.inf
INSTITUTO
DE INFORMÁTICA
UFRGS

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

mwetoolkit:
A tool for automated extraction of multi-word expressions

Vítor De Araújo
Carlos Ramisch

Multi-word expressions

Combinations of words that present linguistical or statistical idiosyncrasies

- Phrasal verbs: *carry up, consist of*
- Support verbs: *take a walk, make a decision*
- Compounds: *computer science, washing machine*
- Idiomatic expressions: *raining cats and dogs, on the other hand*

mwetoolkit (*mwetoolkit.sf.net*)**:**

- Automated tool for MWE extraction from corpora
- Linguistical methods (morphosyntactic patterns)
- Statistical methods (association measures)

mwetoolkit

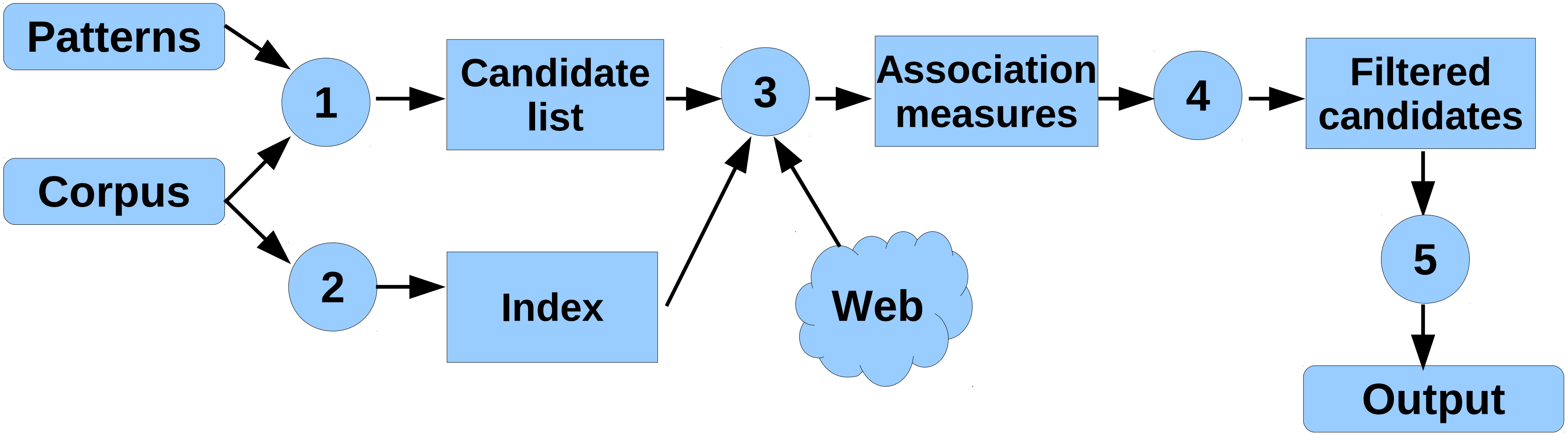
MWEs in natural language processing

- MWEs are **ubiquitous** in natural language
  - Everyday expressions
  - Technical terminology

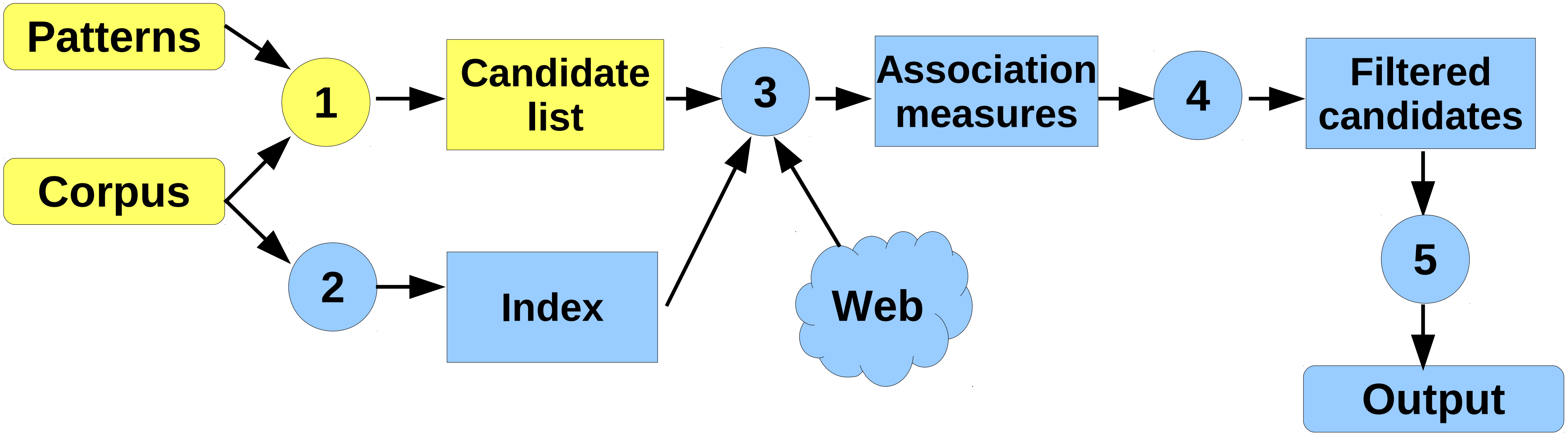
- MWEs are **hard to deal with**
  - Non-compositional: *give up*
  - Conventional/arbitrary: *computer science*
  - Domain-specific: *binary tree, angiosperm tree*

- A challenge to any NLP system requiring **semantic processing**
  - E.g., machine translation: *give up* → **dar para cima*

How does it work?



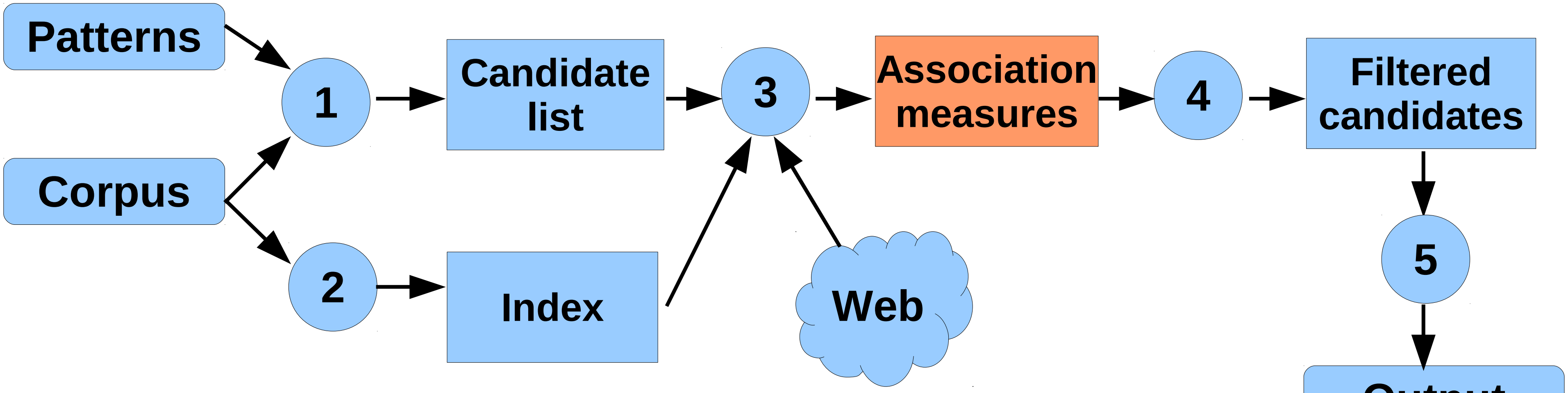
How does it work?



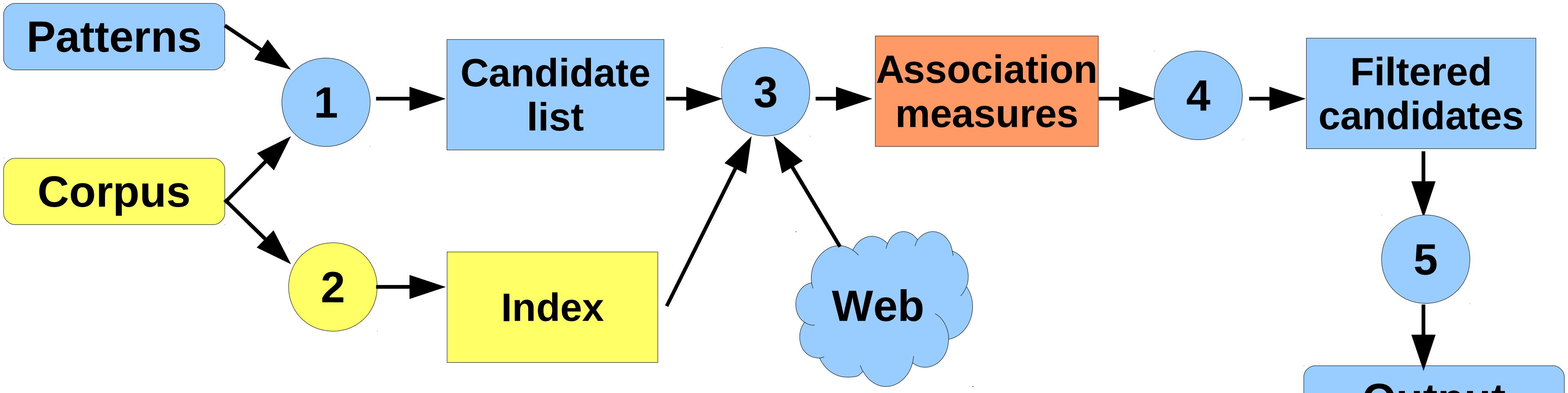
```
<pat>
   <w pos=”A”/>
   <w pos=”A”/>
   <w pos=”N”/>
   <w pos=”N”/>
</pat>
```

```
<ngram>
    <w surface="human" pos="A"/>
    <w surface="cd4+" pos="A"/>
    <w surface="t" pos="N"/>
    <w surface="cells" pos="N"/>
    <freq name="corpus" value="2"/>
</ngram>
```

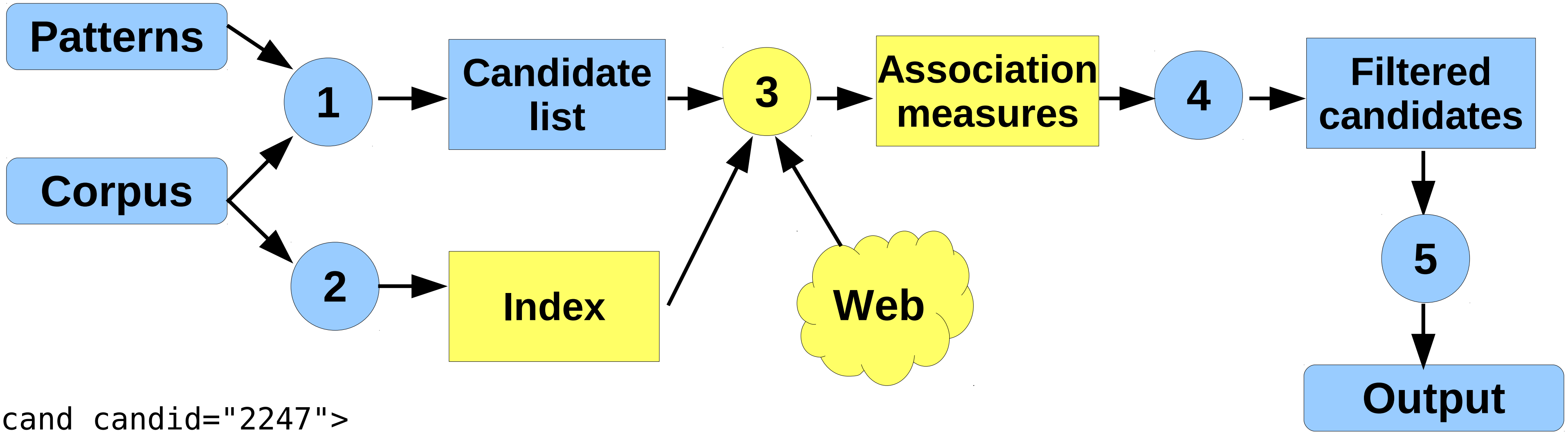
How does it work?



How does it work?

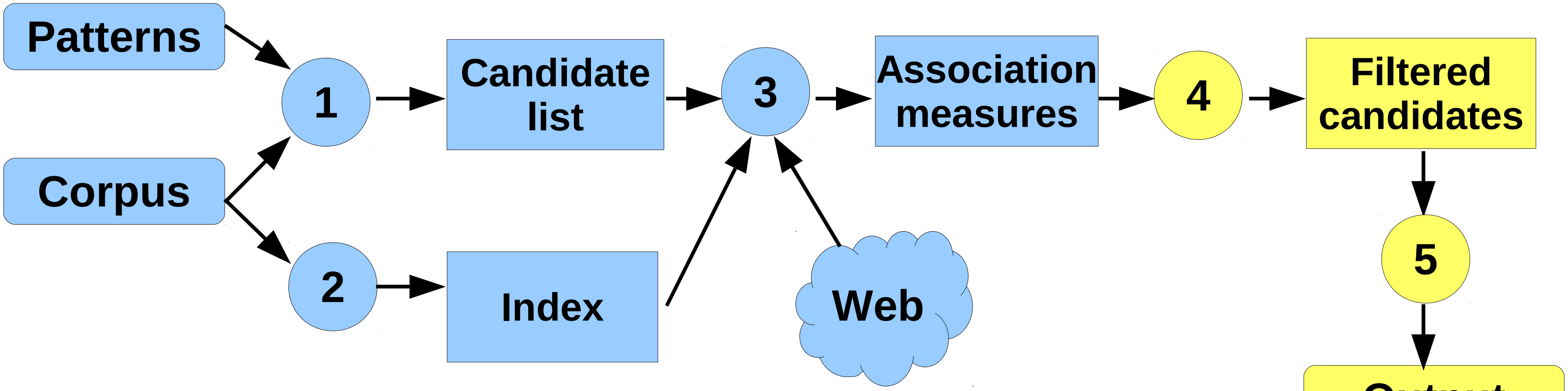


How does it work?



```
<cand candid="2247">
   <ngram>
      <w surface="human" pos="A"><freq name="corpus" value="78" /></w>
      <w surface="cd4+" pos="A"><freq name="corpus" value="5" /></w>
      <w surface="t" pos="N"><freq name="corpus" value="75" /></w>
      <w surface="cells" pos="N"><freq name="corpus" value="152" /></w>
    <freq name="corpus" value="2" /></ngram>
    <occurs>
    ...
    <features>
        <feat name="mle_corpus" value="0.000156201187129" />
        <feat name="pmi_corpus" value="19.8491824326" />
        <feat name="t_corpus" value="1.41421206505" />
        <feat name="dice_corpus" value="0.0258064516129" />
        <feat name="ll_corpus" value="0.0" />
    </features>
</cand>
```

How does it work?



Patterns

- Literal pattern

```
<pat>
   <w pos=”A”/>
   <w pos=”N”/>
   <w pos=”N”/>
</pat>
```

E.g., modern computer science

```
<pat>
   <w lemma="take" pos="V"/>
   <w pos=”Det”/>
   <w pos=”N”/>
</pat>
```

E.g., take a walk

Patterns

- **Regular expressions**
  - Repetitions, optional items

```
<pat>
   <pat repeat=”?”><w pos=”Det”/></pat>
   <pat repeat=”*”><w pos=”A”/></pat>
   <pat repeat=”+”><w pos=”N”/></pat>
</pat>
```

  - Backreferences

```
<pat>
   <w pos=”N” id=”n1”/>
   <w pos=”Prep”/>
   <w pos=”N” lemma=”back:n1.lemma”/>
</pat>
```

E.g., day after day, step by step, hand in hand

Patterns

- **Non-contiguous MWEs**

```
<pat>
   <w pos=”VT”/>
   <pat repeat=”*” ignore=”true”><w/></pat>
   <w pos=”Adv”/>
</pat>
```

E.g. throw *whatever* away

- **Syntactic dependencies**

E.g., verb and its object

```
<pat>
   <w pos=”VT” id=”v1”/>
   <pat repeat=”*” ignore=”true”><w/></pat>
   <w pos=”N” syndep=”dobj:v1”/>
</pat>
```

Index

- Suffix array
- Per-attribute
- Automatic **attribute fusion**
  - E.g., *lemma+pos* (verb "like" vs. noun "like")
  - On-the-fly index generation from *lemma* and *pos*

- C indexing routines
  - British National Corpus
    - 110 million words
    - ~5min per attribute (lemma, surface, pos)
    - ~1GB memory

Other improvements

- Unified **command-based interface**
- Use of **Web 1 Trillion 5-gram** as a source of frequencies
- **LocalMaxs** algorithm: extraction without filtering
- **Preliminar evaluation:** MWE extraction in the discourse of children for study on language acquisition (CHILDES)

Conclusions

- Demo paper

  V. de Araújo, C. Ramisch, A. Villavicencio. Fast and Flexible MWE Candidate Generation with the mwetoolkit. In Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), pages 134–136, Portland, Oregon, USA, 23 June 2011.

  http://aclweb.org/anthology-new/W/W11/W11-0822.pdf

- Improvement, optimization and evaluation of a MWE extraction tool: a challenge for NLP
  - Difficulties in MWE identification
    → Flexible patterns, syntactic information
    → New identification algorithms
  - Consumption of computing resources
    → More efficient algorithms and routines

Future work

- Compare *mwetoolkit* with other tools
- Handling of nested MWEs
  - E.g., [inverse [kappa B [transcription factor]]]
- Improve the performance of candidate extraction