Learning to categorize nouns and verbs on the basis of a few known examples: A computational model relying on *n*-grams

Pascal Amsili[2]

Université Paris Diderot

Dublin Computational Linguistics Research Seminar Series, 17 april 2015

Joint work with

Perrine Brusini[1] Emmanuel Chemla[3] Anne Christophe[3] Olga Seminck[2]

[1] Language, Cognition and Development Lab, Int. School for Advanced Studies (SISSA), Trieste
[2] Laboratoire de Linguistique Formelle, CNRS & Université Paris Diderot
[3] Laboratoire de Sciences Cognitives et Psycholinguistique (CNRS & ENS, EHESS)

Roadmap

1 Motivation

2 Manipulation
- Corpus
- Learning
- Projection
- Prediction

3 Results

4 Discussion

Motivation

- Children can tell apart nominal *vs.* verbal contexts from 18 m. on.

[Cauvet et al., *2014; Bernal* et al., *2010]*

Motivation

- Children can tell apart nominal *vs.* verbal contexts from 18 m. on.
  [Cauvet et al., *2014; Bernal* et al., *2010]*

Example : [Bernal, 2007]

- Unknown words (jabberwocky)
- presented in situations where they can either refer to
  - an action (a puppet bouncing), or
  - an object (new puppet or unknown animal)
- if they are presented with desambiguating auditory stimuli

(1) a. Regarde, le dase
    b. Regarde, il dase
    Look, the/it DASE

- children at 18m. adopt the new word as referring to an action (1b) or an object (1a) accordingly.

Motivation

- Children can tell apart nominal *vs.* verbal contexts from 18 m. on.

  [Cauvet et al., *2014; Bernal* et al., *2010]*

- Which clues are they using?
  - prosody *[Gutman* et al., *2014]*
  - pragmatics *[Tomasello, 2002]*

  ⇒ function words

Function words

- short, unaccented

⇒ bad clues, according to Pinker [1984]

Function words

- short, unaccented

⇒ bad clues, according to Pinker [1984]

But

- very frequent
- often located at prosody boundaries

⇒ easy to notice [Shi *et al.*, 1998]

Function words

- short, unaccented

⇒ bad clues, according to Pinker [1984]

But

- very frequent
- often located at prosody boundaries

⇒ easy to notice [Shi *et al.*, 1998]

Besides

- known to be recognized by toddlers before 12m. [Shi, 2014]
- used by children to select correct category by 18m. [Cauvet *et al.*, 2014; Zangl and Fernald, 2007].

Aim

- Can statistic properties of children-directed language be exploited?
- Feasability study

⇒ No claim as to what toddlers actually do

Hypotheses

- limited lexicon ("semantic seed")
  [Bergelson and Swingley, 2012, 2013] : between 6 and 9 m. toddlers already know a number of verbs and nouns.

- two "semantic" categories:
  - actions,
  - objects (and agents)

  [Carey, 2009] : children have different representations for agents and artifacts on one side and (causal) actions on the other side.

- Word segmentation

Comparison with POS-tagging

POS-tagging in NLP:

- makes use of morphology
- makes use of a larger set of POS
- typically uses HMM techniques

State of the art

[Redington *et al.*, 1998] are the first to demonstrate the usefulness of immediate distributional information to acquire syntactic categories.

However their model is mostly concerned with the discovery of the syntactic category of (relatively) frequent words — and they do not consider specifically function words (see experiment 8, though).

[Mintz, 2003] show that very local recurring patterns ("frequent frames") are extremely good predictors to ascribe a category to unknown words. For instance the frame `you ____ it` only contains verbs in the child-directed corpora he worked on.

It should be noted though that only extremely frequent frames are considered, which provides only a small number of (accurate) predictors.

Roadmap

1 Motivation

2 Manipulation

- Corpus
- Learning
- Projection
- Prediction

3 Results

4 Discussion

Corpus

- Corpus taken from CHILDES 4 database [MacWhinney, 2000]

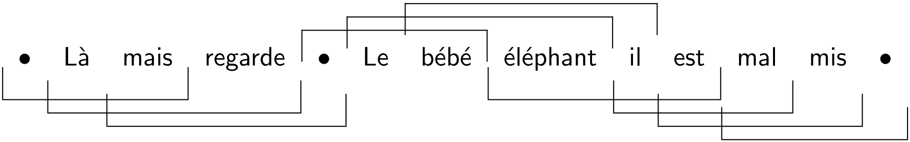
- Written transcriptions of spontaneous speech
- Two mother-child pairs (Marie & Timothée)
- 133 948 tokens
- Only child directed speech (from adult) was selected.

- POS-tagging performed by the French tagger Cordial (part of a word-processing corrector)

- Semi-automatic post treatment to deal with POS-tagging error (10% on nouns and verbs during the first pass)

Learning

- All word *n*-grams found in the corpus are counted.
- Strong punctuation are considered as (border) words,
- and *n*-grams comprising a border word are counted only if the border is the first or the last word of the *n*-gram.



Projection

We assume that the learner already knows some nouns and verbs, and we project known verbs/nouns to their category, so that learning is performed on a single flow

⇒ Very different from classical HMM approaches to POS tagging

We take as already known the most frequent nouns/verbs in the corpus.
Starting point: 10% of the occurrences of V/N, which corresponds to 6 N and 2 V.
Then 5 other "vocabulary states" (V_i : 6×2^i N and 2×2^i V are known).

			—	Là	mais	regarde	!	Le	bébé	éléphant	il	est	mal	mis	!
V_0	6 N	2 V	•	Là	mais	regarde	•	Le	N	éléphant	il	est	mal	mis	•
V_1	12 N	4 V	•	Là	mais	V	•	Le	N	éléphant	il	est	mal	mis	•
V_2	24 N	8 V	•	Là	mais	V	•	Le	N	éléphant	il	est	mal	mis	•
V_3	48 N	16 V	•	Là	mais	V	•	Le	N	N	il	est	mal	mis	•
V_4	96 N	32 V	•	Là	mais	V	•	Le	N	N	il	est	mal	V	•
V_m	1310 N	1253 V	•	Là	mais	V	•	Le	N	N	il	est	mal	V	•

Vocabularies

V_0	6N	doudou	bébé	livre	chose	micro	histoire
	2V	aller			faire		
V_1	V_0+ 6N	pied	poisson	peu[1]	main	lait	nez
	V_0+ 2V	mettre			regarder		
V_2	V_1+12N	caméra	fleur	tête	eau	heure	côté
		oeil	bouche	biberon	assiette	éléphant	fois
	V_1+ 4V	voir			pouvoir		
		dire			falloir		

Prediction

- left context: $n-1$ words preceeding the target
- right context: $n-1$ words following the target
- nested context: $n-1$ words surrounding the target (n odd and $\geqslant 3$)

for a given target w in a given context $(w_1, \dots w_{n-1}, w)$,
the prediction is

$$w_p = \arg\max_{w} \mathsf{freq}(w_1, \dots w_{n-1}, w).$$

Smoothing and backup

- frequencies taken as such (and not as probabilities)
  $\Rightarrow$ no smoothing required
- for unseen contexts, usual backup:
  if $(w_1, \ldots w_{n-1})$ was never met, try with $(w_2, \ldots w_{n-1})$

Test

- Unseen portion of the corpus
- Target positions:
  - not-too-frequent forms (freq $\leqslant 0.05\%$)
  - closest context word already not unknown

Intuition: when the context contains known words, it can be used to make a prediction about an unknown (ie rare) word.

Example

● mais viens, je vais la réparer ta voiture ●

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>

Categorisation taken as a reference

Example

●	mais	viens,	je	vais	la	réparer	ta	voiture	●
<s>	CJ	V	P	V	P	V	D	N	<s>

Categorisation taken as a reference

Targets : frequency $\leqslant 0.05\%$

Example

●	mais	viens,	je	vais	la	réparer	ta	voiture	●
<s>	CJ	V	P	V	P	V	D	N	<s>
		V							

Categorisation taken as a reference

Targets : frequency $\leqslant$ 0.05%

Predictions : 1 No backup

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>
		V							
		h_{V}							

Categorisation taken as a reference

Targets :	frequency $\leqslant 0.05\%$
Predictions :	1
Measures :	1 hit !

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>
		V				N			
		h_V							

Categorisation taken as a reference

Targets : frequency $\leqslant 0.05\%$

Predictions : 1 2 No backup

Measures : 1

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>
		V				N			
		h_{V}				$m_{\mathrm{V}}, f_{\mathrm{N}}$			

Categorisation taken as a reference

Targets : frequency $\leqslant 0.05\%$

Predictions : 1 2

Measures : 1 2 miss + false alarm

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>
		V				N		N	
		h_V				m_V, f_N			

Categorisation taken as a reference

Targets :	frequency $\leqslant 0.05\%$		
Predictions :	1	2	3 Backup!
Measures :	1	2	

Example

●	mais	viens,	je	vais	la	réparer	ta	voiture	●
<s>	CJ	V	P	V	P	V	D	N	<s>
		V				N		N	
		h_V				m_V, f_N		h_N	

Categorisation taken as a reference

Targets : frequency $\leqslant 0.05\%$

Predictions : 1 2 3

Measures : 1 2 3 hit

Example

•	mais	viens,	je	vais	la	réparer	ta	voiture	•
<s>	CJ	V	P	V	P	V	D	N	<s>
		V				voir		N	
		h_V				m_V, f_N		h_N	

Categorisation taken as a reference

Targets :	frequency $\leqslant 0.05\%$		
Predictions :	1	2	3
Measures :	1	2	3

Alternative prediction: same results

Measures

- 3 “categories”: N, V, O (other)
- for each category X:
    - hits h_X
    - misses m_X
    - false alarms f_X
- for each category X:

$$\mathrm{prec}_X = \frac{h_X}{h_X + f_X} \quad \mathrm{recall}_X = \frac{h_X}{h_X + m_X}$$

Baseline + 10-fold

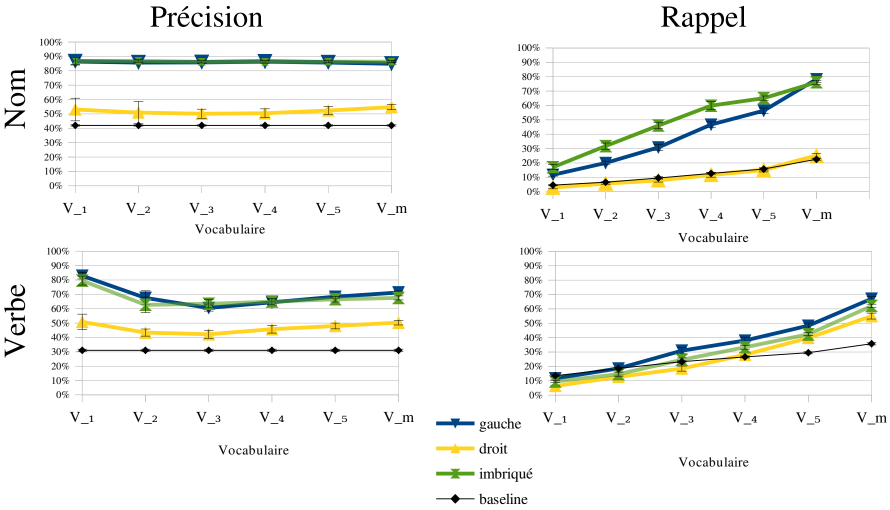
- Baseline
  - Predictions based not on the context, but on the frequences of N, V and the rest in the training corpus
- Ten-fold cross validation
  - 10 trials with $^1/_{10}$ of the corpus divided into
    - $^2/_3$ for training
    - $^1/_3$ for test
  - standard deviation represented as error bars on the graphs

Roadmap

1. Motivation
2. Manipulation
   - Corpus
   - Learning
   - Projection
   - Prediction
3. Results
4. Discussion

Results

Performances of the 3 models (precision/recall) for each category N and V ($n = 3$).



Results

- All the models are better with context than the baseline
- The model relying on right contexts is the least efficient
- Results for N better than for V

- No growth of precision with the size of the vocabulary
- Growth of the recall with vocabulary size
- Very small variability, robustness (similar results with the whole corpus)

Roadmap

1. Motivation
2. Manipulation
   - Corpus
   - Learning
   - Projection
   - Prediction
3. Results
4. Discussion

Discussion

- Very good **precision** even with a small semantic seed
  ⇒ immediate contexts are very informative

Discussion

- Very good **precision** even with a small semantic seed
  ⇒ immediate contexts are very informative
- Categorisation performed without considering the word itself:
  ⇒ avantages for language acquisition:
  - unknown words can be categorized
  - no problem coming from homonymy and morphological ambiguïty

  ⇒ it makes it plausible that morphological analysis come as a later step

Discussion

- Very good **precision** even with a small semantic seed
  ⇒ immediate contexts are very informative
- Categorisation performed without considering the word itself:
  ⇒ avantages for language acquisition:
  - unknown words can be categorized
  - no problem coming from homonymy and morphological ambiguïty

  ⇒ it makes it plausible that morphological analysis come as a later step
- **Recall** strongly dependant on the size of the semantic seed
  ⇒ pertinent for language acquisition:
  - at the beginning only a small number of reliable contexts are known, and no prediction is made with unreliable contexts

  → confirmed by error analysis

Discussion

- Very good **precision** even with a small semantic seed
  ⇒ immediate contexts are very informative
- Categorisation performed without considering the word itself:
  ⇒ avantages for language acquisition:
  - unknown words can be categorized
  - no problem coming from homonymy and morphological ambiguïty

  ⇒ it makes it plausible that morphological analysis come as a later step
- **Recall** strongly dependant on the size of the semantic seed
  ⇒ pertinent for language acquisition:
  - at the beginning only a small number of reliable contexts are known, and no prediction is made with unreliable contexts

  → confirmed by error analysis
- **function words** emerge in useful contexts, while no a priori hypothesis was made, simply because of their frequency and distribution.

Most frequent useful contexts (N/V)

Contexts with the highest number of V (resp. N)

context		#N	#V	tot	Ans
●	un/a	78	1	121	N
est/is	un/a	62	0	82	N
V	le/the	60	31	105	N
N	de/of	56	15	138	N
V	un/a	52	0	94	N
V	des/some	52	0	67	N
●	le/the	46	1	62	N
●	une/a	45	0	57	N
de/of	la/the	41	4	53	N
V	la/the	39	19	68	N
V	les/the	34	12	57	N
●	la/the	33	0	54	N
V	du/of the	33	0	33	N
à/to	la/the	32	1	35	N
V	une/a	32	0	55	N

context		#N	#V	tot	Ans
●	tu/you	1	603	1060	V
●	on/we	0	225	335	V
●	je/I	0	187	276	V
●	V	3	110	446	V
●	il/he	0	101	252	V
●	ça/it	0	95	224	V
que/that	tu/you	1	81	156	V
tu/you	V	5	58	309	V
on/we	V	2	52	227	V
tu/you	as/have	6	46	107	V
V	pas/not	1	45	190	V
qu'/that	il/he	0	45	93	V
qu'/that	on/we	0	44	70	V
V	V	6	42	386	V
V	le/the	60	31	105	N

Ambiguous forms (*le*, *de*, *la*, *des*) “solved” in bigrams

Most frequent useful contexts (N/V)

Contexts with the highest number of V (resp. N)

context		#N	#V	tot	Ans
●	un/a	78	1	121	N
est/is	un/a	62	0	82	N
V	le/the	60	31	105	N
N	de/of	56	15	138	N
V	un/a	52	0	94	N
V	des/some	52	0	67	N
●	le/the	46	1	62	N
●	une/a	45	0	57	N
de/of	la/the	41	4	53	N
V	la/the	39	19	68	N
V	les/the	34	12	57	N
●	la/the	33	0	54	N
V	du/of the	33	0	33	N
à/to	la/the	32	1	35	N
V	une/a	32	0	55	N

context		#N	#V	tot	Ans
●	tu/you	1	603	1060	V
●	on/we	0	225	335	V
●	je/I	0	187	276	V
●	V	3	110	446	V
●	il/he	0	101	252	V
●	ça/it	0	95	224	V
que/that	tu/you	1	81	156	V
tu/you	V	5	58	309	V
on/we	V	2	52	227	V
tu/you	as/have	6	46	107	V
V	pas/not	1	45	190	V
qu'/that	il/he	0	45	93	V
qu'/that	on/we	0	44	70	V
V	V	6	42	386	V
V	le/the	60	31	105	N

Only one context significantly ambiguous

Most frequent useful contexts (N/V)

Contexts with the highest number of V (resp. N)

context		#N	#V	tot	Ans
●	un/a	78	1	121	N
est/is	un/a	62	0	82	N
V	le/the	60	31	105	N
N	de/of	56	15	138	N
V	un/a	52	0	94	N
V	des/some	52	0	67	N
●	le/the	46	1	62	N
●	une/a	45	0	57	N
de/of	la/the	41	4	53	N
V	la/the	39	19	68	N
V	les/the	34	12	57	N
●	la/the	33	0	54	N
V	du/of the	33	0	33	N
à/to	la/the	32	1	35	N
V	une/a	32	0	55	N

context		#N	#V	tot	Ans
●	tu/you	1	603	1060	V
●	on/we	0	225	335	V
●	je/I	0	187	276	V
●	V	3	110	446	V
●	il/he	0	101	252	V
●	ça/it	0	95	224	V
que/that	tu/you	1	81	156	V
tu/you	V	5	58	309	V
on/we	V	2	52	227	V
tu/you	as/have	6	46	107	V
V	pas/not	1	45	190	V
qu'/that	il/he	0	45	93	V
qu'/that	on/we	0	44	70	V
V	V	6	42	386	V
V	le/the	60	31	105	N

Articles play a (nominal) role...

Most frequent useful contexts (N/V)

Contexts with the highest number of V (resp. N)

context		#N	#V	tot	Ans
●	un/a	78	1	121	N
est/is	un/a	62	0	82	N
V	le/the	60	31	105	N
N	de/of	56	15	138	N
V	un/a	52	0	94	N
V	des/some	52	0	67	N
●	le/the	46	1	62	N
●	une/a	45	0	57	N
de/of	la/the	41	4	53	N
V	la/the	39	19	68	N
V	les/the	34	12	57	N
●	la/the	33	0	54	N
V	du/of the	33	0	33	N
à/to	la/the	32	1	35	N
V	une/a	32	0	55	N

context		#N	#V	tot	Ans
●	tu/you	1	603	1060	V
●	on/we	0	225	335	V
●	je/I	0	187	276	V
●	V	3	110	446	V
●	il/he	0	101	252	V
●	ça/it	0	95	224	V
que/th	tu/you	1	81	156	V
tu/you	V	5	58	309	V
on/we	V	2	52	227	V
tu/you	as/have	6	46	107	V
V	pas/not	1	45	190	V
qu'/th	il/he	0	45	93	V
qu'/th	on/we	0	44	70	V
V	V	6	42	386	V
V	le/the	60	31	105	N

Articles play a (nominal) role... and personal pronouns (nom.) a (verbal) role.

Perspectives

- Comparison with other types of “texts”
- Comparison with other languages: among other things, to tell whether the better performance of the left models comes from a language specific property or from a universal property
- Incrementality
- Work on the least plausible hypothesis, namely that of a recording of frequencies for all *n*-grams encountered

Conclusion

This study shows the relevance of a simulation approach with constraints coming from experimental results.

In the present case, it shows that the use of function words as POS predictors for unknown words does not require any a priori knowledge on their categories, probably because their distribution and frequency suffices to highlight them.

It is an interesting result because, in French, homophony of function words makes their categorisation difficult.

We hope that such simulation models will also give predictions that can be tested empirically,

so that we may manage (one day) to build computationel models of the acquisition of categories that are psychologically plausible.

Thank you!

co-authors:
Perrine Brusini
Anne Christophe
Emmanuel Chemla
Olga Seminck

and thanks to
Benoît Crabbé (Paris Diderot),
ANR (grants # ANR-2010-BLAN-1901,
ANR-13-APPR-0012,
ANR-10-IDEX-0001-02 PSL*
ANR-10-LABX-0087 IEC)
Fondation de France.

References I

Elika Bergelson and Daniel Swingley. At 6-9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences of the United States of America*, 109(9):3253–3258, 2012. doi:10.1073/pnas.1113380109.

Elika Bergelson and Daniel Swingley. The acquisition of abstract words by young infants. *Cognition*, 127(3):391–397, 2013. doi:10.1016/j.cognition.2013.02.011.

Savita Bernal, Ghislaine Dehaene-Lambertz, Séverine Millote, and Anne Christophe. Two-years-olds compute syntactic structure on-line. *Developmental Science*, 12:69–76, 2010.

Savita Bernal. *De l'arbre (syntaxique) au fruit (du sens): Interactions des acquisitions lexicale et syntaxique chez l'enfant de moins de 2 ans*. PhD thesis, Université Pierre et Marie Curie, 2007.

Perrine Brusini, Pascal Amsili, Emmanuel Chemla, and Anne Christophe. Simulation de l'apprentissage des contextes nominaux/verbaux par n-grammes. In Philippe Blache, editor, *Actes de TALN 2014 (Traitement automatique des langues naturelles)*, Marseille, July 2014. ATALA.

Perrine Brusini. *Découvrir les noms et les verbes : Quand les classes sémantiques initialisent les catégories syntaxiques*. PhD thesis, Université Pierre et Marie Curie, 2012.

Susan Carey. *The origin of concepts*. Oxford University Press, 2009.

Elodie Cauvet, Rita Limissuri, Severine Millotte, Katrin Skoruppa, Dominique Cabrol, and Anne Christophe. Syntactic context constrains lexical access in French 18-month-olds. *Language Learning and Development*, 10(1):1–18, 2014.

Ariel Gutman, Isabelle Dautriche, Benoît Crabbé, and Anne Christophe. Bootstrapping the syntactic bootstrapper: Probabilistic labeling of prosodic phrases. *Language Acquisition*, (just-accepted), 2014.

Brian MacWhinney. *The CHILDES Project: Tools for analyzing talk*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000. Third Edition.

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

References II

Steven Pinker. *Language Learnability and Language Development*. Harvard University Press, Cambridge, MA, 1984.

Martin Redington, Nick Chater, and Steven Finch. Distributional information : A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22(4):425–469, 1998.

Rushen Shi, James L Morgan, and P Allopenna. Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language*, 25:169–201, 1998.

Rushen Shi. Functional morphemes and early language acquisition. *Child Development Perspectives*, 8(1):6–11, 2014.

Michael Tomasello. Some facts about primate (including human) communication and social learning. In *Simulating the evolution of language*, pages 327–340. Springer, 2002.

Renate Zangl and Anne Fernald. Increasing flexibility in children's online processing of grammatical and nonce determiners in fluent speech. *Language Learning and Development*, 3(3):199–231, 2007.

Détail des prédictions

Notation: “ni N ni V” est noté Z. Les compteurs colorés sont les seuls qu’on prend en compte dans les calculs (autrement dit les “bonnes réponses” ni N ni V ne sont pas comptées).

—	Le	bébé	éléphant	il	regarde	!
•	Le	N	éléphant	il	V	•
•	Le	N	N	il	V	•

		prédiction	décompte	prédiction	décompte		prédiction	décompte	
no predict. f$>$.05%	no predict. f$>$.05%	N	BR_N	N	BR_N	no predict. f$>$.05%	V	BR_V	no predict f$>$.05%
		V	MA_N FA_V	V	MA_N FA_V		N	MA_V FA_N	
		chat	MA_N	éléphant	MA_N		dort	MA_V	
		très	MA_N	très	MA_N		petit	MA_V	