Linguistic Annotation of Old Swedish Texts

Yvonne Adesam

Uppsala, March 2015

MAÞiR project

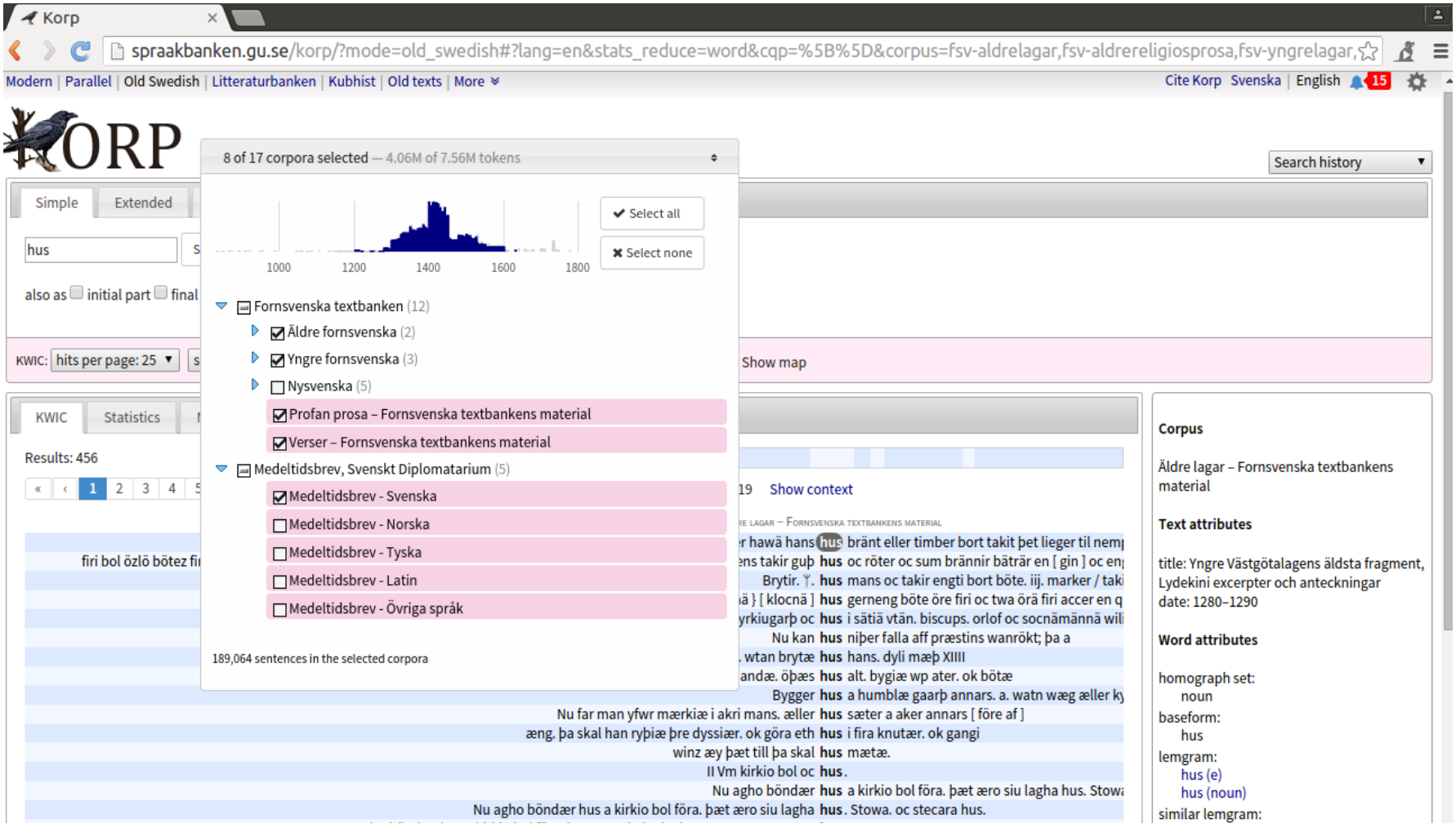
- **M**ethods for the automatic **A**nalysis of **T**ext in digital **Hi**storical **R**esources
  maþir (*mather*, *mandr*, *man* etc.)
  OSw: man, person
- Gerlof Bouma (PI), Yvonne Adesam
- Investigate and develop natural language processing tools for Old Swedish (1225–1526)
- 2014–2016/7, funded by the Marcus and Amalia Wallenberg Foundation

Outline

MAÞiR data

- The Old Swedish (*fornsvenska*) period starts with latin script manuscripts of law text, ends with (pre-)publication of the Gustav Vasa bible
- Fornsvenska textbanken, 3.1M tokens
  - Early (ca 1225-1375)
  - Late (ca 1375-1526)
- Svenskt Diplomatarium, 1M tokens (3.5M)

MAÞiR data


Korp
spraakbanken.gu.se/korp/?mode=old_swedish#?lang=en&stats_reduce=word&cqp=%5B%5D&corpus=fsv-aldrelagar,fsv-aldrereligiosprosa,fsv-yngrelagar,
Modern | Parallel | Old Swedish | Litteraturbanken | Kubhist | Old texts | More
Cite Korp Svenska | English 15
KORP
8 of 17 corpora selected — 4.06M of 7.56M tokens
Search history
Simple
Extended
hus
also as initial part final
Select all
Select none
1000 1200 1400 1600 1800
Fornsvenska textbanken (12)
Äldre fornsvenska (2)
Yngre fornsvenska (3)
Nysvenska (5)
Profan prosa – Fornsvenska textbankens material
Verser – Fornsvenska textbankens material
Medeltidsbrev, Svenskt Diplomatarium (5)
Medeltidsbrev - Svenska
Medeltidsbrev - Norska
Medeltidsbrev - Tyska
Medeltidsbrev - Latin
Medeltidsbrev - Övriga språk
189,064 sentences in the selected corpora
KWIC: hits per page: 25
Show map
KWIC
Statistics
Results: 456
Show context
Corpus
Äldre lagar – Fornsvenska textbankens material
Text attributes
title: Yngre Västgötalagens äldsta fragment, Lydekini excerpter och anteckningar
date: 1280–1290
Word attributes
homograph set:
noun
baseform:
hus
lemgram:
hus (e)
hus (noun)
similar lemgram:

Other resources

- Schlyter (1877), *Ordbok till Samlingen af Sweriges Gamla Lagar*, 10 000 entries
- Söderwall (1884–1973), *Ordbok över svenska medeltids-språket*, 25 000 entries
- Seed morphology

Historical Material in Computational Linguistics?

- Low resource / high-variability material
- Explore methods for such material
- Understand the material better by looking at what methods work
- Results may carry over to other types of material, e.g. social media

Challenges of OSw text

- Lexicon: *aptanbakka, bakvaþi*
- Morphology, syntax: *Nu kan kirkia brinna ælla stulin uarþa [at] ipnum durum*
- Lack of standardization:
  - spelling: *bokstaffua, bokstaffwa, bokstafwa, bokstaua, bokstawa, bogstaffwa*
  - segmentation: Sentence units marked by period, slash, comma, capital, or not at all
- Lack of resources: No annotated data, no complete descriptions, no native speakers

Our approach to NLP for OSw

- Linguistically informed
- Linking
  - lexicon to lexicon
  - text to lexicon
  - between spelling variants
  - between text editions (e.g. bible for annotation projection)
- Manual annotation for training/evaluation data
- Annotation projection (from Icelandic, modern Swedish)

Linking to lemmata

- *bokstaffwa bokstaff bokstawom bokstaf bogstaffwa bokstaua bokstawa bokstafwa bokstaffwom bokstaffua bokstawin bokstaffwane bokstaffwinor(=m?) bokstafuom bokstawane*
- How do we recognize these as occurrences (inflections + spelling variants) of **bokstaver** (Söderwall)

Linking to lemmata 2

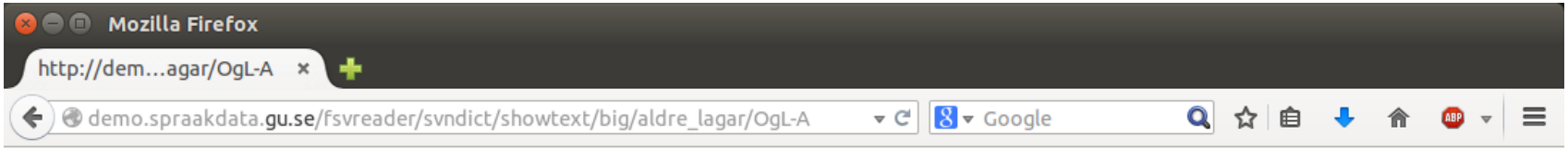
- Using lexicon entries as lemmata
  - language of the period
  - information about PoS, morphology etc
  - Definition and description

A simple algorithm for linking

- List of lemmata + occurrences from dictionary
- b o g s t a f f w a _
  b o k s t a _ _ v e r
  (iterated LD alignment: Wieling et al., 2009)
- Count frequent character group substitutions, turn into rules, weigh according to frequency
- Link a token to the lemma that requires the cheapest substitutions

A simple algorithm for linking

- On *Abota*:
  direct lookup 2/10 correct on first guess
  simple linking 5/10 on first guess
- Split inflection from spelling variation?
  use explicitly coded human knowledge
  (computational morphology)?



Meny Tillbaka

XIII.

Nu sitær bonde kuar mæþ tiunde um ar. böte firi þrea öra. ok ut tiundan præstinum. ælla þöm sum tiundan a. sua firi annat ok sua firi þriþia. Sitær kuar um þry böte biskupe þrea markær. Nu före bonde egh tiunda til firi paska. þa halde þöghine hans rezskap [up] utan kalle han firi kirkiu dyr æpte paska. ok uiti þæt mæþ eþe fiughurtan manna at han sat mæþ luui kuar. ælla ok at han förþe han. ælla han böþ han ut. ok þe wltu sialuir at þe uildu egh uiþær taka. orka egh eþe böte þrea öra. æn han sat minna kuar æn um þry ar. sat han um {þru} [þry] ar böte þre markær sum för uar saght. §. I. Nu ær biskupær skyldughær firi tiunda sin. wighia krismu ok klærka. kalk mæssu klæþe: huart þriþia ar {tik} [til] sokn koma tuæggia natta gingærþ af præste taka. mæþ mannum tolf ok sialuær han þrattande. þa a han folk færma ok næmdir sea. §. 2. þa biskupær uill næmd sea. þa skal han fa manaþa buþ firi sik. præstær a næmd gæra sanna mæn i næmd sætia ok egh uiþærdelu mæn. ok egh soknara. ok egh þæn i eþumin stoþ. ok þe skulu sitia i næmdinne sum þe sighia ia uiþær sum uiþærdulu mænnini æru. siþan a soknarin eþa til næmd bæra. þa skal næmdin sitia sik ensamin ok talas uiþær. alla þa eþa sum hon uær þa uarin uarþi. alla þa eþa* hon fælle þa uarin fælde. utan næmdin uili nakuara eþa undi kunungx dom læggia. ælla laghmanzs. þa biþin þe eþa þær til þe uarþa fællde ælla uarþe. §. 3. all laghkallaþ mal aghu a biskups næmd kuma ekke huma mal. Alli eþa aghu a biskups næmd kuma. þæssin mal aghu a biskups næmd koma. hor. meneþa.

demo.spraakdata.gu.se/fsvreader/lexseasy/þrattande–þrättande–þrätta–þrætta–þrättan–þrættan

þrattande 5 träffar. (*þrättande/þrätta/þrætta /þrättan/þrættan*)

- þrättande nl
  - ***Söderwall*** *nl* trettonde. " thässe fäm konunga hafdho thiänt tolff aar mz skat vnder kodorlaamor konung, oc thrättanda aarith gingo the aff honum " MB 1: 180 . Bir 4: 23 . VKR 69 . Lg 554 .
  - ***Söderwalls Supplement*** *nl* trettonde. " j the kirkio standa tolff stola oc thrättande war then som gudh siäff sat a Prosadikter (Karl M) 251. ib. ther waro tolff sänga . . . oc thrättande i midhio " ib 255 . - þrättande dagher iula, trettonde dag jul, trettondedag(en). trydie torsdagen nest epter trettendhe ... more
- þrätta vb nn
  - ***Söderwall*** *vb* L.
  - ***Söderwall*** *vb* envisas, anstränga sig,

Manual annotation

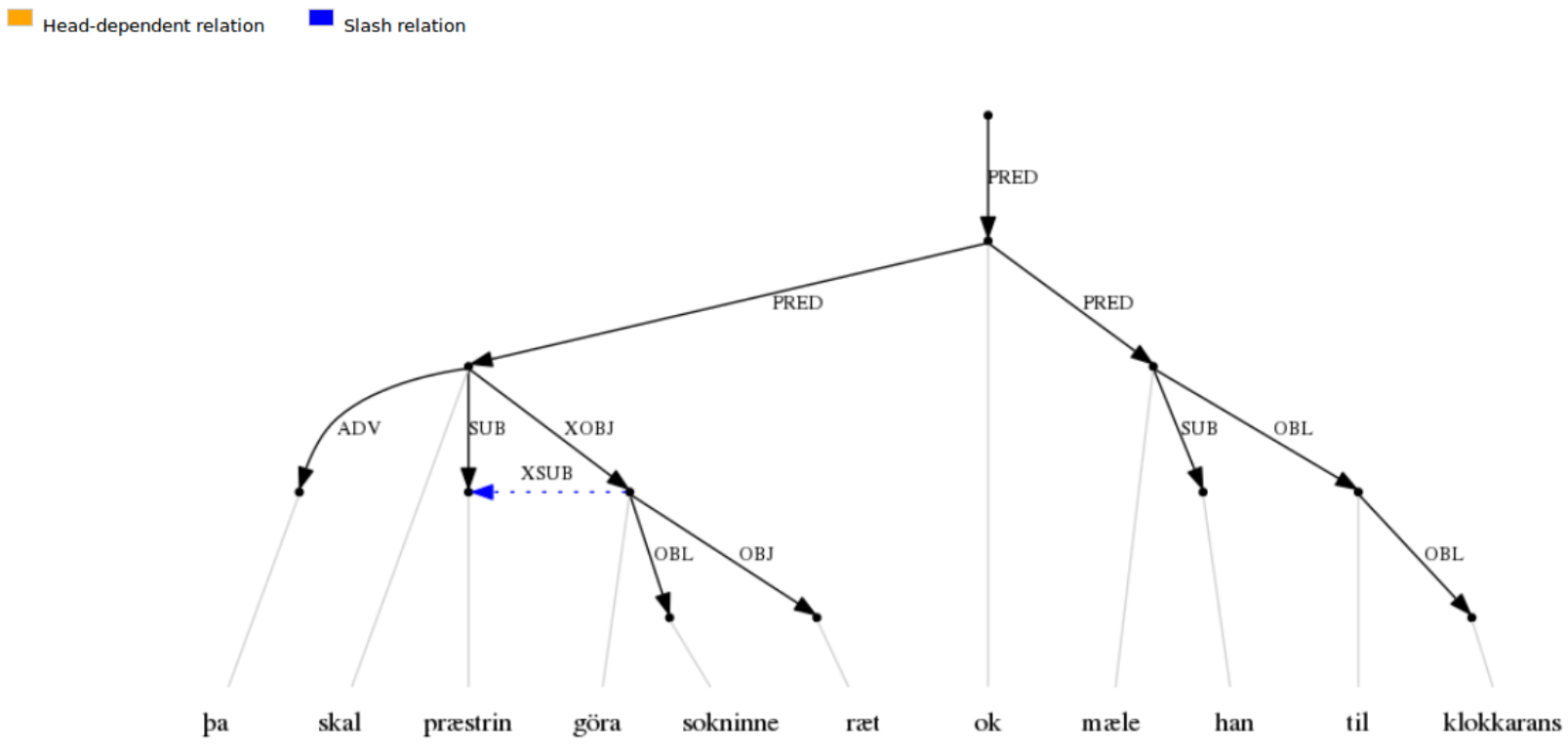
- Mainly following guidelines for Old Norwegian from Menotec project (Haugen & Øverland 2014)
- Lemmata, morphological features, syntactic structures using the PROIEL scheme (Eckhof & Haug 2010) and annotation environment – also used for 16(!) other historic language annotated corpora.
- Lemmata: Söderwall entries
- Status: almost 20k tokens annotated

Morphology (Edit)

þa	***skal***	***præstrin***	***göra***	***sokninne***	***ræt***	***ok***	***mæle***
adv.	verb	subst.	verb	subst.	subst.	sam. knj.	verb
oböj.	ind., pres., akt., 3rd p., sg.	nom., sg., m., def.	inf., pres., akt.	dat., sg., f., def.	ack., sg., m., ind.	oböj.	konj., pres., akt., 3rd p., sg.
þa	*skola*	*präster*	*göra*	*sokn*	*rätter*	*ok*	*mäla*

han	***til***	***klokkarans***
pers. pron.	prep.	subst.
nom., sg.	oböj.	gen., sg., m., def.
han	*til*	*klokkare*

Syntax (Edit)



Data

- Östgötalagen (A Holm B50), ca 1290/1350
  18’269
- Ä. Västgötalagen (A Holm B59), ca 1220/1280
  491
- Abota (Klemming), ca 1448
  541
- Pentateuchparafr. (B Holm A1), ca 1330/1526
  469

Data

- Caveats:
  - no punctuation
  - manual sentence splitting
  - manual tokenization

Data

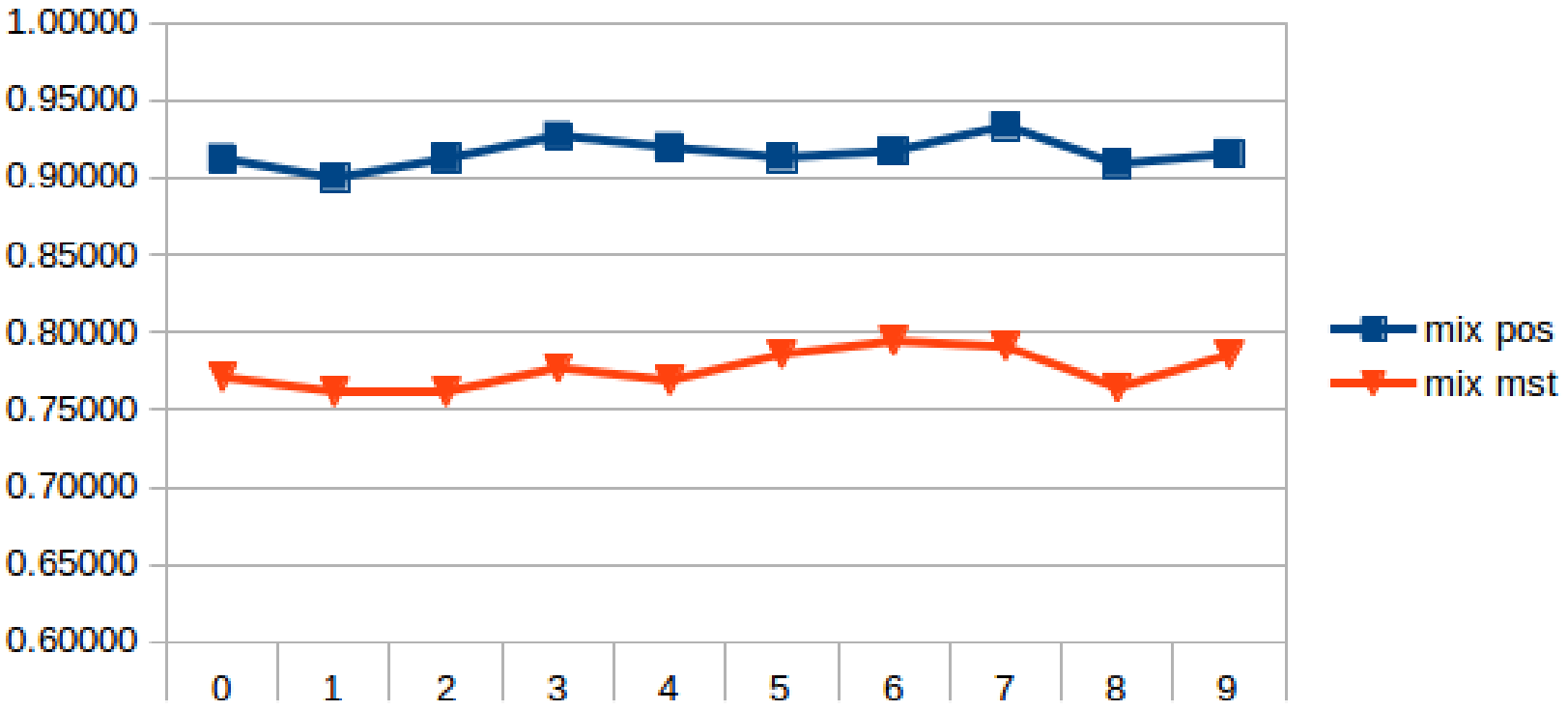
- Caveats:
  - no punctuation
  - manual sentence splitting
  - manual tokenization

Uerder maþer .i. kyrkiu dræpin þet ær niþings værk. þa er kyrkia al vuighz.
If a man is killed in church, this is infamy, then the whole church is unhallowed.

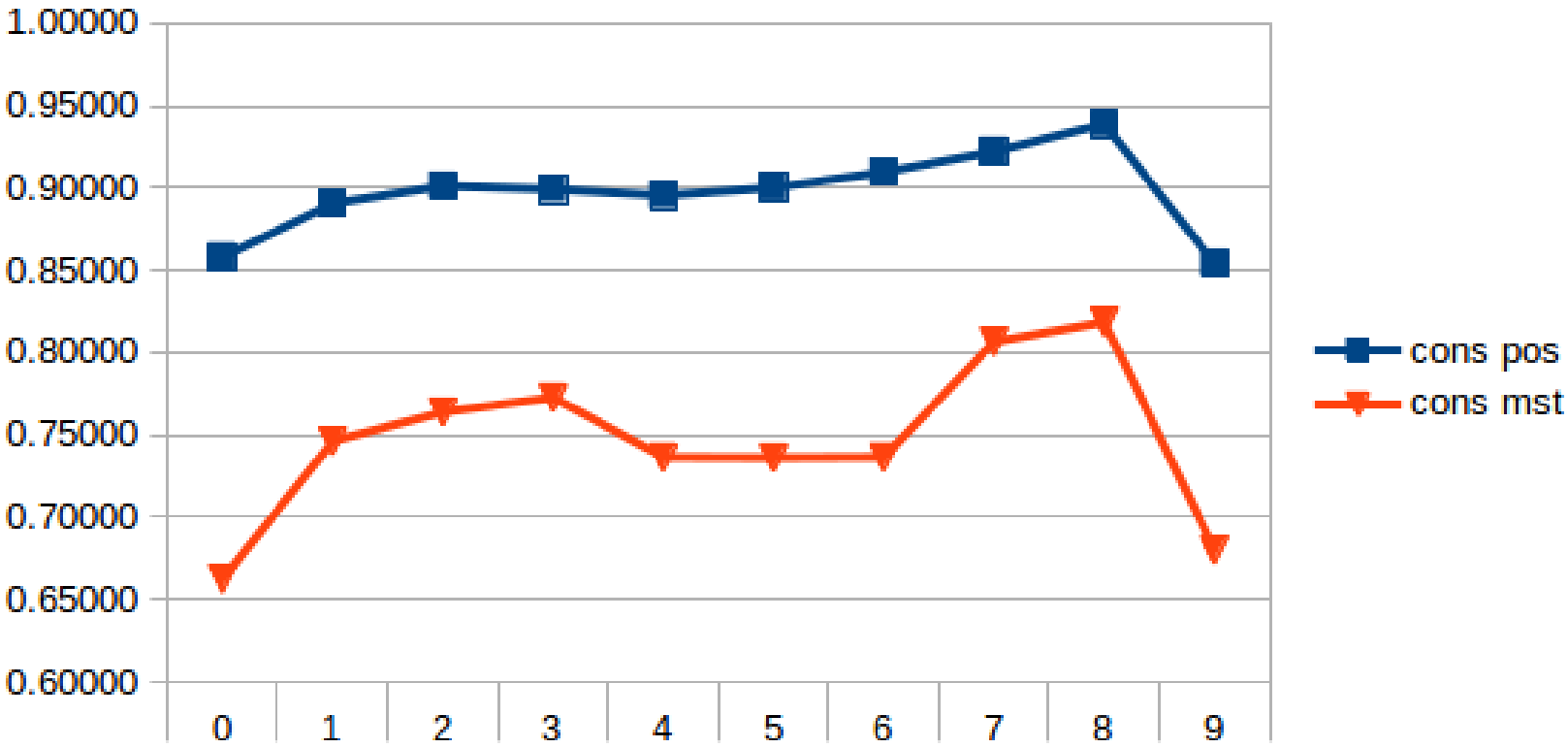
Östgöta crossvalidation

- Stagger (Östling 2013)
- Mixed 10-fold
- Consecutive 10-fold
- 4-fold by book (*balk*, 3500-5000 tokens)

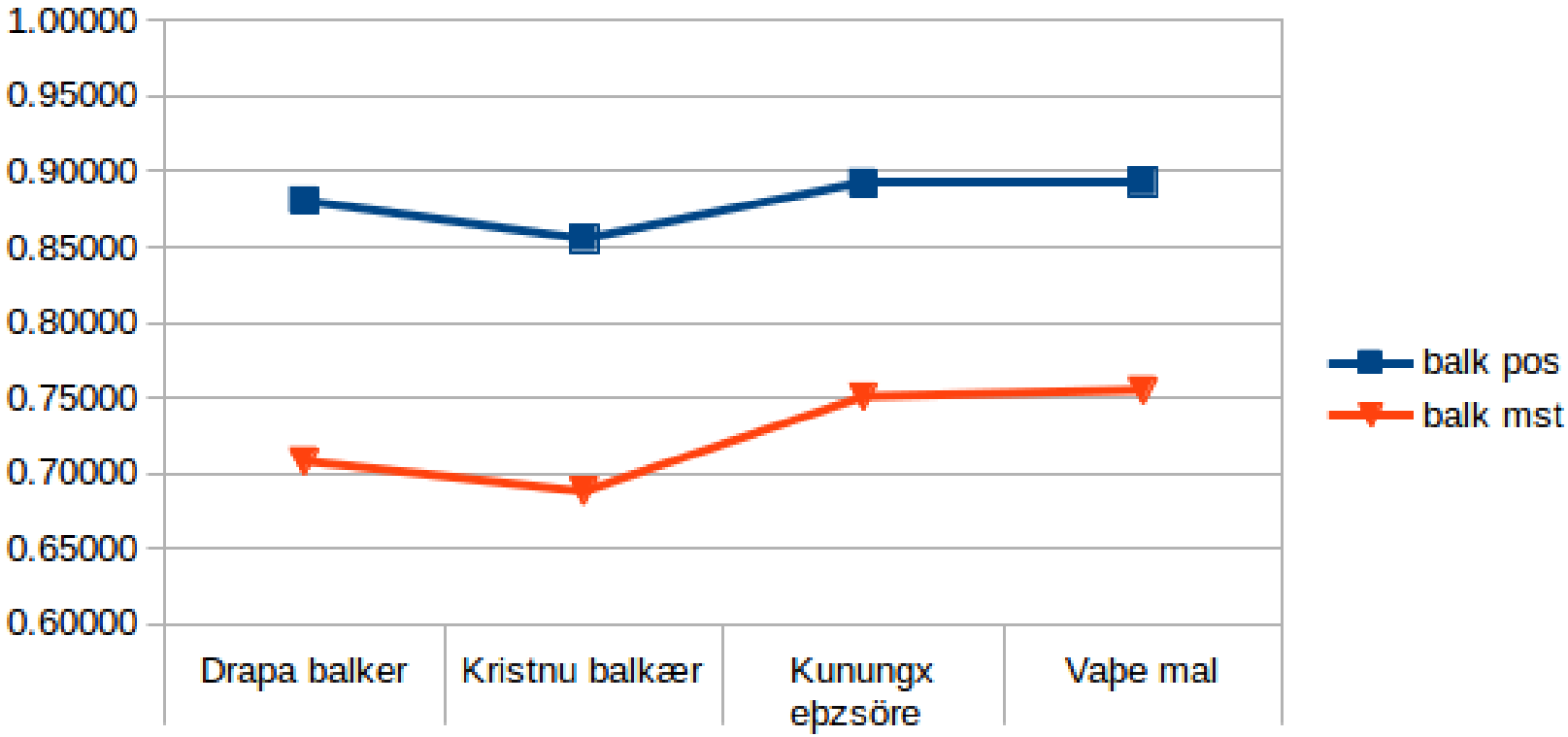
Data splitting



Data splitting



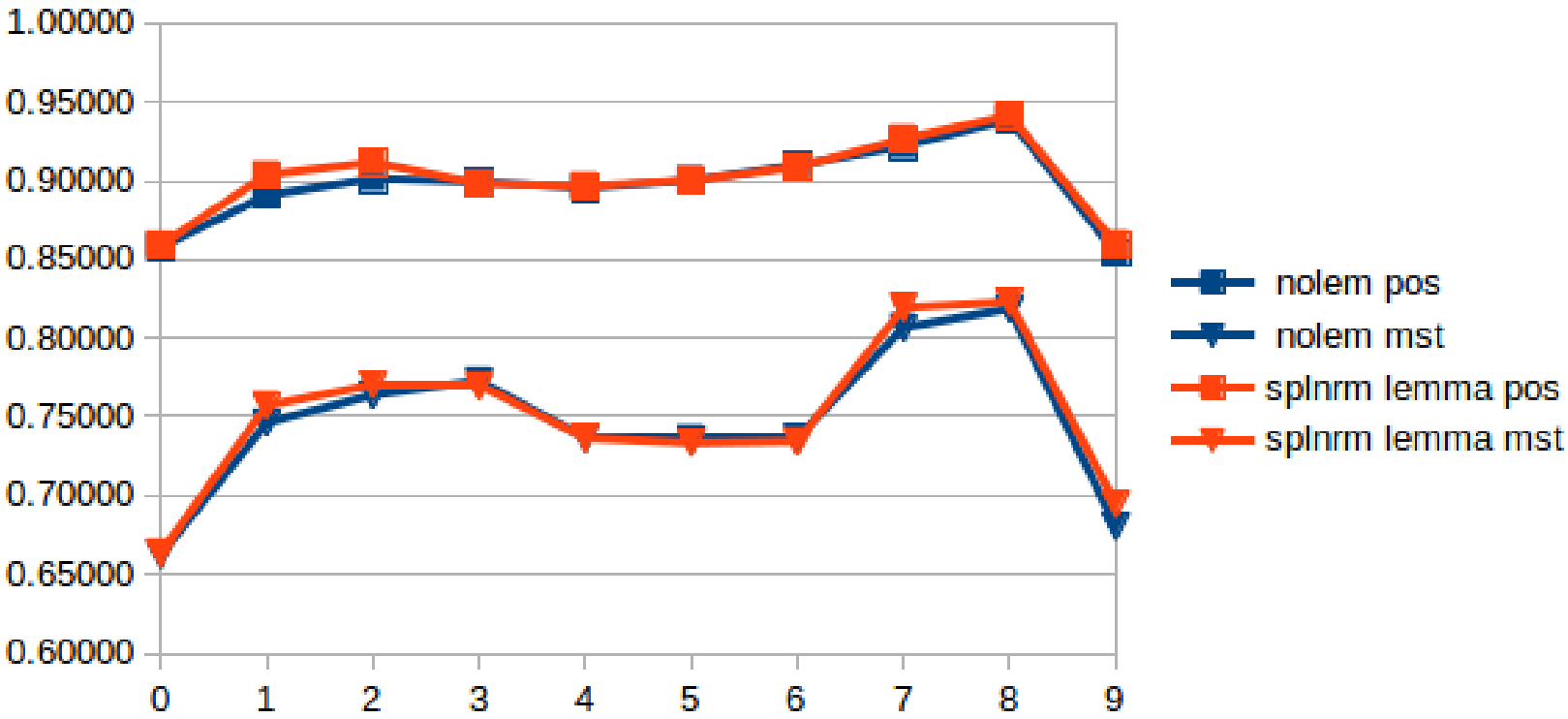
Data splitting



Lemmas and spelling

- Using lemmas as extra clues in tagging
- Simple spelling normalization
  aa $\rightarrow$ a,
  ch/c/q $\rightarrow$ k,
  th $\rightarrow$ t,
  etc.

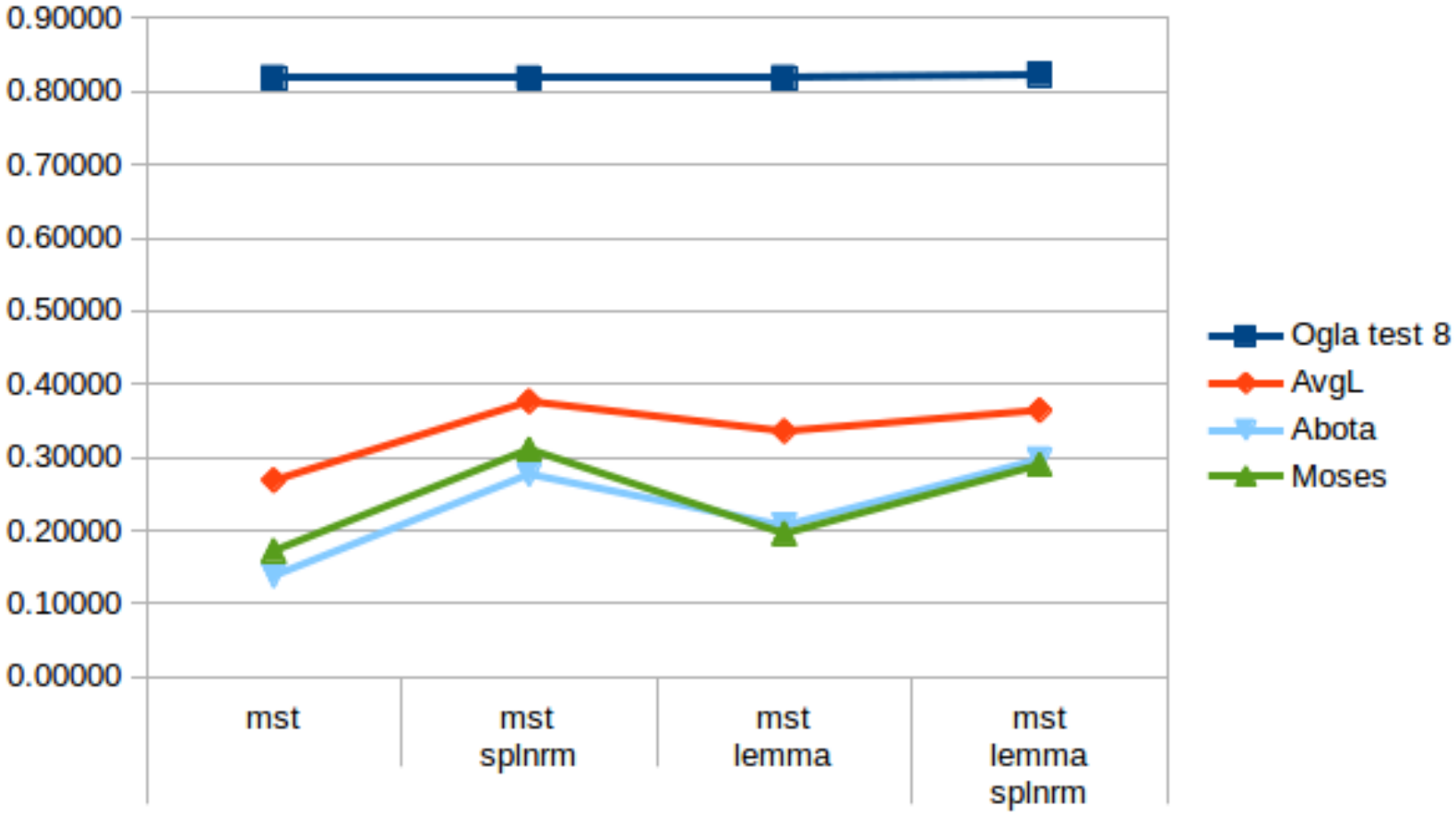
Lemmas and spelling



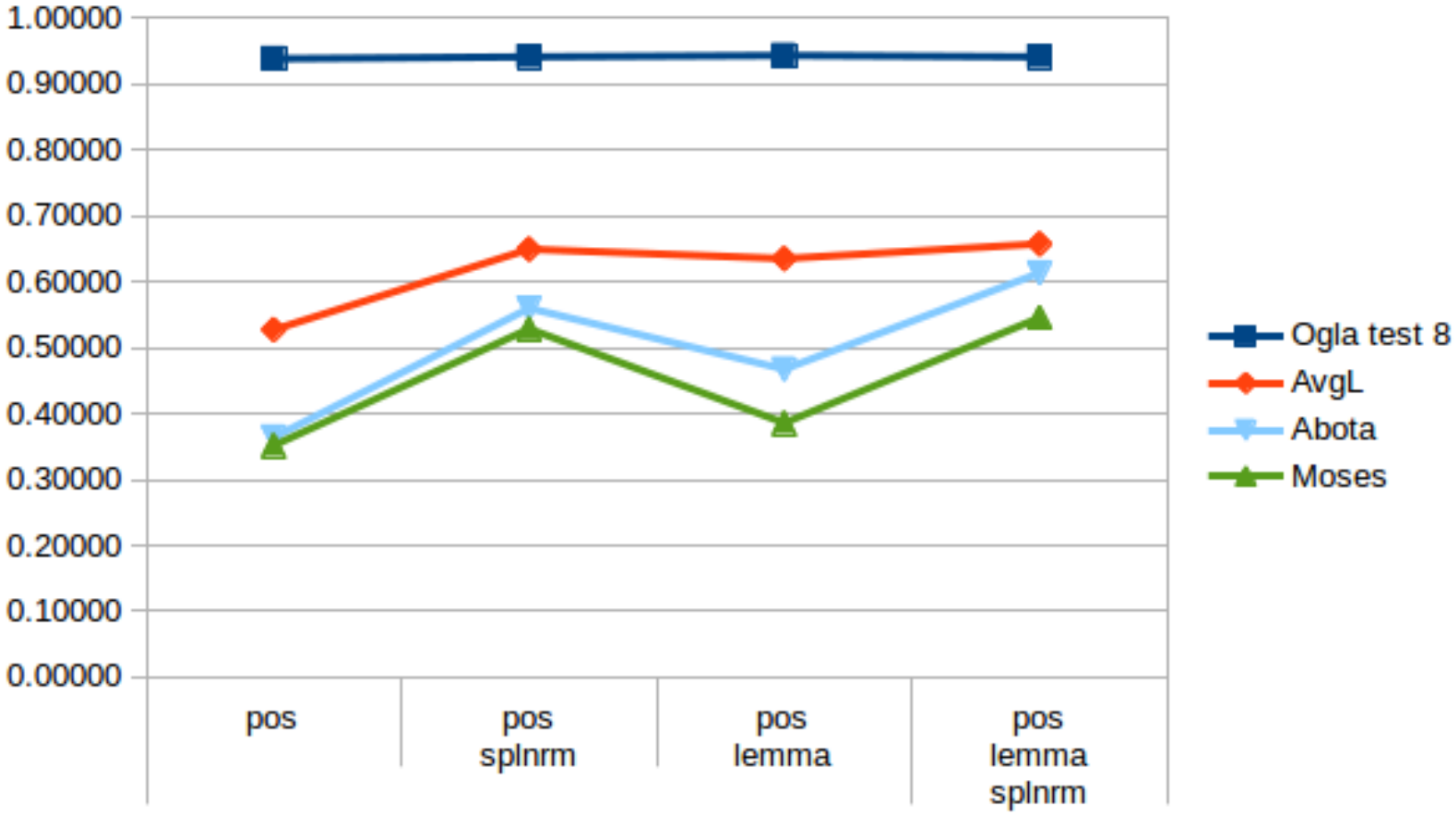
Other texts

- Östgötalagen homogenous
- Testing on other texts?
- Training on other texts?

Other texts


0.90000
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
mst
mst splnrm
mst lemma
mst lemma splnrm
Ogla test 8
AvgL
Abota
Moses

Other texts


1.00000
0.90000
0.80000
0.70000
0.60000
0.50000
0.40000
0.30000
0.20000
0.10000
0.00000
pos
pos splnrm
pos lemma
pos lemma splnrm
Ogla test 8
AvgL
Abota
Moses

OOV

	Östgl	ÄVästgl	Abota	Mose
Tokens	7%	66%	78%	79%
Types	22%	75%	85%	86%

With spelling normalization:

	Östgl	ÄVästgl	Abota	Mose
Tokens	7%	51%	60%	54%
Types	21%	66%	76%	72%

Lemmas and spelling

Improvements

- ÄVästgl
  - spelling: conj, verbs
  - lemma: verbs, conj
- Abota
  - spelling: conj, verbs, adverbs, nouns, prep
  - lemma: verbs, adverbs, conj
- Mose
  - spelling: conj, verbs
  - lemma: verbs

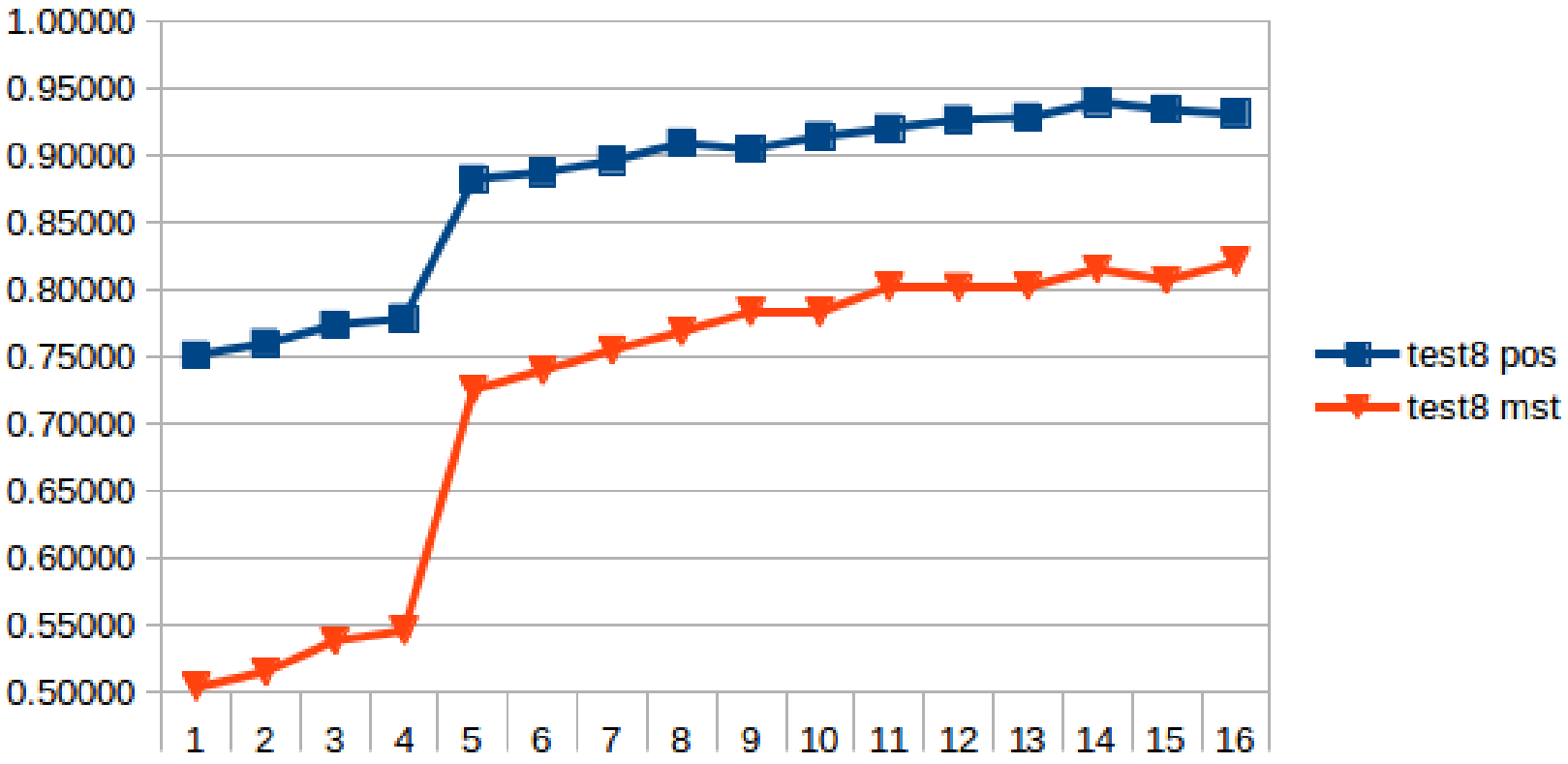
Other texts

- Training 5k Östgöta + 500 other text
- Inconclusive results
- Adding another text sometimes beneficial (compared to just using a lot more Östgöta)
- A lot more Östgöta always better when applying spelling normalization

Training size

- How much data is needed?
- Increase training size in steps

Training size



Conclusions

- Explore linguistically informed methods for annotation of OSw
- PoS experiments on manually annotated data
- Spelling normalization and lemmas improve results (autom. lemmas?)
- Fairly small amounts of annotated data necessary
- Large variation between texts