Semantic Technologies in Computational Terminology Management

Jody Foo, jody.foo@liu.se

Department of Computer and Information Science (IDA)
Linköping university, SWEDEN

LiU expanding reality

Overview

Computational what?

Automatic Term Extraction

Future research

Terminology

RIKSTERMBANKEN
Alla termer på ett ställe

About Rikstermbanken | Contact | svenska

- Front page
- Simple search
- Advanced search
- Sources
- Latest news (Swedish)
- Q & A (Swedish)
- Help (Swedish)

Simple search > Terminological entry

Terminological entry 4 of 6 << Previous Next >>

SWEDISH TERM (PREFERRED):	**motstånd (2)** *-et -0*
DEFINITION:	komponent som har hög resistans (2) och som har till uppgift att minska den elektriska strömmen i en strömkrets
CLASSIFICATION:	Pc [SAB]
SOURCE:	Terminologicentrum TNC: Tekniska basord | 1995 [Ordlistan är under revidering.]

Comments on this entry?

Back to Search Results | New search << Previous Next >>

TNC TERMINOLOGICENTRUM

Rikstermbanken is developed and managed by **Terminologicentrum TNC** (The Swedish Centre for Terminology).

Latest update: 19 december
System version 0.97.41J © Terminologicentrum TNC

LiU expanding reality

Terminology

“

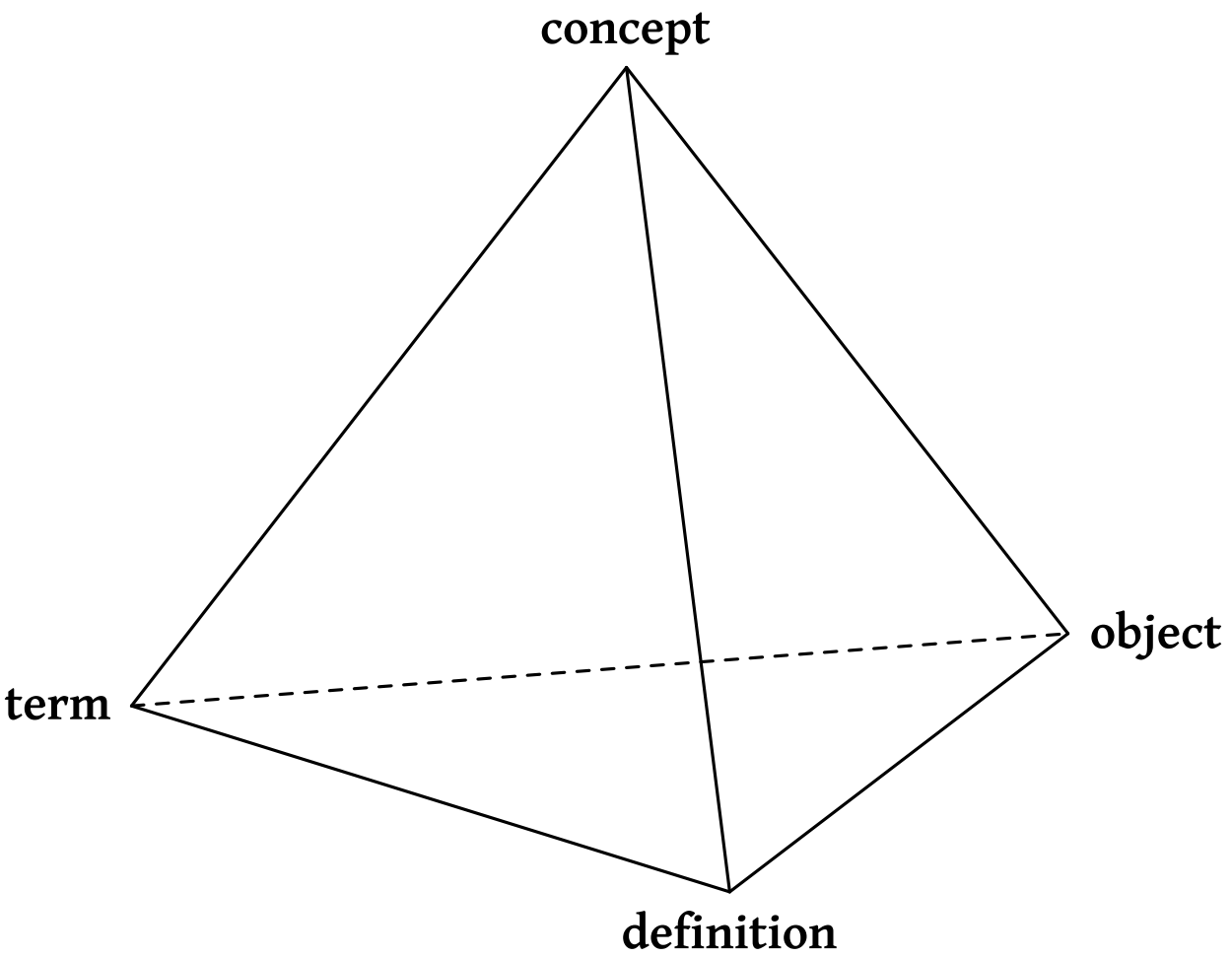
***Terminology Science**, interdisciplinary field of knowledge dealing with concepts and their representations*

*an **aggregate of terms which represent the system of concepts** of an individual subject field*

***a publication** in which the system of concepts of a subject field is represented by terms*

Felber, 1984, p. 1

”



coniferous tree

tree with needle-formed leaves and exposed or naked seeds

fir

coniferous tree of the genus abies

firewood

wood for fuel

light-demanding tree

tree preferring sunny habitat

noble gas

1) helium, neon, argon, krypton, xenon or radon *2)* gas that in the natural state is chemically inactive

tolerant tree

tree preferring shade

tree

tall plant with hard self-supporting trunk and branches that lives for many years

wood

hard fibrous material obtained from the stems and branches of trees or shrubs

wood bin

box for holding firewood

coniferous tree

tree with needle-formed leaves and exposed or naked seeds

fir

coniferous tree of the genus abies

firewood

wood for fuel

light-demanding tree

tree preferring sunny habitat

noble gas

1) helium, neon, argon, krypton, xenon or radon *2)* gas that in the natural state is chemically inactive

tolerant tree

tree preferring shade

tree

tall plant with hard self-supporting trunk and branches that lives for many years

wood

hard fibrous material obtained from the stems and branches of trees or shrubs

wood bin

box for holding firewood

5.5. kaj

Definition: till strandlinjen anslutande konstruktion för förtöjning av fartyg

Engelsk term: wharf, quay

Anmärkning: Konstruktionen har en eller flera lodräta sidor mot djupt vatten och en vågrät ovansida för godshantering.

Källa: Specifikation – Nationell Strandlinje, version 1.5, Sjöfartsverket och Lantmäteriet | 2008

5.6. hamnbassäng

Definition: av kajer, pirar eller vågbrytare helt eller delvis innesluten del av hamns vattenområde

Engelsk term: basin

Anmärkning: Hamnbassängen är avsedd för manövrering och vändning av fartyg.

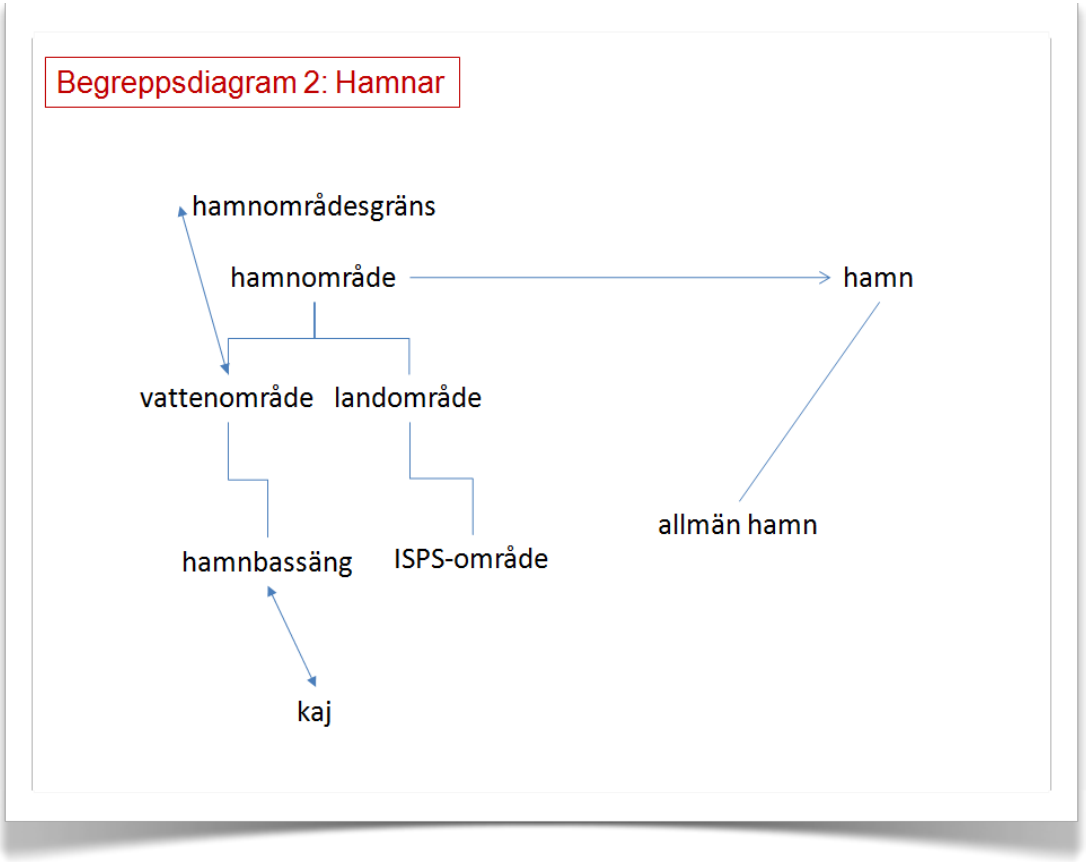
Källa: Baserad på Terminologicentrum TNC, SIS - Allmänna standardiseringsgruppen (SIS-STG): Transportordlista | 1992

5.7. allmän hamn

hamn som inrättats som allmän hamn och som finns förtecknad i Sjöfartsverkets kungörelse med tillkännagivande om beslut om allmänna farleder och allmänna hamnar (SJÖFS 1988:5)

Anmärkning: En allmän hamn får inrättas om hamnen är av väsentlig betydelse för den allmänna samfärdseln. En allmän hamn är öppen för allmän sjötrafik, vilket innebär att alla fartyg i princip har rätt att i mån av plats anlöpa hamnen och utnyttja dess resurser.

Källa: Baserad på Lag (1983:293) om inrättande, utvidgning och avlysning av allmän farled och allmän hamn samt Sjöfartsverkets kungörelse med tillkännagivande om beslut om allmänna farleder och allmänna hamnar (SJÖFS 1988:5).



Ordlista och begreppsdiagram - Farleder och hamnar (2011)

ER-diagram

Finite State Automaton

UML

Class diagram

Semantic Web

Flow chart

Ontology

Network topology

Computational Terminology Management

Terminology Creation

Terminology Maintenance

Terminology Use

Automatic Term Extraction

Monolingual ATE, Bilingual ATE, Evaluation of ATE

LiU expanding reality

These are not the terms we are looking for

Information Retrieval (Manning, 2008)

a ***term*** is a ***type*** included in the IR system's dictionary

a ***type*** is the ***class*** *of* all ***tokens*** containing the same character sequence

a ***token*** is an instance of a ***sequence of characters*** in some particular document that are grouped together as a useful semantic unit for processing

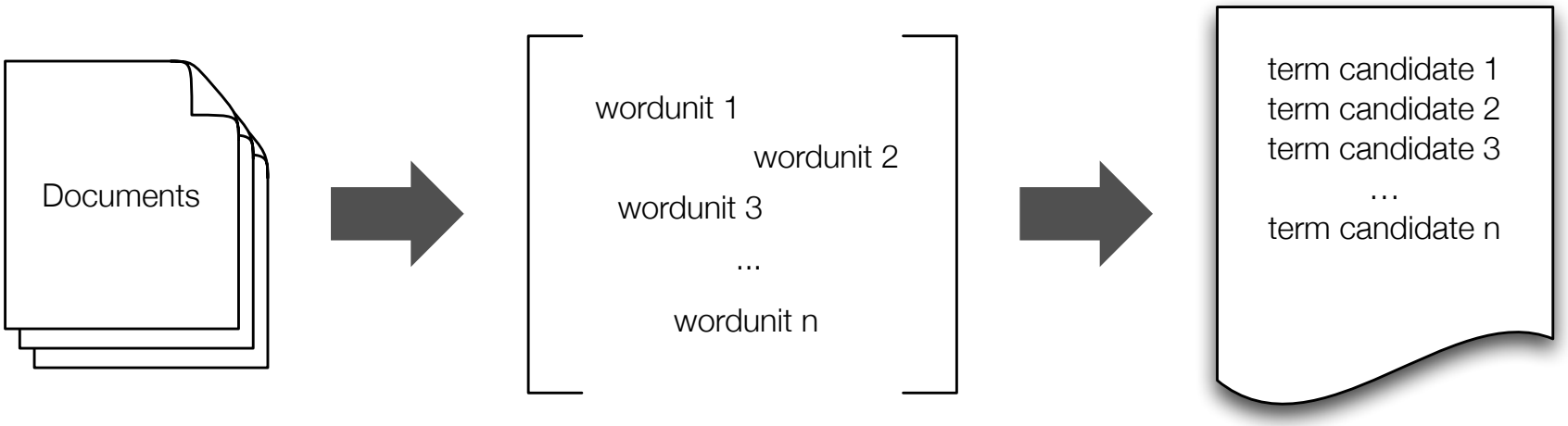
These are the terms we are looking for

Termhood: "*The degree to which a stable lexical unit is related to some domain-specific concepts*"

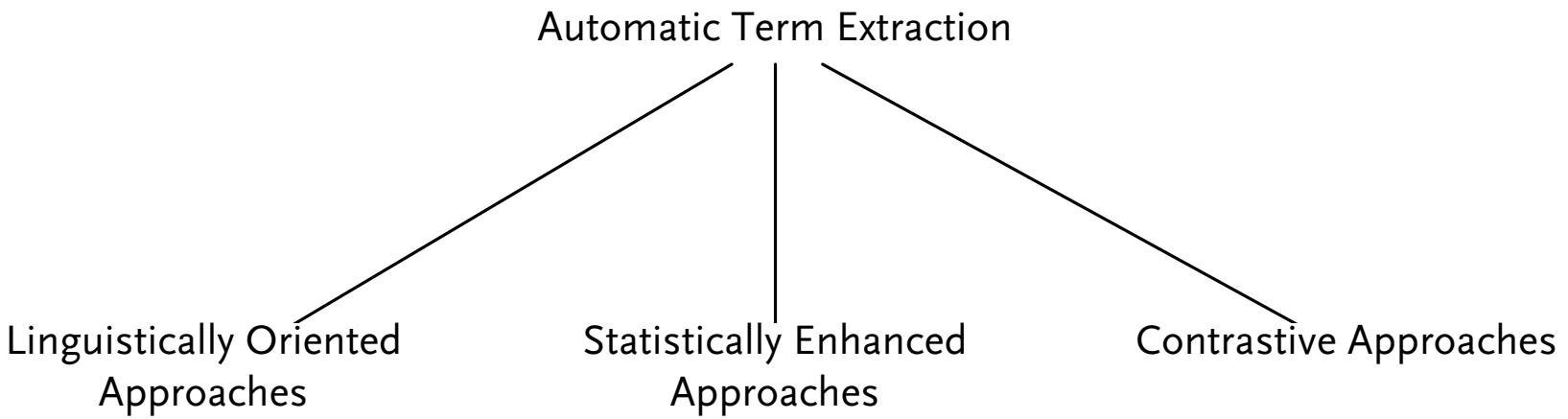
Unithood: "*the degree of strength or stability of syntagmatic combinations and collocations*"

Kageura, 1998

General monolingual ATE process



Automatic Term Extraction (ATE)



Linguistic Approaches

Morphological patterns (Ananiadou 1994)

morphological analysis to identify neoclassical morphological elements (latin, greek) in the medical domain

Part-of-speech based patterns

Daille 1994 (fr): e.g. `N ADJ, N`$_1$ `de N`$_2$

Justeson & Katz 1995 (en): `((Adj|Noun)+|(Adj|Noun)*`

Statistically enhanced approaches

Linguistically oriented approaches + statistical filtering

Association measures

- Log-Likelihood (Daille 1994)
- Mutual Information (Daille 1994)
- Dice Factor (Smadja, 1996)

Specialized measures

C-Value/NC-Value

Frantzi, Ananiadou, Tsujii 1998

Multi-word term extraction method

POS patterns + stop lists

Nested terms: real time {real time clock, real time data ...}

Statistical filtering

C-Value/NC-Value

Frantzi, Ananiadou, Tsujii 1998

$$C-value(a) = \begin{cases} \log_2 |a| \cdot f(a) & \text{if } a \text{ is not nested} \\ \log_2 |a| \cdot \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & \text{otherwise} \end{cases} \quad (4.4)$$

a = 'real time'
f(a) = 300
T_a = {real time clock, real time output}
b_1 = real time clock
$P(T_a) = |T_a| = 2$
$f(T_a) = 500$
C(a) = log(2) * (300 - 250)

C-Value/NC-Value

Frantzi, Ananiadou, Tsujii 1998

Contextual clues

$$Weight(w) = \frac{t(w)}{n}$$

w = adj, n or v
t(w) = number of terms that w appears with
n = total number of terms

C-Value/NC-Value

Frantzi, Ananiadou, Tsujii 1998

NC-Value

$$NC\text{-}value(a) = 0.8 \cdot C\text{-}value(a) + 0.2 \cdot \sum_{b \in C_a} f_a(b) weight(b)$$

Contrastive approaches

Possible data sources for statistical calculations

- internal corpus
- external corpus, same domain
- external corpus, “opposite domain”
- external corpus, “closely related domains”
- external corpus, “other specialized domains”
- external corpus, general language

Weirdness

Ahmad, Gillam, & Tostevin, 1999

$$Weirdness = \frac{w_s/t_s}{w_g/t_g}$$

Contrastive Weight

Basili, Moschitti, and Pazienza (2001)

Contrastive Weight

$$F_t = \sum_j f_t^j$$

$$IWF(t) = log(\frac{N}{F_t})$$

$$w_t^i = log(f_t^i) * IWF(t)$$

TermExtractor

Sclano and Velardi (2007)

Domain Relevance

Domain Consensus

Lexical Cohesion

Domain relevance

$$DR_{D_i}(t) = \frac{\hat{P}(t/D_i)}{max\left(\hat{P}\left(t/D_j\right)\right)} = \frac{freq(t, D_i)}{max\left(freq\left(t, D_j\right)\right)}$$

Domain Consensus

$$DC_{D_i}(t) = -\sum_{d_k \in D_i} \hat{P}\left(t/d_k\right) log\left(\hat{P}\left(t/d_k\right)\right)$$

$$= -\sum_{d_k \in D_i} f_n\left(t, d_k\right) log\left(f_n\left(t, d_k\right)\right)$$

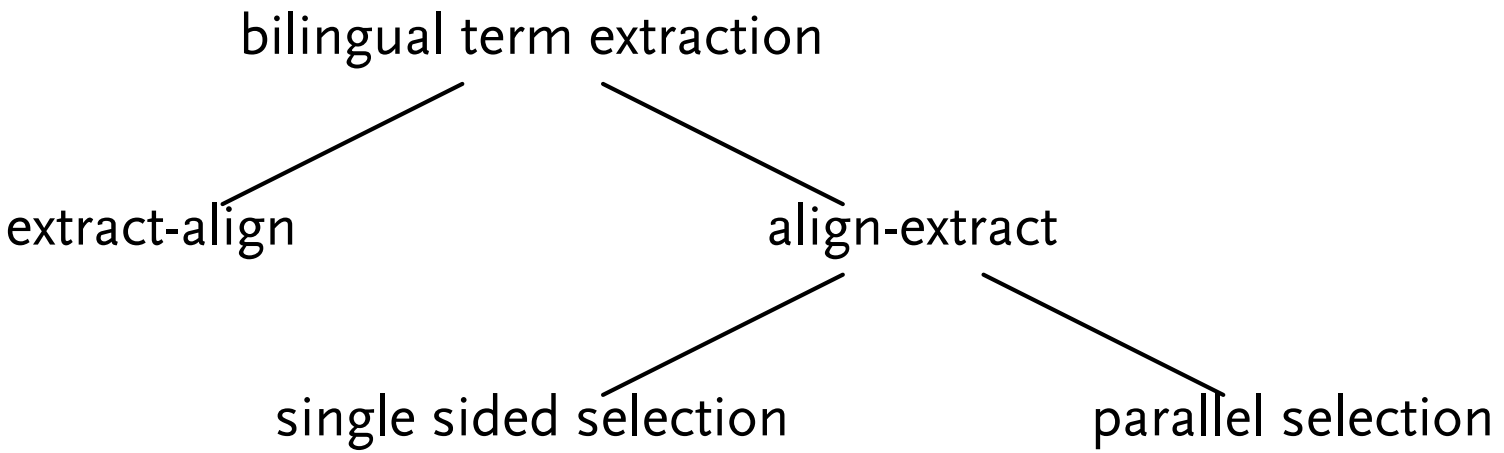
Lexical Cohesion

$$LC_{D_i}(t) = \frac{n \cdot freq\,(t, D_i) \cdot log\,(freq\,(t, D_i))}{\sum_j freq\,(w_j, D_i)}$$

Combine all

$$w(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC$$

Bilingual ATE



Synergy between monolingual and bilingual ATE

Extract-Align

same approach as monolingual ATE

Align-Extract

using monolingual ATE to extract a second language

Future research

Ontology Engineering meets CTM meets the Semantic Web?

Creation

Maintenance

Use

Ontology Engineering meets CTM meets the Semantic Web?

Creation

Maintenance

Use

Intelligent Tools for Terminology Maintenence

Semi-automatic environments based on existing ontologies, terminologies and corpus data.

Supported by visual representations that make relationships

FCA - Formal Concept Analysis

Demo!