Distributional Bootstrapping: a Memory-Based Learning Approach

Giovanni Cassani
CLiPS @ Uantwerpen
18th December 2015
CLIN26@Amsterdam

Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

Breaking into language

A chicken-and-egg problem: understanding a system when you only have access to the system itself.

What information might provide the starting point? How does a child **bootstrap** her language learning?

Breaking into language

A chicken-and-egg problem: understanding a system when you only have access to the system itself.

What information might provide the starting point? How does a child **bootstrap** her language learning?

- Semantics? (Pinker, 1984)

Breaking into language

A chicken-and-egg problem: understanding a system when you only have access to the system itself.

What information might provide the starting point? How does a child **bootstrap** her language learning?

- Semantics? (Pinker, 1984)
- Prosody? (Christophe et al, 2008)

Breaking into language

A chicken-and-egg problem: understanding a system when you only have access to the system itself.

What information might provide the starting point? How does a child **bootstrap** her language learning?

- Semantics? (Pinker, 1984)
- Prosody? (Christophe et al, 2008)
- Syntax? (Gleitman, 1990)

Breaking into language

A chicken-and-egg problem: understanding a system when you only have access to the system itself.

What information might provide the starting point? How does a child **bootstrap** her language learning?

- Semantics? (Pinker, 1984)
- Prosody? (Christophe et al, 2008)
- Syntax? (Gleitman, 1990)
- Distributional co-occurrences? (Maratsos, 1980)

Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

From co-occurrences…

Distributional bootstrapping proposes that children track **co-occurrences among linguistic units** in order to understand a great deal of how language works.

Distributional information has been shown to be useful to

segment words (Saffran et al. 1996, Goldwater et al., 2009),

identify phonemes (Maye et al., 2002),

build a semantic space (Baroni et al., 2007),

and more.

…to lexical categories

A specific task in which distributional cues are very useful is the **grouping of words** into lexical categories, which are in turn essential to master language.

Different proposals so far have investigated

if certain *types of contexts* are more useful than others

(Mintz, 2002)

if a handful of **lexically specific** cues can do the job

(Mintz, 2003)

what computational mechanisms the child can use

(Parisien, 2008)

Lexical specificity

The interest in lexically specific cues stems from the questionable assumption that children cannot evaluate all possible distributional cues and thus have to focus on just some of them. However, it also gives a very **transparent, usage-based account** of lexical category acquisition.

The question shifts to **what** such cues are and *how* can they be identified.

Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

Frame-based approaches

Frequent Frames (Mintz, 2003)

The most frequent *a_X_b* tri-grams

in which two specific words flank a slot that can be taken by any other word, are a very good cue to group the words occupying the empty slot.

Mintz defined most frequent as 45, called such cues **frequent frames**, and showed that they are extremely precise in grouping words. However, precision comes at the expense of recall.

Frame-based approaches

Flexible frames (St Clair et al, 2010)

the most frequent *a_X* and *X_b* bi-grams are better

St Clair and colleagues took the 45 most frequent words, built 90 (left and right) bi-grams, called them **flexible frames**, implemented a feed-forward NN and showed that such cues were better than frequent frames: a bit less precision, but far better recall.

Frame-based approaches

Unrestricted frames (Cassani et al, 2015)

Where do frames come from?

If the picklock to language is in distributional co-occurrences, frames' saliency should be defined distributionally, with **no a-priori restrictions**.

What frames become salient when no assumptions are built-in and only distributional information is considered?

Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

Experiment

Goal

Evaluate learning of broad lexical categories

(i) using **lexically specific** cues

(ii) under different *combinations of restrictions* on available types of contexts and distributional information

(iii) on typologically different languages, to see how important morphological complexity is.

Experiment

Goal

Evaluate learning of broad lexical categories

(i) using **lexically specific** cues

(ii) under different *combinations of restrictions* on available types of contexts and distributional information

(iii) on typologically different languages, to see how important morphological complexity is.

Contexts can be only bi-grams, *only tri-grams*, or **both**.

Distributional information include token frequency, *type frequency* and average conditional probability of context given word, **P(context|word)**.

Experiment

Methods

We mapped gold-standard PoS tags to a five categories tag-set, consisting of **N**ouns, **V**erbs, **Adj**ectives, **Adv**erbs, and **F**unction words, to focus on content words categorization.

We evaluate **recall and precision** on the words that were uttered by all children and all caregivers.

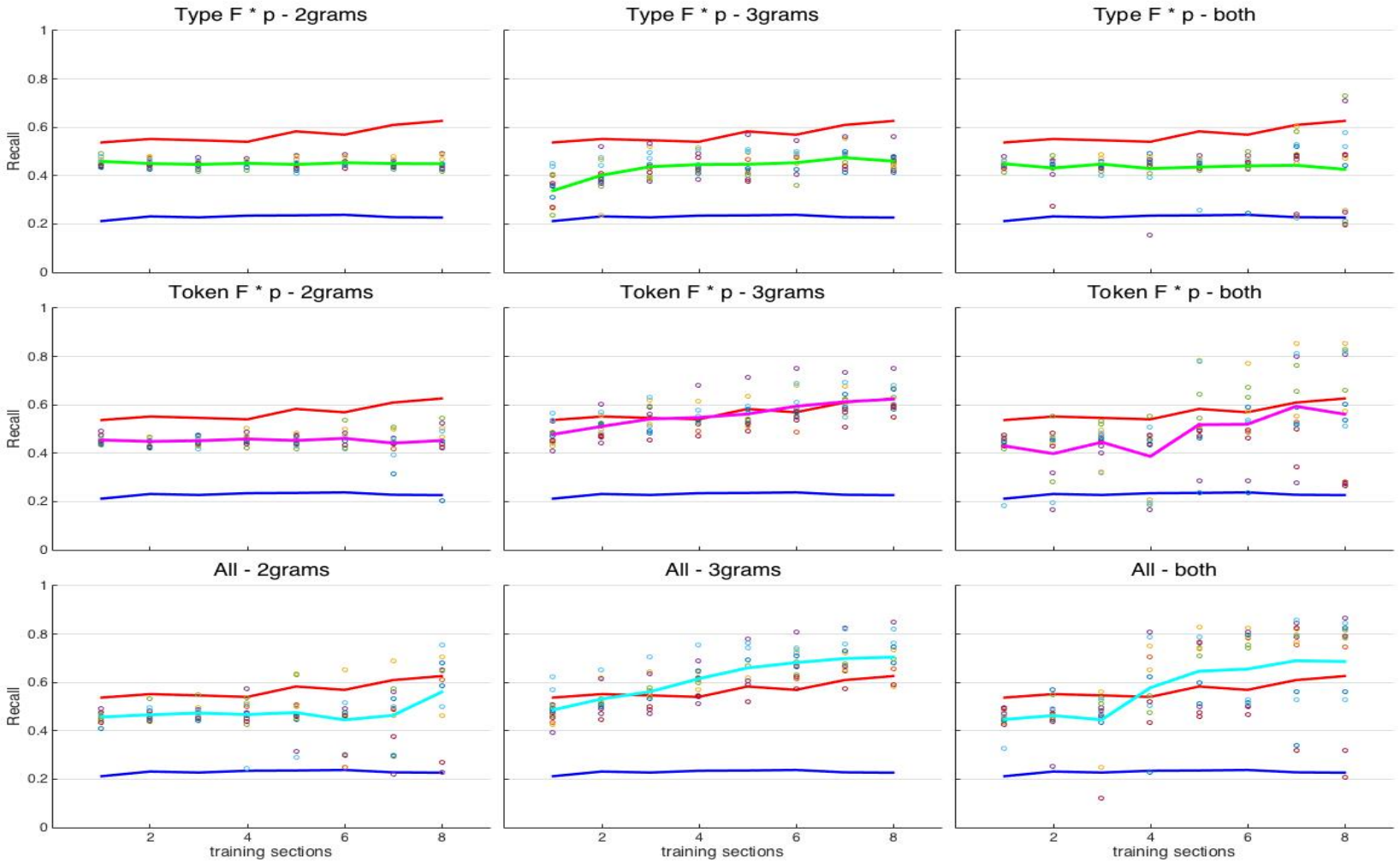
Experiment

Data

Each corpus was divided in sections according to the age in months; we **selected cues and trained the MBL incrementally** on the first *n* months and tested on the last *m* months (*n* and *m* vary depending on the language)

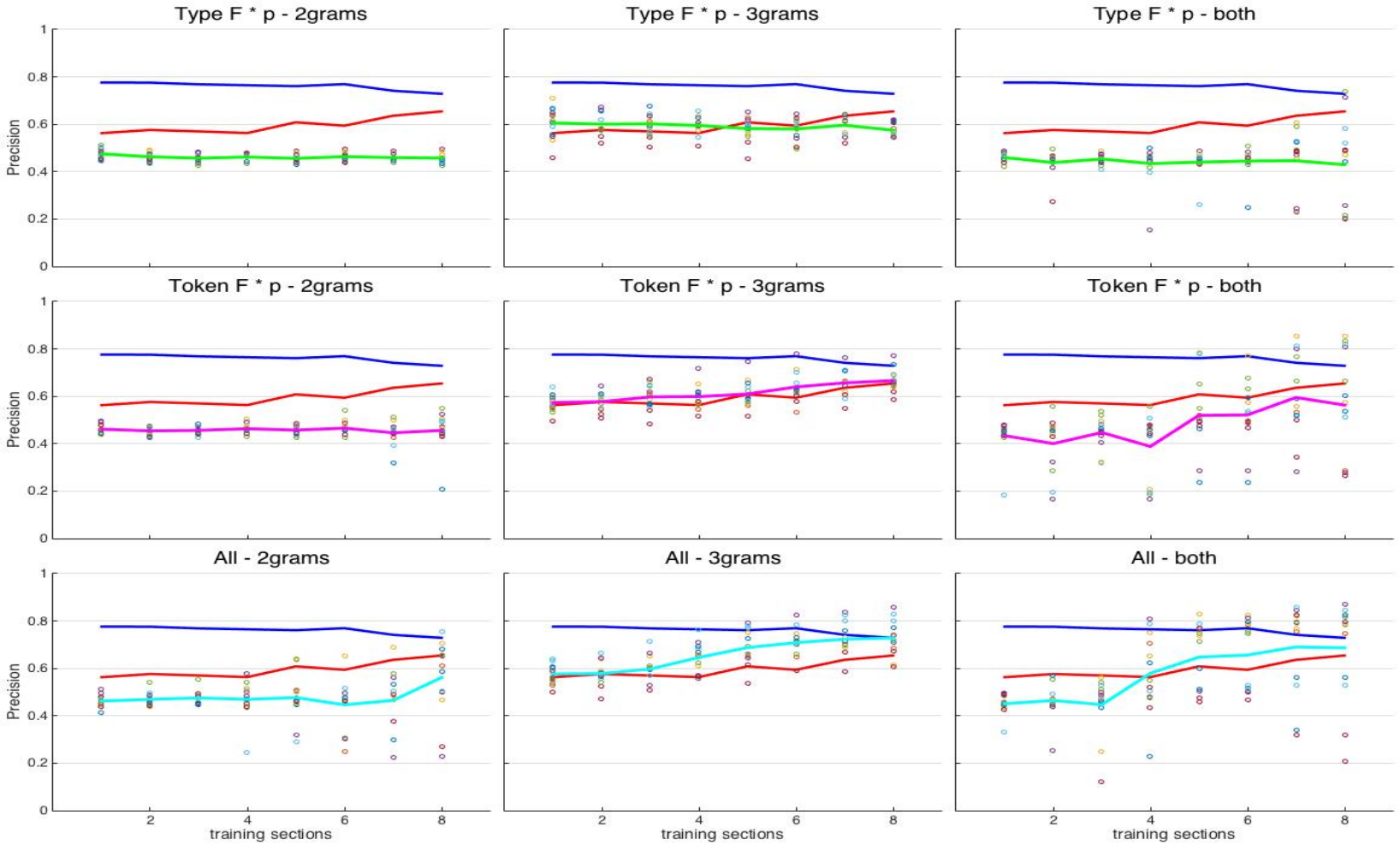
Experiment

Results – English, recall



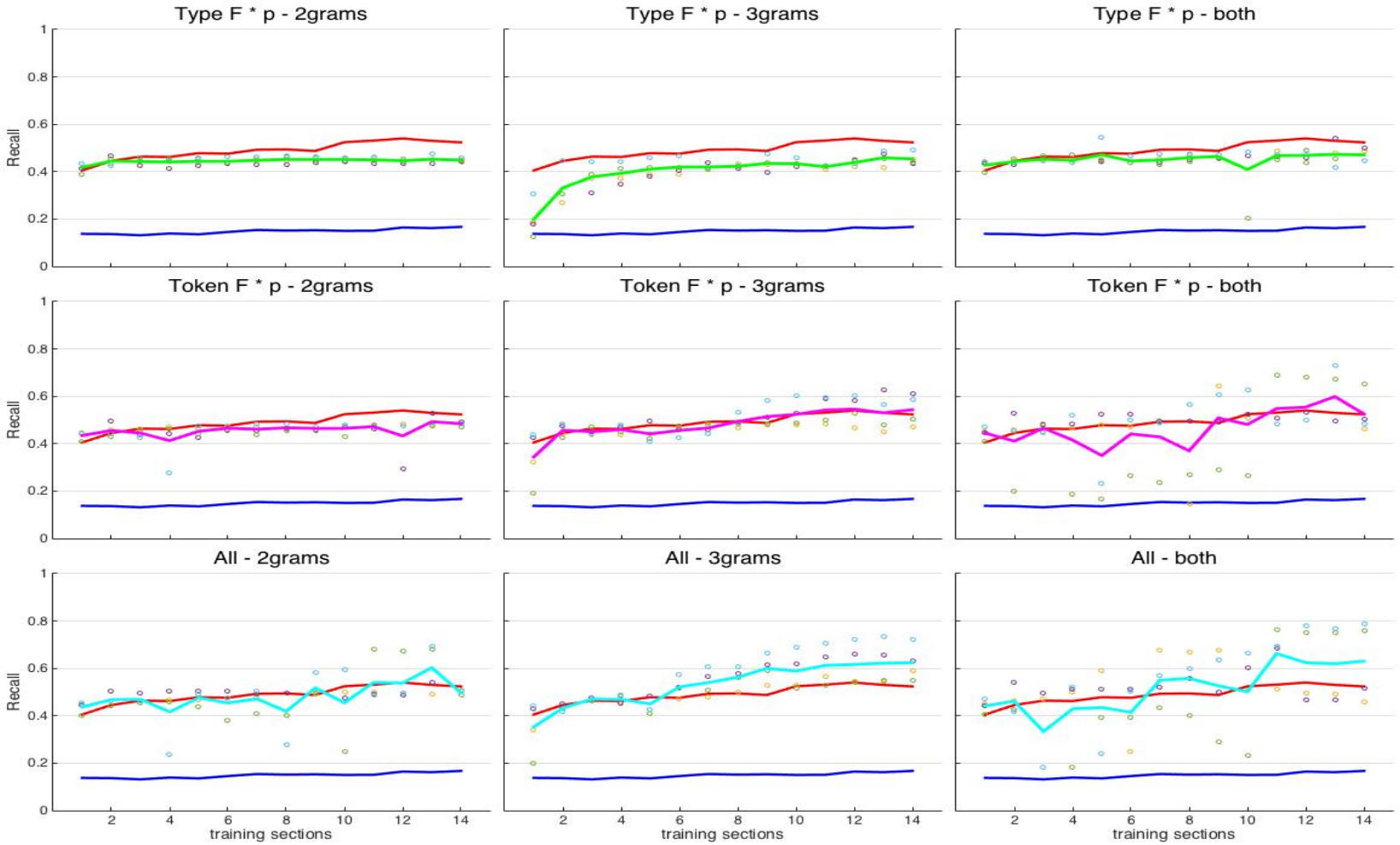
Experiment

Results – English, precision



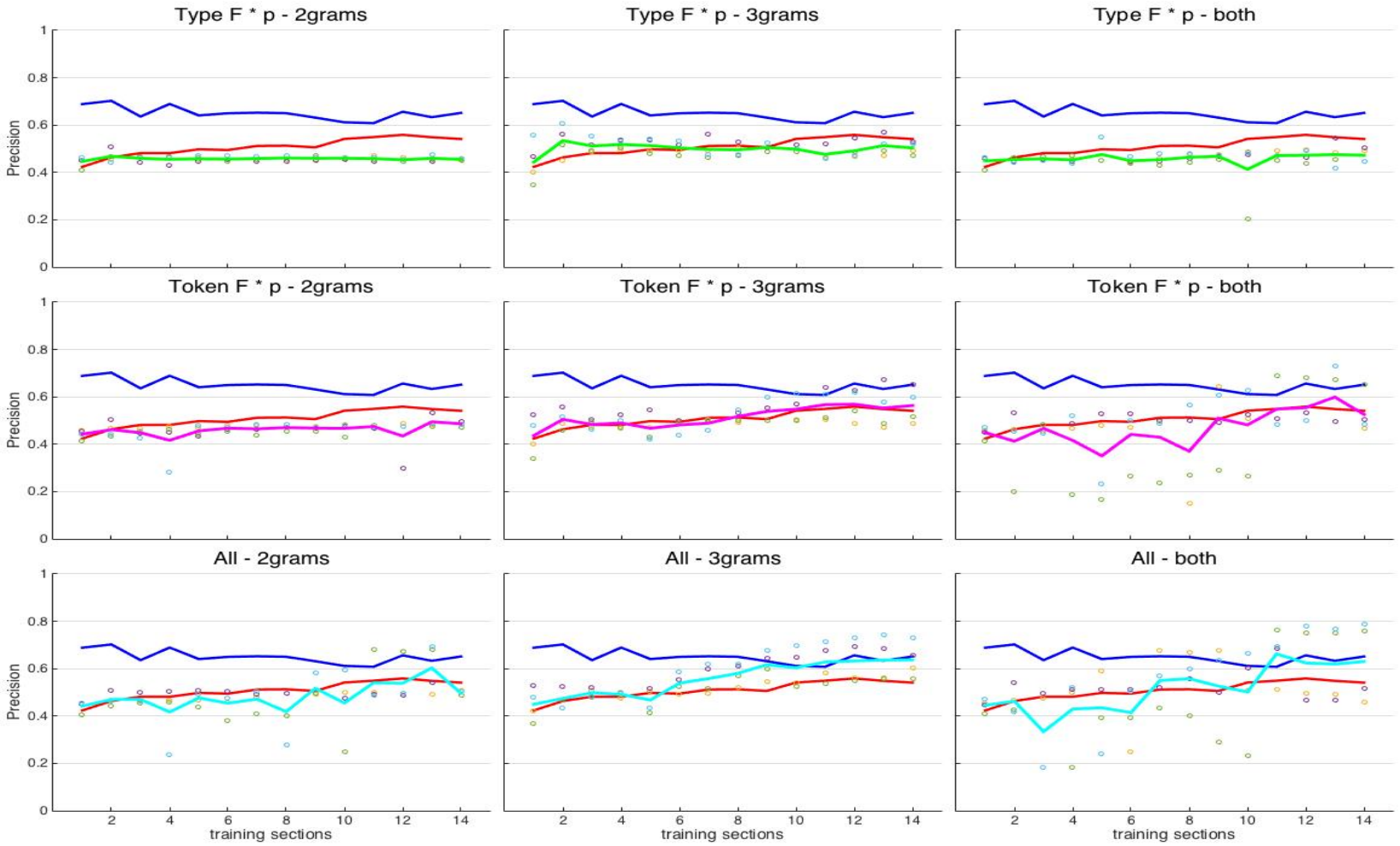
Experiment

Results – French, recall



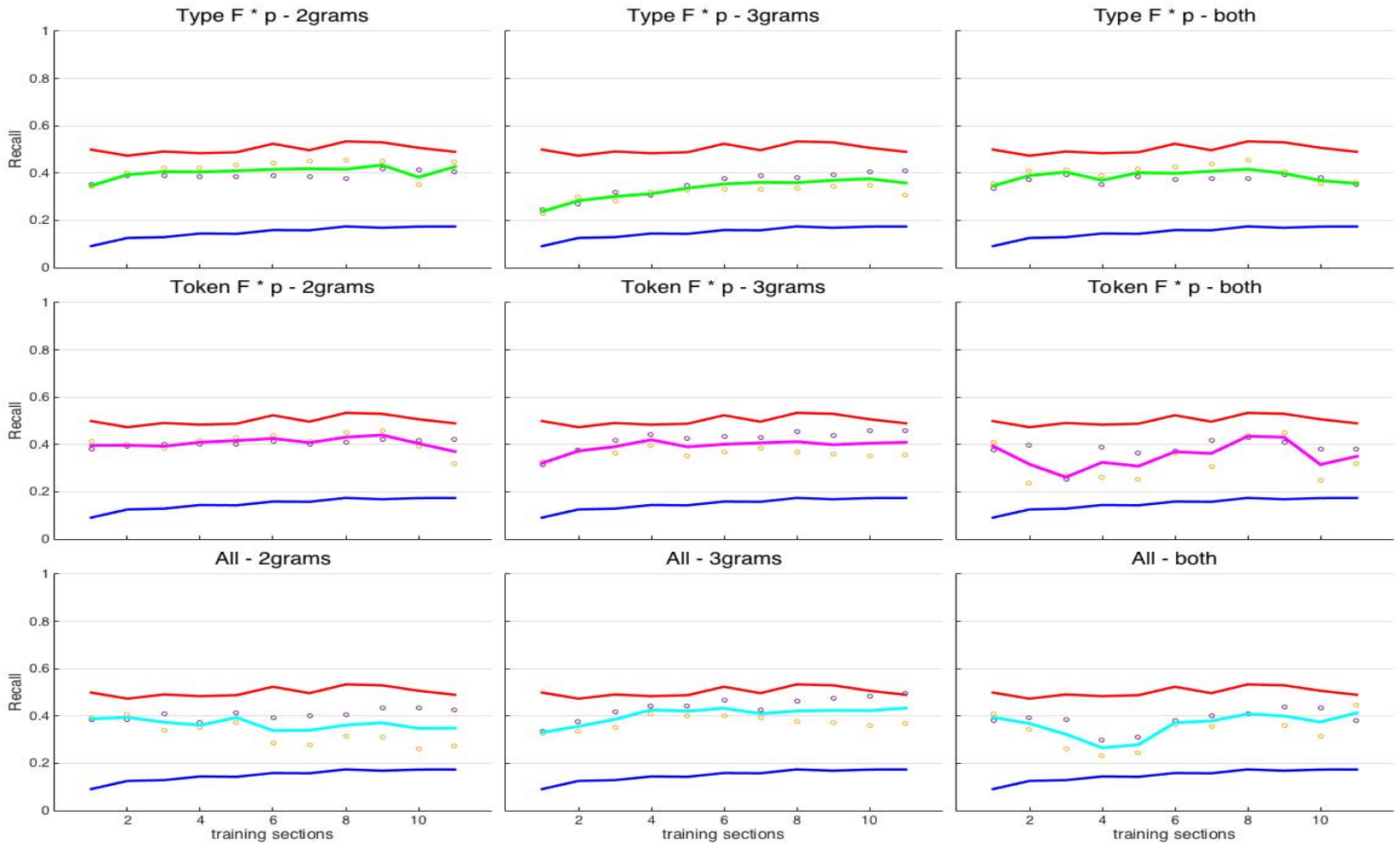
Experiment

Results – French, precision



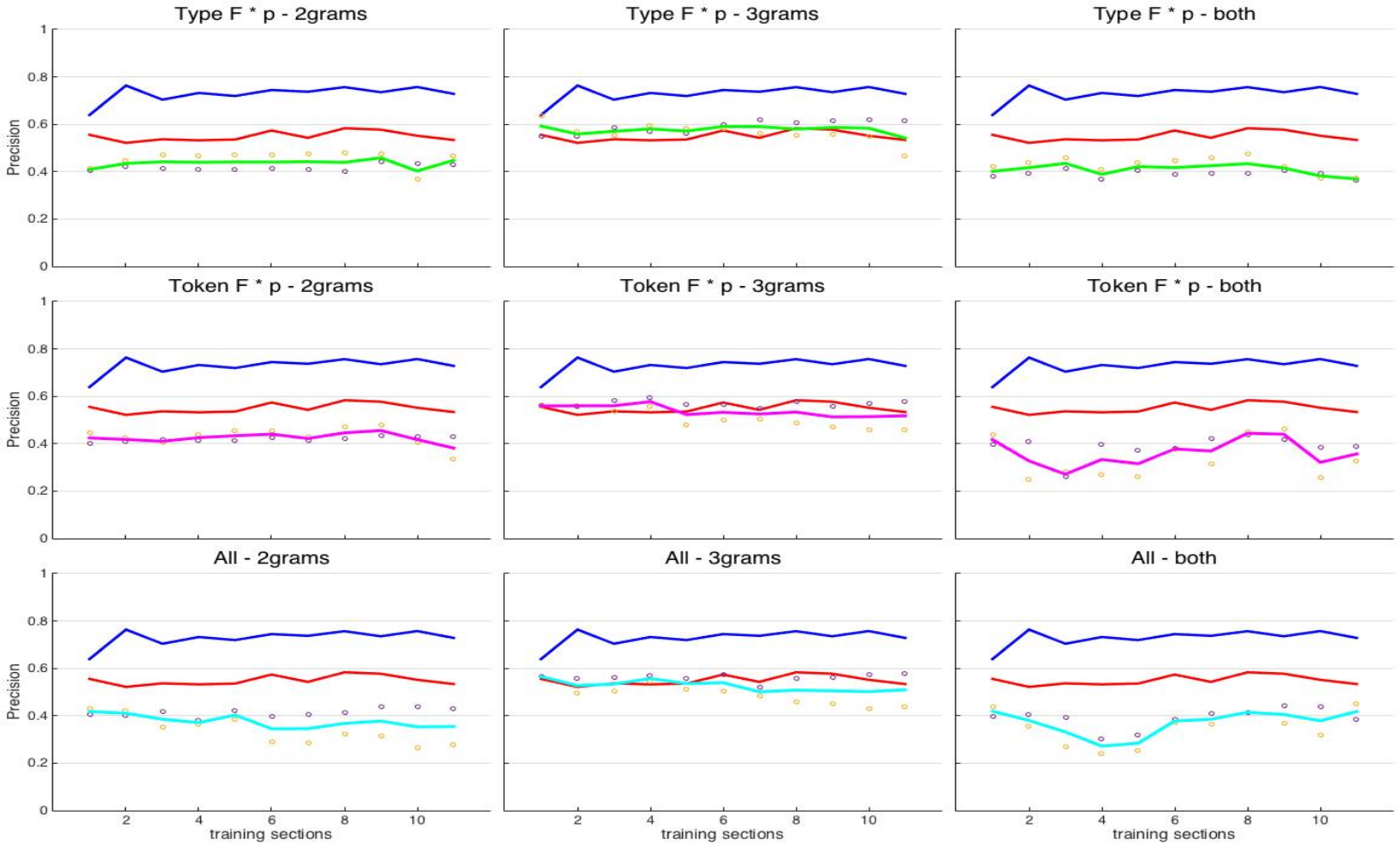
Experiment

Results – Hebrew, recall



Experiment

Results – Hebrew, precision



Outline

- Background
- Distributional bootstrapping
- Frame-based approaches
- Experiment
- Conclusions

Context restrictions

When bi-grams and tri-grams are both considered, the curve is similar to bi-grams at first and become more similar to tri-grams later, suggesting that it might take more time for longer regularities to emerge when no restriction is imposed.

Distributional information

Generally, ***when more information is made available learning is better***, both relatively – stronger increase from first to last stage, and absolutely – higher figures at the end of learning.

However, more information comes with more salient cues being stored, challenging the idea of lexical specificity as a solution to reduced capacities.

Conclusions

A matter of cues

We clearly see that irrespective of the model **performance goes down and learning fades when the morphological complexity of the language increases.**

One can always run the same models using morphemes, but then it's not clear how the child decides upon the right level of analysis.

Conclusions

So…what?

We should stop

X assuming language-specific linguistic knowledge in models of bootstrapping (words *vs* morphemes)

X assuming that higher-order tasks are tackled only after lower-order tasks are accomplished (categorization after segmentation)

We should

✓ move to **perceptually motivated cues** and linguistically motivated outcomes (Baayen et al, 2015).

✓ isolate which pieces of information makes cues salient, and test their sufficiency and necessity

References

Maratsos, M. P., & Chalkley, M. A. (1980). The Internal Language of Children Syntax: The nature and ontogenesis of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. 2, pp. 127-213). New York, NY: Gardner Press.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Gleitman, L. R. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1), 3-55.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35(4), 606-621.

Maye, Jessica, Janet F. Werker, and LouAnn Gerken. "Infant sensitivity to distributional information can affect phonetic discrimination." *Cognition* 82.3 (2002): B101-B111.

Mintz, T. H., Newport, E. L., & Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26(4), 393-424.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1), 91-117.

Baroni, M., Lenci, A., & Onnis, L. (2007). ISA meets Lara: An incremental word space model for cognitively plausible simulations of semantic learning. Paper presented at the *Workshop on Cognitive Aspects of Computational Language Acquisition (CACLA 07)*, Prague, Czech republic.

Christophe, A., Millotte, S., Bernal, S., & Lidz, J. (2008). Bootstrapping Lexical and Syntactic Acquisition. *Language and Speech*, 51(1-2), 61-75.

References

Parisien, C. (2008, 16-17 August 2008). An Incremental bayesian Model for Learning Syntactic Categories. Paper presented at the *12th Conference on Computational Natural Language Learning*, Manchester, UK.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian Framework for Word Segmentation: Exploring the Effects of Context. *Cognition*, 112(1), 21-54.

St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116(3), 341-360.

Baayen, R. H., Shaoul, C., Willits, J., & Ramscar, M. (2015). Comprehension without segmentation: a proof of concept with naive discriminative learning. *Language, Cognition and Neuroscience*, 1-23.

Cassani, G., Grimm, R., Daelemans, W., & Gillis, S. (2015). Which distributional cues help the most? Unsupervised contexts selection for lexical category acquisition. Paper presented at the *6th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, Lisbon.

Conclusions

Thank you!

- Questions?
- Suggestions!
- Criticisms!?