



# Language-neutral Term Extraction in the KYOTO-project

Wauter Bosma <[w.bosma@let.vu.nl](mailto:w.bosma@let.vu.nl)>

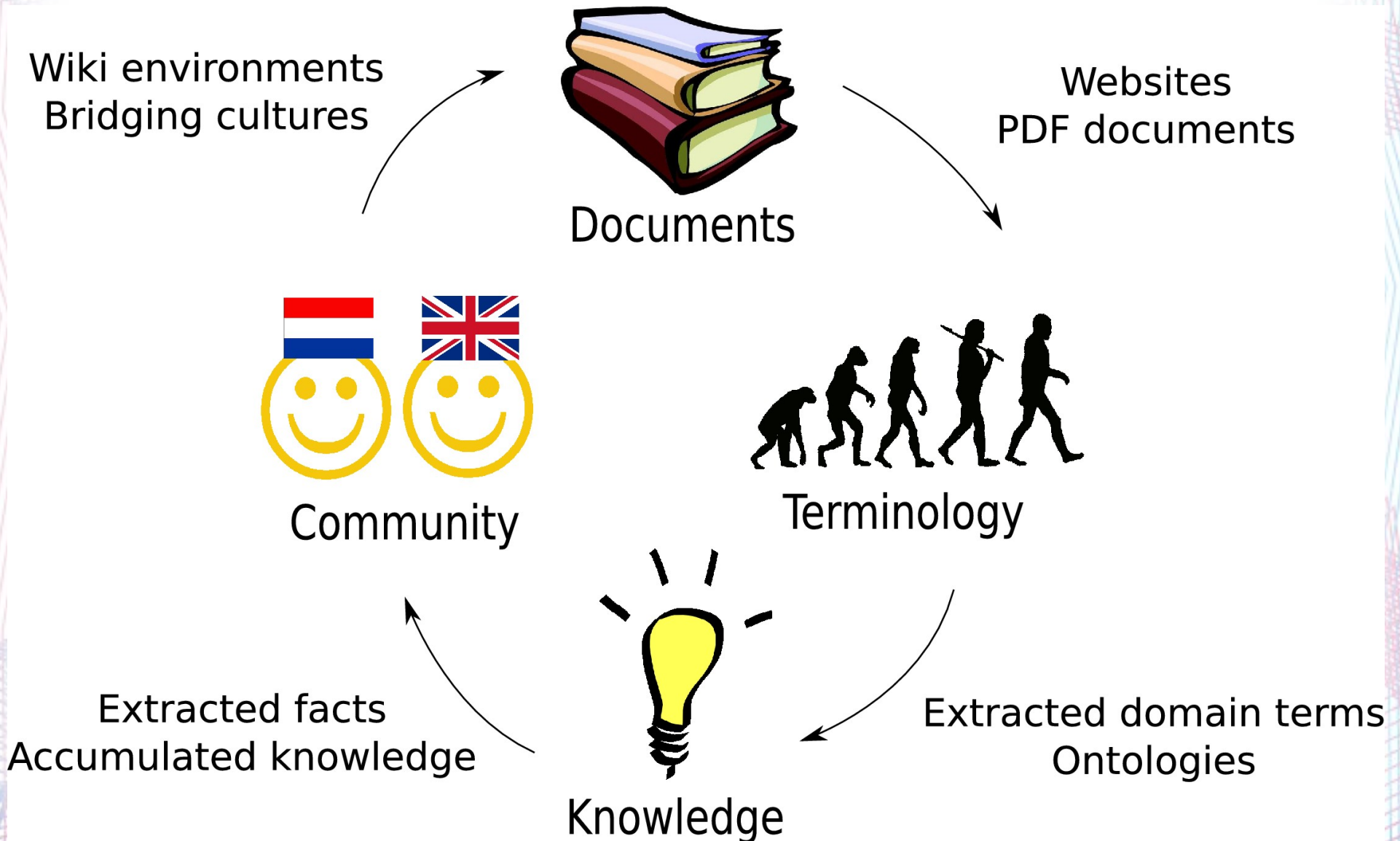
Piek Vossen <[p.vossen@let.vu.nl](mailto:p.vossen@let.vu.nl)>

Computational Lexicology and Terminology Lab  
Faculty of Arts  
VU University Amsterdam

Website: [www.kyoto-project.eu](http://www.kyoto-project.eu)



# The KYOTO Knowledge Cycle



# Semantics in Text

- Goal: domain modelling (*facts* & *concepts*)
- Example: *terrestrial species declined by 55%*
- Terms are **components** of facts:
  - Decline
  - 55%
  - Terrestrial species

# Term extraction at the VU University

- We learn concepts from text rather than just terms, e.g. *lime tree, linden tree, basswood* are different terms for the same concept.
- A concept is defined by:
  - The set of terms that can refer to the same concept
  - The relations with other concepts, e.g. *migratory locust is a locust is a tree*
- Concept-structures are large semantic networks representing the full relational structure underlying a collection of domain text

# Term extraction at the VU University

- We first extract all potential terms from text: maximizing recall;
- We then extract semantic relations between these terms to derive a coherent semantic model;
- Next, we select terms on the basis of relatedness and frequency

# Strategies of Automatic Term & Relation Extraction

- **Morpho-syntactic analysis** (e.g., *terrestrial species c species*);
- **Pattern-based analysis** (e.g., *amphibious species such as frogs*);
- **Distributional statistics** (terms used similarly are similar, e.g. *soup* and *sandwich* are the object of *eat*);
- **Concept merge** (synonymous terms are represented once);
- **Language alignment** by means of wordnet mappings;
- Our strategy: use a combination of the above for **extracting relations** and **ranking terms**.

# Step 1: Candidate Terms

- Nouns (or other **POS**) are candidate terms (e.g., *species*);
- The head of **compound** nouns are candidate terms (e.g. *landbouwbeleid*, *beleid*);
- Noun **phrases** are candidate terms (e.g., *vertebrate terrestrial species*);
- **Reduced** noun phrases are candidate terms. Modifiers are stripped one by one, towards the head:
  - vertebrate terrestrial species  $\rightsquigarrow$  terrestrial species  $\rightsquigarrow$  species
  - migration of species  $\rightsquigarrow$  migration of  $\rightsquigarrow$  migration

# Step 2:

## Morpho-syntactic Analysis

- A noun phrase is a hyponym of derived **reduced** noun phrases (e.g., *terrestrial species*  $\subset$  *species*);
- A **compound** is a hyponym of its head (e.g., *landbouwbeleid*  $\subset$  *beleid* – *agricultural policy*  $\subset$  *policy*).

# Step 3: Pattern-based Analysis

- Learning patterns from existing resources, eg. wordnets, species2000.
- Wordnet: hyponym(frog,amphibian)
- Corpus: ... *amphibians such as frogs* ...
- Pattern: X such as Y
- Corpus: ... *habitat for wading birds such as golden plover, lapwing and redshank;*

# Enumerations

- ... *golden plover, lapwing and redshank.*
- ... *limiting the use of fertilisers, manures and pesticides;*
- Share a syntactic function;
- Share a common hypernym or attribute;
- Usually disjoint (*LREC attracted over 1000 researchers and people*);

# Step 4: Distributional Statistics

WE SKIP THIS ONE

# Step 5: Term concept merge

- Wordnets, term database and other resources provide relations **within a language**;
- For each term we determine the possible meaning in wordnet, e.g. is **bird** an **animal** or a **woman**?
- Infer which terms are synonyms within the document collection:
  - *lime tree, linden tree, basswood*
  - If *endangered = threatened*  
then *endangered species = threatened species*
- Group terms according to the wordnet hierarchy:
  - *Tundra swan, mute swan, common swan* and *whistling swan* are *swans* which are *aquatic birds*

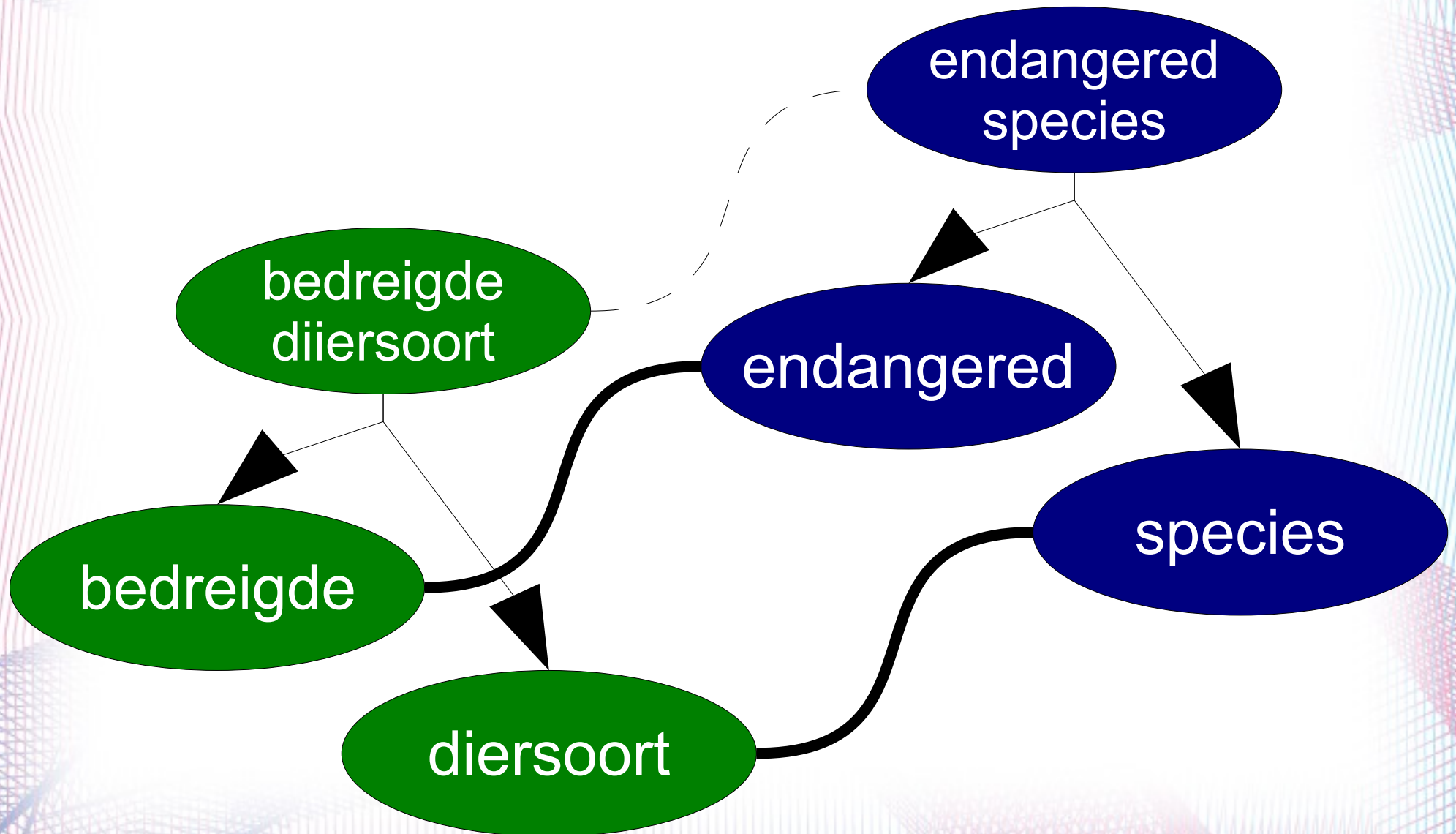
# Step 6: Ranking Terms

- Distinguish **domain-relevant** terms from non-terms;
- (As opposed to distinguishing domain terms from generic terms;)
- No clear boundary;
- A confidence value is assigned to each candidate term, representing its 'termness';
- The confidence value is calculated from the term relation **graph** and occurrence frequency;
- Candidate terms above a certain confidence **threshold** may be regarded terms.

# Step 7: Language Alignment

- Wordnet mappings provide relations **between languages**;
- Infer new relations **between languages**;

# Language Alignment: Example



# Advantages of concept approach

- Synonymous terms are represented as a single concept
- Opposition relations across semantically related concepts become clear
- Any semantic structure can be imposed on a set of target concepts with related terms
- Translation of concepts is easier than translating individual terms
- Term-concept structure can be used in Semantic web applications of the future
- Term-concept structures derived from different sources can be aligned more easily

THANK YOU FOR YOUR ATTENTION