Distributional Information: A Powerful Cue for Acquiring Syntactic Categories

Levindo Gabriel Taschetto Neto
22.11.2017

Probabilistic models of language and cognition
Universität Stuttgart

Agenda

1. Introduction.
2. Learning Syntactic Categories' Problem.
3. Available Information.
4. Utility of Information Sources' Access.
5. Relevant Distributional Approaches.
6. New Distributional Approaches.
7. Experiments.
8. Conclusion.

1. Introduction

- Distributional information is a potentially important source of data for identifying the syntactic categories of words.
- Distributional information provides a powerful cue for acquiring syntactic categories.

2. Learning Syntactic Categories' Problem

- A completely unconstrained search with *n* items and m syntactic categories (assuming, for simplicity, that each item has a single syntactic category), would involve considering *m* in the power of *n* possible mappings.

2. Learning Syntactic Categories' Problem

- A completely unconstrained search with *n* items and m syntactic categories (assuming, for simplicity, that each item has a single syntactic category), would involve considering *m* in the power of *n* possible mappings.

Example:

- Given 3 syntactic categories and 15 different items.

2. Learning Syntactic Categories' Problem

- A completely unconstrained search with <u>*n* items</u> and <u>m syntactic categories</u> (assuming, for simplicity, that each item has a single syntactic category), would involve considering *m* in the power of *n* possible mappings.

Example:

- Given 3 syntactic categories and 15 different items.
  $n = 15$
  $m = 3$

2. Learning Syntactic Categories' Problem

- A completely unconstrained search with <u>*n* items</u> and <u>m syntactic categories</u> (assuming, for simplicity, that each item has a single syntactic category), would involve considering *m* in the power of *n* possible mappings.

Example:

- Given 3 syntactic categories and 15 different items.
  n = 15
  m = 3
  Possible mappings = 3^15 = 14.348.907

3. Available Information

Based on:

3. Available Information

Based on:

1. Distributional analysis of linguistic input.

3. Available Information

Based on:

1. Distributional analysis of linguistic input.
2. Relation of the linguistic input to the situation.

3. Available Information

Based on:

1. Distributional analysis of linguistic input.
2. Relation of the linguistic input to the situation.
3. Phonological cues to syntactic category.

3. Available Information

Based on:

1. Distributional analysis of linguistic input.
2. Relation of the linguistic input to the situation.
3. Phonological cues to syntactic category.
4. Analysis of prosody.

3. Available Information

Based on:

1. Distributional analysis of linguistic input.
2. Relation of the linguistic input to the situation.
3. Phonological cues to syntactic category.
4. Analysis of prosody.
5. Natural knowledge of syntactic categories.

3.1 Distributional analysis of linguistic input

- Information about the linguistic context in which a word occurs.

3.1 Distributional analysis of linguistic input

- Information about the linguistic context in which a word occurs.
- Words of the same category tend to have a large number of distributional regularities in common can be used as a cue to syntactic category.

3.1 Distributional analysis of linguistic input

- Information about the linguistic context in which a word occurs.
- Words of the same category tend to have a large number of distributional regularities in common can be used as a cue to syntactic category.

Example (Maratsos and Chalkley, 1980)

Words which take the suffix *-ed* typically also take the suffix -*s*, and are verbs. Also, words which take the suffix -*s*, but not the suffix -*ed*, are typically count-nouns.

3.2 Relation of the linguistic input to the situation

- A mechanism for the initial classification of words makes use of a correlation between prior semantic categories (such as object and action) in terms of which the child already perceives the world and syntactic categories.

3.3 Phonological cues to syntactic category

- Myriad regularities between the phonology of words and their syntactic categories may be utilized in order to acquire these categories (Kelly, 1992).

3.3 Phonological cues to syntactic category

- Myriad regularities between the phonology of words and their syntactic categories may be utilized in order to acquire these categories (Kelly, 1992).

Example:

English disyllabic nouns tend to have stress on the initial syllable, while verbs have final syllable stress.

3.4 Analysis of prosody

- Learners exploit the mutual predictability between the syntactic phrasing of a sentence, and the way it is said (Morgan and Newport, 1981).

3.5 Natural knowledge of syntactic categories

1. Learning mechanisms that exploit information of any kind in the input may be innately specified.

3.5 Natural knowledge of syntactic categories

1. Learning mechanisms that exploit information of any kind in the input may be innately specified.
2. Innate knowledge or constraints may specify, for instance, the number of syntactic categories or the relationships between them.

4. Utility of Information Sources' Access

- Distributional analysis can be conducted over electronically stored texts, represented purely as sequences of distinct words, and these are (at least for English) available to researchers in almost unlimited supply.

5. Relevant Distributional Approaches

1. Distributional Analysis in Linguistics.

5. Relevant Distributional Approaches

1. Distributional Analysis in Linguistics.
2. Neural Networks.

5. Relevant Distributional Approaches

1. Distributional Analysis in Linguistics.
2. Neural Networks.
3. Statistical Approaches to Language Learning.

5.1 Distributional Analysis in Linguistics

- Distributional linguists were interested in the discovery of language structure from corpora, purely from the point of view of providing a rigorous methodology for field linguistics.

5.1 Distributional Analysis in Linguistics

- Distributional linguists were interested in the discovery of language structure from corpora, purely from the point of view of providing a rigorous methodology for field linguistics.
- They conceived of language as an external cultural product, and did not consider it in a psychological or computational context.

5.1 Distributional Analysis in Linguistics

- Distributional linguists were interested in the discovery of language structure from corpora, purely from the point of view of providing a rigorous methodology for field linguistics.
- They conceived of language as an external cultural product, and <u>did not</u> consider it in a psychological or computational context.
- They were <u>unable to test</u> their methods except with very small samples of language.

5.2 Neural Networks

- The most influential for learning the structure of a sequential material uses SRNs.

5.2 Neural Networks

- The most influential for learning the structure of a sequential material uses SRNs.

Simple Recurrent Networks

Assign similar hidden unit patterns to items which have the same syntactic category in a simple grammar.

5.2 Neural Networks

- Another approach for learning the linguistic categories of small artificial languages uses a competitive network in order to produce a topographic mapping between the distribution of contexts in which an item occurs and a 2-dimensional space.

5.2 Neural Networks

- Another approach for learning the linguistic categories of small artificial languages uses a competitive network in order to produce a topographic mapping between the distribution of contexts in which an item occurs and a 2-dimensional space.

  Results

  The results show that items with the same linguistic category tend to lie in neighboring regions of the space.

5.2 Neural Networks

Limitations (SRNs and Competitive Networks)

- Scaling up still not being possible from very small artificial data sets in order to deal with real linguistic data.

5.2 Neural Networks

Limitations (SRNs and Competitive Networks)

- Scaling up still not being possible from very small artificial data sets in order to deal with real linguistic data.
- The linguistic categories can only be revealed using a subsequent cluster analysis.

5.3 Statistical Approaches to Language Learning

Main aim:

- Practical utility.

5.3 Statistical Approaches to Language Learning

Main aim:

- Practical utility.

Problem:

- It has not demonstrated utility of distributional information concerning syntactic categories for a very large and rich corpora.

6. New Distributional Approaches

1. Measuring the Distribution of Each Word.

6. New Distributional Approaches

1. Measuring the Distribution of Each Word.
2. Comparing the Distributions of Pairs of Words.

6. New Distributional Approaches

1. Measuring the Distribution of Each Word.
2. Comparing the Distributions of Pairs of Words.
3. Grouping Together Words with Similar Distributions.

6.1 Measuring the Distribution of Each Word

Context for a word

6.1 Measuring the Distribution of Each Word

Context for a word

- It may be defined simply in terms of the distribution of words which occur near the target word.

6.1 Measuring the Distribution of Each Word

Context for a word

- It may be defined simply in terms of the distribution of words which occur near the target word.

Measurements' Records

- A record of such statistics can be viewed as a contingency table.

6.1 Measuring the Distribution of Each Word

Example:

6.1 Measuring the Distribution of Each Word

Example:

- **Input:** *The cow jumped over the moon.*
- **Target:** *jumped*.

6.1 Measuring the Distribution of Each Word

Example:

- **Input:** *The cow jumped over the moon.*
- **Target:** *jumped.*

Indexed cells which would be incremented in the contingency table:

6.1 Measuring the Distribution of Each Word

Example:

- **Input:** *The cow jumped over the moon.*
- **Target:** *jumped*.

Indexed cells which would be incremented in the contingency table:

(jumped, the), (jumped, cow), (jumped, over), (jumped, the), (jumped, moon).

6.2 Comparing the Distributions of Pairs of Words

- The more similar the words' distributions, the more likely that they are members of the same category.

6.3 Grouping Together Words with Similar Distributions

Goal due to the syntactic categories' boundaries:

6.3 Grouping Together Words with Similar Distributions

Goal due to the syntactic categories' boundaries:

- Assigning words to discrete categories.

6.3 Grouping Together Words with Similar Distributions

Goal due to the syntactic categories' boundaries:

- Assigning words to discrete categories.

Requirement for it:

6.3 Grouping Together Words with Similar Distributions

Goal due to the syntactic categories' boundaries:

- Assigning words to discrete categories.

Requirement for it:

- Identifying clusters of similarly distributed target words.

7. Experiments

1. Corpus
2. Benchmark Classification
3. Scoring
4. Qualitative Description of Results
5. Experiment 1
6. Experiment 2
7. Experiment 3
8. Experiment 4
9. Experiment 5
10. Experiment 6
11. Experiment 7
12. Experiment 8
13. Experiment 9

7.1 Corpus

- The experiments described below were performed using transcribed speech taken from the CHILDES database (MacWhinney & Snow, 1985).

7.1 Corpus

- The experiments described below were performed using transcribed speech taken from the CHILDES database (MacWhinney & Snow, 1985).
- CHILDES is a machine-readable collection of corpora of child and child-related speech, transcribed by a number of investigators, and largely recorded in informal North American domestic settings.

7.1 Corpus

- The experiments described below were performed using transcribed speech taken from the CHILDES database (MacWhinney & Snow, 1985).
- CHILDES is a machine-readable collection of corpora of child and child-related speech, transcribed by a number of investigators, and largely recorded in informal North American domestic settings.
- The resultant corpus contained several million words of speech, from approximately 6,000 speakers.

7.2 Benchmark Classification

- Used in order to have some categorisation for each word.

7.2 Benchmark Classification

- Used in order to have some categorisation for each word.

Category	Example	n
noun	truck, card, hand	407
adjective	little, favorite, white	81
numeral	two, ten, twelve	10
verb	could, hope, empty	239
article	the, a, an	3
pronoun	you, whose, more	52
adverb	rather, always, softly	60
preposition	in, around, between	21
conjunction	cos, while, and	9
interjection	oh, huh, wow	16
simple contraction		0
complex contraction	I'll, can't, there's	58

7.3 Scoring

Accuracy:

7.3 Scoring

Accuracy:

- Proportion of pairs of items that are grouped together in the derived groups which are also grouped together in the benchmark groups.

7.3 Scoring

Accuracy:

- Proportion of pairs of items that are grouped together in the derived groups which are also grouped together in the benchmark groups

Completeness:

7.3 Scoring

Accuracy:

- Proportion of pairs of items that are grouped together in the derived groups which are also grouped together in the benchmark groups

Completeness:

- Proportion of pairs of items which are grouped by the benchmark that are also grouped together in the derived groupings.

7.3 Scoring

Equations:

7.3 Scoring

Equations:

$$Accuracy = \frac{hits}{hits + falseAlarms}$$

7.3 Scoring

Equations:

$$Accuracy = \frac{hits}{hits + falseAlarms}$$

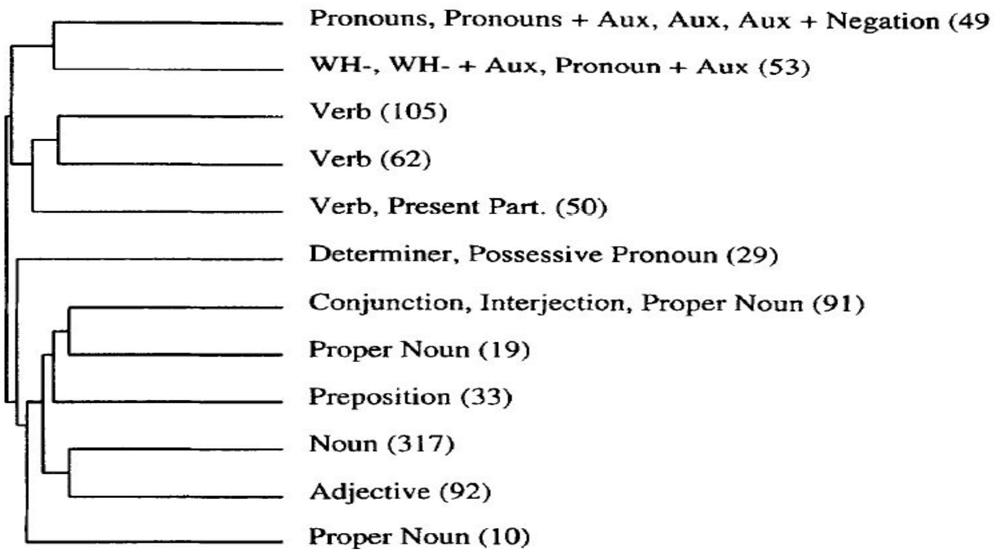
$$Completeness = \frac{hits}{hits + misses}$$

7.4 Qualitative Description of Results

- **Target words:** Most frequent 1000 ones.
- **Context words:** Most frequent 150 ones.

7.4 Qualitative Description of Results

- **The discrete clusters at a similarity level of 0.8 from the CHILDES corpus' analysis**



7.5 Experiment 1: Different Contexts

- How the <u>informativeness of context</u> words changes with distance from the target word, in order to determine which contextual information would be useful to the child.

7.5 Experiment 1: Different Contexts

- How the informativeness of context words changes with distance from the target word, in order to determine which contextual information would be useful to the child.

When similarity increases:

7.5 Experiment 1: Different Contexts

- How the informativeness of context words changes with distance from the target word, in order to determine which contextual information would be useful to the child.

When similarity increases:

- Accuracy decreases.

7.5 Experiment 1: Different Contexts

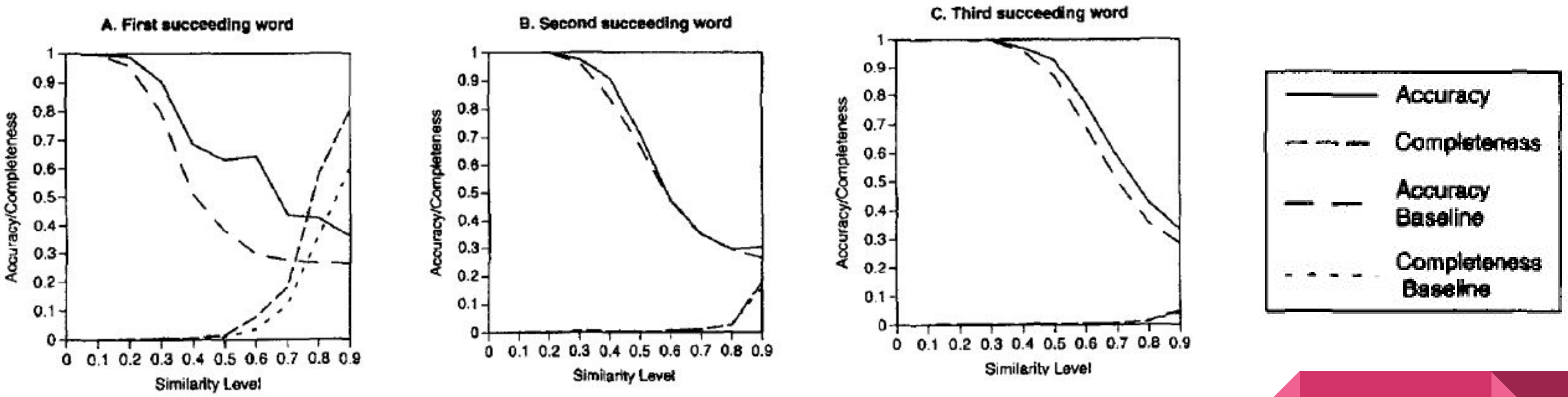
- How the informativeness of context words changes with distance from the target word, in order to determine which contextual information would be useful to the child.

When similarity increases:

- Accuracy decreases.
- Completeness increases.

7.5 Experiment 1: Different Contexts

Accuracy and completeness when the 1st (A), 2nd (B) and 3rd (C) succeeding words are used as contexts:



7.6 Experiment 2: Varying the Number of Target and Context Words

- What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?

7.6 Experiment 2: Varying the Number of Target and Context Words

- What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?

Low amount of target words:

7.6 Experiment 2: Varying the Number of Target and Context Words

- What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?

Low amount of target words:

- Poor performance.

7.6 Experiment 2: Varying the Number of Target and Context Words

- What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?

Low amount of target words:

- Poor performance.

Do the increase of the number of target words always produce more completeness and accuracy?

7.6 Experiment 2: Varying the Number of Target and Context Words

- What number of target (and context) words is required for the distributional method to be effective, and is this number realistic for the child?

Low amount of target words:

- Poor performance.

Do the increase of the number of target words always produce more completeness and accuracy?

- Decreasing between 1000 and 2.000 target words.

7.7 Experiment 3: For Which Classes is Distributional Information of Value?

- The major open classes (noun and verb) are more likely to be acquired first (Tomasello, 1992).

7.7 Experiment 3: For Which Classes is Distributional Information of Value?

The major categories from the Collins Cobuild Lexical Database:

		Observed		Baseline	
Class	n	Accuracy	Completeness	Accuracy	Completeness
noun	407	.90	.53	.43	.14
adjective	81	.38	.45	.09	.16
numeral	10	.09	.82	.02	.27
verb	239	.72	.24	.25	.14
article	3	.10	1.00	.01	.51
pronoun	52	.25	.24	.06	.14
adverb	60	.17	.18	.07	.16
preposition	21	.33	.53	.03	.16
conjunction	9	.06	.33	.02	.24
interjection	16	.18	.67	.02	.20
complex contraction	58	.55	.47	.07	.17
Overall	956	.72	.47	.27	.17

7.7 Experiment 3: For Which Classes is Distributional Information of Value?

The major categories from the Collins Cobuild Lexical Database:

		Observed		Baseline	
Class	n	Accuracy	Completeness	Accuracy	Completeness
noun	407	.90	.53	.43	.14
adjective	81	.38	.45	.09	.16
numeral	10	.09	.82	.02	.27
verb	239	.72	.24	.25	.14
article	3	.10	1.00	.01	.51
pronoun	52	.25	.24	.06	.14
adverb	60	.17	.18	.07	.16
preposition	21	.33	.53	.03	.16
conjunction	9	.06	.33	.02	.24
interjection	16	.18	.67	.02	.20
complex contraction	58	.55	.47	.07	.17
Overall	956	.72	.47	.27	.17

7.8 Experiment 4: Corpus Size

- A conservative minimum of one hour of free play per day would result in 1.5 million words of child-directed speech annually.

7.8 Experiment 4: Corpus Size

- A conservative minimum of one hour of free play per day would result in 1.5 million words of child-directed speech annually.
- The child will be exposed to a very much larger amount of *non-child-directed* speech.

7.8 Experiment 4: Corpus Size

- A conservative minimum of one hour of free play per day would result in 1.5 million words of child-directed speech annually.
- The child will be exposed to a very much larger amount of *non-child-directed* speech.
- More given input

7.8 Experiment 4: Corpus Size

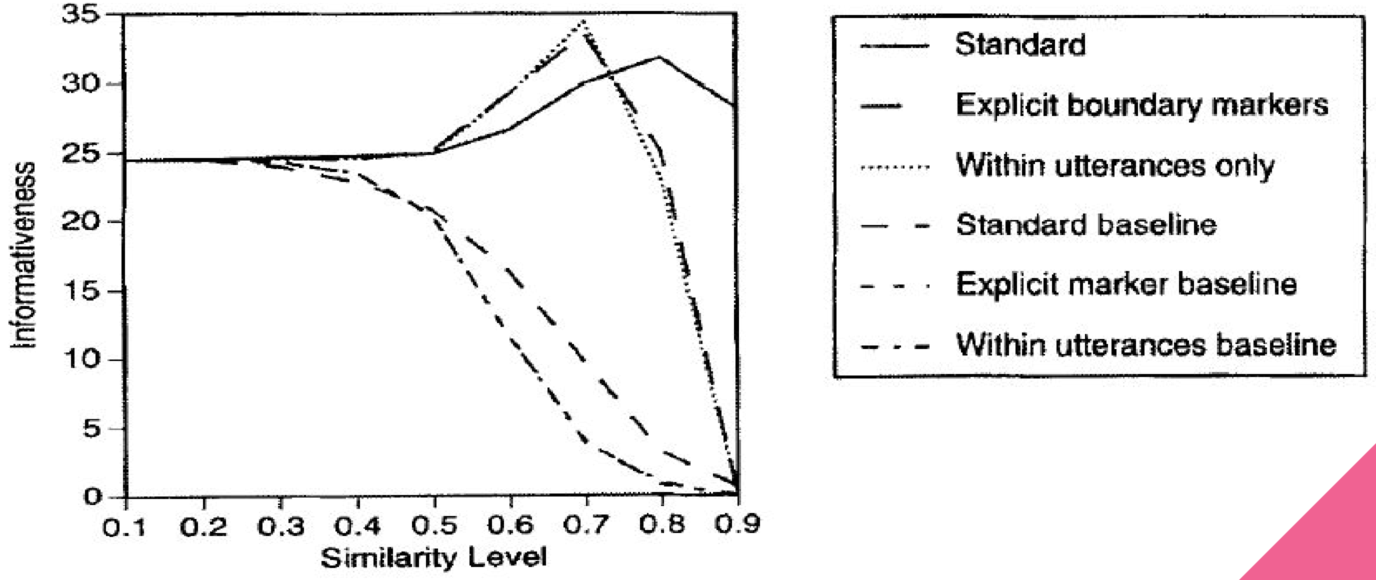
- A conservative minimum of one hour of free play per day would result in 1.5 million words of child-directed speech annually.
- The child will be exposed to a very much larger amount of *non-child-directed* speech.
- More given input → Moderate increase in performance.

7.9 Experiment 5: Utterance Boundaries

- Language can be concatenated from different speakers into a single undifferentiated speech stream.

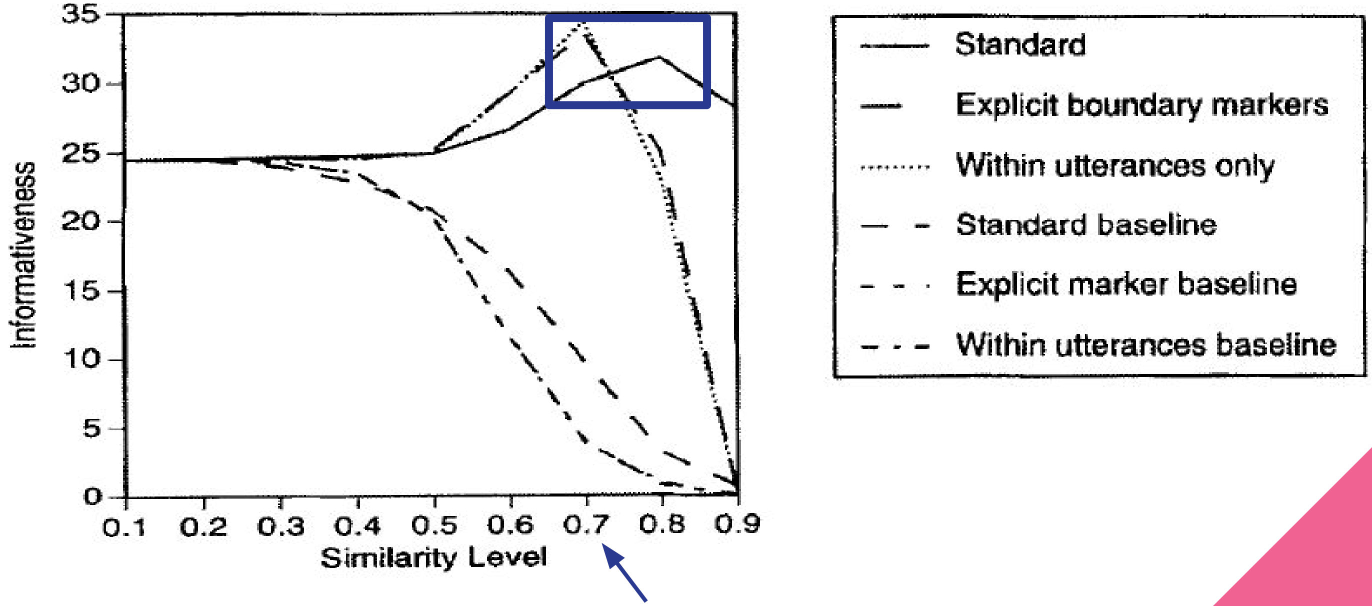
7.9 Experiment 5: Utterance Boundaries

- Language can be concatenated from different speakers into a single undifferentiated speech stream.



7.9 Experiment 5: Utterance Boundaries

- Language can be concatenated from different speakers into a single undifferentiated speech stream.

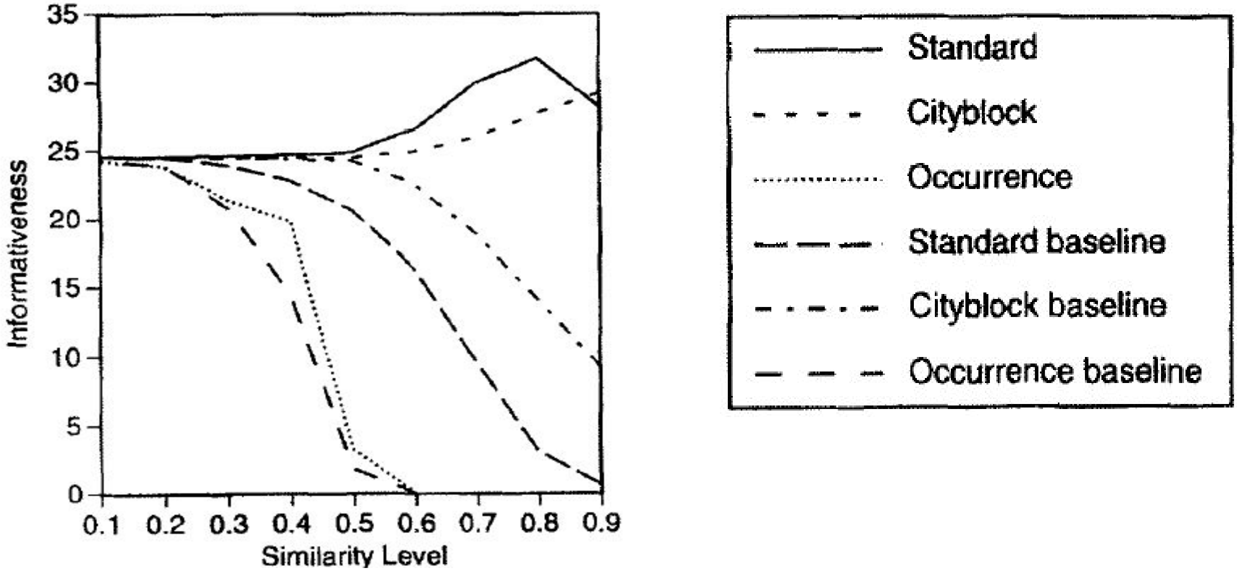


7.10 Experiment 6: Frequency Versus Occurrence

- Replace ranking for a binary notation.

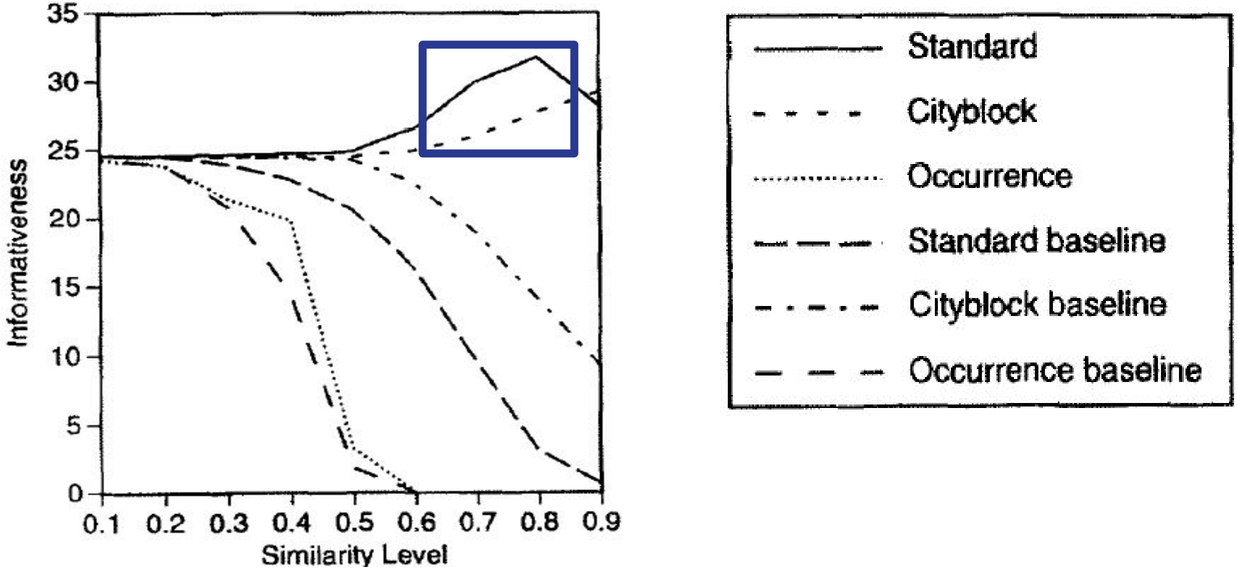
7.10 Experiment 6: Frequency Versus Occurrence

- Replace ranking for a binary notation.



7.10 Experiment 6: Frequency Versus Occurrence

- Replace ranking for a binary notation.



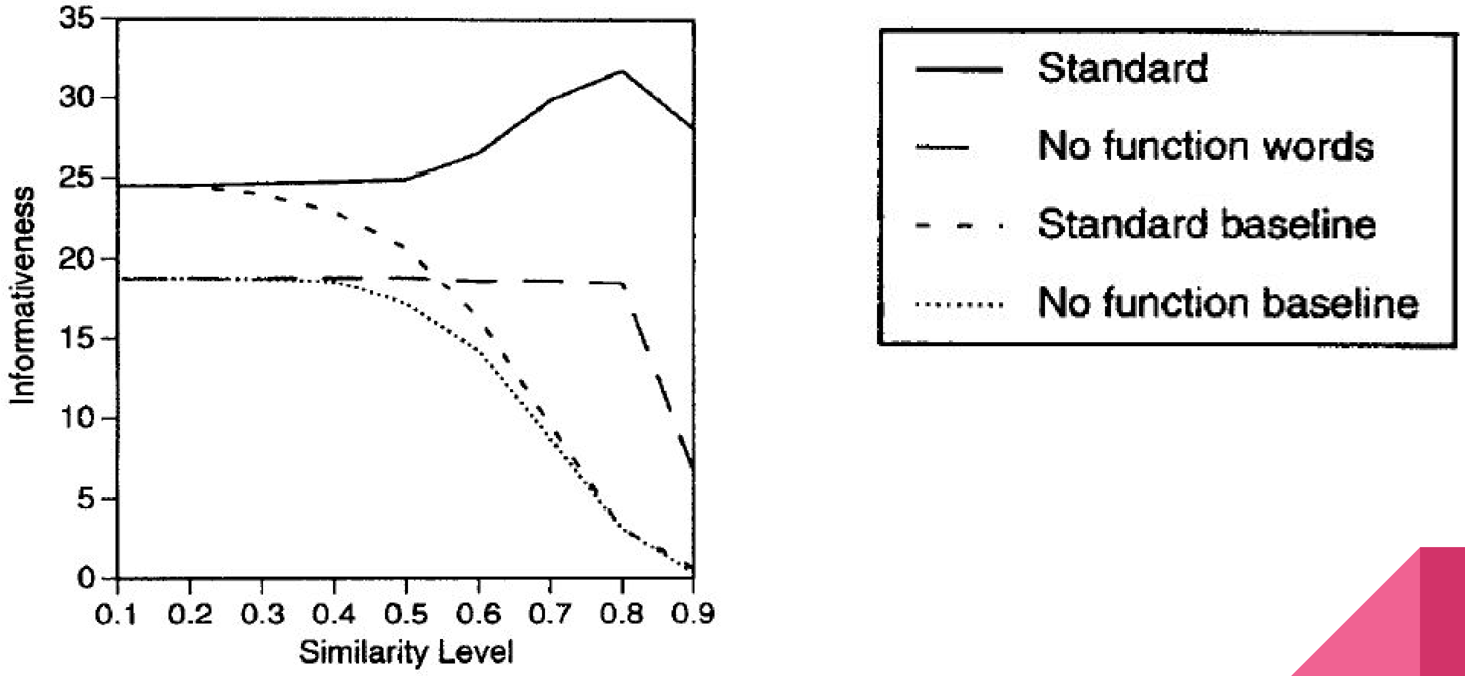
7.11 Experiment 7: Removing Function Words

- Removing function words does have a considerable impact on the amount of information provided by the method.

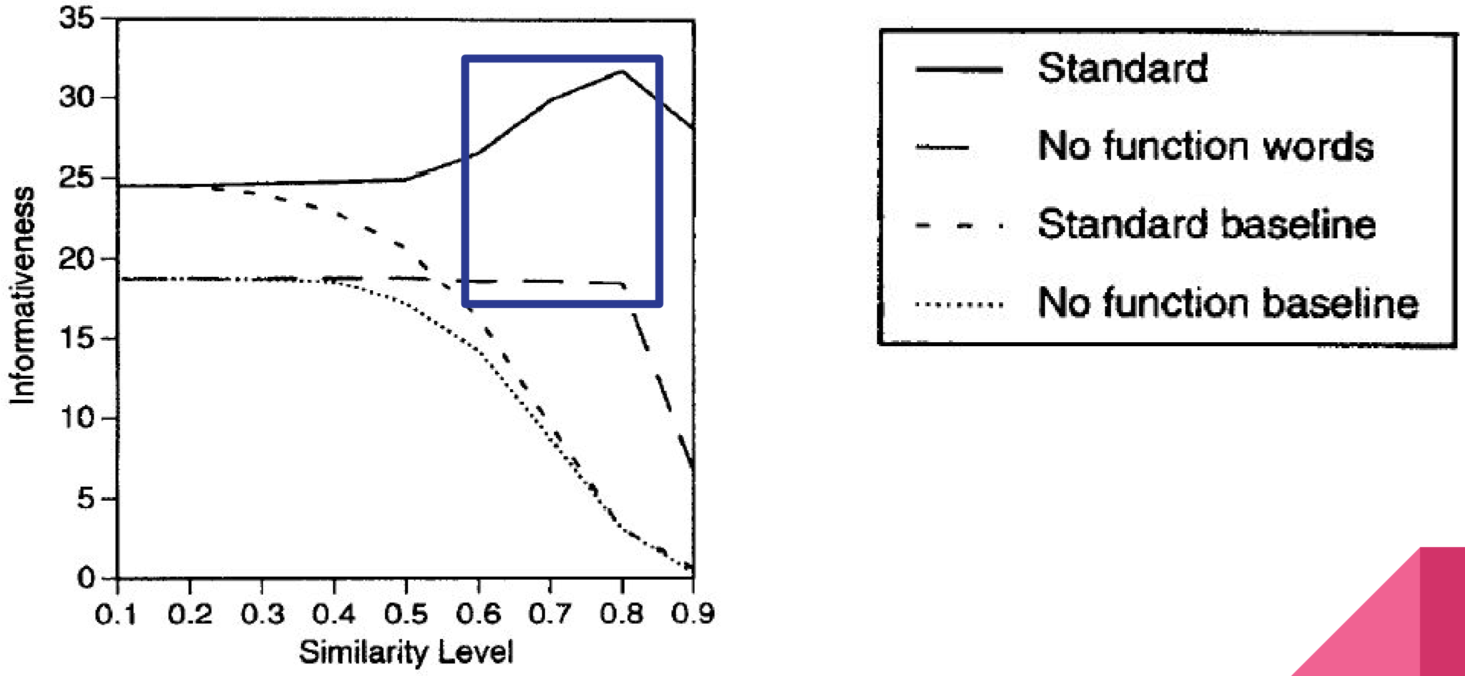
7.11 Experiment 7: Removing Function Words

- Removing function words does have a considerable impact on the amount of information provided by the method.
- Although, the analysis still provides a considerable amount of useful information without it.

7.11 Experiment 7: Removing Function Words



7.11 Experiment 7: Removing Function Words



7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

- The child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis.

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

- The child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis.

Way of using information about a particular class in our distributional analysis:

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

- The child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis.

Way of using information about a particular class in our distributional analysis:

- Replacing all words of a particular category with a symbol.

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

- The child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis.

Way of using information about a particular class in our distributional analysis:

- Replacing all words of a particular category with a symbol

All the nouns
All the verbs
...

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

- The child is likely to be able to exploit a variety of other cues, and these may be used in concert with distributional analysis.

Way of using information about a particular class in our distributional analysis:

- Replacing all words of a particular category with a symbol

All the nouns → NOUN
All the verbs → VERB
… → …

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

Advantage:

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

Advantage:

- Contexts concerning different nouns can be identified as having the same syntactic significance.

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

Advantage:

- Contexts concerning different nouns can be identified as having the same syntactic significance.

Disadvantage:

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

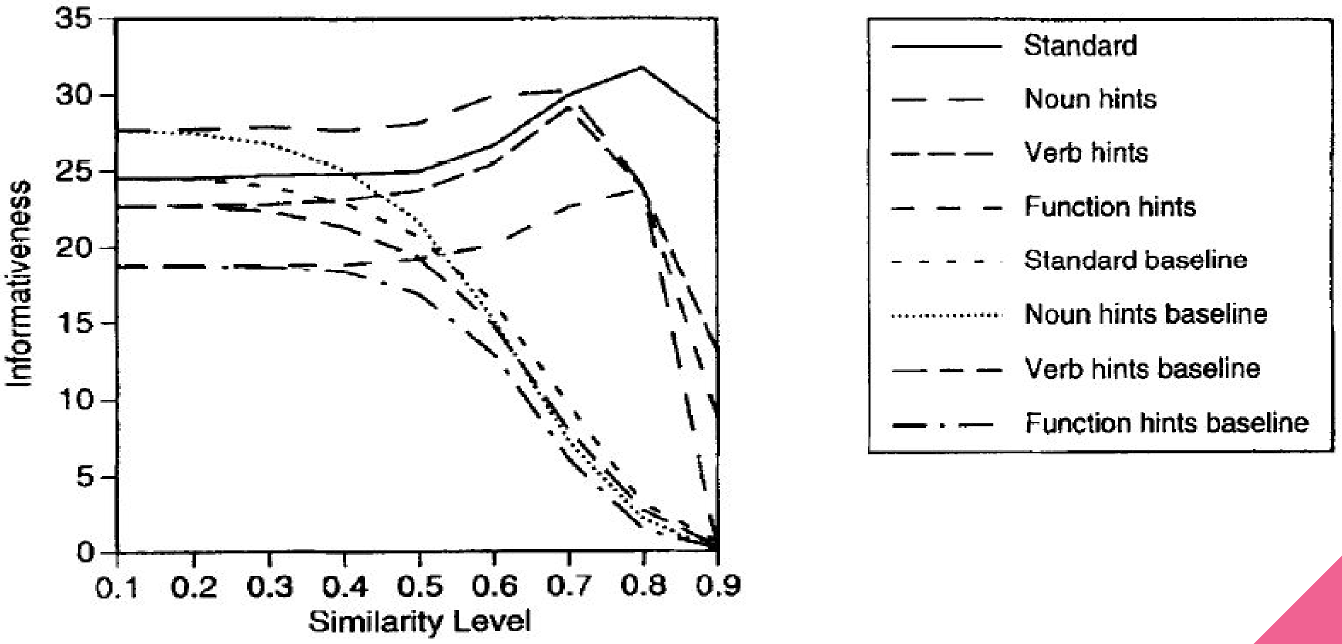
Advantage:

- Contexts concerning different nouns can be identified as having the same syntactic significance.

Disadvantage:

- Information about differences between nouns is lost.

7.12 Experiment 8: Does Information About One Category Help the Acquisition of the Others?

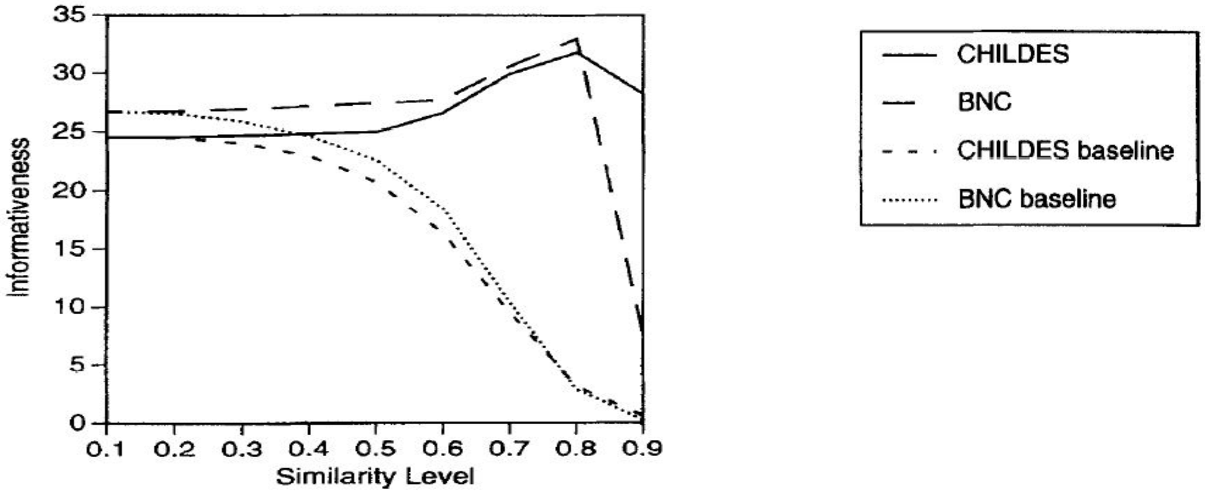


7.13 Experiment 9: Is Learning Easier with Child-Directed Input?

- Analysing whether the present distributional learning analyses are sensitive to the difference between adult speech to children, and adult-adult speech.

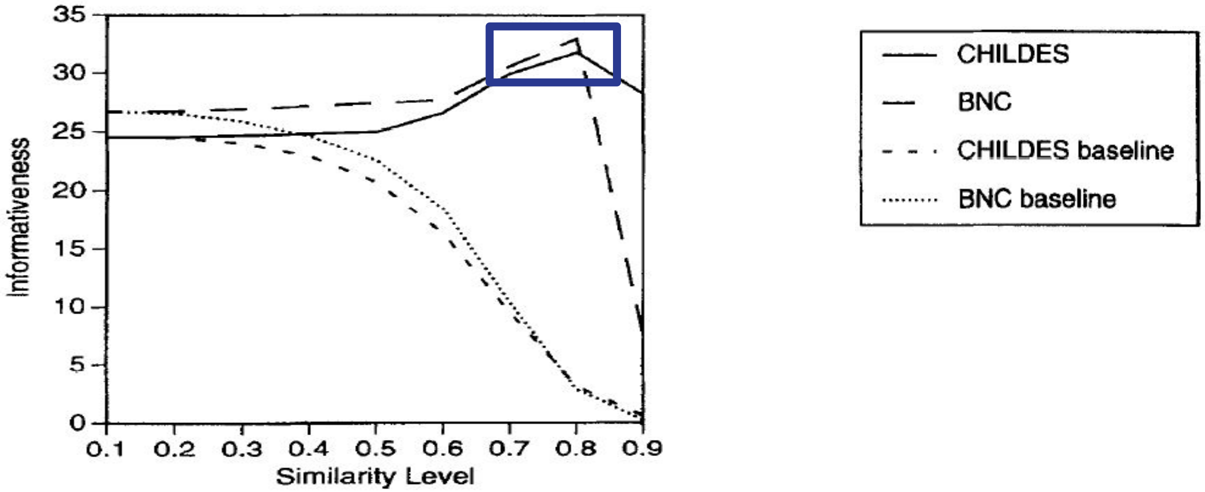
7.13 Experiment 9: Is Learning Easier with Child-Directed Input?

- Analysing whether the present distributional learning analyses are sensitive to the difference between <u>adult speech to children</u>, and <u>adult-adult speech</u>.



7.13 Experiment 9: Is Learning Easier with Child-Directed Input?

- Analysing whether the present distributional learning analyses are sensitive to the difference between adult speech to children, and adult-adult speech.



8. Conclusion

- A model of how children may use distributional information in acquiring syntactic categories has been shown.

8. Conclusion

- A model of how children may use distributional information in acquiring syntactic categories has been shown.
- Distributional information is a potentially powerful cue for learning syntactic categories.

8. Conclusion

- A model of how children may use distributional information in acquiring syntactic categories has been shown.
- Distributional information is a potentially powerful cue for learning syntactic categories.
- The use of distributional methods is often associated with empiricist approaches to language acquisition.

Thank you!

Distributional Information:
A Powerful Cue for Acquiring Syntactic Categories

Get the slides from this presentation on **hyperurl.co/distinfo**

Levindo Gabriel Taschetto Neto
22.11.2017

Probabilistic models of language and cognition
Universität Stuttgart