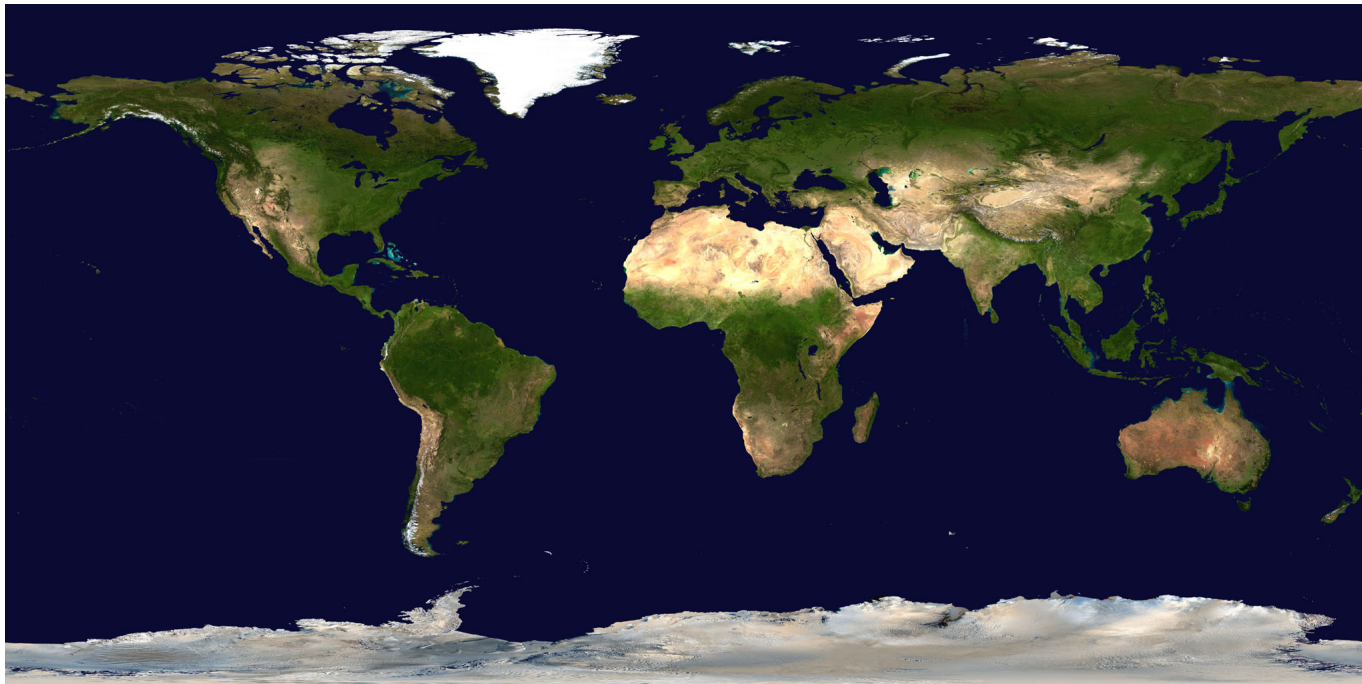


Computational models for comparative and historical linguistics

Jelle Zuidema



Principle(s) of Compositionality

The meaning of a sentence, is a (systematic) function of the meaning of its parts

(Gotlob Frege, 1892, Richard Montague, 1973)

E.g. “The dog that chases the cat that chases the mouse”

Natural Languages:

- Morphology
- Constituent order / Phrase structure

Isolating languages

- No morphological variation for tense, case or plurality.
- Each word typically consists of a single morpheme.

E.g. Vietnamese:

Khi tôi đến nhà bạn tôi, chúng tôi bắt đầu làm bài.
when I come house friend I PLURAL I begin do lesson
“When I came to my friend’s house, we began to do lessons.”

Agglutinating languages

- A word may consist of more than one morpheme
- Boundaries between words are always clear-cut
- A morpheme has a reasonably invariant shape

E.g. Turkish (*adam*: “man”):

	singular	plural
nominative	<i>adam</i>	<i>adam-lar</i>
accusative	<i>adam-i</i>	<i>adam-lar-i</i>
genitive	<i>adam-in</i>	<i>adam-lar-in</i>
dative	<i>adam-a</i>	<i>adam-lar-a</i>
locative	<i>adam-da</i>	<i>adam-lar-da</i>
ablative	<i>adam-dan</i>	<i>adam-lar-dan</i>

Fusional languages

- No clear-cut boundary between morphemes
- Expression of different categories within the same word is fused together to give a single, unsegmentable morph.

E.g. Russian (*stol*: “table”, *lipa*: “lime-tree”):

	singular I	plural I	singular II	plural II
nominative	<i>stol</i>	<i>stol-y</i>	<i>lip-a</i>	<i>lip-y</i>
accusative	<i>stol</i>	<i>stol-y</i>	<i>lip-u</i>	<i>lip-y</i>
genitive	<i>stol-a</i>	<i>stol-ov</i>	<i>lip-y</i>	<i>lip</i>
dative	<i>stol-u</i>	<i>stol-am</i>	<i>lipea</i>	<i>lip-am</i>
instrumental	<i>stol-om</i>	<i>stol-ami</i>	<i>lip-oj</i>	<i>lip-ami</i>
prepositional	<i>stol-e</i>	<i>stol-ax</i>	<i>lip-e</i>	<i>lip-ax</i>

Polysynthetic languages

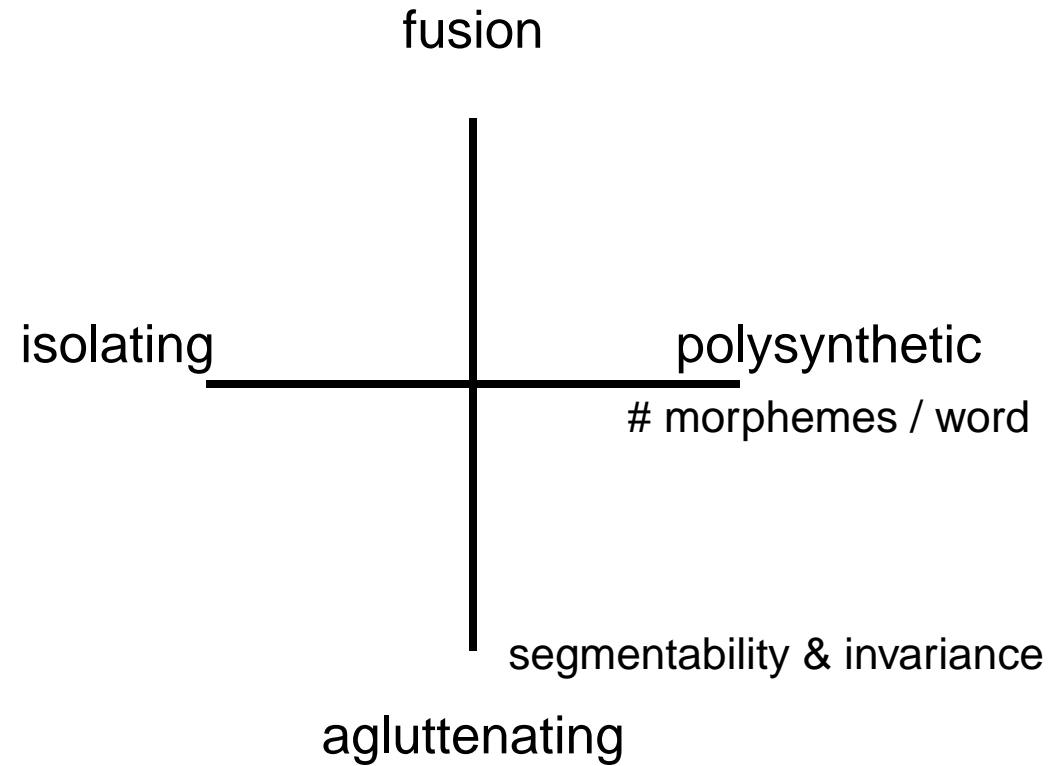
- Many lexical morphemes combined in a single word
- E.g. Chukchi (Siberia): *temeyηεlevtepeγterken*

te- *meyηε-* *levte-* *peγt-* *erken*
great head ache 1stSINGULAR IMPERFECT
“I have a fierce head-ache”

- Not necessarily incorporating. E.g. Eskimo (Siberian Yupik):
angyaghllangyugtuq

angya- *ghlla-* *ng-* *yug-* *tuq*
boat AUGMENTATIVE ACQUIRE DESIDERATIVE 3dSINGULAR
“He wants to acquire a big boat”

Morphological typology



Constituent Order

Je vois la couleur rouge sur la face de Tony

I see the red color on Tony's face

Variables:

- Subject – Verb – Object
- Prepositions / Adpositions
- Adjective – Noun
- Genitive – Noun

Universal Tendency:

- VSO/Pr/NG/NA
- SVO/Pr/NG/NA
- SOV/Po/GN/AN
- SOV/Po/GN/NA

How do we model language change and language variation?

a meaning space that defines all possible meaning representations. It should include e.g. the fundamental semantic roles that comparative linguists use to describe the different ways of encoding argument structure in different languages.

a world model that determines which meanings are relevant and which are not. This can be anything from a probability distribution over meanings to a full-blown (simulated) world dynamics, causality and intentions.

a form space that defines all possible expressions. This can be simply strings of characters from a small alphabet, but to study the interactions between grammar and phonology it might be a full-blown model of human articulation and acoustics.

a production and interpretation procedure that provides the mapping from meanings to forms and vice versa. Many formalisms from more traditional linguistics can be used.

a learning procedure that builds up a useful lexicon and or grammar from for example form–meaning pairs.

an interaction model that defines how individuals learn language from each other. Typically, individuals will learn from all other individuals in the population, and new individuals enter the population gradually.

Grammatical Language

Compositionality: Simply storing complete sentences and their meaning is not enough, nor is simply adding up the meaning of different words (as in the multiple word games, Instead, the formalism should be able to build-up sentences by combining lexicon entries, and attribute the *argument structure* (the “who did what to whom”) based on word order or morphological markers. However, the formalism should not be restricted to only build-up sentences from words: it should be able to deal with larger units such as complete idioms as well.

Phrase structure: To attribute argument structure correctly (and in a later stage deal with e.g. stress patterns), the formalism should be able to recognize the phrase structure of sentences. I.e. it should identify

“the block” as a phrase (a noun phrase) in the sentence “the circle approaches the block”, and observe that “approaches the” is not such a phrase.

Recursion: It should be possible to nest phrases in other phrases, as in “the block approaches the block that just hit the triangle” (the triangle’s phrase is nested in the block’s phrase). Only a system that is both compositional and recursive can “make infinite use of finite means”, which is seen as a fundamental property of human language chomsky99handbook.

L-R-A Grammar

Basic unit: a <form,meaning,category> association

A form is simply a string, e.g. “approach” or “the block bounces on the table”.

A category is either

- of the basic categories n or s
- of the structure (yields needs constraint),
 - yields and needs are categories
 - constraint is l (left), an r (right) or an a (anywhere)

A meaning is a list of predicates, with a head and optionally a list of lamda's.

E.g.:

```
(?x | λ?x λ?y | (approach ?z) (arg1 ?z ?x) (arg2 ?z ?y))
```

When applied to the following semantic description

```
(?p || (circle ?p)
```

the resulting description is as follows:

```
(?p | λ?y | (approach ?z) (arg1 ?z ?p) (arg2 ?z ?y)) (circle ?p))
```

Morphology based languages: Latin

```
(("cub-" (?X NIL ((BLOCK ?X))) MASCULIN)
("circ-" (?X NIL ((ROUND ?X))) MASCULIN)
("triangul-" (?X NIL ((TRIANGLE ?X))) MASCULIN)
("-us" (?X (?X) NIL) (NOMINATIVUS MASCULIN L))
("-i" (?X (?X) NIL) (NOMINATIVUS-PLURAL MASCULIN L))
("-o" (?X (?X) NIL) (ABLATIVUS MASCULIN L))
("mov-" (?X (?X) ((MOVING ?Y) (AGENT-OF ?Y ?X))) VERBSTEM)
("-et" (?X (?X) NIL) ((S NOMINATIVUS A) VERBSTEM L))
("-ent" (?X (?X) NIL) ((S NOMINATIVUS-PLURAL A) VERBSTEM L))
("ad" (?X (?Y ?X) ((TOWARDS ?X ?Y))) ((S S A) ABLATIVUS R)))
```

Evaluating the sentence

```
(parse-sentence "ad triangul- -o cub- -us mov- -et" *latin*)
```

yields the following result:

```
((("ad" ("triangul-" "-o")) ("cub-" "-us") ("mov-" "-et")))
  (?46 NIL
    ((TOWARDS ?46 ?42) (TRIANGLE ?42) (MOVING ?Y) (AGENT-OF ?Y ?46)
      (BLOCK ?45)))
  S))
```

The sentence “momet cubus ad triangulo” gives the same interpretation, with just a different phrase structure:

```
((“mov-” “-et”) (“cub-” “-us”)) (“ad” (“triangul-” “-o”))
```

- *momet cubo ad triangulus
- *momet cubi ad triangulus
- movent cubi ad triangulo

Constituent Order based languages: English

```
((("block" (?X NIL ((BLOCK ?X))) N-SINGULAR)
("triangle" (?X NIL ((TRIANGLE ?X))) N-SINGULAR)
("move-" (?X (?X) ((MOVING ?Y) (AGENT-OF ?Y ?X))) VERBSTEM)
("-s" (?X (?X) NIL) ((S NP-SINGULAR L) VERBSTEM L))
("-" (?X (?X) NIL) ((S NP-PLURAL L) VERBSTEM L))
("-s" (?X (?X) NIL) (N-PLURAL N-SINGULAR L))
("to" (?X (?Y ?X) ((TOWARDS ?X ?Y))) ((S S L) NP-SINGULAR R))
("to" (?X (?Y ?X) ((TOWARDS ?X ?Y))) ((S S L) NP-PLURAL R))
("the" (?X (?X) ((DETERMINED ?X))) (NP-SINGULAR N-SINGULAR R))
("the" (?X (?X) ((DETERMINED ?X))) (NP-PLURAL N-PLURAL R))
("-" (?X (?X) ((UNDETERMINED ?X))) (NP-PLURAL N-PLURAL R))
("a" (?X (?X) ((UNDETERMINED ?X))) (NP-SINGULAR N-SINGULAR R)))
```

Constituent Order

(parse-sentence "the block move- -s to the triangle -s" *english*)

yields

```
(((((("the" "block") ("move-" "-s")) ("to" ("the" ("triangle" "-s")))))
  (?1111 NIL
    ((TOWARDS ?1111 ?1110) (DETERMINED ?1110) (TRIANGLE ?1110)
      (MOVING ?Y) (AGENT-OF ?Y ?1111) (DETERMINED ?1107) (BLOCK ?1107)))
  S))
```

- *the block move to the triangles
- *the blocks moves

Model Overview

a meaning space Predicate logic, lambda calculus

a world model probability distribution over meanings / simulated blocks world

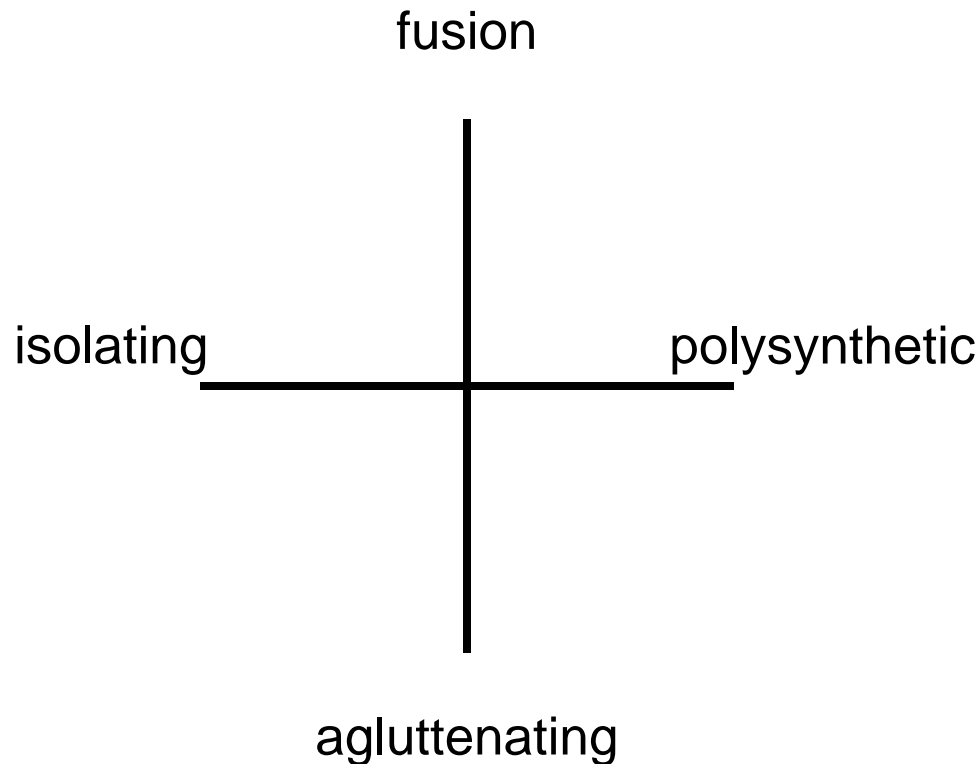
a form space strings of characters from a small alphabet

a production and interpretation procedure l-r-a grammar, depth-first search

a learning procedure semantics not yet implemented

an interaction model language game with flux

Morphology & Constituent Order



- Subject – Verb – Object
 - Prepositions / Adpositions
 - Adjective – Noun
 - Genitive – Noun
-
- What is a word?
 - What is a morpheme?
 - How much of the compositionality is historical rather than productive?