Automatic acquisition of lexical information for low-frequent words

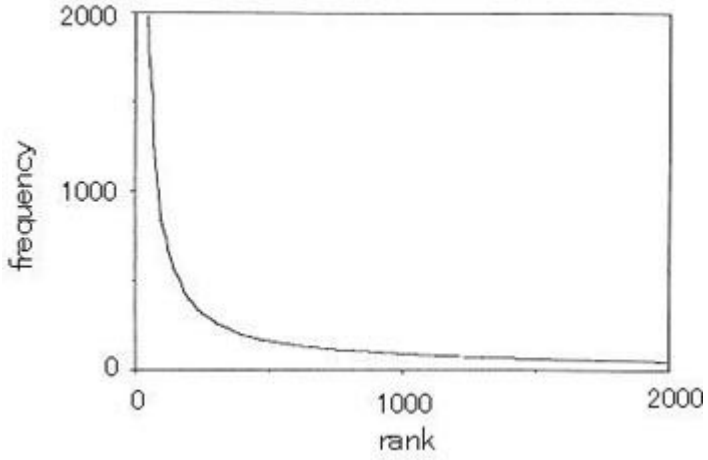
Silvia Necşulescu
Universitat Pompeu Fabra, Barcelona

INSTITUT UNIVERSITARI DE LINGÜÍSTICA APLICADA

UNIVERSITAT POMPEU FABRA

Objective

- To find a new similarity scheme able to handle low-frequency words in order to discover other possible co-occurring words for low-frequent words besides the co-occurrences observed in the corpus.



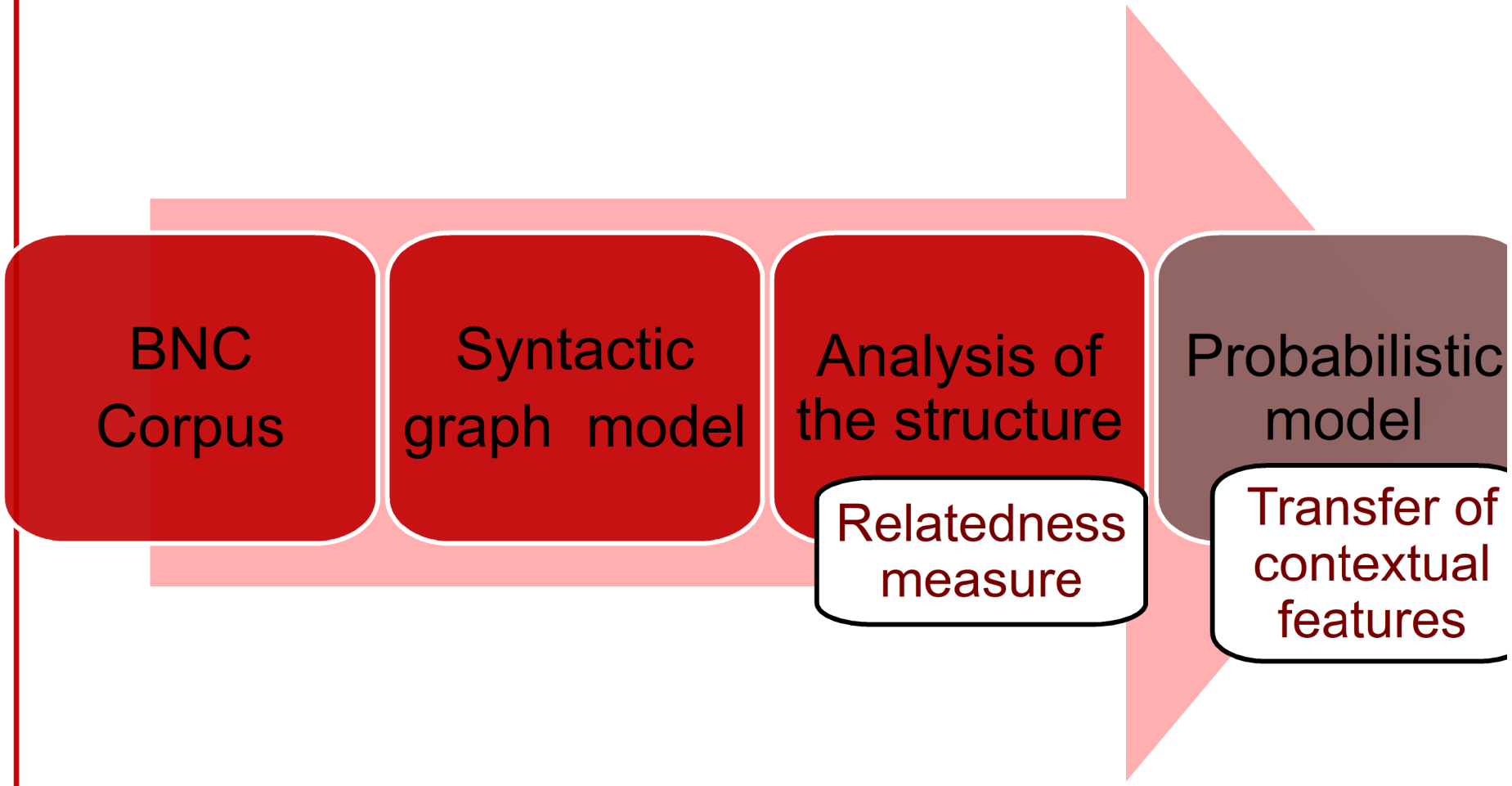
Word classes

- The words are grouped in different sets of classes based on the set of features chosen for classification
  - Semantic: Synonyms, near-synonyms, hyponyms/hypernyms, antonyms, etc.
  - Concepts: (land)mammals, fish, vegetables, fruit, trees, vehicles, clothes, tools, kitchenware
  - Grammatical: Nouns, Verbs, Adjectives, Adverbs
- The classes may differ in granularity

General hypothesis

- Based on the word context we aim at inducing semantic similarities among words
  - semantic similarities that must be abstracted and generalized into word classes.
- A word class should be defined as an open set of words bounded by restrictions over their **syntagmatic** and **paradigmatic** relations.

Workflow


BNC Corpus
Syntactic graph model
Analysis of the structure
Relatedness measure
Probabilistic model
Transfer of contextual features

Word class feature detection

- Detect automatically:
  - Words that describe a given class:
    - General features for the class
    - Specific features for at least one seed

Syntagmatic relations

  - Words that belong to the same class

Paradigmatic relations

Evaluation

- Class defined by an initial set of seeds
- Objective: detect if the high-ranked words are
  - Class related:
    - Syntagmatic
    - Paradigmatic
  - General terms or unrelated with the class

Evaluation: TOOLS class

- Seeds: *screwdriver, chisel, scissors, kettle, hammer, spoon, pencil, pen, bowl, knife, telephone, cup, bottle*
- 59036 words in graph
- Only 2259 nodes are ranked above the threshold

Evaluation: TOOLS class

- Evaluation
  - Human annotation of ~2000 words in 3 classes
  - Human annotation of a *random sample* of 1000 words (**only 36 are above the threshold).**

Thank you!