Multilingual and Parallel Corpora

Alignment beyond Translation

aligned historical corpora (ctd)

Amir Zeldes

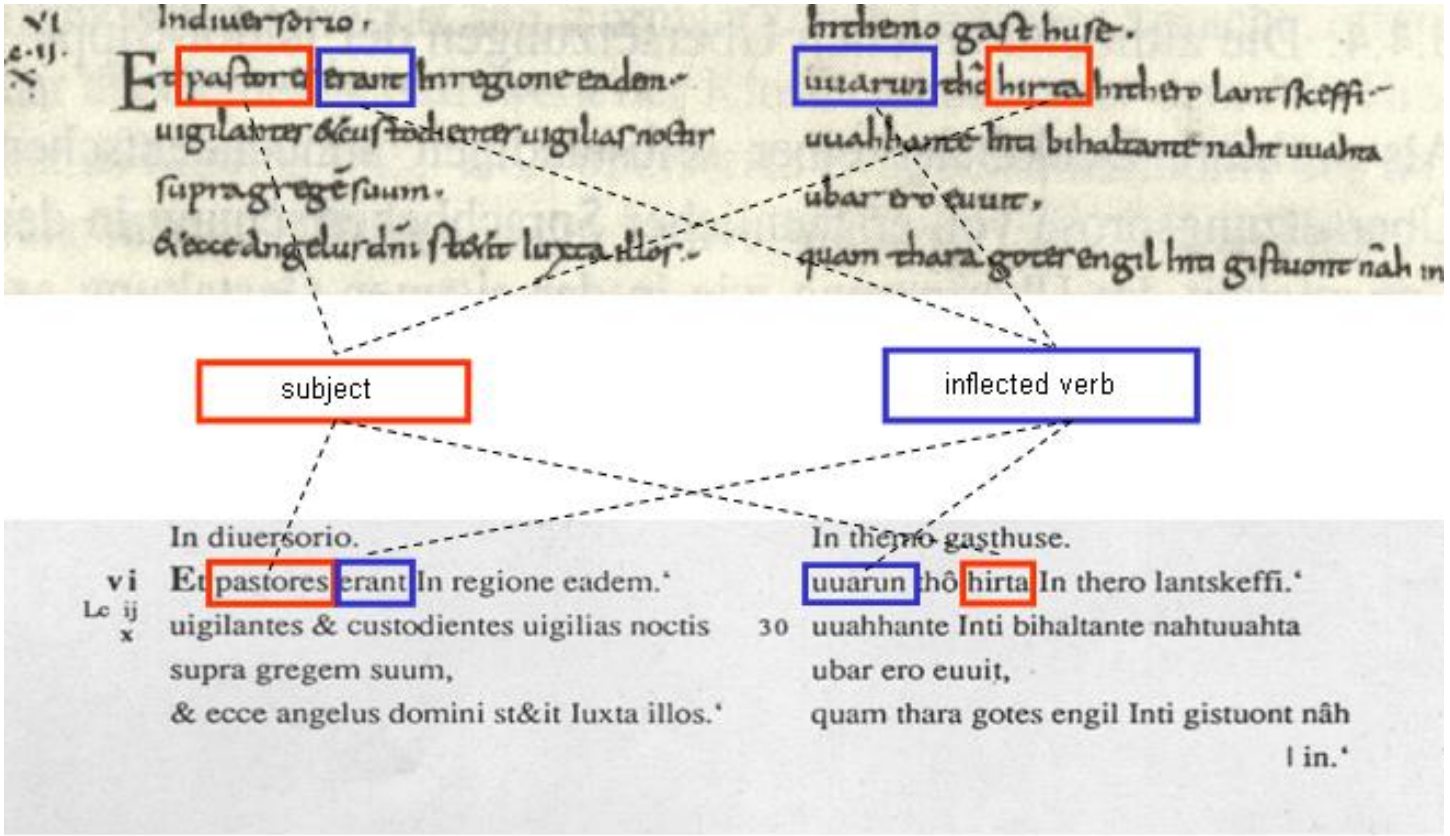
amir.zeldes@georgetown.edu

Translation effects in historical data

- 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…

Corpora of deviations - T-CODEX

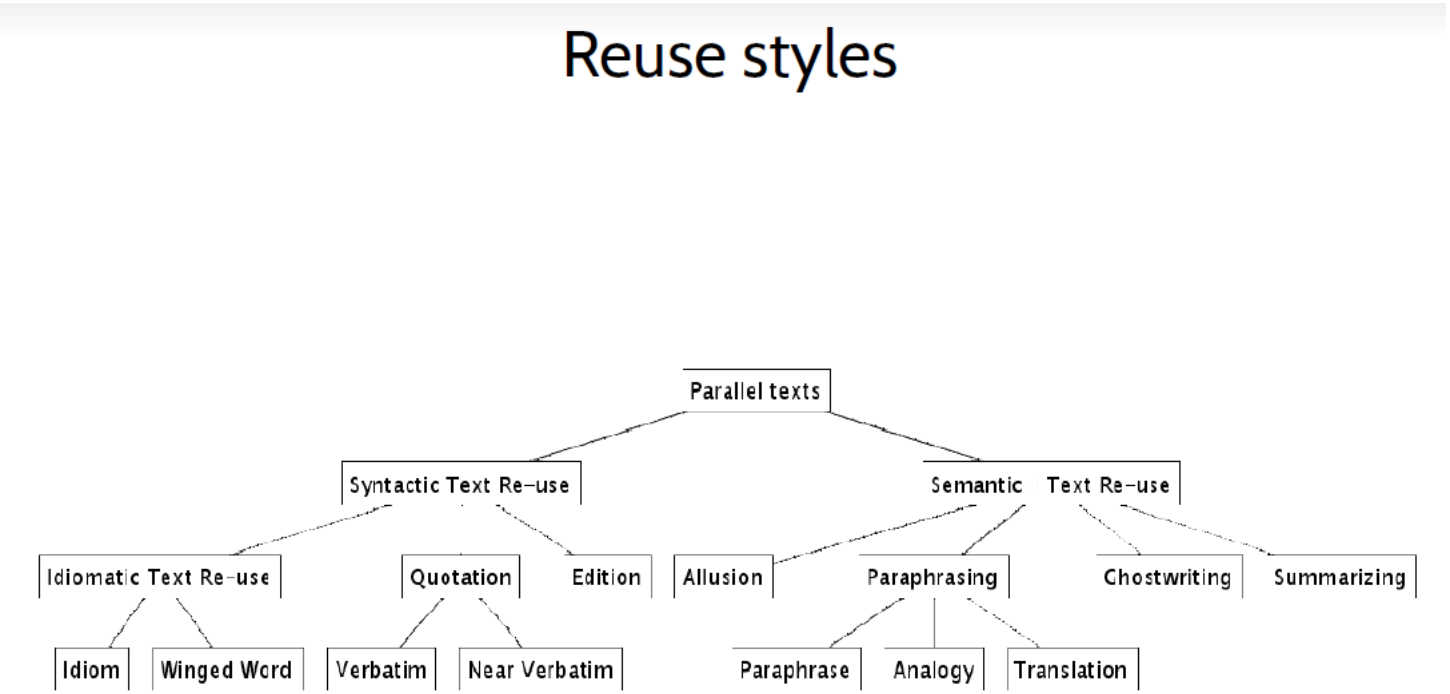


Other types of corpora

- Deviation corpora
- Textual reuse annotation
- Full-fledged parallel corpora (often: complete Bible)

Text reuse

- Miyagawa (2016), eTrap/Tracer project



Text reuse

- Miyagawa (2016), eTrap/Tracer project

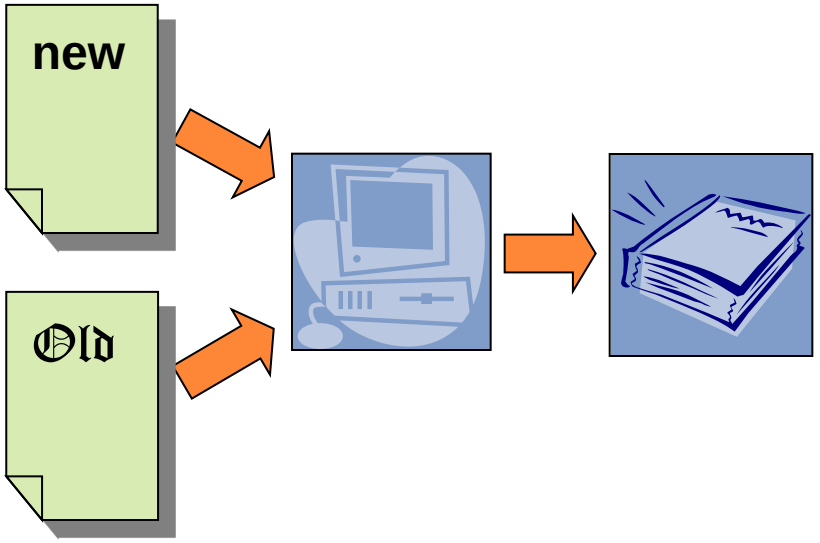
Besa
ⲙⲡⲣ ⲕⲧⲉ ⲡ ϩⲁⲡ ⲉ ⲩ ⲭⲟⲗⲏ ⲁⲩⲱ ⲡ ⲕⲁⲣⲡⲟⲥ ⲛ ⲧ ⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲛ ⲟⲩ ⲥⲓⲱⲉ
Shenoute XF
ⲉ ⲩ ⲡⲱⲱⲛⲉ ⲙ ⲡ ϩⲁⲡ ⲉ ⲩ ⲭⲟⲗⲏ ⲁⲩⲱ ⲡ ⲕⲁⲣⲡⲟⲥ ⲛ ⲧ ⲇⲓⲕⲁⲓⲟⲥⲩⲛⲏ ⲉ ⲩ ⲥⲓⲱⲉ ⲉ ⲩ ⲱⲡ ⲙ ⲡ ⲕⲁⲕⲉ ⲛ ⲟⲩⲟⲉⲓⲛ ⲁⲩⲱ ⲡ
ⲟⲩⲟⲉⲓⲛ ⲛ ⲕⲁⲕⲉ ⲉ ⲩ ⲭⲱ ⲁⲉ ⲟⲛ ⲙⲙⲟ ⲥ ⲉ ⲡ ⲉⲧ ⲥⲁⲱⲉ ⲭⲉ ⲫ ϩⲟⲗⲇ ⲁⲩⲱ ⲡ ⲉⲧ ϩⲟⲗⲇ ⲭⲉ ⲫ ⲥⲁⲱⲉ

Amos 6:12

"But you have turned justice into poison and the fruit of righteousness into wormwood" (ESV)

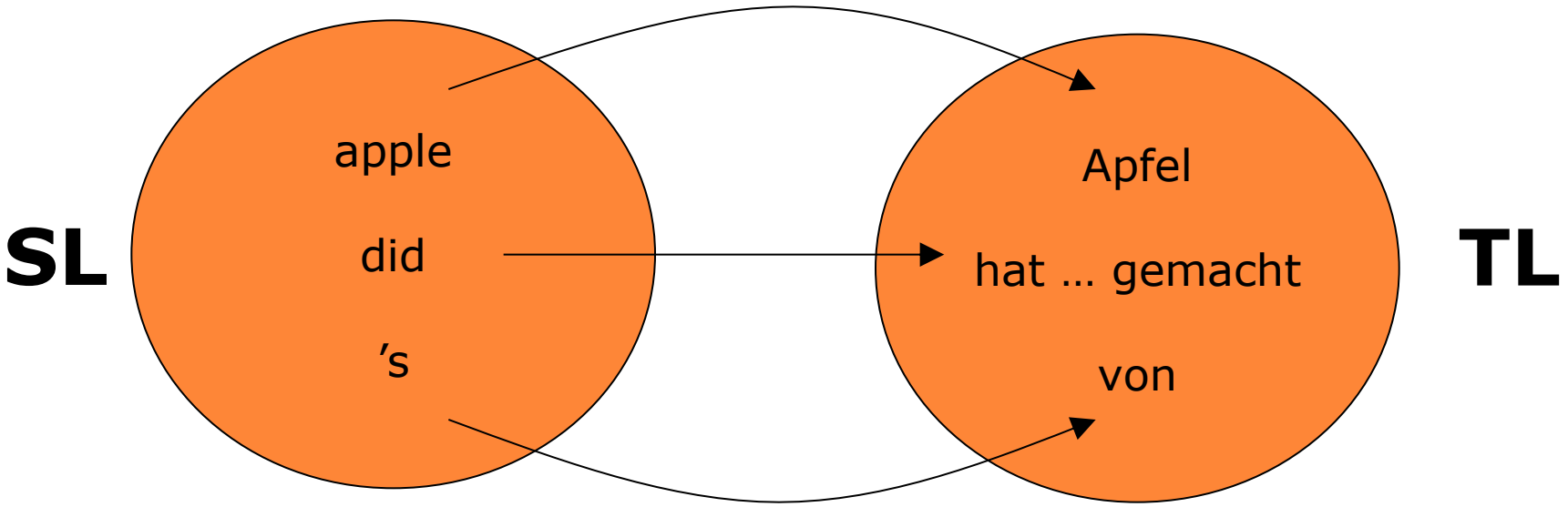
Parallel corpus > Auto grammar?

- Can we derive a historical grammar from a parallel corpus?
- Wanted: correspondence rules on all levels
  - Phonology
  - Morphology
  - Vocabulary
  - Syntax
  - …
- Can we use an MT approach?



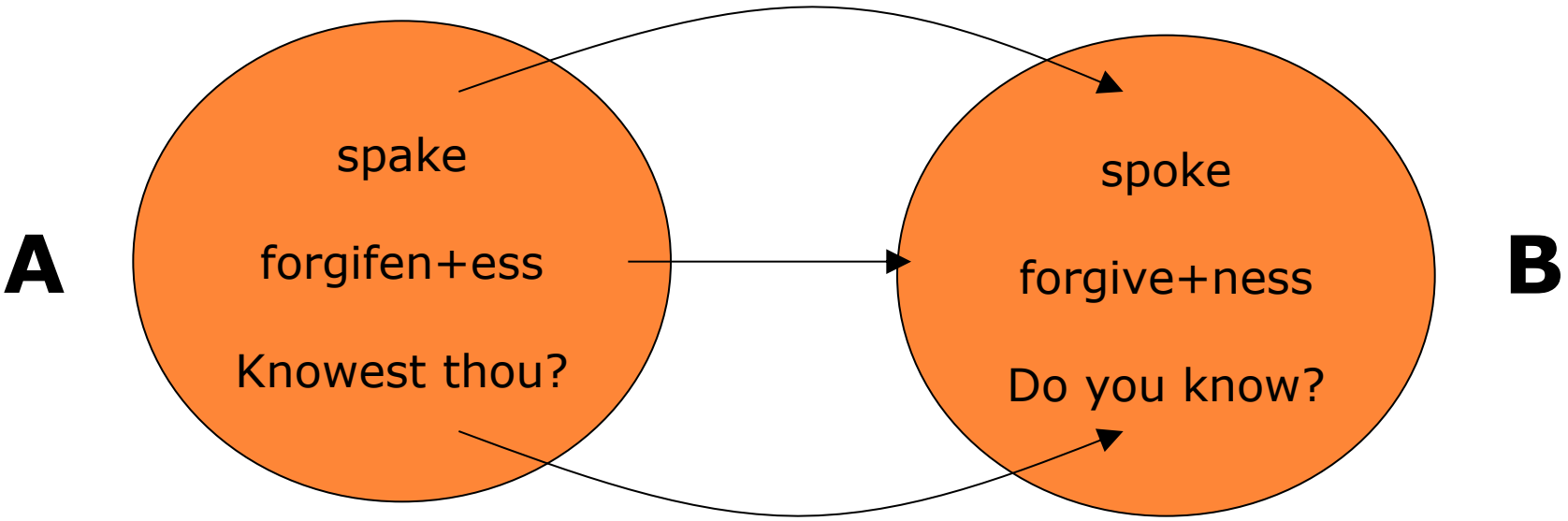
Translation rules and historical grammar

- X in SL = Y in TL



Translation rules and historical grammar

- X in stage A = Y in stage B



Historical grammar ≈ translation model probabilities?

Case study – Polish (Zeldes 2007)

- Parallel Bible translations, from 1606 and 1975
- Fully automatic extraction of correspondences
  - Inflection
  - Derivation
  - Lexical change
  - Syntactic change

Data

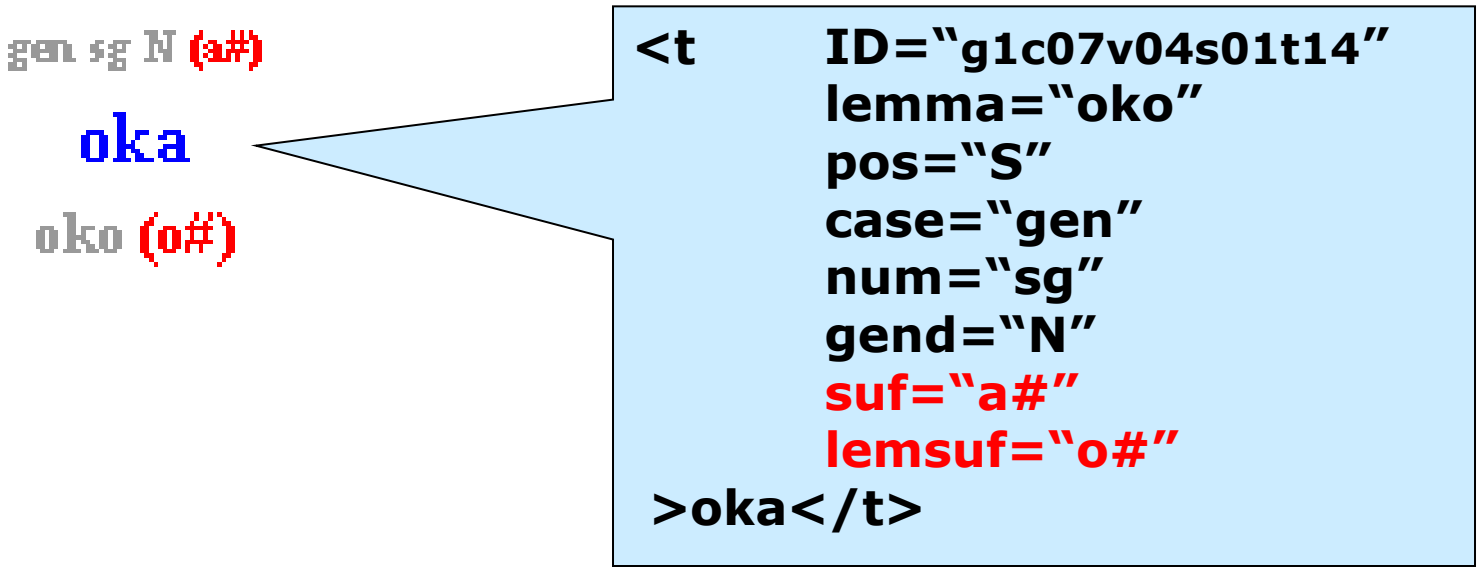
- The Gospel of Matthew: (protestant versions)
  - Biblia Gdańska (1606)
  - Biblia Warszawska (1975)
- 46,000+ tokens in 1071 parallel verses

Drugie podobieństwo przełożył im, mówiąc: Podobne jest królestwo niebieskie człowiekowi, rozsiewającemu dobre nasienie na roli swojej.	Inne podobieństwo podał im, mówiąc: Podobne jest Królestwo Niebios do człowieka, który posiał dobre nasienie na swojej roli

(Württembergische Landesbibliothek Stuttgart)

Annotation

- Lemmatization
- Morphological analysis – case, number, tense, aspect…
- Affixes and lemma affixes



Same annotations – different affixes

category	suffix pair		examples		
acc pl MP	R4y#	ów#	anioły	aniołów	angels
	R4e#	ów#	króle	królów	kings
	R4e#	R4y#	nauczyciele	nauczycieli	teachers
	#	R4y#	sług	sługi	slaves
gen pl MP	ów#	#	poganów	pogan	Pagans
inst pl M/N	R4y#	ami#	duchy	duchami	spirits
…					

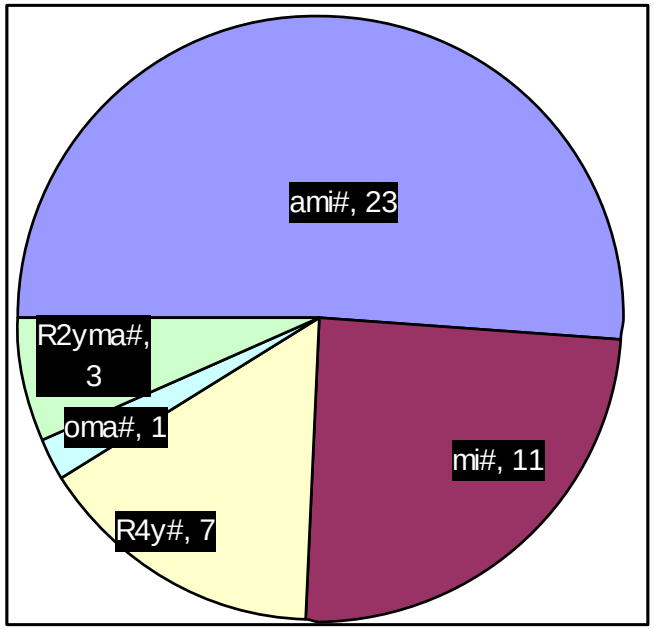
Prep	inst pl MA (R4y#)	Adj inst pl nV (R4ymi#)
nad	duchy	nieczystymi
nad	duch (#)	nieczysty (R4y#)

Prep	inst pl MA (ami#)	Adj inst pl nV (R4ymi#)
nad	duchami	nieczystymi
nad	duch (#)	nieczysty (R4y#)

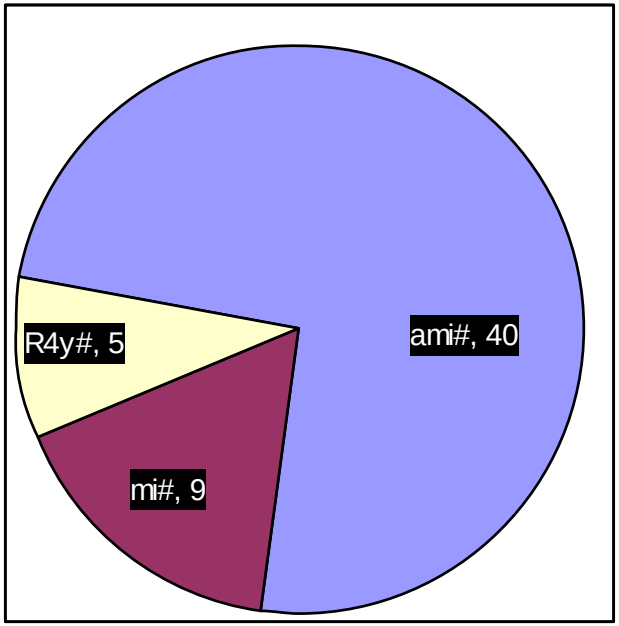
Affix distribution – example: *-ami#*

- Parallel distribution: masc./neut. inst. pl.

Middle Polish



Modern Polish



Extracting parallel elements - association

- MI3 (Daille 1995):
  Given *a*, how likely is *b*?

$$MI3 = \log\left(\frac{pairs(a\&b)^3 \cdot N}{pairs(a) \cdot pairs(b)}\right)$$

- Supports **negative** association (beyond MT models)
- Allow *n* to *m* pairs – here, bi-grams to unigrams

a	b	translation	pairs(a)	pairs(b)	pairs(a&b)	MI3
przedni kapłan	arcykapłan	high priest	441	237	19	13.931
słowo	słowo	word	343	585	24	14.001
…						

What else can we do with these pairs?

- A data-based historical dictionary
  - Items missing in newer corpus > obsolete?
  - Do some items get significantly more/less frequent?
  - In what contexts?
  - (Linked) attestation, examples
- Enables detailed studies of words, phrases, morphological categories…

What else can we do with these pairs?

- Example study – verbal word formation
- What can happen to a verb across versions?
  - Stay the same
  - Get replaced by a non-verb
  - Get replaced by a different verb

Verb substitution / change

- Prefix change:
  na-śmiać : wy-śmiać ‘*ridicule*’
  at-laugh *out-laugh*

- Stem change:
  wy-gnać : wy-pędzić ‘expel, exile’
  out-drive *out-rush*

- Stem alteration:
  za-bieżeć : za-biec ‘*run at*’
  behind-run1 *behind-run2*
  (root= √bieg)

Prefix change

- *a* und *b* are corresponding verbs based on MI3
- **Edit distance** 1-4 on the left side

a	b	Sense	pairs(a)	pairs(b)	pairs(a&b)	MI3	EDis
uwinąć	owinąć	wrap	6	6	1	12.691	1
nagotować	przygotować	prepare	4	8	5	12.031	4
zwołać	przywołać	call	78	76	3	9.903	3
…							
zmiłować	zlitować	pity	9	3	6	13.065	2

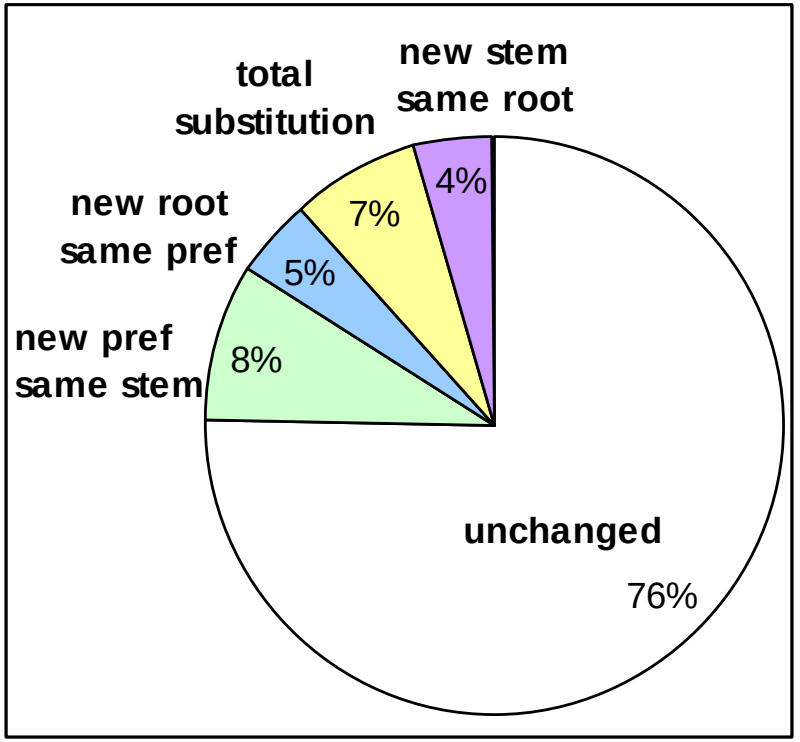
- Recall = 52/53 ≈ 98%

(missed 1 - poprzewracać : powywracać ‘overturn’)

$$F = 2 \cdot \frac{Pr \cdot Rc}{(Pr + Rc)} \approx 94\%$$

Verb-Verb pairs

- 76% identical
- 7% replaced
- 8% prefix change
- 5% root change
- 4% stem alteration



Syntactic change

- Annotation n-grams as elements:

TEXT TEXT TEXT TEXT TEXT
TEXT [VPPA] [NN nom] TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT [VPPA] [NN nom]
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT

TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT [VFIN past] [NN nom]
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
[VFIN past] [NN nom] TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT
TEXT TEXT TEXT TEXT TEXT

Possessive adjectives

- Syn dawidowy “Davidian son”
  Son David-ian
- Matka jakóbowa “Jacob’s mother”
  Mother Jacobian

- Search for possessive adjectives with congruent agreement:

```
  Syn    dawidowy → [NN nom agr] [AdjPos nom agr]
Sohn     Davidian
```

Possessive adjectives

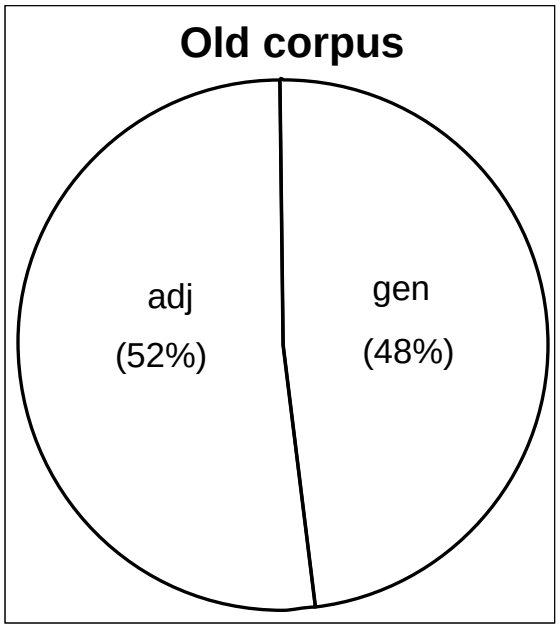
a (old corpus)	b (new corpus)	(a)	(b)	(a&b)	MI3
[NN acc agr] [AdjPos acc agr]	[NN acc] [NNP gen]	63	194	3	8.88
[NN voc agr] [AdjPos voc agr]	[NN voc] [NNP gen]	97	77	5	11.81
[NN nom agr] [AdjPos nom agr]	[NN nom] [NNP gen]	42	199	6	12.43
[NN dat agr] [AdjPos dat agr]	[NN dat agr] [AdjPos dat agr]	33	31	2	10.71
[NN gen agr] [AdjPos gen agr]	[NN gen agr] [AdjPos gen agr]	142	72	8	13.39
Total: [N.* agr] [AdjPos agr]	*Total:* [N.*] [NNP gen]	4779	2169	31	9.26
Total: [N.* agr] [AdjPos agr]	*Total:* [N.* agr] [AdjPos agr]	2169	696	14	8.60

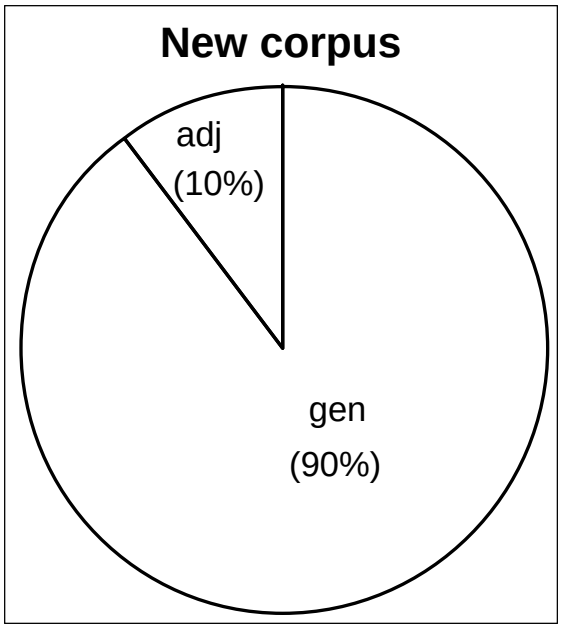
- Frequently (but not always) replace by proper names in the genetive (cf. Rospond, 2003: 195)

  Davidian son > David's son

Possessive adjective

- Non-aligned distribution is **misleading**!





Discussion

- How can we maximize advantages and minimize disadvantages of parallel data?
- What kinds of other texts could we have available?
- What types of annotation can we add?
- Can we integrate our more complex theories of alignment and equivalence into the field of language comparison?