The Conceptual Aspects of Terminographic Definitions: Towards Automatic Genus Tagging

Selja Seppälä

Terminology
Multilingual Information Processing Department

Modules M1-M2, Approches formelles et cognitives du langage
Outline

• Definitions
• Background and open questions
• Methodological approach
• Case study: towards automatic genus tagging
• Conclusions and perspectives
What is terminology?

• An **applied activity**: writing mono- or multilingual dictionaries for specialised domains (sciences, activities, practices, etc.)

• A **scientific discipline**: study of terminological phenomena on a linguistic or a conceptual level, or both
What is a terminographic definition?

- A linguistic representation of a concept of a specialised domain
  
  \textit{peptidyltransferase} = An enzyme located on the large ribosomal subunit that catalyzes peptide bond formation.

- A synthesis of knowledge rich contexts (Meyer, 2001)

- It reflects the structure of the concept

- Is subject to some formal restrictions
What is a genus?

• Relates defined & superordinate concepts
  \(\textit{peptidyltransferase} = \textit{An enzyme}...\)

• Different types:
  – Conceptual relation to defined concept
    • \textbf{IS\_A} (Rebeyrolle 2000)
    • \textbf{PART\_OF} (Iris, et al. 1988; L'Homme 2003)
    • \textbf{SET\_OF} (L'Homme 2003)
  – Conceptual category \(\textit{An enzyme} = \text{MATERIAL ENTITY}\)

• Different forms:
  – simple unit \(\textit{Protein}...\)
  – complex unit \(\textit{A chemical}..., \textit{The set of modifications}...\)

• In French, often the 1st word (65% of test corpus)
Open questions for work

• What makes a context more relevant than another to define a certain concept?
• Since the definition is rather limited in space, what are the characteristics of the concept to be included in the definition?
Hypotheses

• Property selection related to:
  – type of conceptual category
  – type of domain
  – language

• Can be studied through conceptual structure of definitions
Conceptual structure of a definition

peptidyltransferase
{An enzyme} [located on the large ribosomal subunit] [that catalyzes peptide bond formation.]

[PEPTIDYLTRANSFERASE] IS_A
{GEN MATERIAL_ENTITY} [SPE LOCATION]
[SPE FUNCTION]
Methodological approach

- **Corpus study** of terminographic definitions
  - complying to generally accepted definition writing rules
  - conceptually annotated
    - **conceptual category**
      \[(A \text{ peptidyltransferase} = \text{MATERIAL\_ENTITY})\]
    - **conceptual relations**
      - between the genus and the defined concept
        \[(\text{peptidyltransferase IS\_A enzyme})\]
      - between the specific and the genus
        \[(\text{that catalyzes peptide bond formation = FUNCTION})\]
- Requires **conceptual parsing**
Case study: Identification of the task

• Find the **genus element** in a definition (semi-)automatically
• Mark it with an XML tag including the corresponding **relation to the defined concept**

*peptidyltransferase*

<GEN relation_VE="IS_A">An enzyme</GEN>

located on the large ribosomal subunit that catalyzes peptide bond formation.
Genus extraction: State of the art

**Objectives:** information retrieval to build:
- lexical resources for NLP, ontologies, knowledge bases… (Alshawi 1987, Ide & Véronis 1993, Markowitz et al. 1986)
- terminological resources (Pozzi & Medina 2005, Rebeyrolle 2000)

**Text types:**
- lexicographic definitions (Barnbrook 2002, Ide & Véronis 1993, Markowitz et al. 1986)
- terminological definitions (L’Homme 2003, Pozzi & Medina 2005)

**Methods:**
- based on (boundary) markers (Barnbrook 2002, Rebeyrolle 2000)
- statistical methods (Pozzi & Medina 2005)
- use of external resources (POS tagger…) (Vossen et al. 1989)
Case study: How to find a genus?

By taking advantage of:

1. the **definition sublanguage** (Barnbrook 2002)
   - fixed phrases
   - present & past participles
   - characteristic morphosyntactic & lexical items

Search for its **boundaries** on the basis of:

- fixed phrases (part of, group of...)
- morphosyntactic patterns (present & past participles, relative pronouns...)

Finding **regularities** to form extraction patterns

- in literature (Iris et al. 1988; L’Homme 2003, Rebeyrolle 2000...)
- with a concordancer
  
  (GEN + end of following word;
  "Type of" + 1 to 3 words...)

Selja Seppälä  
Case study: How to find a genus?

By taking advantage of:
2. the **hierarchical structure** of the domain

**molecule** = A structural unit of matter consisting of one or more atoms

**nucleic acid** = A **molecule** formed by nucleotides located on one strand or two.

Search for **terms** of the domain (**genus proximus**) → Creating a **list of terms**
Case study: Implementation

- **Training corpus**: 500 definitions from different domains
- **Test corpus**: 92 definitions from the *terminology of protein biosynthesis in eukaryotic cells* (Bourjault, 2005)
- **Perl** program
- **Regular expressions**
Case study: The processing method

Four ordered steps:

1. Insert **opening tag** ⇒ Search for probable specific **elements preceding** the genus
   
   `<SPE>The first</SPE><GEN>phase of translation</GEN>that…`

2. Find GEN including **terms** from the domain

   `<GEN>An enzyme</GEN> located on…`
Case study: The processing method

3. Search for **closing boundary** markers:
   - specific rules: fixed phrases
     
     <GEN>The set of</GEN> rules</GEN> used for…
   - general rules: morphosyntactic & lexical markers
     
     <GEN>An enzyme</GEN> located for…
     <GEN>A protein</GEN> that acts…

4. Tag **1st word** of unmarked definitions
Case study: Performance evaluation

• Baseline: tag the 1st word (65%)
• Performance of the method: 78/92 (85%)
• Raised mainly by:
  – term search
  – fixed phrases
• Errors due to:
  – absence of fixed phrases indicating: 10/14
    • PART relation: 9/10 (*Branch of... [a science]*)
    • WHOLE relation: 1/10 (*An assembly of...*)
  – inexact extraction patterns: 4/14 (too greedy)
Case study: Main challenges

- Adequately refining the extraction patterns (avoid greedyness, find multiple genus…)
- Finding the best ordering of the rules
- Discriminating lexicalized compound words from free word sequences
  (A polypeptide chain… vs. A linear chain…)
- Collecting all forms of fixed phrases for each relation (can be domain specific)
Conclusion

• Manually created and classified rules perform well
• Advantage: does not require external NLP resources
• Method can be used in other languages
• Basis for automating:
  – rule definition
  – rule classification
Perspectives

• Adapt GEN processing method to SPEs
• Apply method to other languages
• Study conceptual regularities in definitions to find patterns according to:
  – concept type
  – domain
  – language
Thank you for your attention!
References


• Ide, N. et Véronis, J. (1993), "Extracting knowledge bases from machine-readable dictionaries : Have we wasted our time?" in *KB&KS'93 Workshop*, Tokyo

• L'Homme, M.-C. (2003), "Indices de relations conceptuelles dans les définitions terminologiques. Application au domaine de l'informatique", in, Bach C. et Martí J. Jornada Internacional sobre la Investigación en Terminología y Conocimiento Especializado, Barcelona, IULA.

• Markowitz, J., et al. (1986), "Semantically significant patterns in dictionary definitions", in Proceedings of the 24th conference on Association for Computational Linguistics, pp. 112-119.

