Anchoring Dutch Cultural Heritage Thesauri to WordNet: two case studies

LATECH 2007, ACL workshop

Veronique Malais, Antoine Isaac

Luit Gazendam, Hennie Brugman

Presented by:
Ehsan Khoddammohammadi

Roadmap

1. **Terminology**

2. Alignment and Anchoring task

3. Discussion about the proposed methodology in this paper

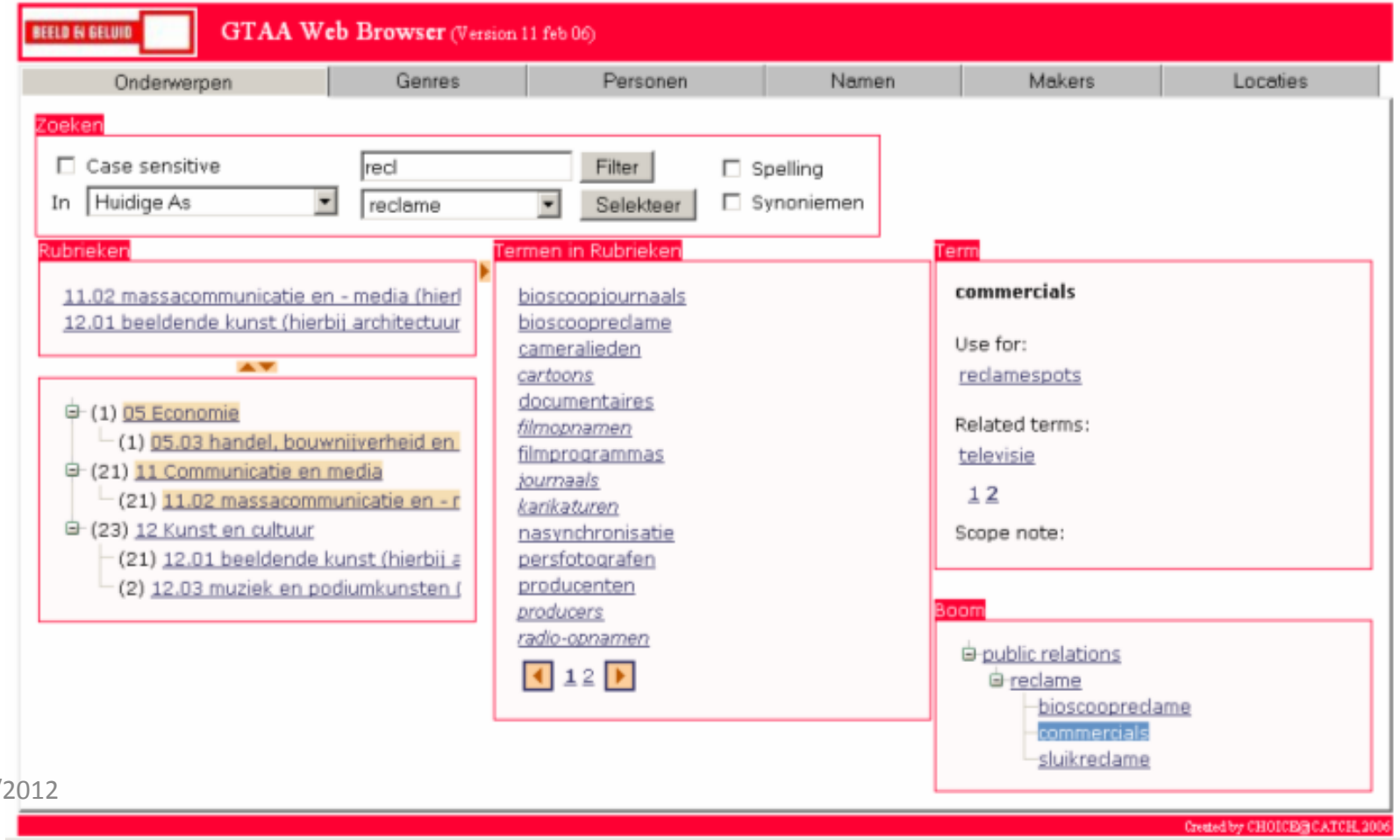
Thesauri

- Definition
- Terminology:
  - Broader Term
  - Narrower term
  - Related Term
  - Preferred Terms Vs. Non-Preferred terms
  - Scope notes
- Benefits for Cultural Heritage data

Gemeenschappelijke Thesaurus Audiovisuele Archieven

Common Thesaurus Audiovisual Archives

- Contains 5800 subjects in hierarchical format.



Goo Trefwoorden Thesaurus

- 3500 concepts
- 50000 terms
- Used in Dutch National Library for indexing and retrieving books.

WordNet

- Lexical database
- Definition of *synset*
- Provides **semantic relation**
- Distinguishes between **Noun**, **Verb**, **Adverb** and **Adjective**
- ~ 117000 sysnsets

Word to search for: love Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Display options for word: word#sense number

Noun

- S: (n) **love#1** (a strong positive emotion of regard and affection) *"his love for his work"; "children need a lot of love"*
- S: (n) **love#2**, passion#6 (any object of warm affection or devotion) *"the theater was her first love"; "he has a passion for cock fighting"*
- S: (n) beloved#1, dear#1, dearest#1, honey#2, **love#3** (a beloved person; used as terms of endearment)
- S: (n) **love#4**, sexual love#1, erotic love#1 (a deep feeling of sexual desire and attraction) *"their love left them indifferent to their surroundings"; "she was his first love"*
- S: (n) **love#5** (a score of zero in tennis or squash) *"it was 40 love"*
- S: (n) sexual love#2, lovemaking#1, making love#1, **love#6**, love life#1 (sexual activities (often including sexual intercourse) between two people) *"his lovemaking disgusted her"; "he hadn't had any love in months"; "he has a very complicated love life"*

Verb

- S: (v) **love#1** (have a great affection or liking for) *"I love French food"; "She loves her boss and works hard for him"*
- S: (v) **love#2**, enjoy#3 (get pleasure from) *"I love cooking"*
- S: (v) **love#3** (be enamored or in love with) *"She loves her husband deeply"*
- S: (v) sleep together#1, roll in the hay#1, **love#4**, make out#6, make love#1, sleep with#1, get laid#1, have sex#1, know#8, do it#1, be intimate#1, have intercourse#1, have it away#1, have it off#1, screw#1, fuck#1, jazz#2, eff#1, hump#2, lie with#1, bed#4, have a go at it#1, bang#5, get it on#1, bonk#1 (have sexual intercourse with) *"This student sleeps with everyone in her dorm"; "Adam knew Eve"; "Were you ever intimate with this man?"*

Roadmap

1. Terminology

2. **Alignment and Anchoring task**

3. Discussion about the proposed methodology in this paper

Vocabulary Alignment
Benefits

- (Semi-)automatic optimization of search processes

- Sharing information across institutions

Vocabulary Alignment

Previous Strategies

- Lexical techniques
- Structural techniques
- Instance-based techniques
- Techniques based on some background knowledge: **Anchoring**

Roadmap

1. Terminology

2. Alignment and Anchoring task

3. **Discussion about the proposed methodology in this paper**

Anchoring Methodology

- Map Dutch terms to English terms using dictionary

- Comparison of thesaurus terms and WordNet glosses → closest gloss pointed to the intended meaning

Anchoring feasibility experiments

- Anchoring GTAA concepts
  1. Finding English definition for GTAA terms
  2. Anchoring to WordNet synsets

- Anchoring GTT concepts

Anchoring GTAA concepts

First Step

- Query:
  - 3800 preferred terms
  - 2000 non-preferred terms
  - 3200 singular forms
- Result:
  - 2222 translations returned
  - Doubles removed → 1784 terms
  - Cleaning wrt POS tag
  - Cleaning wrt thesaurus relationships → 1655

Anchoring GTAA concepts
Second Step

- Stem GTAA terms and WordNet glosses
- Exact matching of GTAA terms to WordNet glosses
- 1060 concepts are anchored
- Only 28% of #concepts

Anchoring GTAA concepts

Qualitative Perspective

- Three different mappings:

1. One to Many, Many to One
2. One to Many, Many to Many
3. Different Terms with no relation mapped to one

Anchoring GTT concepts

- 70% of its concepts already translated → used as gold standard

- associated term matched one of the expert translations of the concept → correct anchoring

Anchoring GTT concepts

- Precision: 25.7% → low lexical coverage of manual translation
- Manual assess of the method → 77% precision
- Disappointing recall: 5.7%

Anchoring GTT concepts

- Reasons of low performance:
  - Special Dutch *ij*
  - Specialized scientific terms
  - Compound words
- How to remedy multi-word problem?
  - Compositional semantics
  - One to many relation
  - Use specialized dictionaries

GTAA and GTT alignment

- gtaa:arbeiders ←→ gtt:arbeiders
- gtaa:bekeringen ←→ gtt:bekering

- gtaa:plant(Plant) ←→ gtt:begroeiing (Excessive growth of vegetation)
- gtaa:drank ←→ gtt:alcohol (not direct match)

Questions?
Thank you for your attention