Teilprojekt: “Sprachliche Beziehungen”
Sprachdatenbank ‘Simplex’

Subproject: “Linguistic Connections”
Language Database ‘Simplex’

Timothy Price, PhD
Digital Humanities
Goethe-Universität Frankfurt

"Jesus Christ = Creamy Josh"?: historical semantic blurring and the need for a computational means to get at the 'root' of etymological relationships, i.e. a call for a web-tool that'll do the work for us

Timothy Price, PhD
Digital Humanities
Goethe-Universität Frankfurt

Underlying goal of DigiHum

- Computation-statistical models for judging linguistic similarity/difference

- Same question: Why Germanic?

  1. Similar thematic material (Biblical)
  2. Translation from common sources (?) ***
  3. Separation 100s of years = diminished multual intelligibility
  4. All great criteria for developing test scenarios to ‘teach’ computers to recognize cognates and genetically-related linguistic features

“Antrag zur Errichtung eines LOEWE-Schwerpunkts: Digital Humanities”

[…]

5.1.1 Teilprojekt „Parallel Corpora“

5.1.1.1. Problemstellung

Gegenstand des Teilprojekts sind die Wechselbeziehungen zwischen Übersetzungstexten und den (mutmaßlichen) Vorlagen (Originalien), von denen sie abhängen.

“At the focus of the subproject are the linguistic connections between translated texts and the (purported) models (originals) from which they derive.”

Use statistical models to see what the translator was copying (or perhaps even just reading).

Forensic Linguistics

- Plagiarism
- Author identification (e.g. *suicide note* or *murderer's red herring*?)
- Google's search algorithms (corrections for misspelling, other related items, etc.)

Hardly a new concept:

- Eduard Sievers. *Ziele und Wege der Schallanalyse* (1924):
  - *Schallanalyse* "sound analysis" to describe, compare, contast poetic/rhythmic patterns

“Antrag zur Errichtung eines LOEWE-Schwerpunkts: Digital Humanities”

[…]

5.1.2 Teilprojekt „Sprachliche Beziehungen“

5.1.2.1. Problemstellung

Die diachrone und die vergleichende Sprachwissenschaft stehen im Grunde genommen vor derselben Aufgabe: sie vergleichen verschiedene Sprachsysteme miteinander. Bei der diachronen Linguistik handelt es sich, nach Saussures Definition des Begriffs *Diachronie*, um unterschiedliche sprachliche Systeme auf der Zeitachse einer einzigen Sprache, deren systematische Übereinstimmungen und Unterschiede es herauszufinden gilt.

“Diachronic and Comparative Linguistics essentially share a common task: they compare different language systems to one another. In accordance with Saussure’s definition of the term ‘diachrony’, Diachronic Linguistics deals with the various linguistic systems along the timeline of an individual language, whose systematic similarities and differences it stands to discover.”

In other words…

A single language evolves over time, eventually changing to a point where speakers of its descendant varieties may be unable to understand the original and/or even one another's varieties.

We know *that* this happens.

The question, then, is *why* and/or *how* it happens.

Part of the answer:

- languages don’t just change, rather
- speakers change language

i.e.,

Language change is at least partly the result of human decision, e.g. in translation.

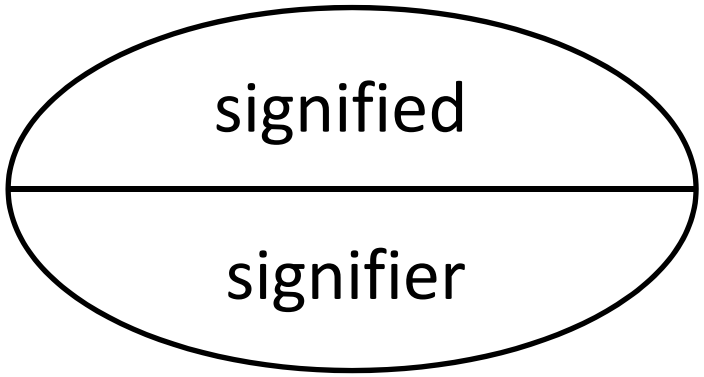
Translator decisions not made in vacuum

Function of culture, therefore influenced by:

- Social interactions (whom s/he knows)
- Intercultural knowledge (travel, education, etc.)
- Fluency
- Economics
  - Speed of translation
  - Intended audience
  - Author intention (theme, style, etc.)

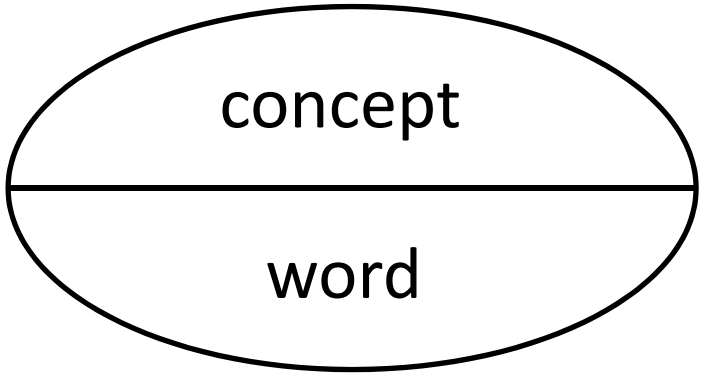
Saussure's *Sign*

De Saussure, Ferdinand. *Cours de linguistique générale* (Course in General Linguistics). 1916



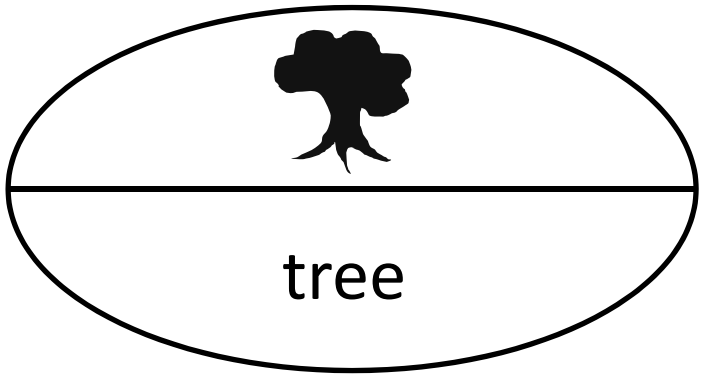
Saussure's *Sign*

De Saussure, Ferdinand. *Cours de linguistique générale* (Course in General Linguistics). 1916



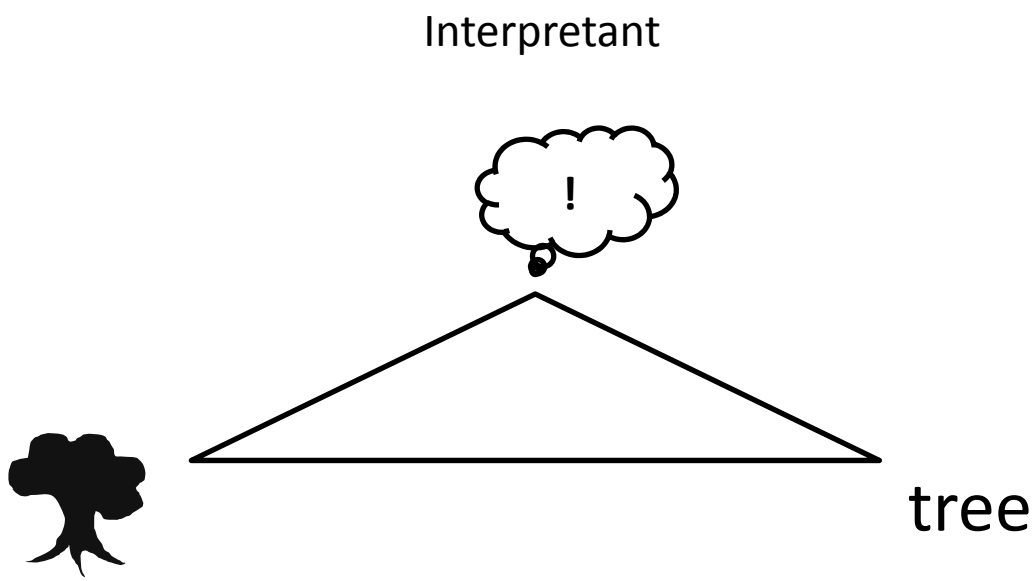
Saussure's *Sign*

De Saussure, Ferdinand. *Cours de linguistique générale* (Course in General Linguistics). 1916



Peirce' *Sign*

Peirce, Charles Sanders. *Collected Papers of Charles Sanders Peirce.* 1931–58.

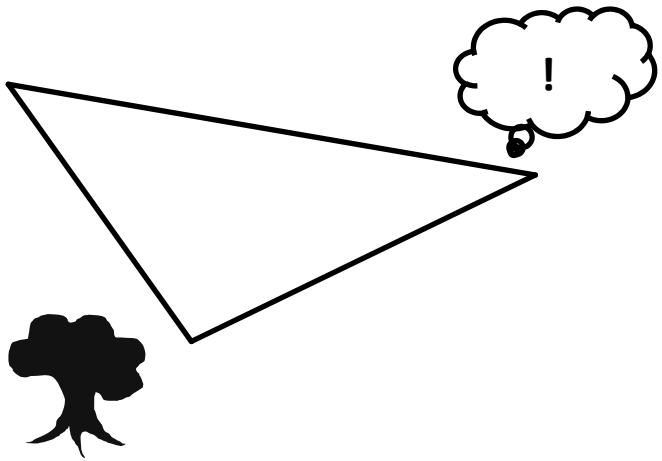


Semantic blurring

The decoupling of the *Sign*

Sometimes with a recoupling with a new Representation (e.g. word):

Baum



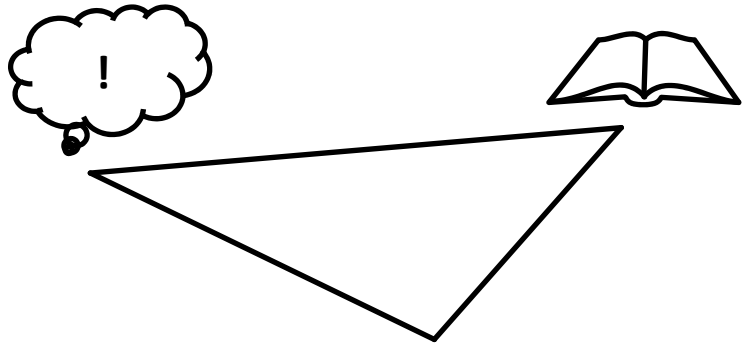
tree

Semantic blurring

The decoupling of the *Sign*

And/or a reassignment of the old representation

With a new entity:



Semantic blurring or even lexical loss through translation.

Translators make decisions about what to:

- translate (e.g. ENG *ice cream* > GER *Eis*)
- calque (e.g. ENG *ice cream* > GER *Eiscreme*)
- borrow (e.g. ENG *ice cream* > GER *das Icecream*)

Translators' decisions are long-lasting

- Output is evident as a new, recorded Peircean *Sign*-relationship
- Input is lost completely
  - Unless s/he records decision-making process, or
  - Source materials used by the translator are known, in which case a set of deciding factors can be recreated

Worst-case scenario is a translation without documentation of the process or the resources.

Given a set of criteria, e.g.

- Time period
- Location
- Other facts related to period and location

Model probabilities of linguistic interference, i.e.

How do you reverse-engineer a 100-yrs.-old decision?

- Etymological dictionaries, based on
  - Historical usage (texts)
  - External accounts (e.g., archaeology)
  - Morpho-phonemic form

Which dictionary? How many? Which is good?

Case in point:

Oxford English Dictionary (OED).

- 1856: conceived
- 1878: Obtained publisher
- 1884: first ‘fascicle’ printed
- 1928: last ‘fascicle’ printed, first supplement
- 1933: second supplement, original proj. finished
- 1957: OED2—project expanded to include language development since 1856 (Published 1972, 1976, 9182, 1986)
- 1983-1989: Digitized
- 1991: OED2 completed
- 2000: OED2 digitized, available online
- 2002: Beginning of OED 3, complete overhaul (as of 2007, anticipated completion in 2037)

Intermediate Goals:

1. Cognate Metacrawler
   - Use already digitized, marked-up corpora with regularized markup
   - Develop coding to 'teach' computer to link between two cognate languages,
     - e.g. ENG *wife* :: GER *Weib*
2. Recognize complex input
   - Expand corpora: digitized with varying markup
   - Develop coding to handle various markup styles
   - Coding to triangulate between word forms that are not immediately identifiable as related, non- or distantly-related languages (e.g. Swadesh lists, etc.)

Long-range Goals:

3. Recongnize complex semantic, simple formal relationships
   - Coding for recognizing morphemic alternations consequent of phonemic change
     1. Synchronic, e.g. /r/ > (/z/ >) /s/ GER *verlie**r**en, Verlu**s**t*
     2. Diachronic, e.g. GMC /VC_0j/ > GMC *St**ä**rke*, ENG *star**ch***
   - Coding to reliably triangulate similar but obscure morphemic roots based semantic relationships (Cf. Visual Thesaurus)

4. Recognize complex phonetic relationships
   - Coding for predicting phonemic forms based on statistical models for phonetic change
   - Application of above to predict obscure root relations, make suggestions that compare above with semantic information

Possible uses:

- Automatic coordination of translations into closely-related languages made from common original text
- New, statistical measurements for determining relatedness of distant languages using non-basic, archaic and obscure vocabulary
- Statistical means of measuring lexical impact of language contact scenarios
- …