Learning Syntactic Categories

Informatics 1 CG: Lecture 10

Mirella Lapata

School of Informatics
University of Edinburgh
mlap@inf.ed.ac.uk

February 2, 2016

Reading:

Redington et al. (1998). Distributional Information: A Powerful Cue for Acquiring Syntactic Categories. Cognitive Science 22, 425-469.

Recap: Word Learning

Word learning is hard, children use multiple sources of support:

- socio-pragmatic skills
- some aspects of child directed speech
- biases towards certain interpretations over others
- linguistic constraints through use of syntax

How Do Children Learn Syntactic Categories?

One of most basic requirements of understanding language is identifying the syntactic categories to which the words belong.

- Is a word a noun, verb, adverb, or adjective?
- How do children learn these categories and which words belong to them?
- Are categories hard-wired in the brain (rationalist view)?
- Or are they learned (empiricist view)?

Open and Closed Classes in Natural Language

- Several broad word classes are found in all Indo-European languages and many others: nouns, verbs, adjectives, adverbs.
- These are examples of open classes. They typically have large membership, and are often stable under translation.
- Other word classes are more specific to particular languages: prepositions (English, German), post-positions (Hungarian, Urdu, Korean), particles (Japanese), etc.
- These are examples of closed classes. They typically have small, relatively fixed membership, and often have structuring uses in grammar. Little correlation between languages.

Parts of Speech

How do we tell what word class (part of speech) a word belongs to?

At least three different criteria can be used:

- Semantic criteria: What does the word refer to?
- Morphological criteria: What does the word look like?
- Distributional (syntactic) criteria: Where is the word found?

We will look at different parts of speech (POS) using these criteria.

Nouns

Semantically, nouns generally refer to living things (*mouse*), places (*Scotland*), things (*harpoon*), or concepts (*marriage*).

Morphologically, *-ness*, *-tion*, *-ity*, and *-ance* tend to indicate nouns. (*happiness, exertion, levity, significance*).

Distributionally, we can examine the contexts where a noun appears and other words that appear in the same contexts.

like a Newfoundland dog just from the water
he was seen swimming like a dog , throwing his long arms
such a deceitful dog ! It was only the last
was mauled to death by her pet dog have described her as their
Adopting an adult dog can be a marvelous alternative

Nouns

Semantically, nouns generally refer to living things (*mouse*), places (*Scotland*), things (*harpoon*), or concepts (*marriage*).

Morphologically, *-ness*, *-tion*, *-ity*, and *-ance* tend to indicate nouns. (*happiness, exertion, levity, significance*).

Distributionally, we can examine the contexts where a noun appears and other words that appear in the same contexts.

like a Newfoundland dog just from the water
he was seen swimming like a dog , throwing his long arms
such a deceitful dog ! It was only the last
was mauled to death by her pet dog have described her as their
Adopting an adult dog can be a marvelous alternative

Verbs

Semantically, verbs refer to actions (*observe*, *think*, *give*).

Morphologically, words that end in *-ate* or *-ize* tend to be verbs, and ones that end in *-ing* are often the present participle of a verb (*automate, calibrate, equalize, modernize; rising, washing, grooming*).

Distributionally, we can examine the contexts where a verb appears and other words that appear in the same contexts, which may include their arguments.

Had he married a more amiable woman , he might have
he was very young when he married , and very fond of his wife .
I am sure she will be married to Mr . Willoughby very soon .
Biddy Henshawe ; she married a very wealthy man .
I widowed that poor girl when I married her , Starbuck ;

Verbs

Semantically, verbs refer to actions (*observe*, *think*, *give*).

Morphologically, words that end in *-ate* or *-ize* tend to be verbs, and ones that end in *-ing* are often the present participle of a verb (*automate, calibrate, equalize, modernize; rising, washing, grooming*).

Distributionally, we can examine the contexts where a verb appears and other words that appear in the same contexts, which may include their arguments.

Had he married a more amiable woman , he might have
he was very young when he married , and very fond of his wife .
I am sure she will be married to Mr . Willoughby very soon .
Biddy Henshawe ; she married a very wealthy man .
I widowed that poor girl when I married her , Starbuck ;

Adjectives

Semantically, adjectives convey properties of or opinions about things that are nouns (*small*, *wee*, *sensible*, *excellent*).

Morphologically, words that end in *-al*, *-ble*, and *-ous* tend to be adjectives (*formal, gradual, sensible, salubrious, parlous*)

Distributionally, adjectives usually appear before a noun or after a form of *be*.

a great pity that such a sensible young man should be so
soaked through , it ' s hard to be sensible , that ' s a fact .
She was sensible and clever ; but eager in everything
I should have been sensible of it at the time , for we always
He was confused , seemed scarcely sensible of pleasure in seeing

Adjectives

Semantically, adjectives convey properties of or opinions about things that are nouns (*small*, *wee*, *sensible*, *excellent*).

Morphologically, words that end in *-al*, *-ble*, and *-ous* tend to be adjectives (*formal, gradual, sensible, salubrious, parlous*)

Distributionally, adjectives usually appear before a noun or after a form of *be*.

a great pity that such a sensible young man should be so
soaked through , it ' s hard to be sensible , that ' s a fact .
She was sensible and clever ; but eager in everything
I should have been sensible of it at the time , for we always
He was confused , seemed scarcely sensible of pleasure in seeing

The Problem of Learning Syntactic Categories

Difficult problem from both nativist and empiricist perspectives on language acquisition.

- Nativists: syntactic categories, are innate; learner must map lexicon of target language into these categories. There must be significant constraints on which mappings are considered.
- Empiricists: finding correct mappings appears more difficult still, since even the number of syntactic categories is not known a priori.
- On both views, learner must make the first steps in acquiring syntactic categories without being able to apply constraints from knowledge of the grammar.

What Information is Available?

Distributional Information

Words of the same category have a large number of distributional regularities in common, i.e., occur in similar linguistic contexts.

Semantic Bootstrapping

Abstract syntactic categories are innately specified, the learner makes a tentative mapping from lexical items to these syntactic categories, using semantic information (Pinker, 1984).

Phonological Constraints

There are regularities between the phonology of words and their syntactic categories which aid acquisition (stress, word duration).

Innate Knowledge

Learning mechanisms which exploit information in the input may be innately specified and used to constrain search space of the learner.

Redington et al. (1998)

Distributional properties can be highly informative of syntactic category. This information can be extracted by psychologically plausible mechanisms:

1. Measuring distribution of contexts within which words occur.
2. Comparing the distributions of contexts for pairs of words.
3. Grouping together words with similar distributions of contexts.

Measuring Distribution for each Word

What should count as a context for a word?

> *. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later . . .*

Measuring Distribution for each Word

What should count as a context for a word?

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their $\boxed{\textit{symbols. these must first be examined in}}$ *the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial* $\boxed{\textit{understanding. to learn the meaning of}}$ *symbols is part of the anthropologist's* $\boxed{\textit{practical semantics: to discover the meaning of}}$ *words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts.* $\boxed{\textit{these must come first; fantasy can come later}}$ *...*

Measuring Distribution for each Word

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their $\boxed{\textit{symbols. these must first be examined in}}$ *the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial* $\boxed{\textit{understanding. to learn the meaning of}}$ *symbols is part of the anthropologist's* $\boxed{\textit{practical semantics: to discover the meaning of}}$ *words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts.* $\boxed{\textit{these must come first; fantasy can come later}}$ *...*

Measuring Distribution for each Word

. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their *symbols. these must first be examined in* *the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial* *understanding. to learn the meaning of* *symbols is part of the anthropologist's* *practical semantics: to discover the meaning of* *words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts.* *these must come first; fantasy can come later* *. . .*

first
learn
discover

Measuring Distribution for each Word

... The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later ...

	these	meaning	to	practical	come
first					
learn					
discover					

Measuring Distribution for each Word

. . . The field anthropologist must gain understanding and start with the explanations and commentaries which his informants themselves offer about their symbols. these must first be examined in the contexts in which they are usually employed, where they occur naturally, although subsequent generalizing discussion helps the anthropologist to improve his initial understanding. to learn the meaning of symbols is part of the anthropologist's practical semantics: to discover the meaning of words, noticing when their use is appropriate and when it is not. all this requires imagination, patience, considerable linguistic skill, but above all a rigorous respect for the facts. these must come first; fantasy can come later . . .

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Measuring Distribution for each Word

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Measuring Distribution for each Word

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

Measuring Distribution for each Word

Context words

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

Measuring Distribution for each Word

Context words

	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

Context vectors

Measuring Distribution for each Word

Context words

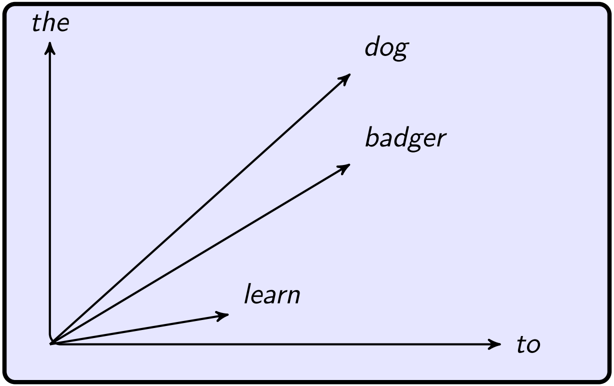
	these	meaning	to	practical	come
first	2	0	0	0	2
learn	0	1	1	0	0
discover	0	1	1	0	1

Target words

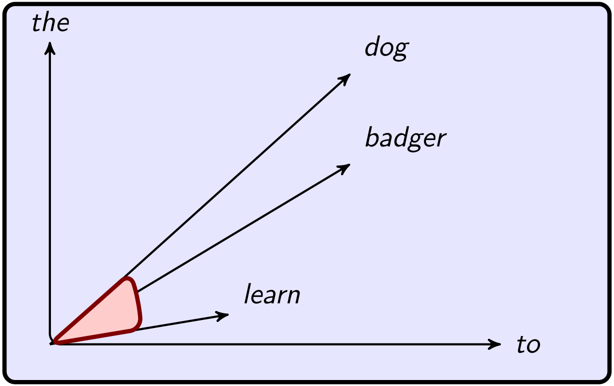
Context vectors

- Words are represented by context vectors.
- Redington et al. obtain such context vectors from CHILDES (a corpus of child directed speech, 2.5 million words).
- An algorithm takes vectors as input and produces clusters.
- Clusters correspond to parts of speech.

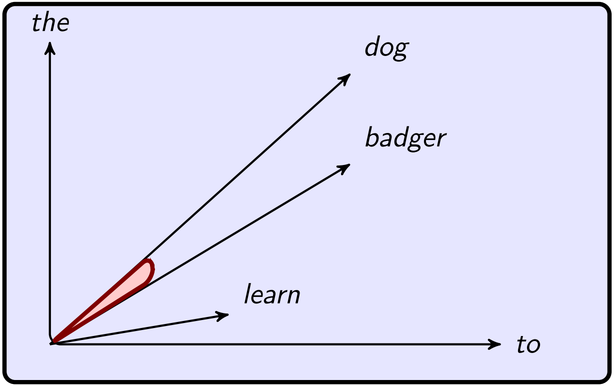
Words as Context Vectors



Words as Context Vectors



Words as Context Vectors



Agglomerative Clustering

Learning Algorithm

1: Place each data point into its own singleton group
2: Repeat: iteratively merge the two closest groups
3: Until: all the data are merged into a single cluster

- Algorithm measures how close two groups are according to a distance or similarity function.
- Redington et al. use Spearman's rank correlation
- Many other choices are possible (e.g., cosine measure)

Agglomerative Clustering

Learning Algorithm

1. Place each data point into its own singleton group
2. Repeat: iteratively merge the two closest groups
3. Until: all the data are merged into a single cluster

- The algorithm results in a sequence of groupings
- It is up to the user to choose "natural" clustering sequence
- Dendrogram: plot each merge at the similarity between two merged groups
- Provides interpretable visualization of algorithm and data

Group Similarity

Given a distance measure between points, the user has many choices for how to define intergroup similarity.

Single-linkage: similarity of the closest pair

$$d_{SL}(G, H) = \min_{i \in G \; j \in H} d_{ij}$$

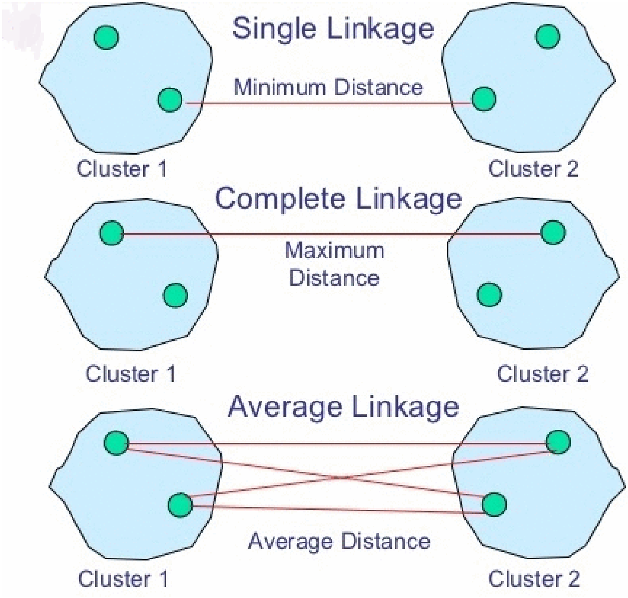
Complete-linkage: similarity of the furthest pair

$$d_{CL}(G, H) = \max_{i \in G \; j \in H} d_{ij}$$

Group average: the average similarity between groups

$$d_{GA}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{j \in H} d_{ij}$$

Group Similarity



Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}

$d(\{A, B\}) = 1, d(\{A, C\}) = 2, d(\{A, D\}) = 2, d(\{A, D\}) = 3$
$d(\{B, C\}) = 2, d(\{B, D\}) = 4, d(\{B, E\}) = 5$
$d(\{C, D\}) = 1, d(\{C, E\}) = 5$
$d(\{D, E\}) = 3$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}

$d(\{A,B\}) = 1, d(\{A,C\}) = 2, d(\{A,D\}) = 2, d(\{A,D\}) = 3$
$d(\{B,C\}) = 2, d(\{B,D\}) = 4, d(\{B,E\}) = 5$
$d(\{C,D\}) = 1, d(\{C,E\}) = 5$
$d(\{D,E\}) = 3$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}

$d(\{A,B\}) = 1, d(\{A,C\}) = 2, d(\{A,D\}) = 2, d(\{A,D\}) = 3$
$d(\{B,C\}) = 2, d(\{B,D\}) = 4, d(\{B,E\}) = 5$
$d(\{C,D\}) = 1, d(\{C,E\}) = 5$
$d(\{D,E\}) = 3$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}

$$
\begin{aligned}
d(\{A,B\},\{C,D\}) &= min\{d(A,C), d(A,D), d(B,C), d(B,D)\} \\
&= min\{2,3,2,4\} \\
&= 2 \\
d(\{A,B\},\{E\}) &= min\{d(A,E), d(B,E)\} \\
&= min\{3,5\} \\
&= 3 \\
d(\{C,D\},\{E\}) &= min\{d(C,E), d(D,E)\} \\
&= min\{5,3\} \\
&= 3
\end{aligned}
$$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}

$$\begin{aligned}
d(\{A,B\},\{C,D\}) &= min\{d(A,C), d(A,D), d(B,C), d(B,D)\} \\
&= min\{2,3,2,4\} \\
&= 2 \\
d(\{A,B\},\{E\}) &= min\{d(A,E), d(B,E)\} \\
&= min\{3,5\} \\
&= 3 \\
d(\{C,D\},\{E\}) &= min\{d(C,E), d(D,E)\} \\
&= min\{5,3\} \\
&= 3
\end{aligned}$$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	*k*	*K*
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}
2	2	{A,B,C,D}, {E}

$$\begin{aligned}
d(\{A,B\},\{C,D\}) &= min\{d(A,C), d(A,D), d(B,C), d(B,D)\} \\
&= min\{2,3,2,4\} \\
&= 2 \\
d(\{A,B\},\{E\}) &= min\{d(A,E), d(B,E)\} \\
&= min\{3,5\} \\
&= 3 \\
d(\{C,D\},\{E\}) &= min\{d(C,E), d(D,E)\} \\
&= min\{5,3\} \\
&= 3
\end{aligned}$$

Single Link Agglomerative Clustering: Example

	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}
2	2	{A,B,C,D}, {E}

$$\begin{aligned} d(\{A, B, C, D\}, \{E\}) &= min\{d(A, E), d(B, E), d(C, E), d(D, E)\} \\ &= min\{3, 5, 5, 3\} \\ &= 3 \end{aligned}$$

Single Link Agglomerative Clustering: Example

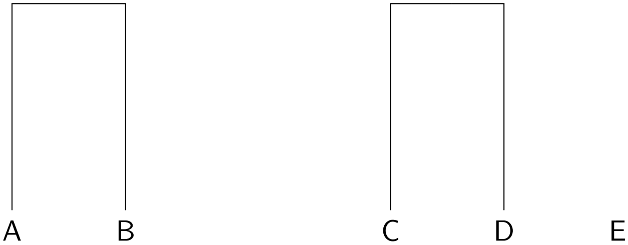
	A	B	C	D	E
A	0	1	2	2	3
B	1	0	2	4	3
C	2	2	0	1	5
D	2	4	1	0	3
E	3	5	5	3	0

d	k	K
0	5	{A}, {B}, {C}, {D}, {E}
1	3	{A,B}, {C,D}, {E}
2	2	{A,B,C,D}, {E}
3	1	{A,B,C,D,E}

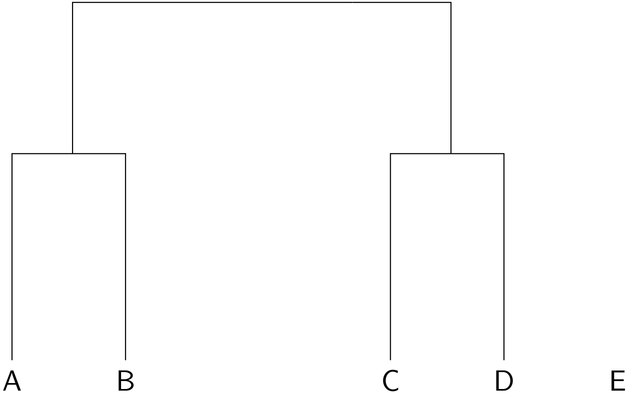
Dendrogram

A B C D E

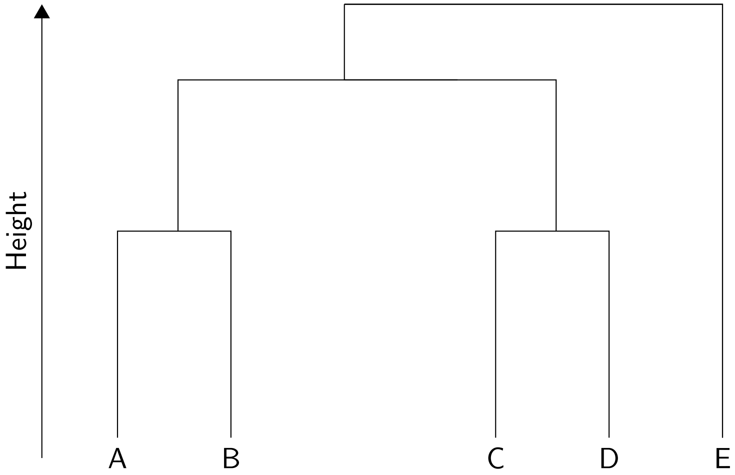
Dendrogram



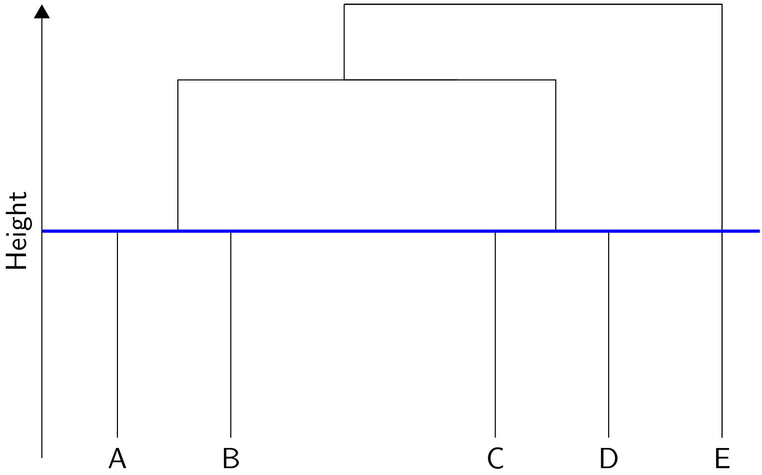
Dendrogram



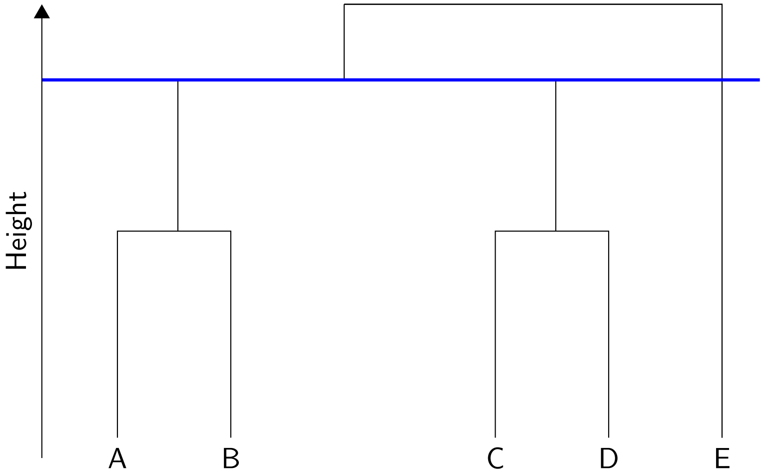
Dendrogram



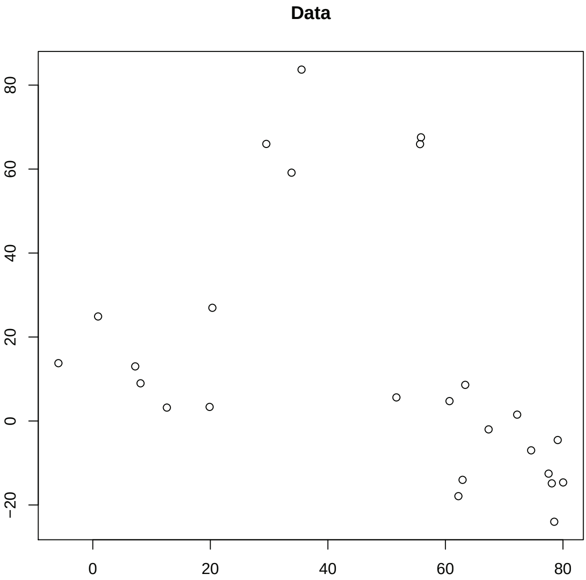
Dendrogram



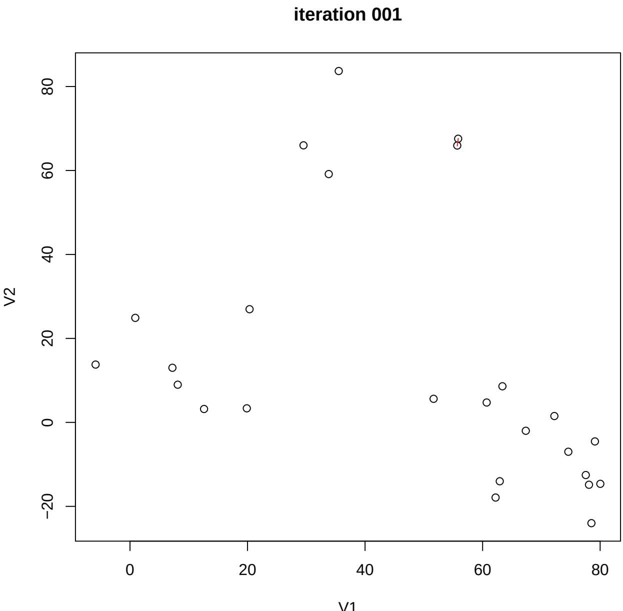
Dendrogram



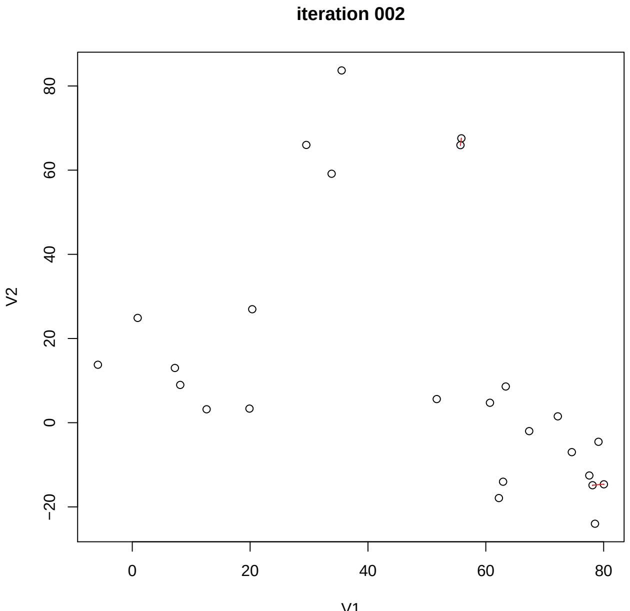
Example



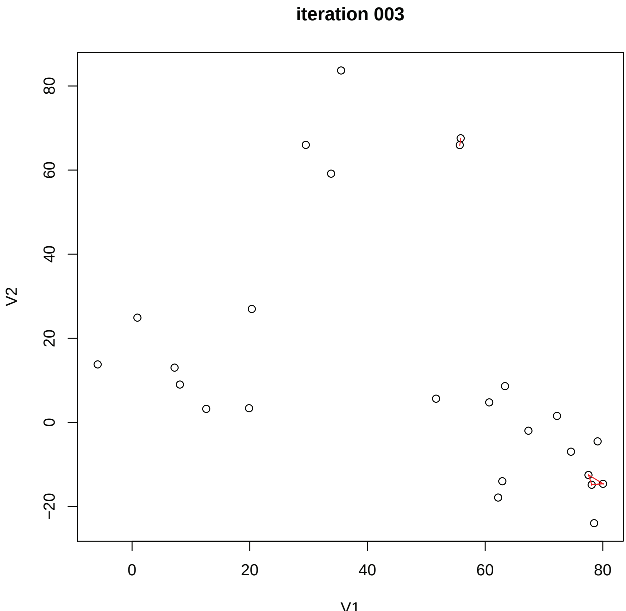
Example



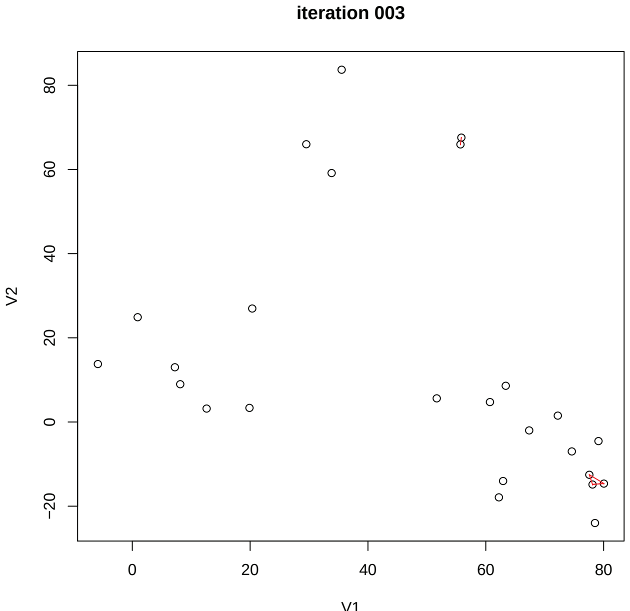
Example



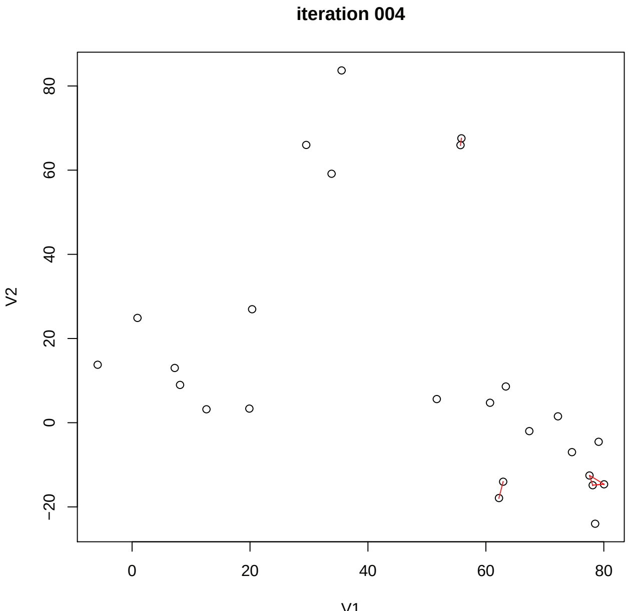
Example



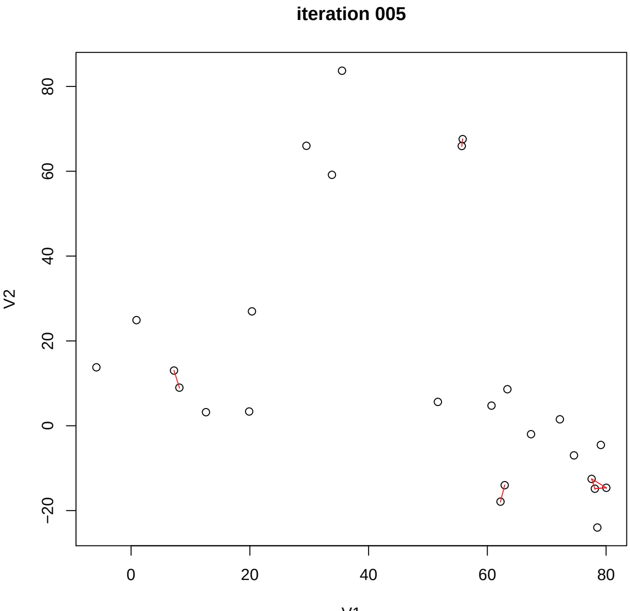
Example



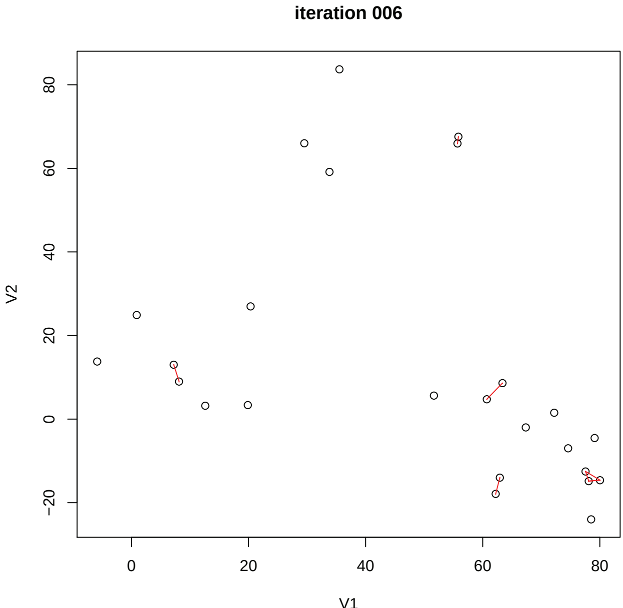
Example



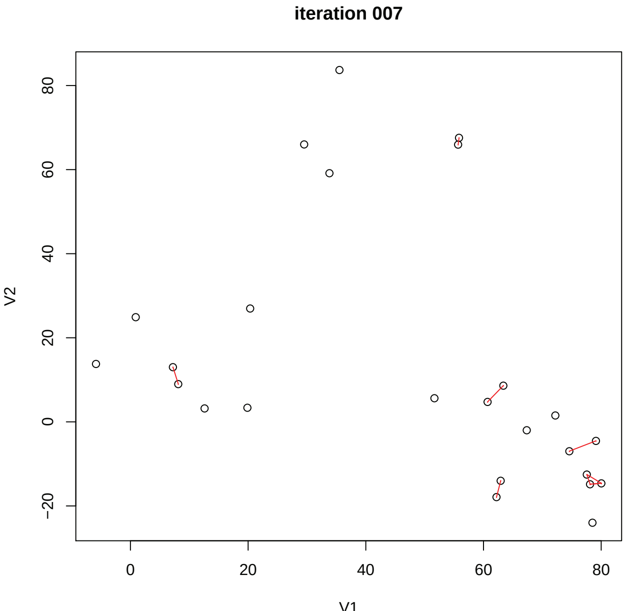
Example



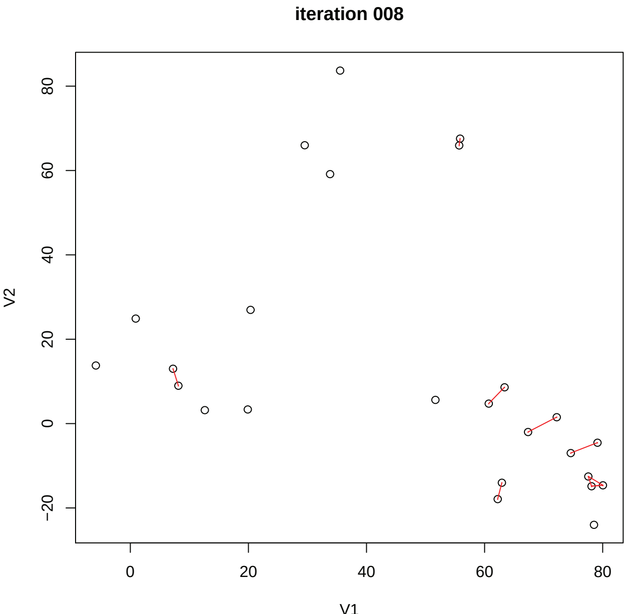
Example



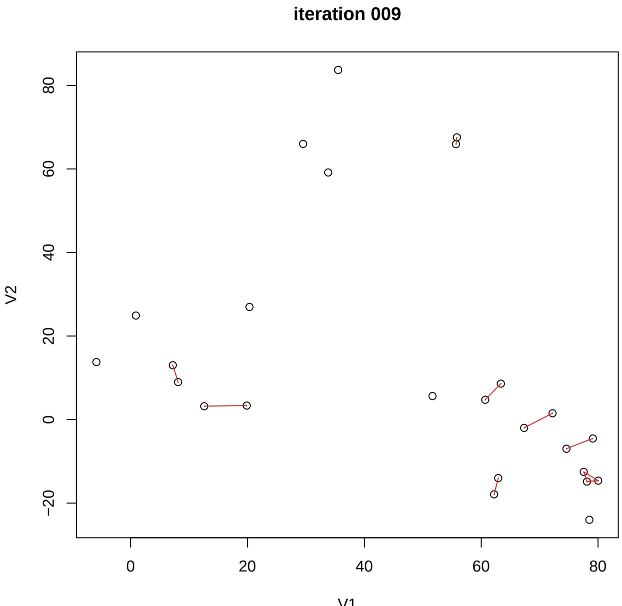
Example



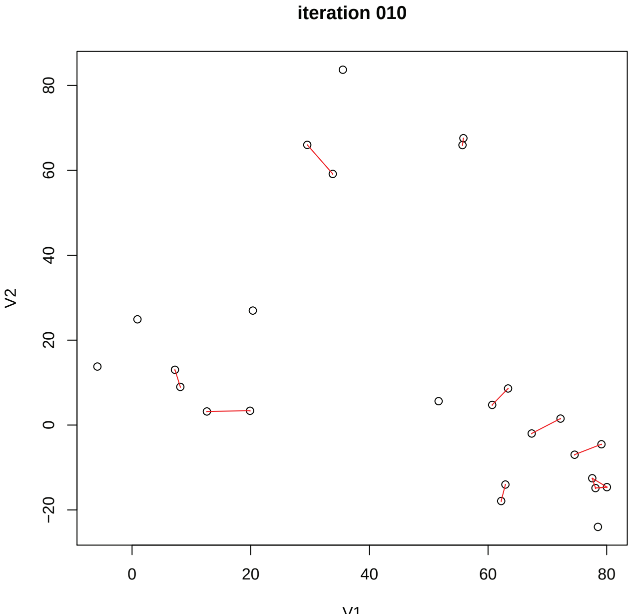
Example



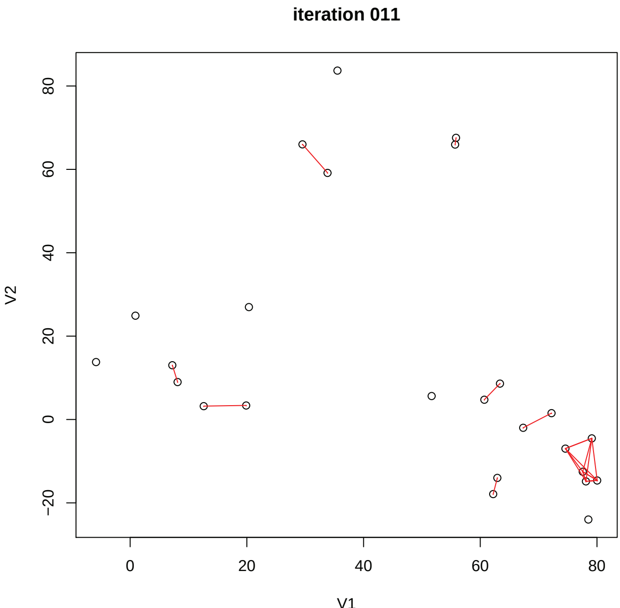
Example



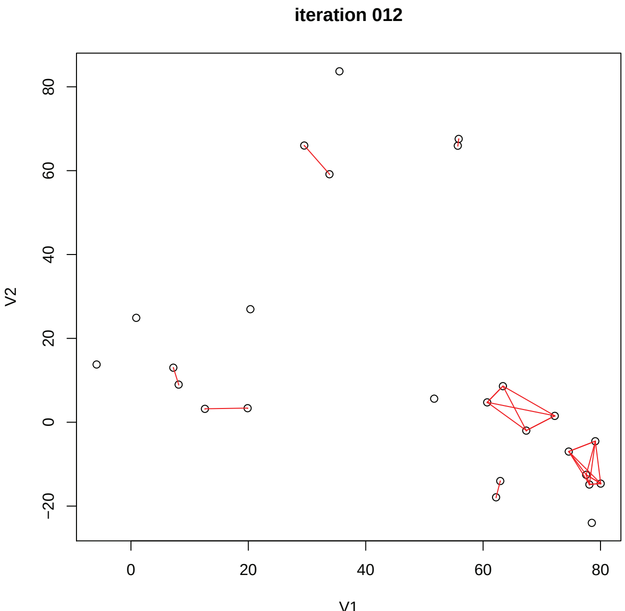
Example



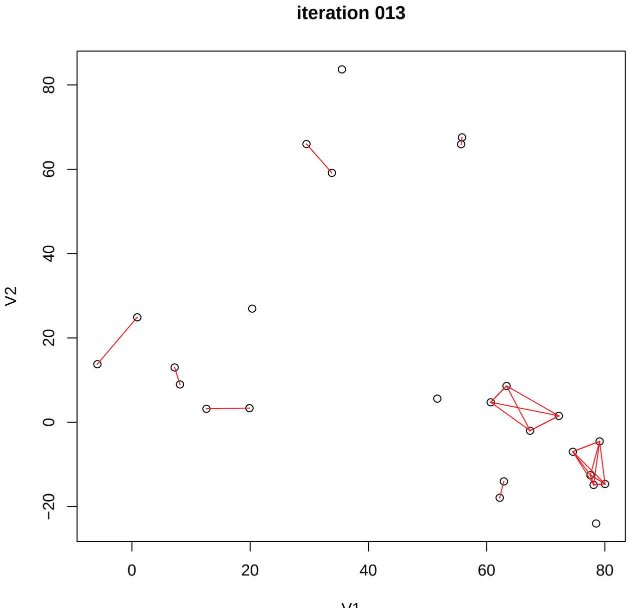
Example



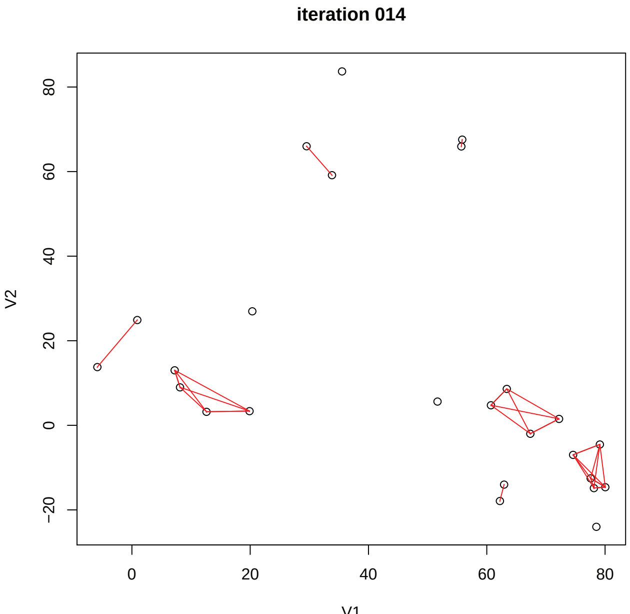
Example



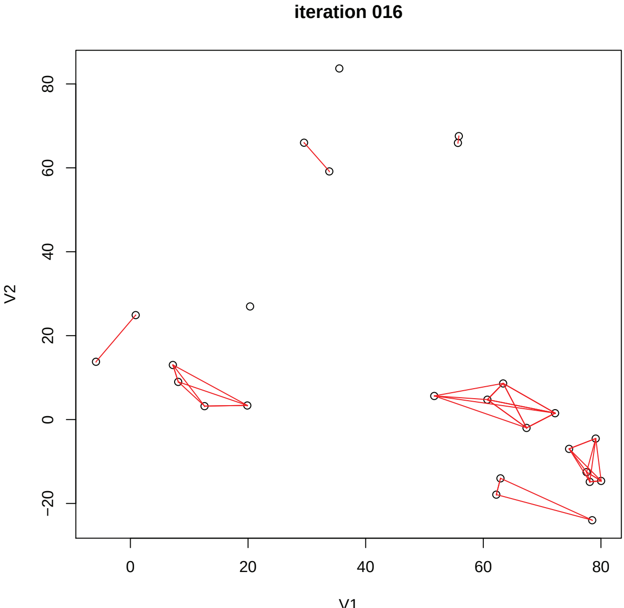
Example



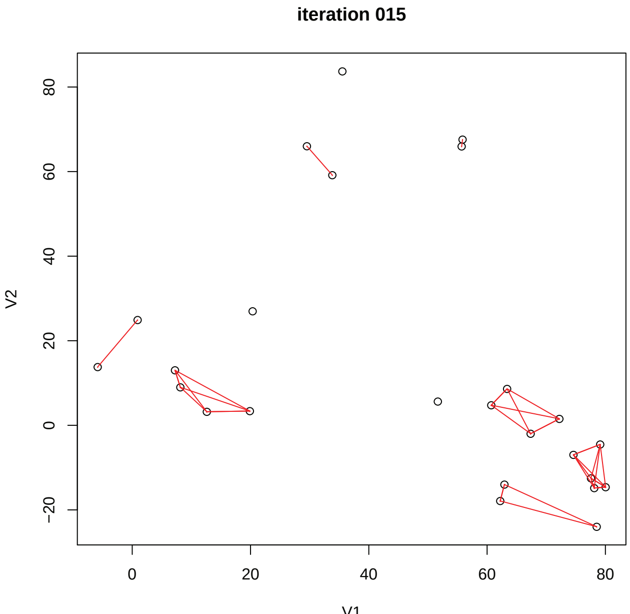
Example



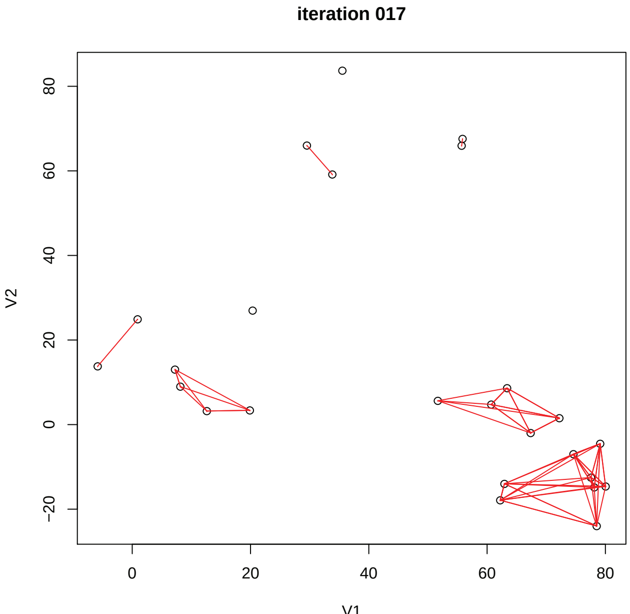
Example



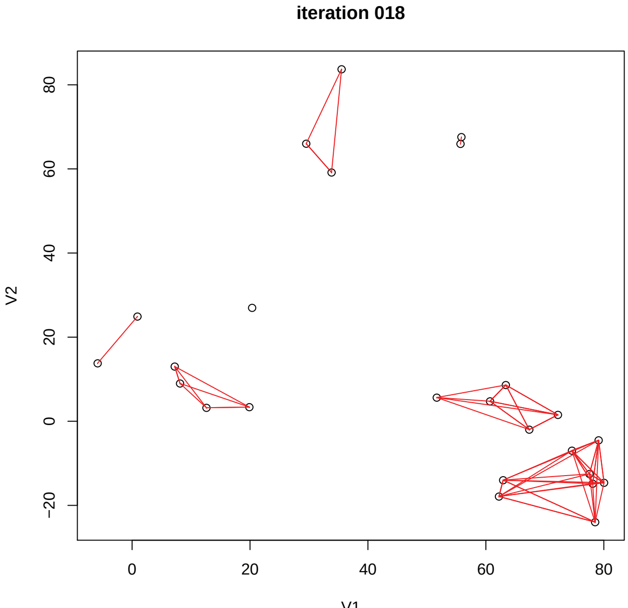
Example



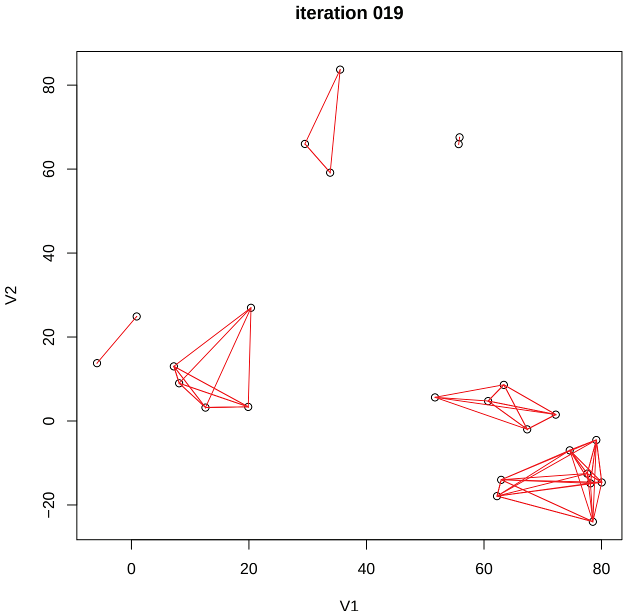
Example



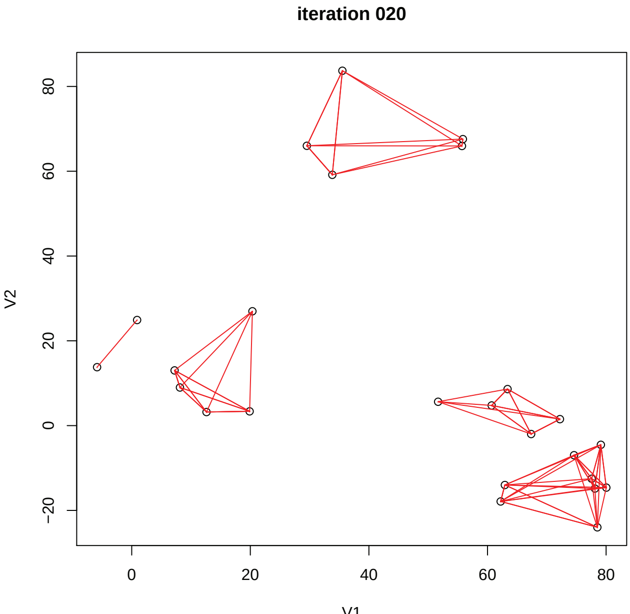
Example



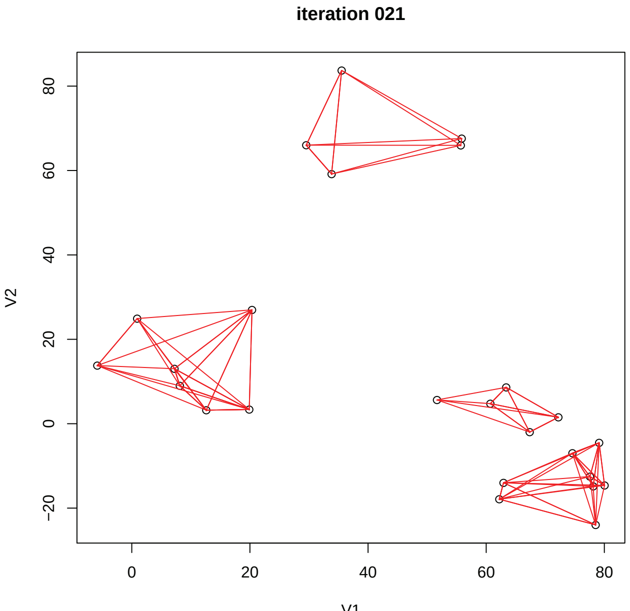
Example



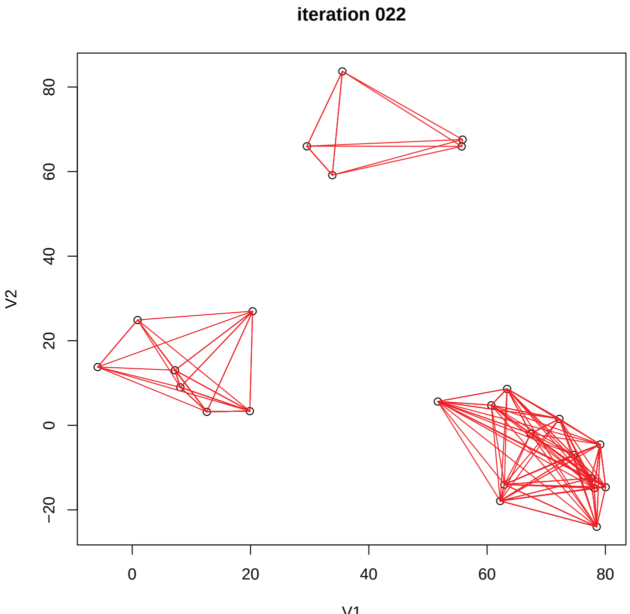
Example



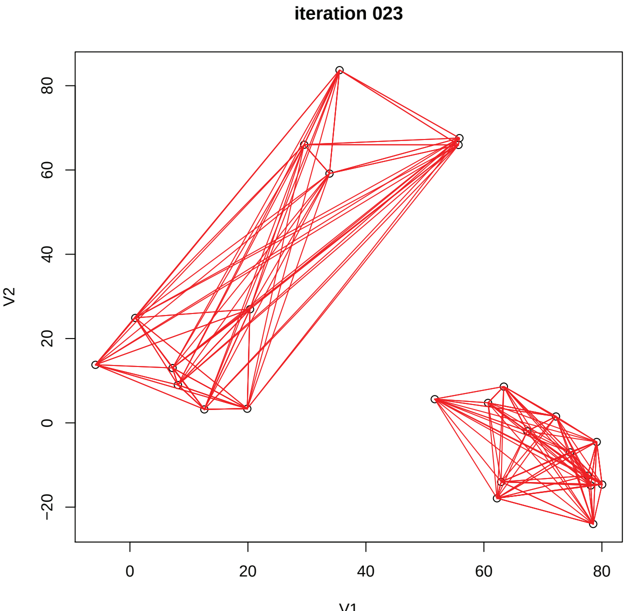
Example



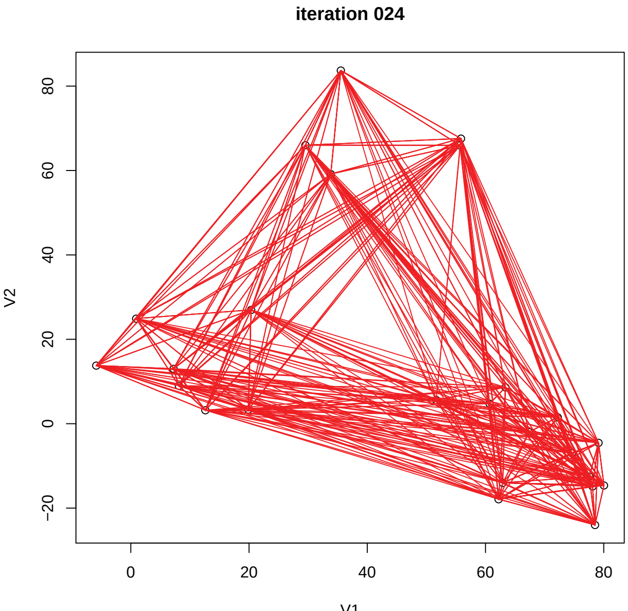
Example



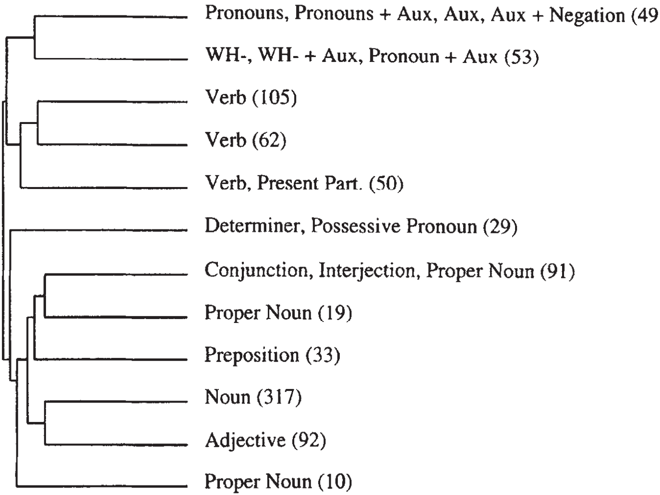
Example



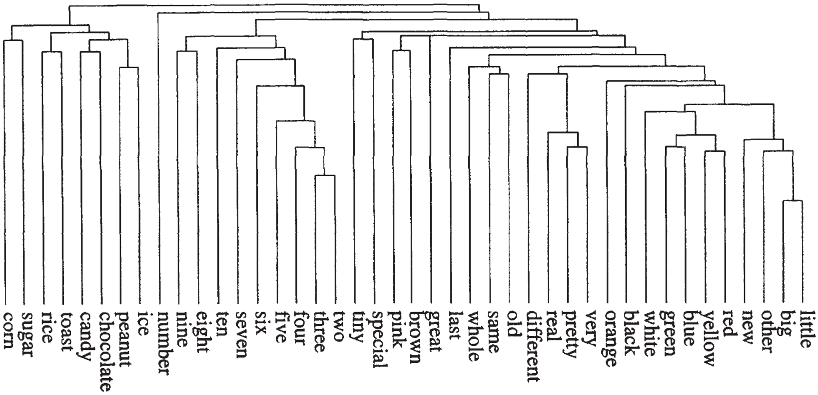
Example



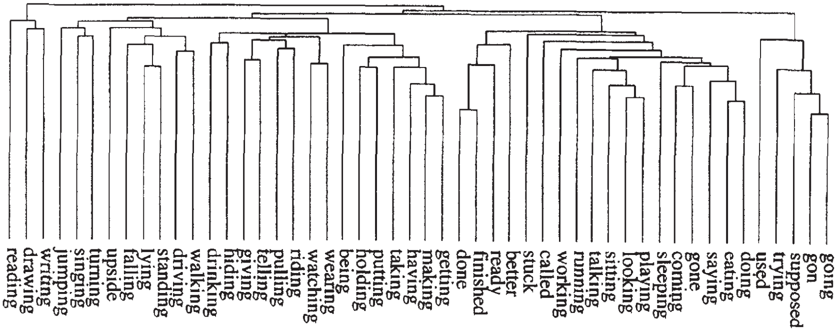
Clusters from Redington et al.



Adjectives Cluster



Present Participles Cluster



Redington et al.'s Results

- The model uses highly local distributional information which is consistent with early vocabulary development
- It is most effective for learning nouns, then verbs, and least effective for function words, mirroring children's syntactic development
- The method learns using the input corpora of the order of magnitude received by the child
- The success of this model suggests that distributional information may make an important contribution to early language development.

Summary

Discussed the problem of learning syntactic categories.

- Model of how children may use distributional information in acquiring syntactic categories.
- Using agglomerative clustering on CHILDES corpus
- Distributional information is a potentially powerful cue for learning syntactic categories and language in general.
- General approach uses computationally explicit model of specific aspects of language acquisition.

Remaining questions:

- Does proposed method apply to languages other than English without strong word order constraints?
- How about integrating other sources of distributional information (e.g., morphological or phonological cues)?
- Induced syntactic categories are not ambiguous (*frank words* vs *frank a stamp*).