MACQUARIE
University
SYDNEY · AUSTRALIA

Morphological Segmentation with Adaptor Grammars

Kairit Sirts
Centre of Language Technology, Macquarie University

WISP, 01.09.2015

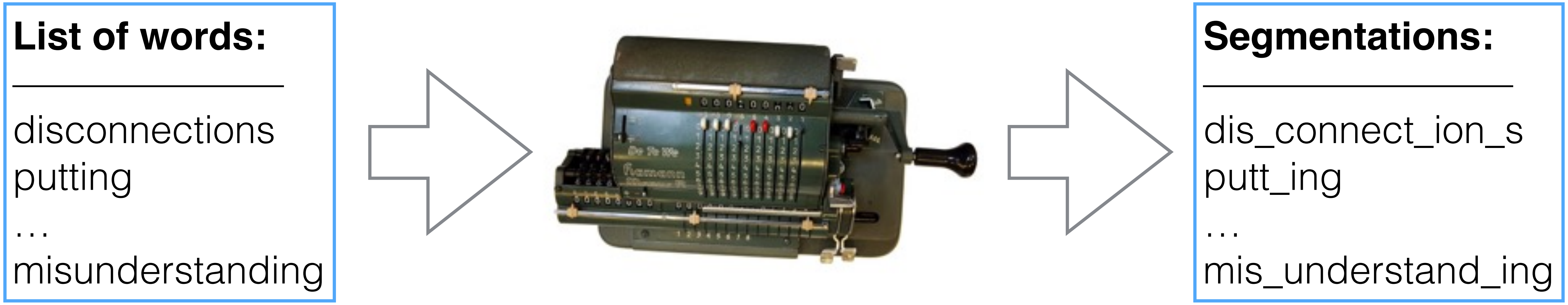
Morphological segmentation

dis_connect_ion_s

- Input is text
- Simplest form of morphological analysis
- Assumes concatenative morphology

putt_ing or **put_ting** ?

Computational modeling



List of words:

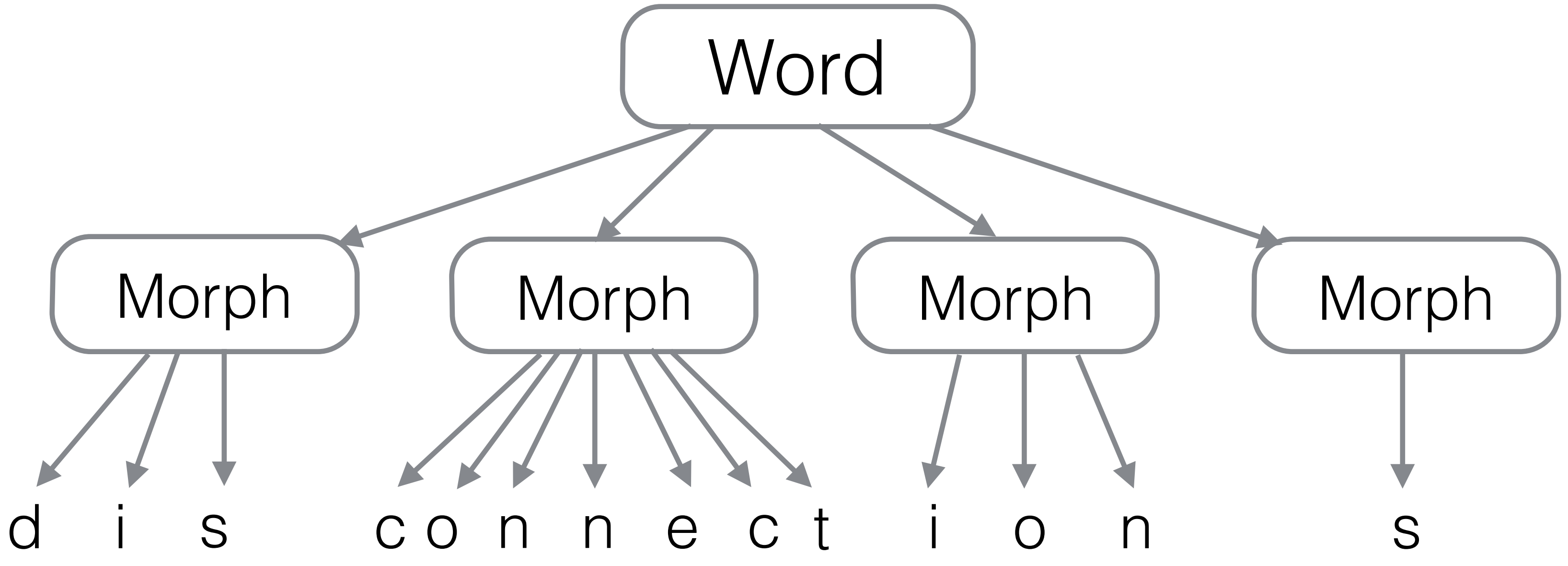
disconnections
putting
…
misunderstanding

Segmentations:

dis_connect_ion_s
putt_ing
…
mis_understand_ing

Adaptor Grammar model

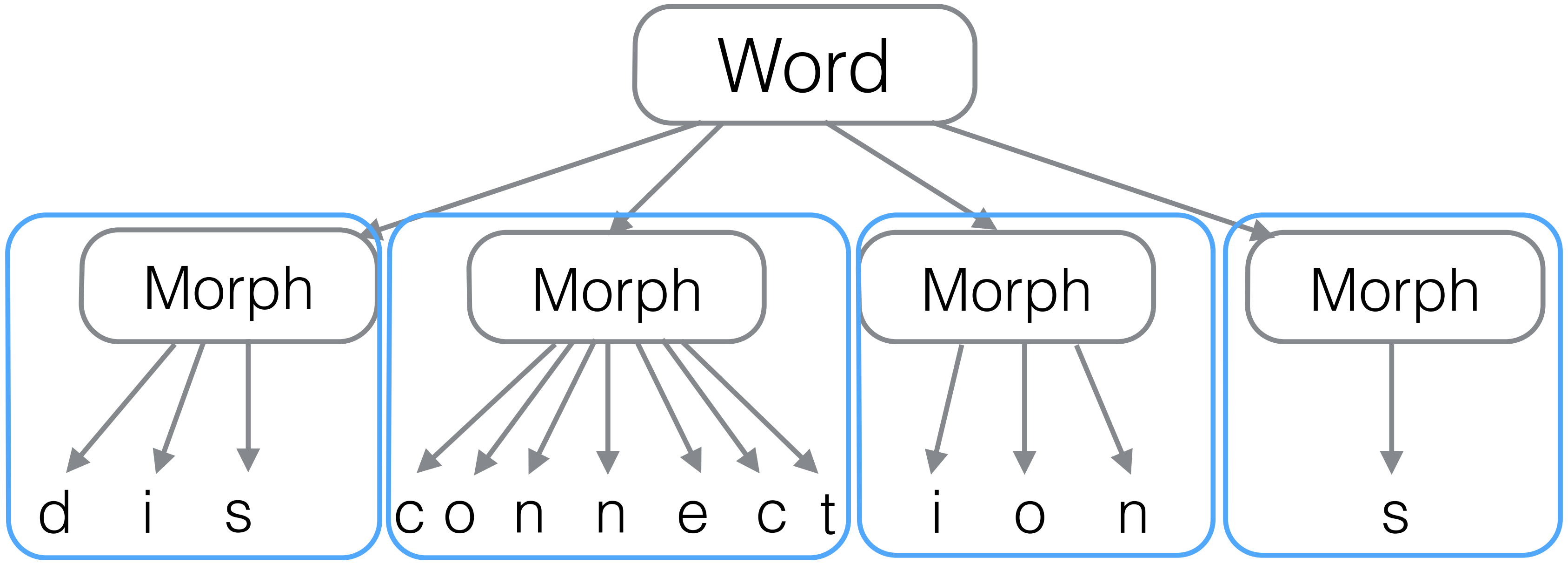
- Parsing model, assuming context-free grammar



- Prefers to reuse the generated subtrees

Adaptor Grammar model

- Parsing model, assuming context-free grammar



- Prefers to reuse the generated subtrees

SubMorph grammar

Word —> Morph$^+$

Morph —> SubMorph$^+$

Compounding grammar

Word —> Compound$^+$

Compound —> Prefix* Stem Suffix*

Prefix, Stem, Suffix —> SubMorph$^+$

CollocMorph grammar

Word —> Colloc$^+$

Colloc —> Morph$^+$

Morph —> SubMorph$^+$

Data and experimental setup

- List of word types from newspaper corpora (lexicon)
- 5 training sets: 10K - 50K most frequent words
- Train different models with all those training sets with all grammars
- Test on a smaller held-out annotated word list
- Experiment on English and Estonian
- The experiments *were not* designed for acquisition research

For the purpose of this talk:

- Assume *as if* it was an acquisition study
- What kind of scenarios could be interesting?
  - Look at certain suffixes
  - How do suffix accuracies vary with the amount of training data?
  - How do the grammars affect the suffix accuracy?

Suffixes

English:

’s	- noun genitive
s, es	- plural noun, 3rd person verbs
ed	- past tense verbs
ing	- present participle verbs
ly	- forming adverbs
ness	- derivational suffix
er	- derivational suffix

Estonian:

ma	- verb base form
da	- *to* (*to look, to play*)
n	- 1st per sg present verb
b	- 3nd per sg present verb
s	- 3rd per sg past verb, sg inessive noun (*in*)
l	- sg adessive noun (*on*)
le	- sg allative noun (*onto*)

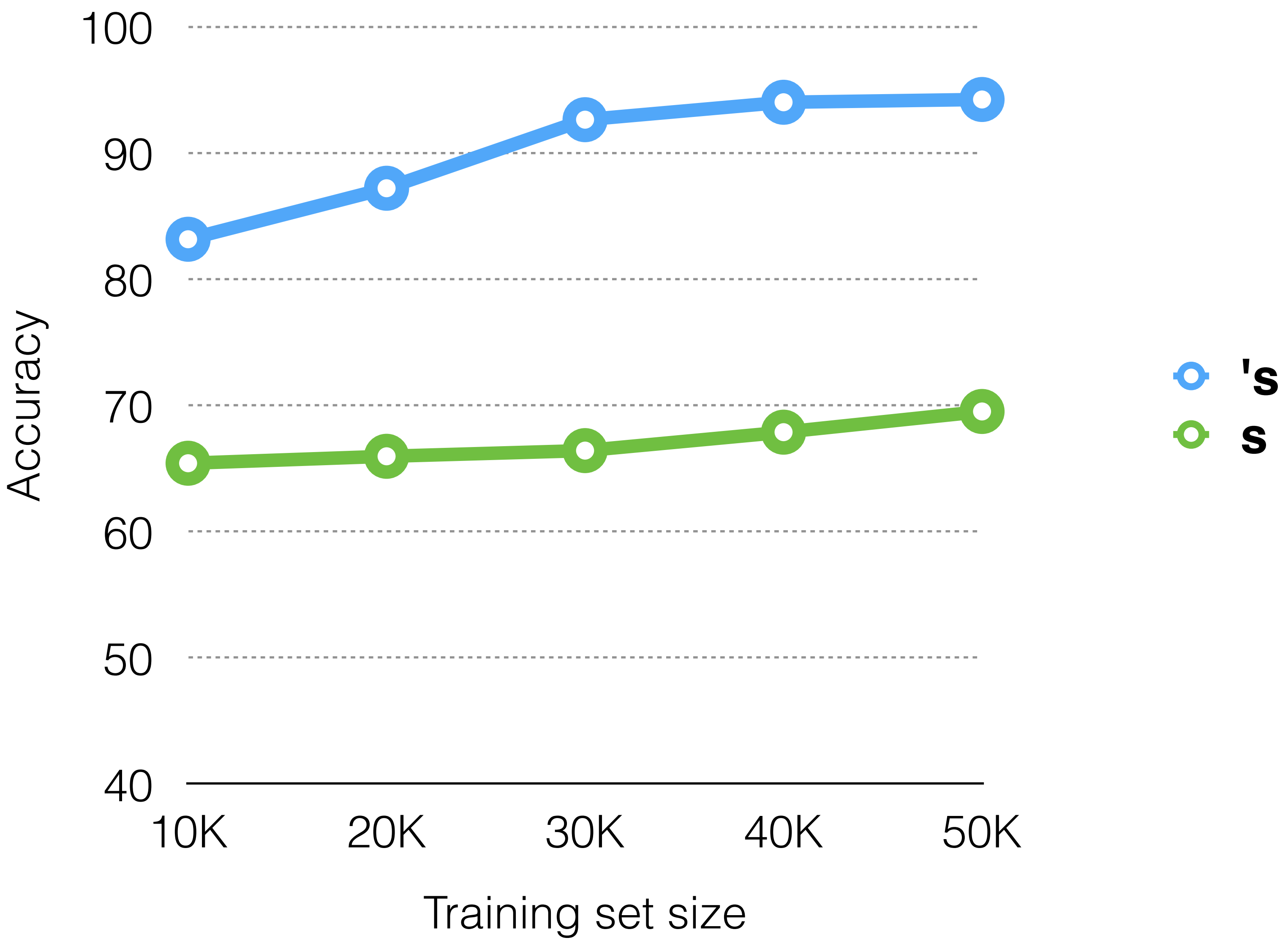
Does the accuracy increase with more training data?

- General segmentation accuracy increases with more training data
- Treat the model as a learner exposed to data
- More data —> more accurate suffixes?

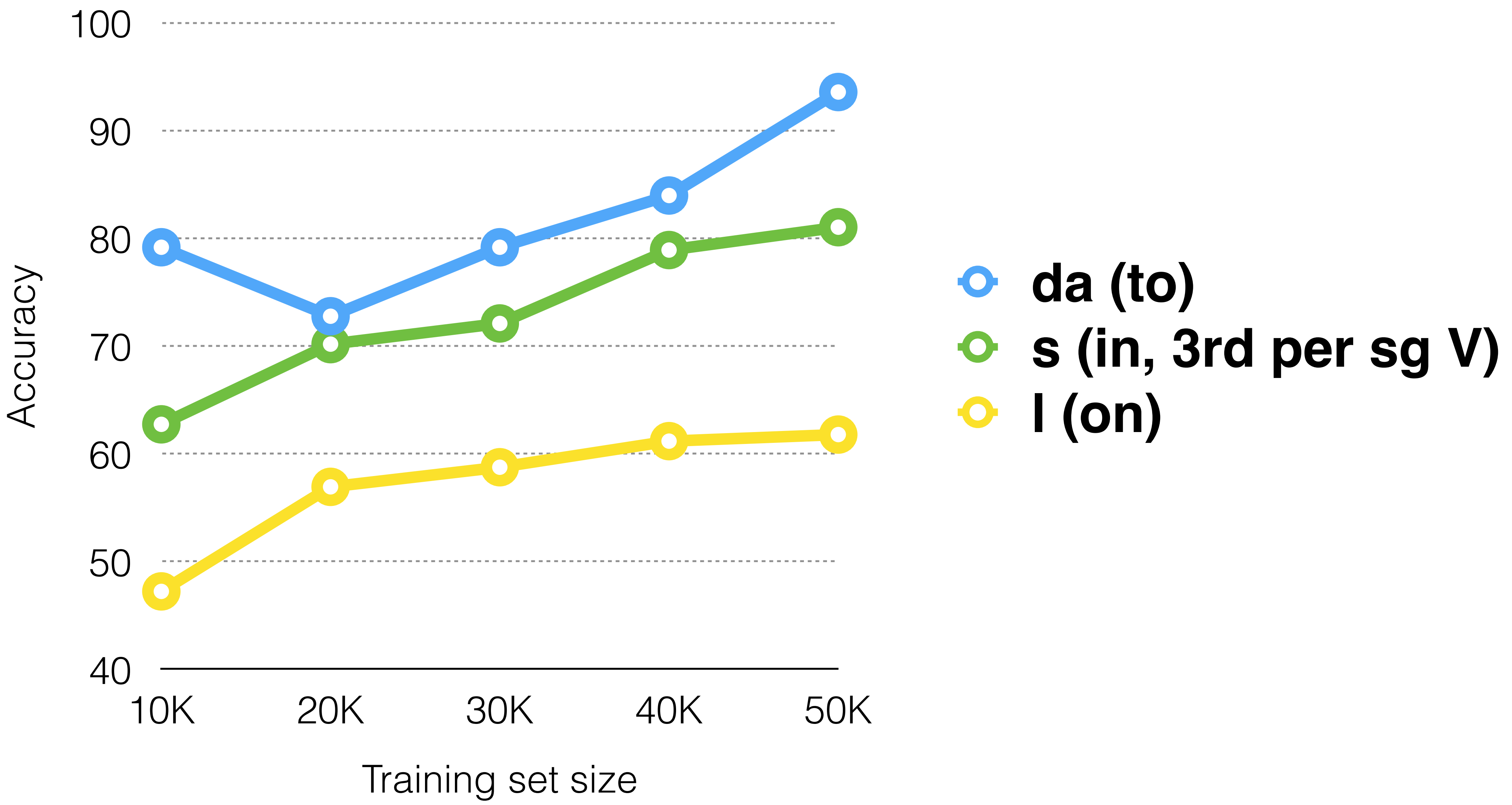
Does the accuracy increase with more training data?

- Not really!
- For most suffixes no consistent improvement
- For some suffixes, things seem to get worse
  - English: **ed**, **ly**
- Some suffixes improve under SubMorph grammar:
  - English: **’s**, **s**
  - Estonian: **da**, **s**, **l**

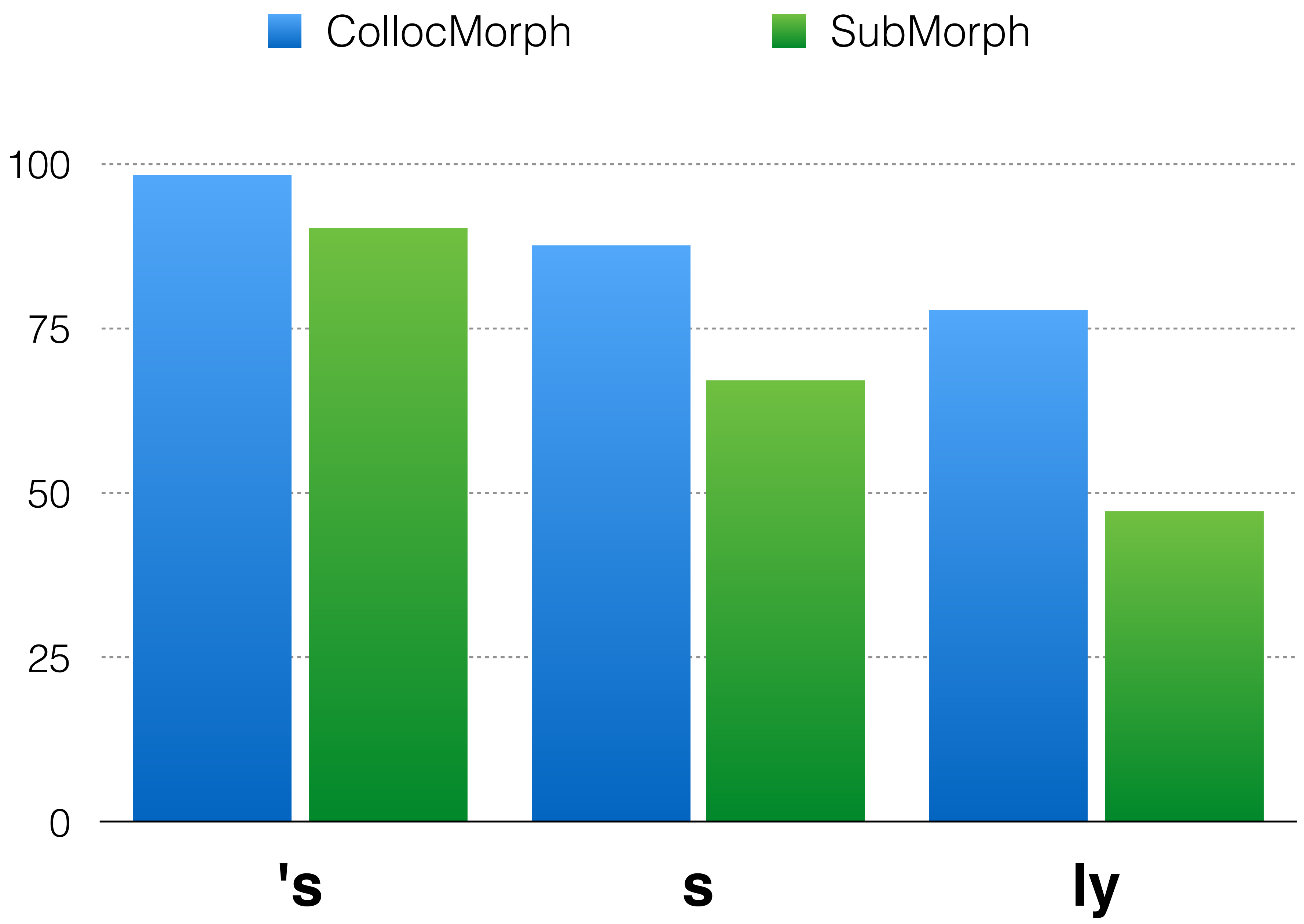
English **’s** and **s**



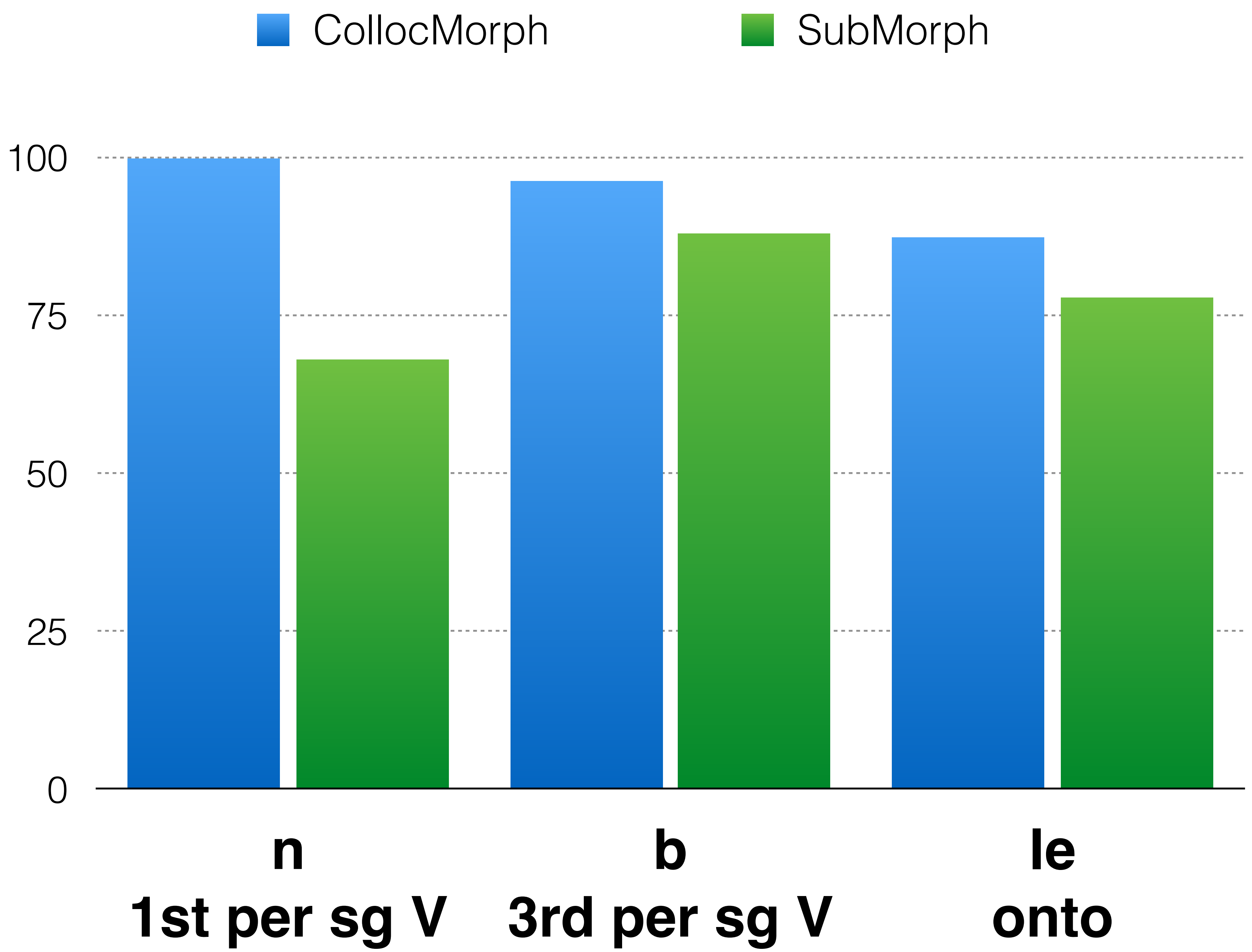
Estonian **da**, **s** and **l**



CollocMorph mostly the best

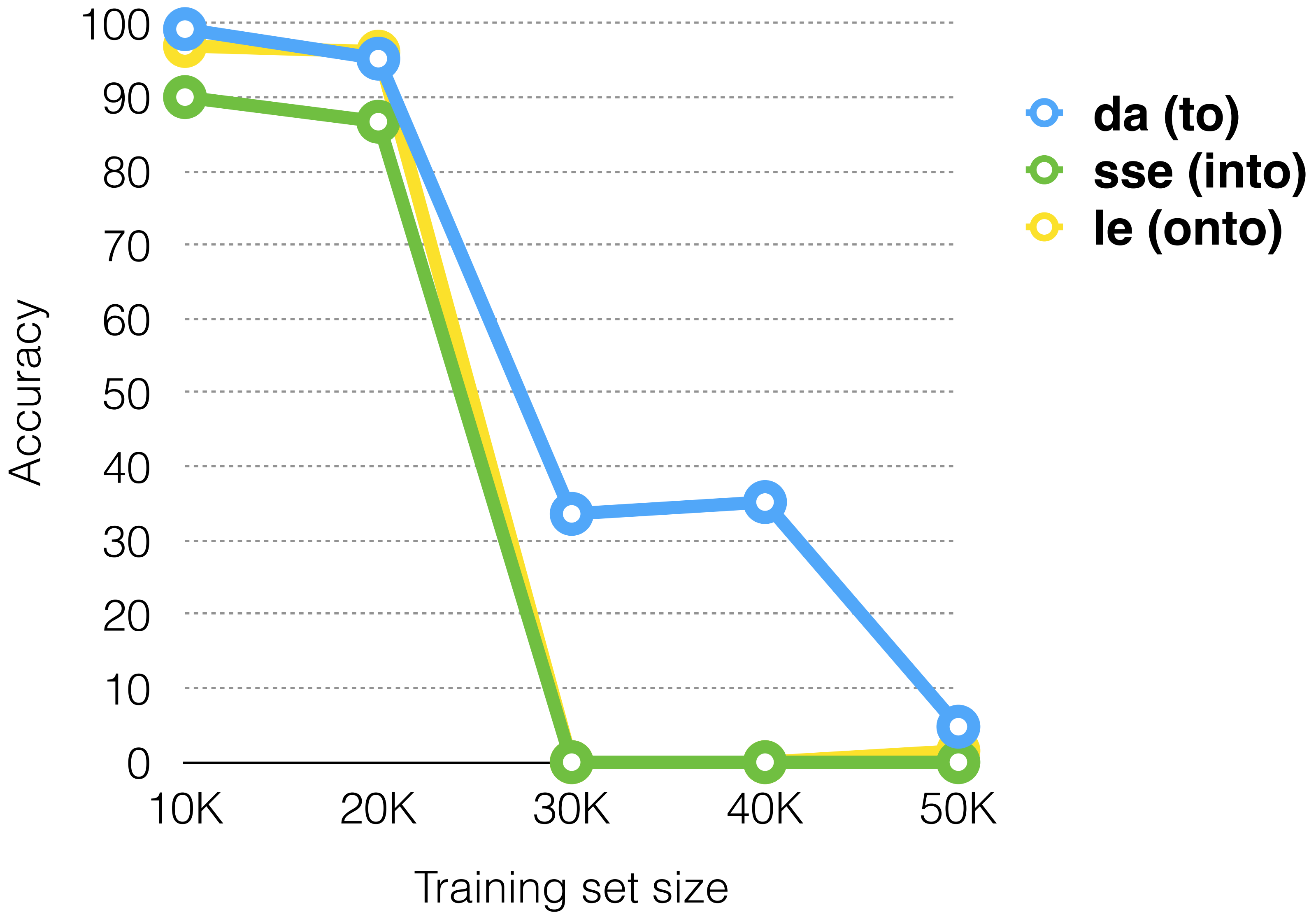


CollocMorph mostly the best



Compounding oscillates between different solutions

da	vs	**d_a**
sse	vs	**s_se**
le	vs	**l_e**



Possibilities for language acquisition research?

- Train on phonetic/speech data
  - deal with suffix allomorphy:
  - /s/ vs /z/ in English noun plural
  - orthographic variation of the stems
  - the stem in *put* and *putting* is phonologically the same
- Train on child directed speech data
  - Apply the model to child's speech data
  - Do the results align in any way with infant research?

Conclusions

- Computational model for morphological segmentation
- Experimental setting *was not* designed for acquisition research
- Searched for interesting results in suffix morphology
- Perhaps provides interesting opportunities for infant speech researchers?