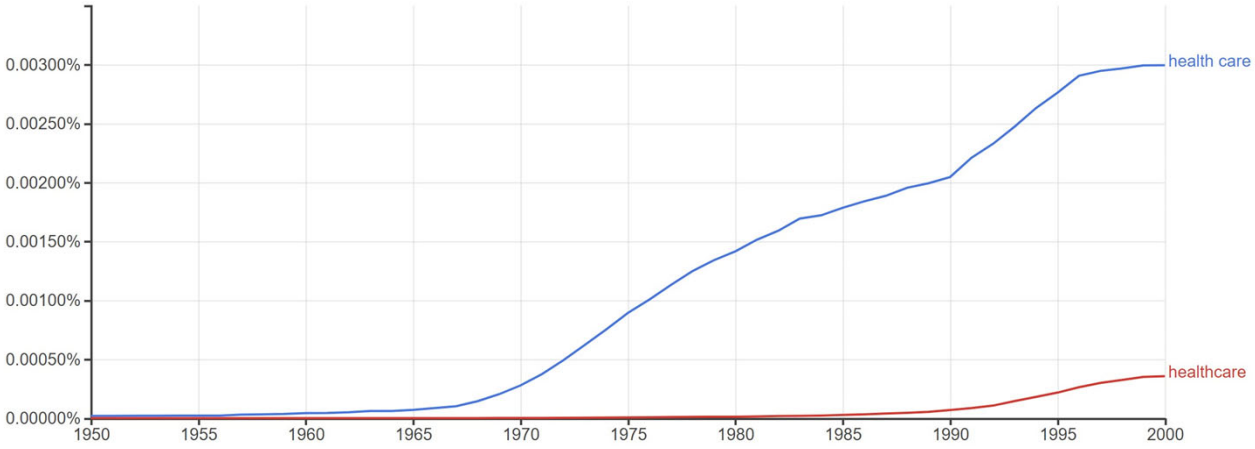
Nouns Modifying Nouns

- Are they compounds?
  - *health care* vs. *healthcare*
  - *friend zone* vs. *friendzone*

UC SANTA BARBARA

Tempted to analyze nouns modifying nouns as compounds. Does this work?

health care vs. *healthcare* (Google Books)



UC SANTA BARBARA

In some cases, nouns modifying nouns do *become* compounds.

But not in every case.

Nouns Modifying Nouns

- Are they compounds?
  - *health care* vs. *health care*
  - *friend zone* vs. *friendzone*
- We don't analyze these as adjectives because:
  - tradition
  - the change is unmarked

UC SANTA BARBARA

friendzone doesn't (yet) appear in the Google Books corpus

Linguists selective about their criteria

Cherry pick to accommodate:
- tradition
- their theoretical perspective

Aside

- *the truth is they **friendzone** everyone who tries to be with them*
- *just ate two slices of veggie pizza for lunch so basically I'm all **healthed** up for at least a month*

UC SANTA BARBARA

friendzone – *does* appear on Twitter; entire phrase can become *lexicalized* (a new meaningful word in itself)

healthed – very unexpected use of this word as a verb

able

- **N:** *that feeling of* ***abling*** *to run 22 miles a week*
- **V:** *always* ***abling*** *and abetting the horses*
- **A:** *an* ***able*** *mind overcomes challenges*

UC SANTA BARBARA

time

- **N:** *still one of my favorite series of all* ***time***
- **V:** *I'm so bored in this class that I'm* ***timing*** *how long I can hold my breath*
- **A:** *2 years ago today (or yesterday depending on your* ***time*** *zone)*

UC SANTA BARBARA

Parts of Speech in English

- How common is flexibility in English?
- *rigid* vs. *flexible* words

UC SANTA BARBARA

lexical vs. *grammatical* words

English is sometimes claimed to be rigid, sometimes flexible

A crosslinguistic problem

Nuuchahnulth (Wakashan; Pacific Northwest)

1. mamuːk-ma **quːʔas**-ʔi
   working-PRES(INDIC) man-DEF
   'the man is working'

2. **quːʔas**-ma mamuːk-ʔi
   man-PRES(INDIC) working-DEF
   'the working one is a man'

UC SANTA BARBARA

Nuuchahnulth (Wakashan; Pacific Northwest)

1. **mamuːk**-ma qu:ʔas-ʔi
   working-PRES(INDIC) man-DEF
   'the man is working'

2. qu:ʔas-ma **mamuːk**-ʔi
   man-PRES(INDIC) working-DEF
   'the working one is a man'

UC SANTA BARBARA

Flexibility is present for both words

Central Alaskan Yup'ik (Eskimo-Aleut)

3.	*angya-**qa***	'my boat'
	*ner'a-**qa***	'I am eating it'
	*angya-**a***	'his/her boat'
	*ner'a-**a***	'he/she/it is eating it'
	*angya-**at***	'their boat'
	*nera-**at***	'they are eating it'

UC SANTA BARBARA

Entire paradigm matches

Central Alaskan Yup'ik

"In the Eskimo mind the line of demarcation between the noun and the verb seems to be extremely vague, as appears from the whole structure of the language, and from the fact that the inflectional endings are, partially at any rate, the same for both nouns and verbs." (p. 1057)

Thalbitzer, W. 1911. Eskimo. In Franz Boas (ed.), *Handbook of American Indian Languages* (Bureau of American Ethnology Bulletin 40), 967–1069.

Riau Indonesian (Austronesian)

4.	*ayam*	*makan*	• The chicken is eating.
	chicken	eat	• The chicken is being eaten.
			• The chicken is making somebody eat.
			• Somebody is eating for the chicken.
			• Somebody is eating where the chicken is.
			• the chicken that is eating
			• where the chicken is eating
			• when the chicken is eating
			• how the chicken is eating

UC SANTA BARBARA

Famously claimed by David Gil to lack parts of speech entirely.

Mundari (Austroasiatic)

5. **buru** = ko bai-ke-d-a
   mountain = 3PL.S make-COMPL-TR-INDIC
   'They made the mountain.'

6. saan = ko **buru**-ke-d-a
   firewood = 3PL.S **mountain**-COMPL-TR-INDIC
   'They heaped up the firewood.'

Flexibility with Fully-Inflected Words

7. **Chitimacha** (isolate; Louisiana)
   dzampuyna
   'they usually thrust/spear (with it)' = 'a spear'

8. **Mohawk** (Iroquoian; Ontario / Quebec)
   ieráhkhwa'
   'one puts things in with it' = 'a container'

UC SANTA BARBARA

Mohawk: verbs show a cline from fully lexicalized to fully productive / analyzable; some words have both uses

I have yet to find a language where flexibility hasn't been observed in sufficiently great amounts that it merits comment in the literature or a grammatical description.

The Problem of Lexical Flexibility

How to analyze lexical flexibility?

- *conversion / zero-derivation* vs. *underspecification*
- *lexical flexibility* is used in a neutral sense here

UC SANTA BARBARA

conversion – traditional approach, favored by generativists / formalists (exception: Distributed Morphology)

underspecification – newer approach, gradually gaining proponents

What is a word?

- *lexeme*
  - abstract representation (cognitive or grammatical) of a group of related *wordforms*
  - whatever it is that's common to those wordforms (usually a stem)
- *lemma* – conventional wordform used to represent a group of wordforms
  - keyword: *conventional*
- *token* – a specific instance of a lexeme in discourse

UC SANTA BARBARA

Determining what two uses of a form count as the same lexeme is tricky.

What we're interested in when we're talking about what counts as instances of the same "word" is actually a *lexeme*.

lexeme – abstract representation (cognitive or grammatical) of related *wordforms*; whatever it is that's common to those wordforms (usually a stem)
- example: *help*, *helps*, *helped*
- example: *eat*, *ate*, *eaten*
- example: *am, is, are, was, were, be*

lemma / *headword* – conventional wordform used to represent this bundle; just a matter of convention

How to determine wordhood?

- words have many senses

UC SANTA BARBARA

If a word has many different senses, where do we draw the line between one lexeme and the next?

Senses of *run*

- Dictionary.com lists 148 senses of *run*, some nouns, some verbs (but again no adjectives)

• fast pedestrian motion:	*I run every day*
• conduct a political campaign:	*he ran a fair campaign*
• come undone:	*these stockings run easily*
• operate or function:	*does it run well?*
• get or become:	*the well ran dry*

UC SANTA BARBARA

Should we count all of these as the same "word" / lexeme? Where do we draw the line?

How to determine wordhood?

- words have many senses
- grammatical categories vs. cognitive associations
- categories are prototypal
- derived words have unpredictable meanings

UC SANTA BARBARA

Cognitive literature suggests that we have cognitive associations between historically related or synchronically similar wordforms, even if they're totally different lexemes.
- response times
- priming effects

We *do* have some association between the many senses of *run* – probably a family network.

prototypal – They cluster around a prototype
- prototypical noun: *man*
- non-prototypical noun: *running*

Prototypicality established through:
- listing experiments
- response / recall time
- corpus frequency
- historical primacy (usually)

Predictability – since the meaning has changed (enough), it must be a new word
- BUT, some languages have cases of conversion which are predictable as well as cases

which are not (Mandinka) (probably most languages)

Semantic Predictability

- *brother* vs. *brethren*
- *cloth* vs. *clothes*
- *new* vs. *news*
- (hunting) *blind* vs. (window) *blinds*
- *custom* vs. *customs*
- *arm* vs. *arms*
- *wood* vs. *woods*

UC SANTA BARBARA

Inflection also can create a significant shift in meaning

brother, *cloth* – historical divergence
blind – window interpretation not available in the singular
custom – international travel sense not available in the singular
arm – military force sense not available in the singular
wood – singular and plural refer to different types of things (a material vs. a collection)

Semantic Predictability

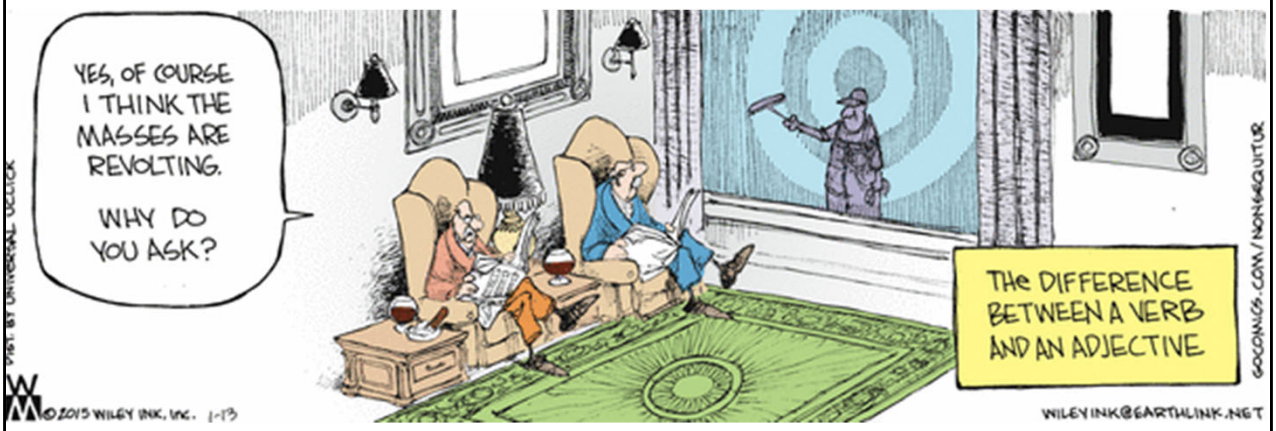
- inflectional vs. derivational uses of the same morpheme
- English *–ing* progressive / gerund
  - *the running man* (inflectional)
  - *the running of the bulls* (derivational)
- Chitimacha *–ma* pluractional
  - *guxma-* 'eat (multiple things)' (inflectional)
  - *haakxtema-* 'design' (from *haakxte-* 'draw') (derivational)

UC SANTA BARBARA

Can't even be sure when a morpheme is acting inflectionally vs. derivationally

That is, we don't know when it becomes a new word

English –*ing*: Inflection vs. Derivation



UC SANTA BARBARA

Note the caption here: *The difference between a verb and an adjective*

~~How to determine wordhood?~~

- bad question
- good questions
  - How common is flexibility / unmarked derivation?
  - Does the frequency / degree of flexibility vary from word to word?
  - Does the frequency / degree of flexibility vary from language to language?
  - Why are some instances of derivation marked, and others not? What's special about the (un)marked ones?

UC SANTA BARBARA

Given what we (don't) know about lexical categories, I think this is an unhelpful question.

We know lexical relatedness is gradient and complex. Can we say something about it anyway?

We should treat lexical flexibility as an object of study in its own right, without assuming anything about the relatedness of different uses of a word.

Yes. These are my long-term research questions. This research project aims to tackle just the first question.

This Study

Research Questions

1. How flexible are the words of English, and English generally?
2. Does flexibility correlate with semantic domain?

UC SANTA BARBARA

Question 2 is an initial attempt to determine what might motivate unmarked conversion.

Determining Degree of Flexibility

1. For a given word, count how often that word is used as a noun, verb, or adjective.

2. Calculate a flexibility score for that word – how evenly distributed its uses are across different categories.

3. Apply this method to each word in the language (or a representative sample of them)

Data & Methods

- Spoken portion of the Open American National Corpus (OANC) (~3.5 million words) (not Twitter)
- Randomly selected wordforms from 100 different frequency bins
- Created a list of every instance of those 100 words (~380,000 tokens total)
- Annotated each token for its function: noun, verb, adjective

UC SANTA BARBARA

frequency vs. corpus dispersion – [mention if you have some time to fill]

Annotated 16 out of the 100 lexemes completely so far

Results

able

- **N:** [none]
- **V:** *Are you **able**?*
- **A:** *most of the **able** bodied Americans*

UC SANTA BARBARA

Noun: We saw a Twitter example earlier, but none appear in the OANC. The form *abling* doesn't appear once.

- Notice that there's already a marked derivation for this: *ability*
- This phenomenon is sometimes called *blocking*, though it's unclear if this is actually what's happening here

Vast majority of instances of *able* are attributive predicates, which is interesting because *able* is historically an adjective

Verbal uses are due to copula constructions, which are structurally equivalent to inflected verbs

- *I am ahead*
- *I am running*

Almost anything can be predicated in English, unmarked

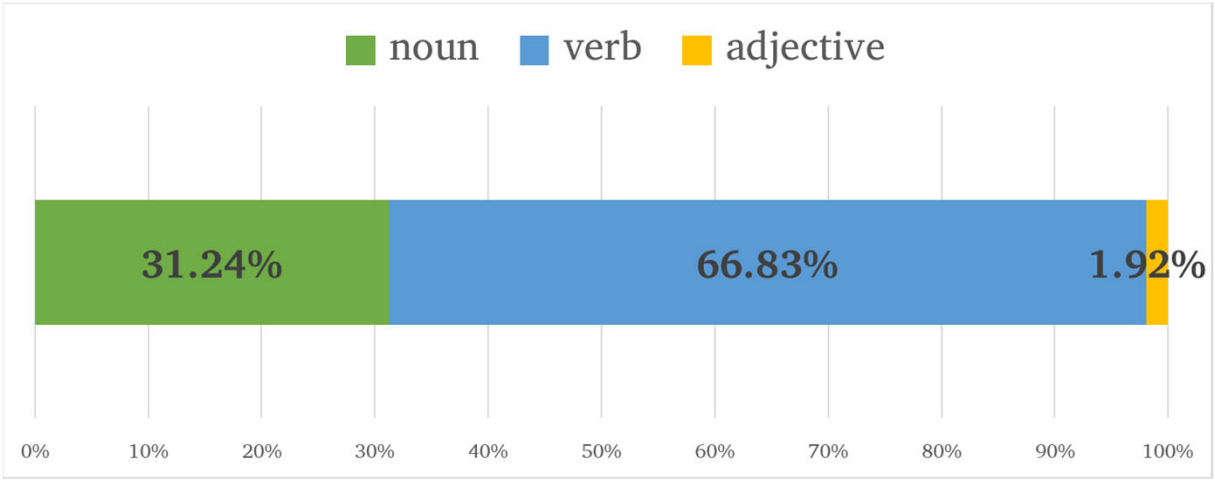
Omnipredicativity – originally proposed for Nahuatl (Aztec)

- appears to be a prevalent feature of all, possibly most, languages

Omnipredicativity

UC SANTA BARBARA

able



UC SANTA BARBARA

ahead

- **N:** [none]
- **V:** *I'm **ahead** of him right now*
- **A:** [none]

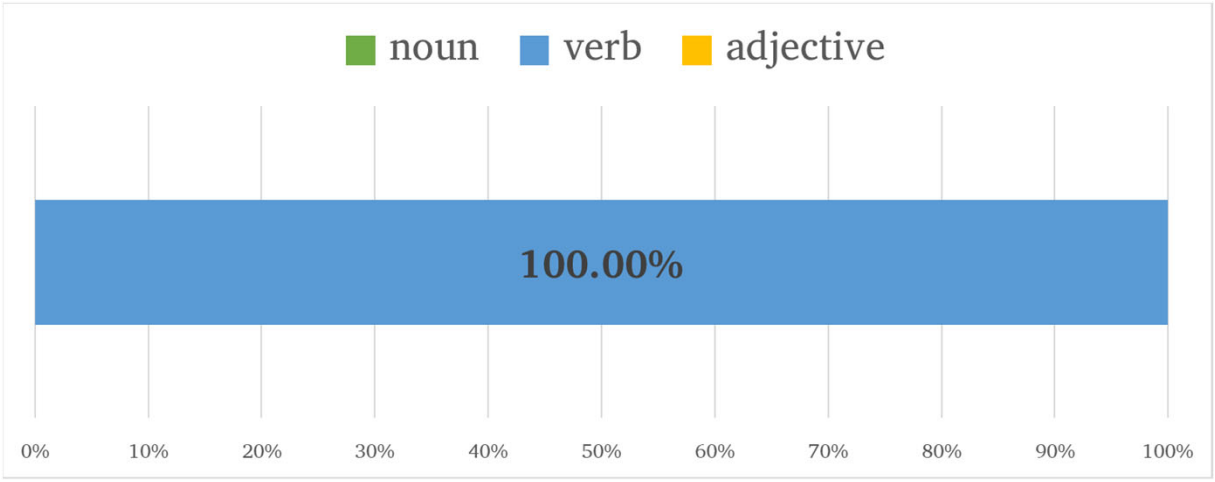
UC SANTA BARBARA

Adverbs like *ahead* were generally tricky. Frequently they seem like nouns *functioning* adverbially or *functioning* to modify. Historically they're often nouns or locative phrases (*at head*)

- *I got ahead of him* (reference)
- *the next guy ahead of me* (modification)

Didn't count cases like these unless they were really clear, but it makes me think adverbs are another area where we're adhering to traditional ways of analyzing terms even when they aren't appropriate to the actual data.

ahead



UC SANTA BARBARA

anything

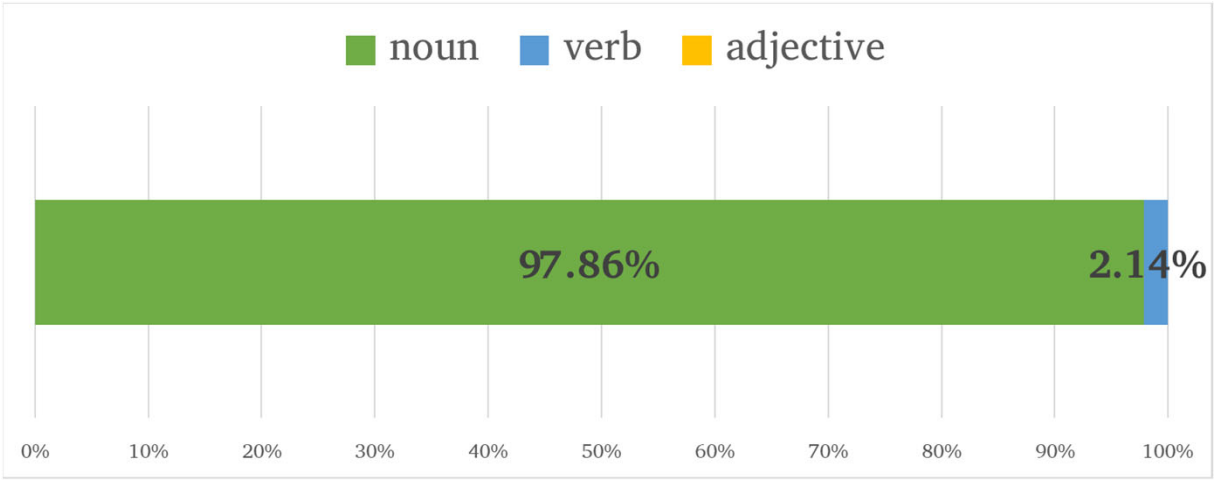
- **N:** *I was never exposed to **anything** of the sort*
- **V:** *it's **anything** in that hobby line*
- **A:** [none]

UC SANTA BARBARA

Not surprising. However it should be noted that you can find modifying and verbal examples online, not just for *anything*, but for just about anything!

V: *She loads me down with goodies that she searches out as not being sprayed, shot, or artificially **anythinged***
A: *that wasn't very country or **very anything**, really*

anything



UC SANTA BARBARA

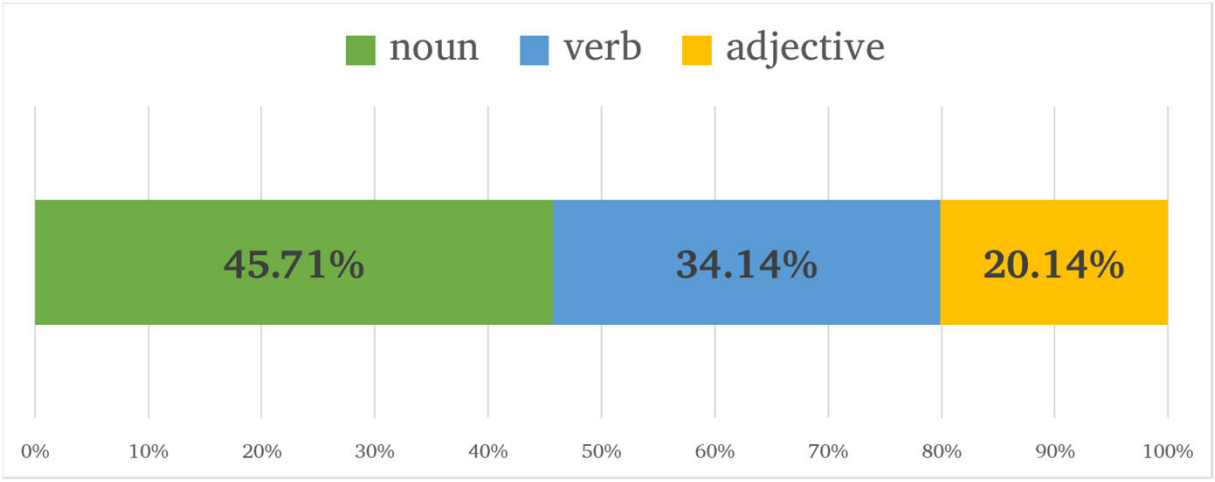
back

- **N:** *hand print on the* ***back*** *of her leg*
- **V:** *as I'm* ***backing*** *off I'm still keeping an eye on it*
- **A:** *when I look out my* ***back*** *door*

UC SANTA BARBARA

Very easy to find instances of all three functions for *back*, even ignore attributive predicative cases

back



UC SANTA BARBARA

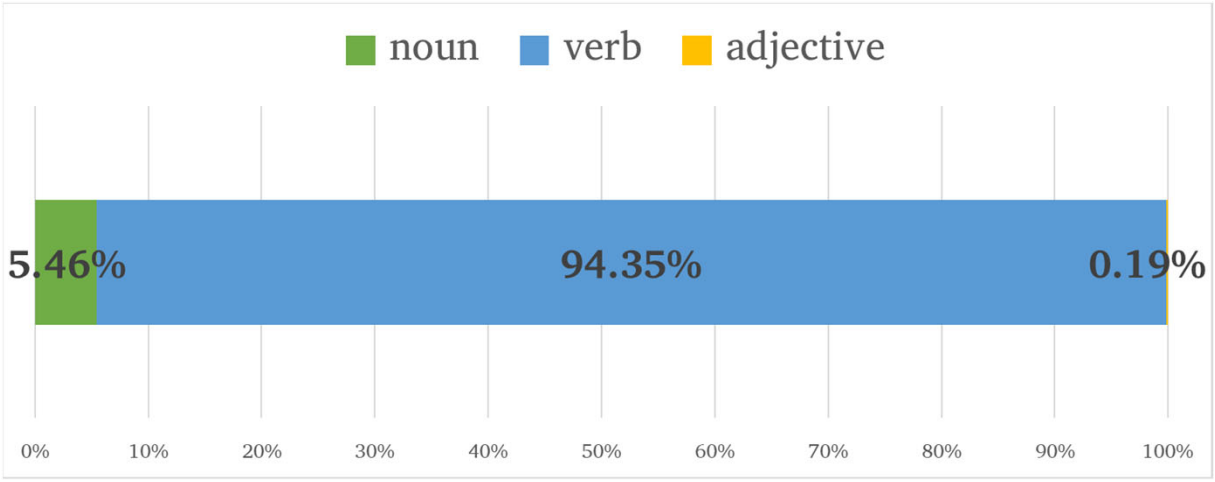
believe

- **N:** *I don't have any choice but* ***to believe*** *it*
- **N:** *all those feelings of* ***believing***
- **V:** *I don't* ***believe*** *she read a lot*
- **A:** *the* ***believing*** *scientist*

UC SANTA BARBARA

N: infinitives are traditionally analyzed as a verbal inflection, but infinitives are typologically noun-like, and they're often considered a nominal form of a verb

believe



UC SANTA BARBARA

best

- **N:** *summer is the* ***best***
- **V:** *the new crew was* ***best***
- **A:** *he is one of the* ***best*** *actors*

UC SANTA BARBARA

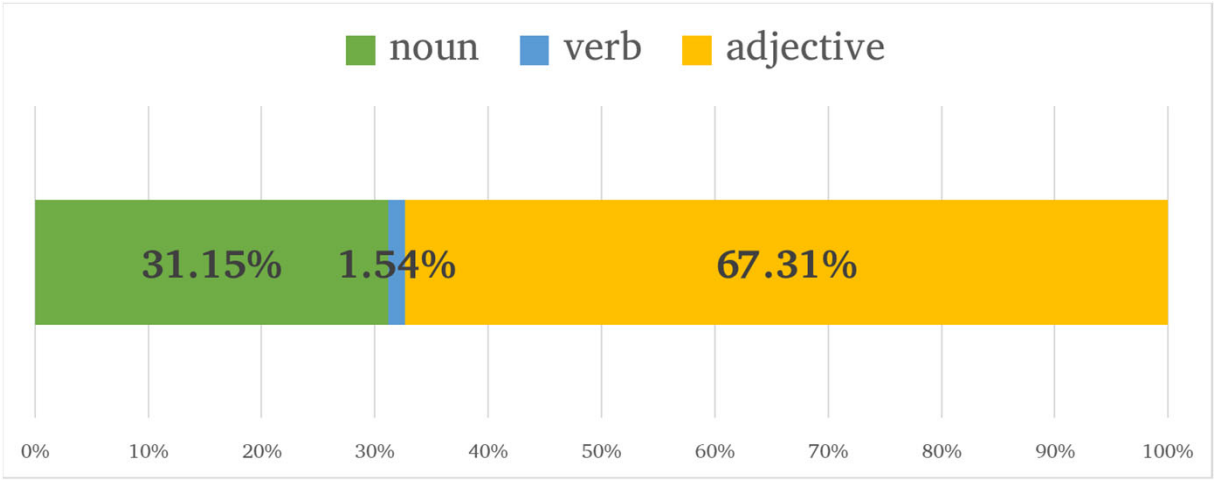
For the attributive predicative uses, I'd like to go back and recode them as a distinct category.

V: could also have been *to best someone*, but that use didn't appear in the OANC

Future research question: Can we determine the prototypical use of a word by the distribution of its functions?

- adjective: primarily adjective, some noun
- noun: primarily noun, some verb, some adjective
- verb: primarily verb, some noun, little to no adjective

best



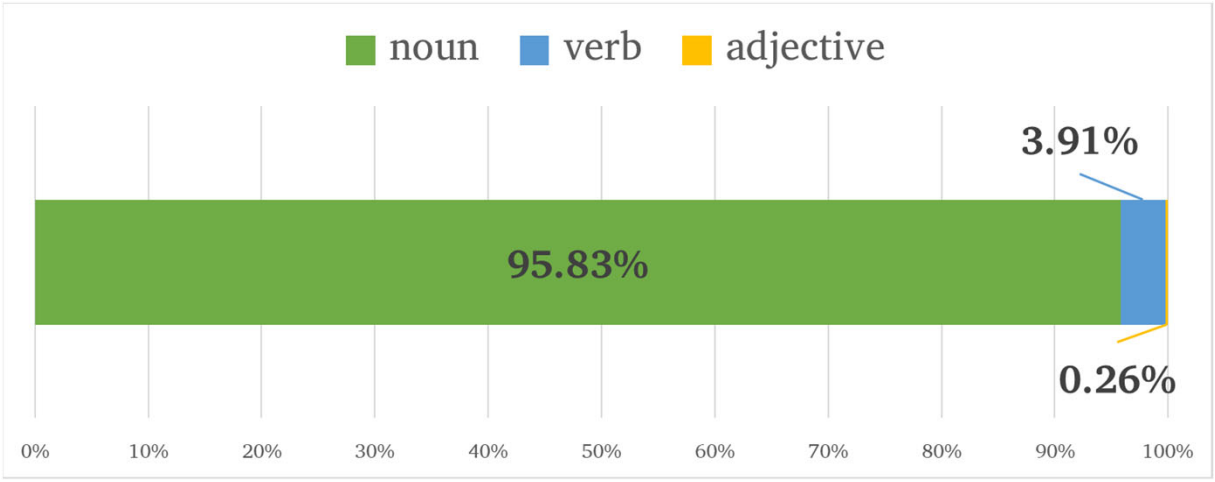
UC SANTA BARBARA

bill

- **N:** *the **bill** always comes in*
- **V:** *they could **bill** Uncle Sam for that hospital care*
- **A:** ***bills** wise we divide everything*

UC SANTA BARBARA

bill



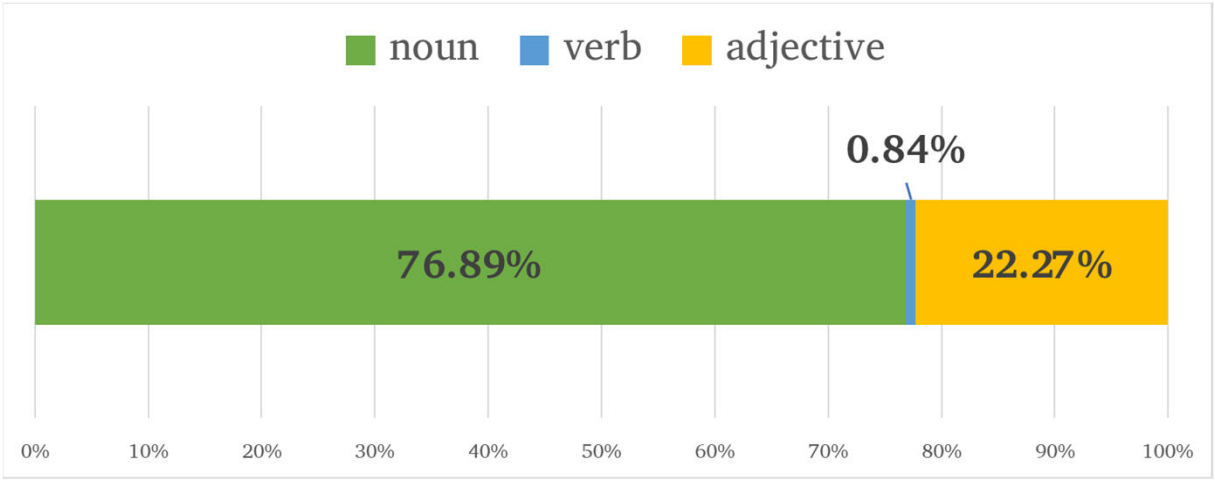
UC SANTA BARBARA

business

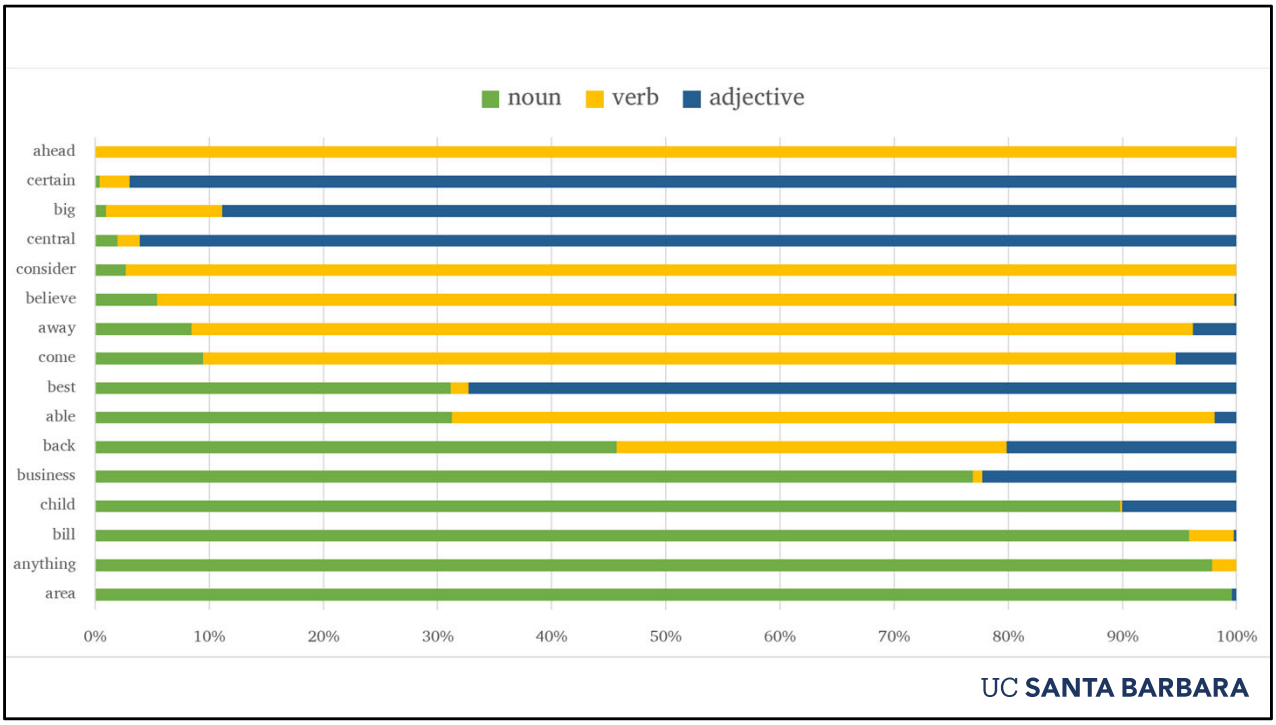
- **N:** *we were in the retail milk* ***business***
- **V:** *it's* ***business*** *and it's serious*
- **A:** *here's my* ***business*** *card*

UC SANTA BARBARA

business



UC SANTA BARBARA

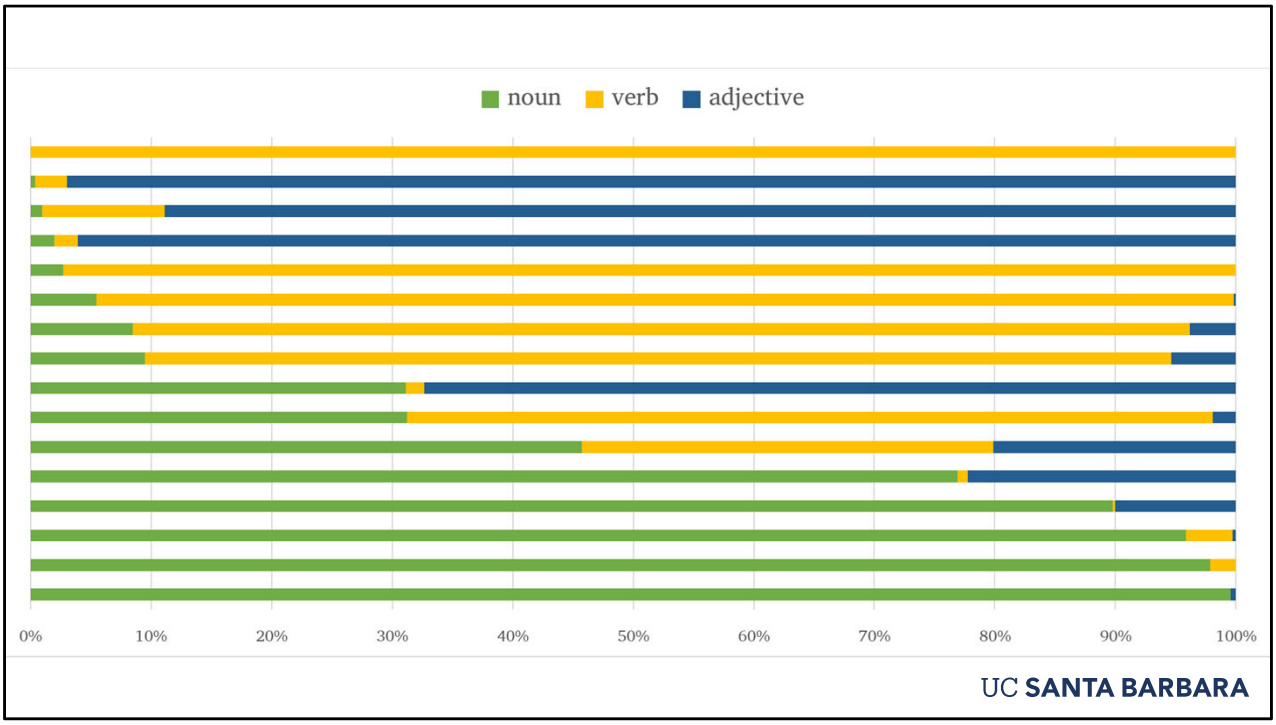


Preliminary Results for English

Most words of English do not exhibit much flexibility – one function predominates
- The results are a little boring! But that in itself is interesting!
- This says something about linguists' perception of English as a flexible language
- Linguists' perceptions seem to be based on striking, standout cases rather than actual data

(Almost) all words of English exhibit *some* flexibility
The only word which is 100% consistent in its distribution is *ahead*, which is typically thought to be an adverb!
- Results would probably look very different if I included adverbial uses
- verbs have nominal forms by default: verb + noun flexibility
- anything can be predicated using a copula construction: omnipredicativity

back is the most evenly distributed between the three functions
- Are body part terms more flexible than other semantic domains? Why?
- Potential answer: The wide range of spatial and instrumental metaphors that body part terms are used for



The categories seem to be gradient – most words are not clear-cut.

Which of these are nouns? Verbs? Adjectives?

Preliminary Results from English

- Most words of English are not especially flexible
  - One function tends to predominate for any given word
- All (?) words of English exhibit *some* flexibility
- Possible blocking effects (e.g. *ability* ↛ *the able*)
- Body part terms may exhibit more flexibility than other semantic domains

UC SANTA BARBARA

Next Steps

- Add data from Nuuchahnulth (and other languages)
- Annotate more than 100 words per language
- Code data for semantic domain, especially body part terms
- Investigate historical development of flexible uses
- Investigate correlations between frequency and flexibility

UC SANTA BARBARA

Diachrony
- Specific *senses* of a word jump the POS boundary
- Consider *friend*: When used as a verb, it refers specifically to social media
- Not all of the senses of *friend* immediately jumped the POS boundary along with this sense

Thanks!

Discussion: How I got interested in this topic

- POS tagging English for Rosetta Stone
- lexical categories course with Elaine Francis at LSA Institute 2011
- I don't typically work with English – this is just a baseline for work with other languages