Terminology and Ontology

Introduction to terminology

Mathieu Constant

2021/2022

Two parts

- Terminology
  - **Lecturer**: Mathieu Constant (Mathieu.Constant@univ-lorraine.fr)
  - **Slot**: Wednesdays afternoon (6 $\times$ 2h)
  - **Period**: September - October 2021
  - **Additional date**: November 24, 2021
  - **Exam**: December 1st, 2021
- Ontologies
  - **Lecturer**: Mathieu D'Aquin
  - **Period**: December 2021 - February 2021
- Course documents on Arche (`https://arche.univ-lorraine.fr`)

Terminology and Ontologies

- **Terminology**: vocabulary specific to a scientific, technical or administrative domain ; each term (of the vocabulary) is associated to a concept of the domain
  e.g. *rotor*, *frequency band*, *morphology*, *supervised learning*

- **Ontology**: computational representation of the knowledge of a specific domain, relating concepts together in the form of a graph

Terminology and Ontologies

- **Terminology**: vocabulary specific to a scientific, technical or administrative domain ; each term (of the vocabulary) is associated to a concept of the domain
  e.g. *rotor*, *frequency band*, *morphology*, *supervised learning*

- **Ontology**: computational representation of the knowledge of a specific domain, relating concepts together in the form of a graph

→ **Terms populate ontologies**

Relation between the two parts

- Part 1 : Term processing, including automatic term acquisition
- Part 2 : Ontology building

Term acquisition is a prerequisite to ontology building

Terminology part program

1. Slot 1 : basic concepts of terminology and related NLP applications/tasks
2. Slot 2 : monolingual term acquisition
3. Slot 3 : multilingual term acquisition
4. Slot 4 to 6 : project

The courses will contain both theoretical and practical parts
Practicals : some preparation exercises for the project

Evaluation

- Project : 40% (Deadline : December 1st, 2021, noon)
  - teams of two people
  - **hard constraint**: "alternants" should be with "alternants" only
  - **preferably**: a "linguist" with a "computer scientist"
  - **subject**: implementing a term tagger for a specific domain
- Exam : 60 % (December 1st, 2021)
  - **no documents allowed**
  - course questions and exercises
  - questions on the project

This course : basic concepts of terminology and term processing

1. Introduction to terminology and term processing
2. Practical : text preprocessing

    - Definition of terminology and term
    - Terminology principles
    - Linguistic properties of a term
    - Some terminological resources
    - Term processing
    - Related linguistic notions and NLP tasks

Main sources for this course

- Béatrice Daille, Benoît Habert, Christian Jacquemin, Jean Royauté. Empirical Observation of Term Variations and Principles for their Description. *Terminology*, 1996, 3 (2), pp.197–257.
- Christian Jacquemin and Didier Bourigault. Term Extraction and Automatic Indexing. *The Oxford Handbook of Computational Linguistics*. Edited by Ruslan Mitkov. 2005
- unknown author, slides "Terminology : theory and practice", `http://er.nau.edu.ua:8080/bitstream/NAU/10325/14/Lecture%2014.pdf`
- D. Cram and B. Daille. TermSuite : Terminology Extraction with Term Variant Detection. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics — System Demonstrations, pages 13–18, Berlin, Germany, August 7-12, 2016.

A problem of communication in a specific domain

A different but familiar language : a sub-language

- When experts of a specific field start talking shop, it sounds they are speaking a different language
- Use of a shared specific vocabulary
- Built upon general language : use of general language words and grammar

Term examples

- *rubric, lesson plan, pop quizz, term paper, student engagement* (education)
- *blood work, CVC, scalpel, set* (medicine)
- *forthwith, heretofore, the part of the first part* (law)

(www.vocabulary.com)

Abstract from *Mechanics Research Communications*

(Gu and Chen 2018)

A failure criterion for isotropic materials is developed in the present study. The fundamental hypotheses of the Mohr-Coulomb criterion are re-examined and re-evaluated. The failure function of Mohr's envelope is firstly expanded into a polynomial in terms of the stress components on the failure plane, then truncated at the second order due to the fine curve-fitting results of quadratic functions with experimental data. The parameters of the nonlinear failure functions are calibrated by the three basic mechanical properties – the uniaxial tensile and compressive strengths, T and C, as well as the pure shear strength, S. Theoretical and experimental evaluation for various isotropic materials demonstrates that the present failure criterion provides no worse results for ductile materials and far better results for brittle materials compared with the linear Mohr-Coulomb criterion. The strength theory shows good agreement with physical reality and has a wide range of applicability.

Abstract from *Mechanics Research Communications*

(Gu and Chen 2018)

A failure criterion for isotropic materials is developed in the present study. The fundamental hypotheses of the Mohr-Coulomb criterion are re-examined and re-evaluated. The failure function of Mohr's envelope is firstly expanded into a polynomial in terms of the stress components on the failure plane, then truncated at the second order due to the fine curve-fitting results of quadratic functions with experimental data. The parameters of the nonlinear failure functions are calibrated by the three basic mechanical properties – the uniaxial tensile and compressive strengths, T and C, as well as the pure shear strength, S. Theoretical and experimental evaluation for various isotropic materials demonstrates that the present failure criterion provides no worse results for ductile materials and far better results for brittle materials compared with the linear Mohr-Coulomb criterion. The strength theory shows good agreement with physical reality and has a wide range of applicability.

Definitions of *terminology* in general dictionaries

Oxford dictionary

The body of terms used with a particular technical application in a subject of study, profession, etc. ‘the terminology of semiotics’

vocabulary.com

Terminology is vocabulary associated with a certain field of study, profession, or activity. Knowing the terminology is an important part of being able to work in a given profession. A system of words used to name things in a particular discipline

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

http ://term-portal.de/en/introduction-to-terminology.html

Terminology is defined as the entirety of all concepts and terms in one specialist field. Therefore, one can equate terminology with specialist vocabulary. Efficient communication with regard to technical language is not possible without the correct use of specialist vocabulary. Terminology work deals with the preparing, the processing, the documenting and the use of specialist vocabulary... Terminology work is concept-oriented and thus, from a methodical point of view, specially suited to solve multilingual communication tasks. Terminology is, as a general rule, very time consuming and cost intensive. For that reason it is important to work with reliable methods, procedures and tools.

More precisions

Wikipedia

Terminology is the study of terms and their use. ... Terminology is a discipline that systematically studies the "labeling or designating of concepts" particular to one or more subject fields or domains of human activity. The discipline of terminology consists primarily of the following aspects :

- analyzing the concepts and concept structures used in a field or domain of activity
- identifying the terms assigned to the concepts
- in the case of bilingual or multilingual terminology, establishing correspondences between terms in the various languages
- compiling the terminology, on paper or in databases
- managing terminology databases
- creating new terms, as required.

More precisions

Wikipedia

Terminology is the study of terms and their use. ... Terminology is a discipline that systematically studies the "labeling or designating of concepts" particular to one or more subject fields or domains of human activity. The discipline of terminology consists primarily of the following aspects :

- analyzing the concepts and concept structures used in a field or domain of activity
- identifying the terms assigned to the concepts
- in the case of bilingual or multilingual terminology, establishing correspondences between terms in the various languages
- compiling the terminology, on paper or in databases
- managing terminology databases
- creating new terms, as required.

General dictionary definitions of *term*

Cambridge dictionary

A word or expression used in relation to a particular subject, often to describe something official or technical : "Without let or hindrance" is a legal term that means "freely".

Merriam-Webster

A word or expression that has a precise meaning in some uses or is peculiar to a science, art, profession, or subject legal terms.

Definition of a term by terminology specialists

Standard ISO 1087

A term is the designation of a defined concept in a special language by a linguistic expression. A term may consist of one or more words.

Ahrenberg (2009)

In classical terminology a term is defined as the expression (or label, or representation) of a concept.

Roche et al. (2009)

A term is a "specialised linguistic unit" which denotes a concept of the domain called the meaning of the term

Definition of terminology and term by NLP researchers

Cram and Daille (2016)

- A terminology is a coherent set of terms that constitutes the vocabulary of a domain.
- It also reflects the conceptual system of that domain.
- A term could be :
  - a single word term (*rotor*)
  - a complex term like compounds (*broadband*), or multi-word terms (*frequency band*)
- Terms are lexical items used in discourse and are subjected to linguistic variations such as modification or coordination.

Two classical defining properties of terms

1. **Unithood**: A linguistic unit with morphological, syntactic and semantic properties as well as variations
2. **Termhood**: A unit specific to a domain (vs. a unit of the general language)

Abstract from *Mechanics Research Communications*

(Gu and Chen 2018)

A failure criterion for isotropic materials is developed in the present study. The fundamental hypotheses of the Mohr-Coulomb criterion are re-examined and re-evaluated. The failure function of Mohr's envelope is firstly expanded into a polynomial in terms of the stress components on the failure plane, then truncated at the second order due to the fine curve-fitting results of quadratic functions with experimental data. The parameters of the nonlinear failure functions are calibrated by the three basic mechanical properties – the uniaxial tensile and compressive strengths, T and C, as well as the pure shear strength, S. Theoretical and experimental evaluation for various isotropic materials demonstrates that the present failure criterion provides no worse results for ductile materials and far better results for brittle materials compared with the linear Mohr-Coulomb criterion. The strength theory shows good agreement with physical reality and has a wide range of applicability.

Examples

Legal terms (http ://www.uscourts.gov/glossary)

Arraignment

A proceeding in which a criminal defendant is brought into court, told of the charges in an indictment or information, and asked to plead guilty or not guilty.

Indictment

The formal charge issued by a grand jury stating that there is enough evidence that the defendant committed the crime to justify having a trial ; it is used primarily for felonies.

Inculpatory evidence

Evidence indicating that a defendant did commit the crime.

Examples

Agriculture terms (`https://agclass.nal.usda.gov`)

Acidulants

An additive added to a product to increase the acidity of the product. Acidulants are often added to food products to increase tartness.

Annual pastures

A pasture consisting of introduced forage species planted for only one year or season's duration.

Pastures

An area devoted to the production of forage (introduced or native) and harvested by grazing.

An interdisciplinary discipline

- "An interdisciplinary but autonomous subject at the service of scientific and technical discipline" (source : unknown author, slides "Terminology : theory and practice", `http://er.nau.edu.ua:8080/bitstream/NAU/10325/14/Lecture%2014.pdf`)
- It combines
  - **philosophy**: logical organization of concepts
  - **linguistics**: terms have linguistic properties
  - **subject field**: need for expertise in the subject field
  - **computer science**: need for computational representation of terminologies

Motivations for working on terminology

In the subject field

Facilitating communication among specialists, thanks to the standardization of the naming of the concepts of the field

Information systems

Indexing of technical documents for better search in collections of technical documents

Computer-aided tasks

- computer-aided translations
- computer-aided writing of documents (with controlled terminology)

Historical general principles of terminology

Vienna School based on the work of E. Wüster

- Work starts from a concept and its goal is to clearly delimit the concepts from each other
- Concepts are clear-cut ; they should not be studied in isolation but as elements in a concept system through an analysis of concept characteristics.
- In order to be able to place concepts within a concept system, definitions must be intensional (define its characteristics).
- Each concept should be designated by only one term and each term should refer to only one concept (univocity principle) ; synonymy and polysemy are excluded.
- Since emphasis is on the concept system, which is the basis of special language, terminology does not study language development/evolution, therefore it takes a synchronic perspective (analysis of terms from a specific point in time).

(source : InMyOwnTerms, by Patricia Brenes, Terminology for Beginners and Beyond)

Some criticisms and evolutions from 2000

A specialized language is a natural language

- Designation of concepts may evolve by usage : possible use of variations/synonymy
- Priority of concepts over designation is not so clear as they are built in discourse

Concept evolution

Never-ending evolution of concept system as the field is evolving

Multilinguality

- Necessity to take multilinguality into account
- ex. countries and international institutions with several official languages : Canada, European Union, United Nation Organization

Corpus-based terminology : a practical approach

(Bourigault and Jacquemin 2003)

General idea

- A term is the output of a procedure of terminological analysis deciding whether a phrase is a term
- Decision may involve a group of researchers or practioners, normalization institutions, single engineer or terminologist
- A terminological resource should be dualy relevant with respect to the corpus and the target application (cf. no unique conceptual representation of a given area)

Computational approach

- if few and short documents, manual discovery of terminology
- In reality, many and long documents $\rightarrow$ need for computational tools to help acquire terminology

Ontoterminology

(Roche et al. 2009)

- "Saying is not modelling" (Roche 2007)
- Two distinct systems :
  - **linguistic system**: terms, term definition in natural language
    $\rightarrow$ lexical structure
  - **conceptual system**: definitions of concepts in a formal (not natural) language
    $\rightarrow$ (not task-related) ontology

Ontoterminology

(Roche et al. 2009)

- "Saying is not modelling" (Roche 2007)
- Two distinct systems :
  - **linguistic system**: terms, term definition in natural language
    $\rightarrow$ lexical structure
  - **conceptual system**: definitions of concepts in a formal (not natural) language
    $\rightarrow$ (not task-related) ontology

$\rightarrow$ In this part of the course, we are only interested in the linguistic system.

Examples of resources

Databases and software

- http://www.cilf.fr/unepage-terminologie-terminologie-zoom-25125-0-1.html
- https://termcoord.eu
- http://www.termsciences.fr/termsciences/?lang=en
- http://iate.europa.eu
- https://unterm.un.org

LINGUISTIC PROPERTIES OF TERMS

Two classical defining properties of terms

1. **Unithood**: A linguistic unit with morphological, syntactic and semantic properties as well as variations
2. **Termhood**: A unit specific to a domain (vs. a unit of the general language)

Some properties of terms

Unithood

- A term is a lexical unit with some morphological, syntactic, semantic properties
- They are made of one or several word
- Most terms are complex

Forms of a term

- Simple term : *cell*
- Complex terms :
  - multiword terms : *blood cell*
  - closed compounds : *Kühleinrichtung (cooling device)* $\rightarrow$ *kühl (cool)* + *einrichtung (device)* [German]

Different patterns for complex terms

Mainly nominal patterns

- phrases having a noun as head (ex. *intermediate frequency*, *information retrieval*, *order of magnitude*)
- many term recognition systems only consider nominal patterns

Other possible patterns

- verbal expressions

  Sitting in. The art of using the peloton or bunch to conserve your energy. Sitting in the middle of the bunch means that you are protected, out of the wind and using less energy than everyone else. *She sat in all day and sprinted for the win.*

- adjectival/adverbial expressions

  Off the back. Similar term to that of getting dropped. When a rider loses touch with the peloton and is effectively off the back.

(Source : https ://cyclingtips.com/2015/04/how-to-talk-like-a-cyclist/)

Variations vs. fixedness

- In principle, standardization of terms (no variations, but morphological), e.g. *intermediate frequency*, *intermediate frequencies*
- Some people say that terms are the idioms of the specialized domains as idioms are mostly fixed
- But... in reality, many term variants (cf. Daille et al. 1996)

Examples of variants from a medicine corpus I

Daille et al. (1996)

Insertion variation

- epithelial tumor cells
- epithelial tumour cells
- epithelial cancer cell
- epithelial respiratory cell
- epithelial cultured cell

Coordination variation

- epithelial and carcinoma cells
- epithelial and myoepithelial cells
- epithelial and glial cells

Examples of variants from a medicine corpus II

Daille et al. (1996)

Use of pivot elements

cells **expressed** epithlial

Morpho-derivational variation (*epithlial* → *epythelium*)

- epithelium specific cell
- cell infiltrate **of** the epithelium
- cell **associated with** breast epithelium
- cell carcinoma **contained** dysplastic epythelium
- cells **were** identical tononpigmentaed ciliary epithelium

Variation types I

Daille et al. (1996)

Graphical and orthographic variations

- Capitalized form : *domestic service* → *Domestic service*
- Use of hyphens : *paquet mode* → *paquet-mode*
- Orthographic variation : tumour cell (British) → tumor variation (US)
- Abbreviations : *amplitude modulation* → *AM*

Inflectional variations

- Plural form : *vitamin deficiency* → *vitamin deficiencies*
- geruntive form acting like a noun : *acoustic test* → *acoustic testing*

Variation types II

Daille et al. (1996)

Syntactic variations

- insertion and juxtaposition variations : two dimensional analysis $\rightarrow$ two dimensional tryptic peptide analysis
- coordination variations : systolic pressure $\rightarrow$ systolic and diastolic blood pressure
- permutation variations around a pivot element : chromosome association $\rightarrow$ association with chromosome

Variation types III

Daille et al. (1996)

Morpho-syntactic variations

- Adjective nominalisation : enzymatic activity $\rightarrow$ enzyme activities, abnormal chromosome $\rightarrow$ abnormalities of chromosome
- Verb nominalisation : tumor promotion $\rightarrow$ promotes degradation of the cellular tumor

Term processing

(Bourigault and Jacquemin 2003)

	Prior terminological data	No prior terminological data
Term discovery	*Term enrichment*	*Term acquisition*
Term recognition	*Controlled indexing*	*free indexing*

- monolingual
- multilingual

Related concepts in NLP

In the general language, two concepts related to terms

- Named entities : real-world object that can be denoted with a proper name
  e.g. Donald Trump, Paris, Los Angeles, Apple, Volkswagen Golf
- Multiword expressions : idiosyncratic combinations of words, e.g. in spite of, light house, burn the midnight oil, above board, piece of cake

Why ?

- They form lexical units
- they are conventional expressions
- with limited variations

Some properties of multiword expressions

- They can be discontiguous : e.g. *John pulled Mary's leg*
- Tendency to some fixedness, but they may vary :
  - in their morphological forms (*hot dog*, *hot dogs*)
  - in their lexical elements (lose one's mind/head)
  - in their syntactic structure (he took a step, the step he took)

Bibliography

1. Béatrice Daille, Benoît Habert, Christian Jacquemin, Jean Royauté. Empirical Observation of Term Variations and Principles for their Description. Terminology, 1996, 3 (2), pp.197–257.
2. Christian Jacquemin and Didier Bourigault. Term Extraction and Automatic Indexing. *The Oxford Handbook of Computational Linguistics*. Edited by Ruslan Mitkov. 2005
3. D. Cram and B. Daille. TermSuite : Terminology Extraction with Term Variant Detection. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics — System Demonstrations, pages 13–18, Berlin, Germany, August 7-12, 2016.
4. Lars Ahrenberg. Term extraction : a review. On-line draft version. 2009
5. C.Roche. Saying is not modelling. NLPCS 2007, Natural Language Processing and Cognitive Science, Funchal (Portugal). 2007, pp. 47-56
6. C. Roche, M. Calberg-Challot, L. Damas, P. Rouard. Ontoterminology : A new paradigm for terminology. KEOD 2009 (International Conference on Knowledge Engineering and Ontology Development). 2009
7. Jiefei Gu and Puhui Chen. A failure criterion for isotropic materials based on Mohr's failure plane theory. Mechanics Research Communications. Volume 87, January 2018, Pages 1-6