Multilingual and Parallel Corpora

Alignment beyond Translation

aligned historical corpora

Amir Zeldes

amir.zeldes@georgetown.edu

Topics

- How can we use parallel corpora to study other relationships between texts?
  - Different versions of the same text?
  - Development of a text?
    - Historical (language change, literary development)
    - Processual (revision processes)

- Here:
  - Parallel historical corpora for language change studies
  - Textual revision alignment for SLA/L2 error analysis

Historical corpora – research questions

- Historical-synchronic: How does a unit behave in older language stages?
  - Words
  - Constructions
  - Grammatical categories
- Diachronic:
  - How do these categories change over time?
- Areas of interest:
  - Language contact
  - Standardization
  - Geographical/dialectal variation

Factors affecting language change

- Language external:
  - Sociolinguistic (social status, relationships between speakers)
  - Textual (genre, register, systems of thought/religion)
  - Dialect and language contact (historically determined)
- Language internal
  - Sound change (phonological)
  - Analogy
  - Reanalysis
  - Grammaticalization

Examples?

Language change and corpora

- Language change is:
  - Quantitative
  - Context sensitive (including intra- and extralinguistic context)
- Data
  - For current/recent language change: monitor corpora, newly collected data:

    "The history of English is, of course, happening every day." (Curzan 2008)
  - For older language change: historical corpora

Monitor corpora

- Dynamically growing data sets
- Internal composition often unstable (hard to collect new data of exactly the same kind/proportions)
- Examples:
  - NOW corpus (and to some extent COCA, as of 2015)
    - http://corpus.byu.edu/now/
  - Bank of English (used by COBUILD dictionary)
  - ? Google ‘corpora’ – Google Books, N-Gram ‘corpora’
- Features:
  - Coverage of latest language
  - Good for lexicography, studying neologisms, short-term language change
  - Involve constant upkeep effort

Historical corpora

- Ideally designed around parameters:
  - Time
  - Place
  - Text type/genre
- We would like a matrix filled with all feature combinations
- Problems
  - Parameters change: e.g. dialect borders grow/shrink…
  - Hard to find all text types in all periods
    - Older language stages often have only religious works
    - Precisely these works are also available in parallel versions
- Recommended reading: Rissanen (2008) in Canvas

Historical corpora - Normalization

- Searching in historical corpora often requires **normalization**
  - *I have not yet bin any were, but at shopes and a veseting; but I believe shall be on Munday at a ball at St. Jeames , where, as they tell me, ther is a famose new danser to apere*

Historical corpora - Normalization

- Searching in historical corpora often requires **normalization**
  - Within a language stage:
    - *Veseting, viseting, visiting -> visiting*
  - Across language stages:
    - What is the normalized form of *a* in *a veseting*?
- Lemmatization and hyperlemmatization

Historical **parallel** corpora - cons

Parallel data has the same issues as other historical corpora, but also:

- Selection of texts highly restricted (especially Bible, other “Classics”)
- Very often not autochthonous works (some exceptions)
- In Europe: very many texts are translations from Latin and Greek – specific translation effects

Translation effects

- *Pater hemōn ho en tois ouranois* Gr.
- *Pater noster, qui es in caelis* Lat.
- *Fæder ure þu þe eart on heofonum* OE (995 CE)
- *Our fadir that art in heuenes* ME (1389)
- *O oure father which arte in heven* EME (1526)
- *Our father which art in heauen* (KJV, 1611)

➢ Morphology?
➢ Word order?

Translation effects

- Sentence initial *and* in narrative typical in Semitic languages: *and it came to pass…*
- Hebr./Aram. pluralia tantum – *the heavens, the waters*
- Hebr.: Particle *hine* → Gr. *idou* → *behold!*
- Hebr.: *stiff-necked* → *Furthermore the LORD spake unto me, saying, I have seen this people, and, behold, it is a stiffnecked people* (Deut. 9:13)

➢Figures of speech, vocabulary, loan translations…

Consequences?

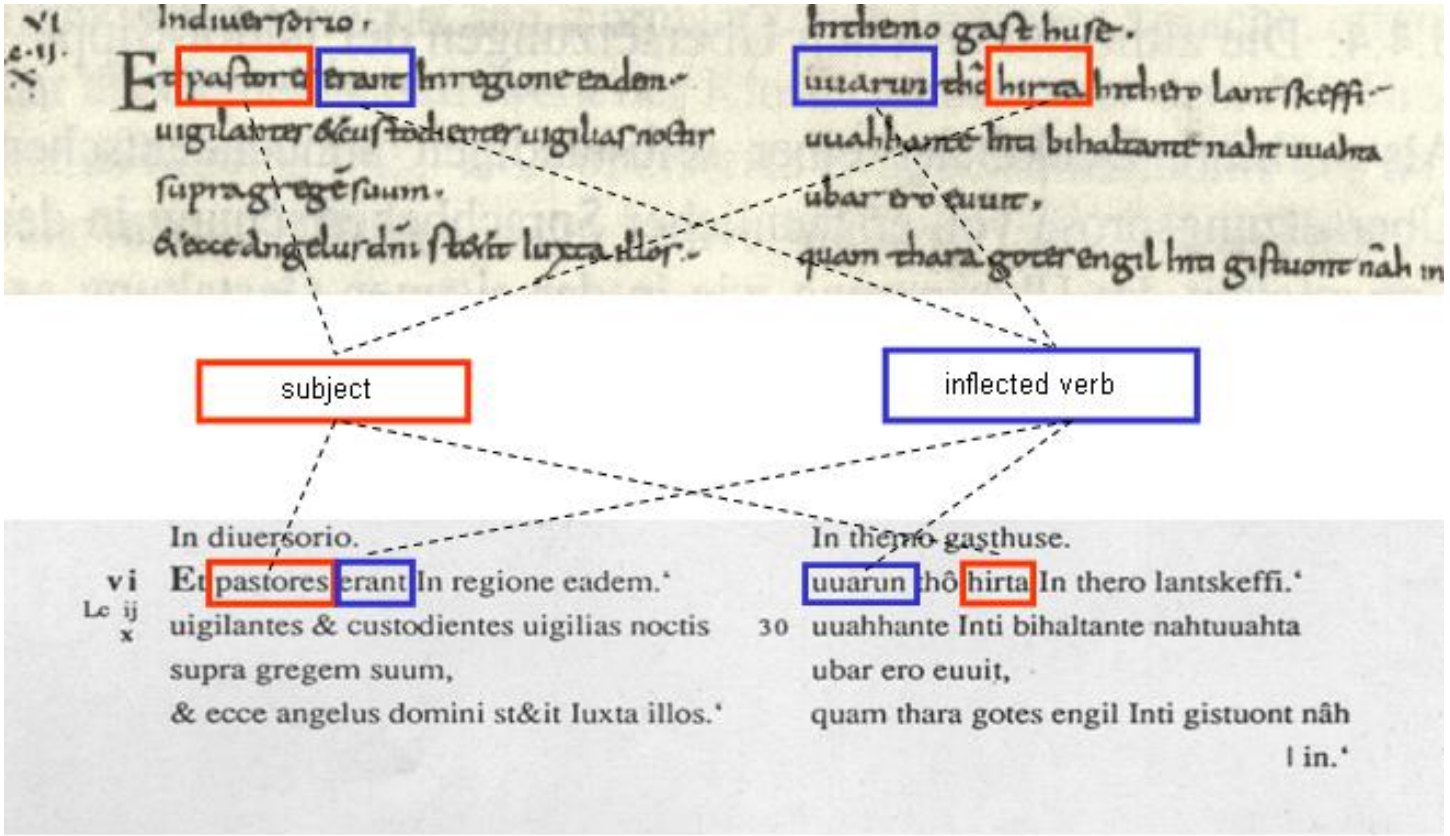
- We must limit our statements in the first instance to translated language in the genre in question:
  - Findings about “Biblical English”
  - Genre? Text type? Language?
  - Same/different for Saints’ Lives? Martyrdoms?

- In “conservative” genres: even greater meaning to **deviating** from expectation
- Support conclusions with evidence from native, comparable corpora, whenever available

Corpora of deviations

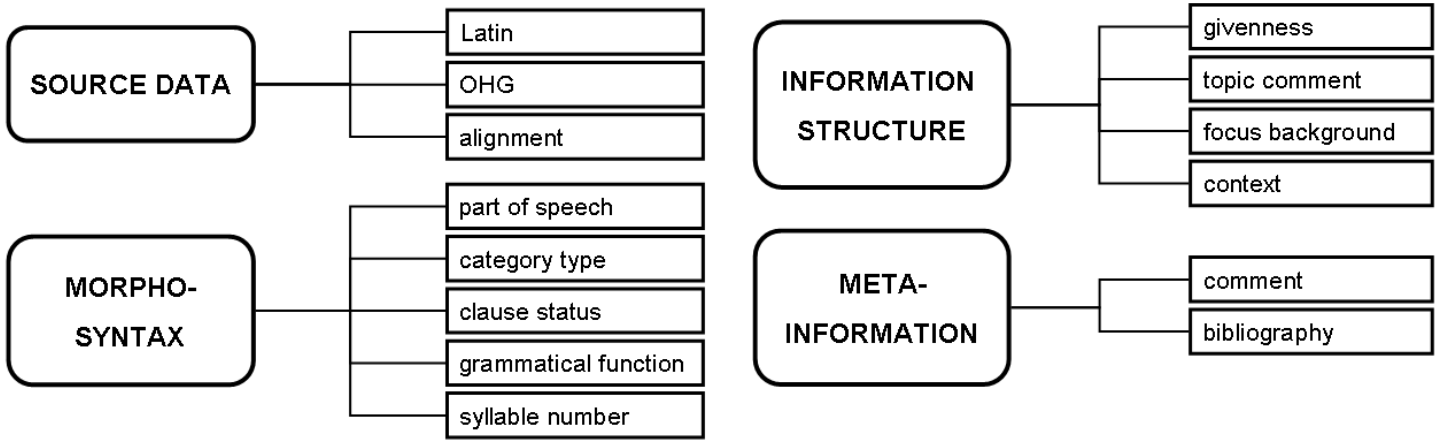
- If we don't believe parallel evidence when it converges with SL….
- Why not look just at deviations?
- Example:
  T-CODEX – Tatian Corpus of Deviating Examples (Petrova et al. 2009)

T-CODEX


subject
inflected verb
In diuersorio.
vi Et pastores erant In regione eadem.'
Lc ij uigilantes & custodientes uigilias noctis
x supra gregem suum,
& ecce angelus domini st&it Iuxta illos.'
In thero gasthuse.
uuarun thô hirta In thero lantskeffi.'
uuahhante Inti bihaltante nahtuuahta
ubar ero euuit,
quam thara gotes engil Inti gistuont nâh
l in.'

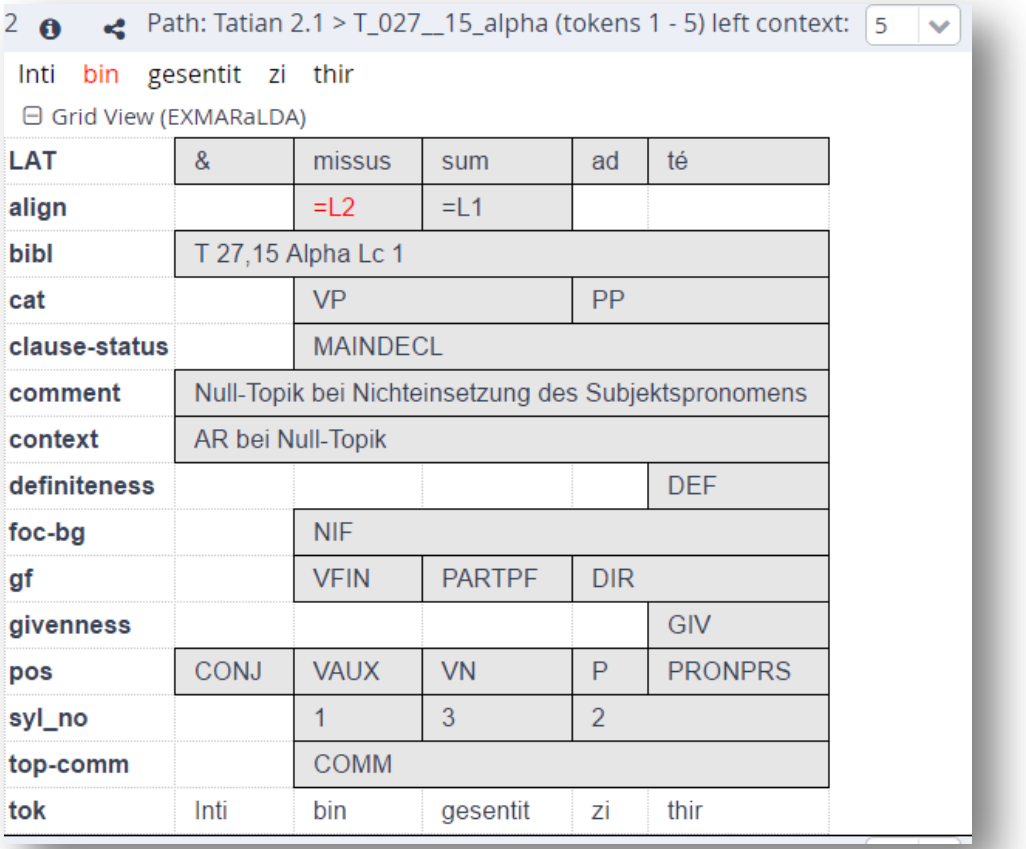
T-CODEX

- Annotation and alignment



Alignment as a span annotation

- https://corpling.uis.georgetown.edu/annis



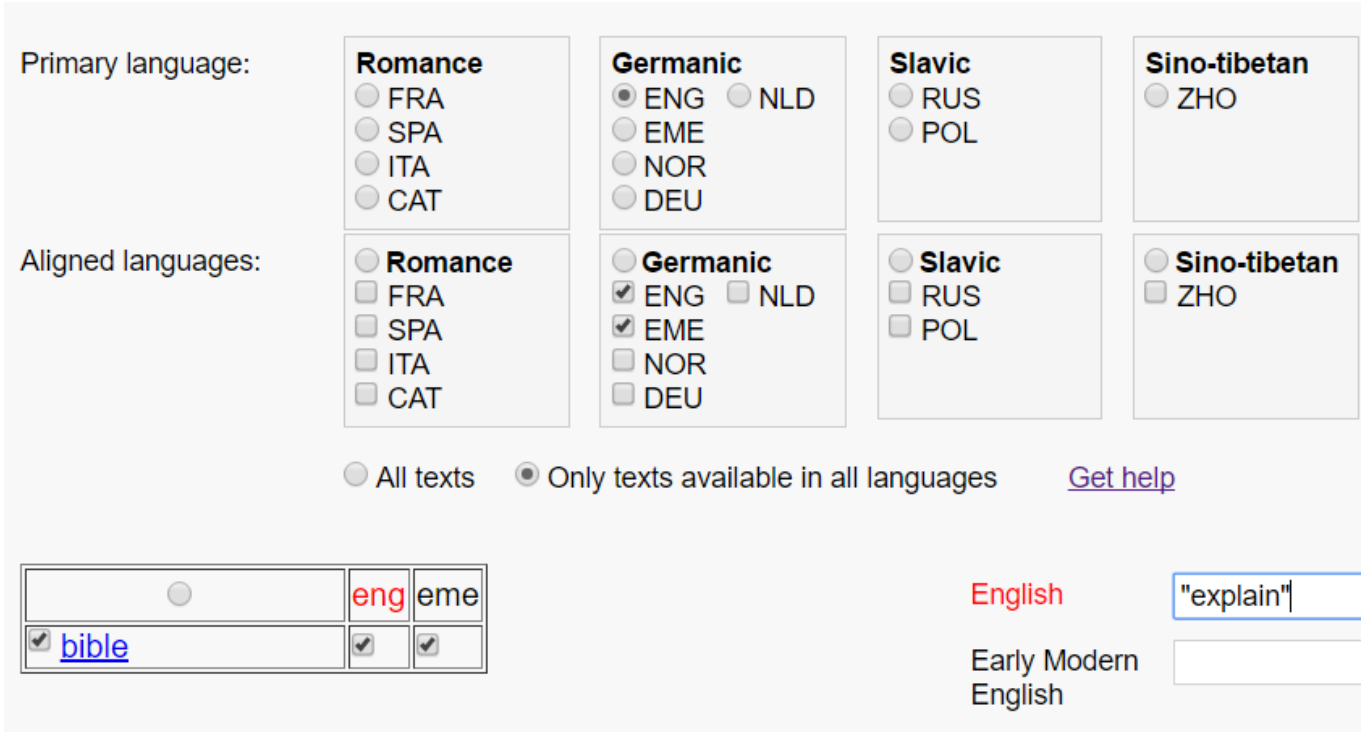
Some example queries

- Subordinate clauses with final verb:
  - clause-status=/SUB.*/ & gf="VFIN" & #2 _r_ #1
- Reordered object before or after verb? Does this relate to information status?
  - clause-status=/SUB.*/ & gf=/.*O/ & gf="VFIN" & align & align & #1 _i_ #2 & #1 _i_ #3 & #2 _i_ #4 & #3 _i_ #5 & #3 .* #2
- More: https://atala.org/IMG/pdf/TAL-2009-50-2-02-Petrova.pdf

Historical parallel corpora - pros

- Completely comparable distribution of virtually all phenomena
- Neutralize content-related variation (and genre, register)
- Greater statistical significance (paired test, repeated measures)
- Alignment – we know more or less exactly what corresponds to what

Let's try it out!



eng	eme
the magicians , but there was no one who could **explain** it to me . " Joseph said to Pharaoh ,	but there was none that could declare it to me .
Whom will he teach knowledge ? To whom will he **explain** the message ? Those who are weaned from the milk	Whom shall he teach knowledge ? and whom shall he make to understand doctrine ? that are weaned from the milk , and drawn from the breasts .

Try to find…

- Content word differences
  - Nouns
  - Verb
  - Adjectives
- Function word differences
  - Prepositions
  - Conjunctions
- Morpho-syntactic differences
  - Inflections
  - Word order
  - Constructions

For Monday

- Continue working with the parallel Bible corpus
- Find an interesting linguistic change between the versions that exhibits **competition**
  - Define the relevant variable and variants
  - Give a distributional analysis (relative frequency of variants)
    - Can be total count if manageable
    - Otherwise base on a subset (e.g. 100 cases)
  - Is this a case of language change outside of the Bible? What could be responsible?