# Cross-lingual Ontology Alignment using EuroWordNet and Wikipedia

**Gosse Bouma**

Information Science, University of Groningen
PO box 716, 9700 AS Groningen, The Netherlands
g.bouma@rug.nl

**Abstract**

This paper describes a system for linking the thesaurus of the Netherlands Institute for Sound and Vision to English WordNet and dbpedia. The thesaurus contains subject (concept) terms, and names of persons locations, and miscalleneous names. We used EuroWordNet, a multilingual wordnet, and Dutch Wikipedia as intermediaries for the two alignments. EuroWordNet covers most of the subject terms in the thesaurus, but the organization of the cross-lingual links makes selection of the most appropriate English target term almost impossible. Precision and recall of the automatic alignment with WordNet for subject terms is 0.59. Using page titles, redirects, disambiguation pages, and anchor text harvested from Dutch Wikipedia gives reasonable performance on subject terms and geographical terms. Many person and miscalleneous names in the thesaurus could not be located in (Dutch or English) Wikipedia. Precision for miscellaneous names, subjects, persons and locations for the alignment with Wikipedia ranges from 0.63 to 0.94, while recall for subject terms is 0.62.

## 1. Introduction

Metadata in the form of classification scheme's, ontologies and tags, folksonomies, tags, etc. is available for a rapidly increasing number of web documents, media files, software products, etc. With it comes the need for automatically aligning metadata. I.e. the same song or band may be categorized according to slightly different categories and tags in Wikipedia and MusicBrainz. In such cases, the question might arise whether the category *bebop* is equivalent, similar, or more general than the MusicBrainz tag *cool jazz*. In a multilingual setting, the task can be even more challenging. The category system of English Wikipedia does not match easily with the category system used for other languages. The Dutch category *Postmodernist Philosophers*, for instance, is a subcategory of *Philosophers*, whereas the English category for *Postmodernists* is a subcategory of both *Philosophers* and *Postmodernism*. No exact equivalent of the Dutch category exists in English Wikipedia, as the English *Postmodernists* includes artists as well. Automatic alignment of taxonomies is rapidly becoming one of the more prominent research topics within research on the Semantic Web (Berners-Lee et al., 2001). An overview of ontology alignment techniques is given in Shvaiko and Euzenat (2005).

This paper describes our system for the very large cross-lingual resources (VLCR) task of the Ontology Alignment Evaluation Initiative (OAEI) workshop 2009[1], which asked for an alignment between the thesaurus of the Netherlands Institute for Sound and Vision (GTAA) and English WordNet and (English) dbpedia (Bizer et al., 2009), a database extracted from Wikipedia. GTAA is a Dutch thesaurus used to index video fragments from news shows. The purpose of aligning the thesaurus with other resources is that it may increase accessibility of the collection, by reducing the lexical gap between user search queries and the existing metadata. (Malaisé et al., 2007; Hollink et al., 2009). In this particular case, it helps to make the collection accessible to an international user group

Our participation in the OAEI VLCR task was motivated by the fact that we wanted to establish to what extent resources and techniques we had used to create an informal but wide-coverage Dutch ontology could be useful for the present task as well. For our work on open domain question answering, information extraction, and coreference resolution, we are interested in creating general, informal, taxonomies of entities encountered in Dutch texts.[2] As part of this work, we created a Dutch counterpart of the Yago system (Suchanek et al., 2007), in which Dutch Wikipedia category labels are aligned with a the Dutch part of EuroWordNet (Vossen, 1998). By linking Wikipedia category labels to WordNet, a taxonomy arises that combines the strengths of WordNet (a carefully designed lexical database, organized around word senses and synsets) with the strengths of Wikipedia (a wide-coverage, rapidly evolving, encyclopedia with loosely organised categories). In Bouma (2009), we show that the Alpino parser (van Noord, 2006) can be used succesfully to determine the syntactic head of the often complex Wikipedia category labels (i.e. *Opgeheven luchtvaartmaatschappij van het Caribisch gebied en Midden-Amerika (former airline company from the Caribics and Central America)* or *Italiaans verzetsstrijder uit de Tweede Wereldoorlog (Italian freedom fighter in the Second World War)*). Approximately 60% of the 20,000 category labels are syntactically complex phrases. If the head of a label matches a wordnet sense entry, it is linked to that entry as a hyponym. If multiple sense entries match, a wide-coverage word sense disambiguation technique for finding predominant word senses (following McCarthy et al. (2007)) is used to select the most probable sense, with an accuracy of 0.62. The techniques used to create this resource (especially stemming and parsing of labels, and using predominant word senses for sense disambiguation) appear to be applicable to the VLCR task as well. Note also that the resources used in Bouma (2009), EuroWordNet and Dutch Wikipedia, can be used to solve part of the current alignment problem, as they provide a cross-lingual map-

[1] http://oaei.ontologymatching.org/2009/

[2] Some results can be found on www.let.rug.nl/gosse/Ontology

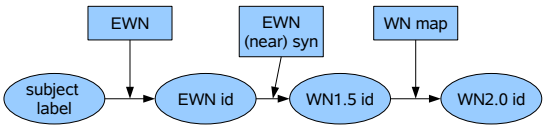


Figure 1: Mapping GTAA to WordNet using EuroWordNet as intermediary

ping from Dutch to English Wordnet and Wikipedia, respectively. The VLCR task also presents novel challenges, as the GTAA thesaurus is less structured than WordNet or Wikipedia, and also, because it is a cross-lingual alignment task.

## 2. Aligning GTAA to WordNet

The multilingual aspect of the VLCR to WordNet alignment task can be approached by using a bilingual lexical resource. The obvious choice in this case is EuroWordNet (Vossen, 1998), a multilingual wordnet which provides, among others, a mapping from Dutch synsets to English synsets.

We restricted ourselves to aligning subject (concept) labels, as the Dutch part of EWN contains few proper names, so we expected the overlap between EWN and other parts of the thesaurus (covering various types of names) to be minimal. The alignment procedure is schematically represented in figure 1.

Entries in the thesaurus are often plurals (*afgevaardigden (representatives), spoorwegen (rail roads), autobussen (buses)*), whereas dictionary entries in EWN are typically singular. To ensure coverage of these cases, all subject labels in the thesaurus and all EWN entries were stemmed using the morphological analyzer of the Alpino parser (van Noord, 2006). As the analyzer also performs compound analysis (ie. *autobussen* is analyzed as *auto_bus*), we also parsed all EWN entries with Alpino. After stemming an compound analysis of both resources, we can find a subject label for a GTAA concept in EWN by simply comparing stems.

Each EWN synset is linked to one or more inter language index ids (ILIs). ILIs in turn are linked to WordNet 1.5 ids. Links can express among others a *synonym, near-synonym, hyponym* or *hypernym* relation. We used only the *synonym* and *near-synonym* relations to find English WordNet ids. In the final step, we mapped WordNet 1.5 ids to WordNet 2.0 ids (which was the target of the mapping), using the WordNet mappings described in Daude et al. (2000).[3]

### 2.1. Results

Table 1 gives an indication of the coverage of GTAA concepts in EWN. 67% of the concepts can be linked to at least one synset in EWN.

Where no link could be found, this is mostly due to multiword subject labels such as *alternatieve energie (alternative energy)* or *bedreigde diersoorten (endangered species)* and compounds. Of the 1261 subject concept labels that could not be linked to EWN, 1030 (82%) were multiword

[3]available from from `www.lsi.upc.es/~nlp/tools/mapping.html`

| | | |
|---|---|---|
| subject labels | 3878 | |
| linked to Dutch EWN | 2,617 | (67%) |
| unique ILIs | 3,703 | |
| avg. ambiguity | 1.4 | |
| linked to WN2.0 | 2,392 | (62%) |
| unique synsets | 3,676 | |
| avg. ambiguity | 1.5 | |

Table 1: Coverage of the mapping from GTAA to the Dutch part of EuroWordNet and WordNet 2.0

phrases or compounds. The GTAA thesaurus contains 324 multiword subject labels. Multiword phrases are generally absent from EWN (except for some foreign expressions such as *accent grave*), and we made no attempt to search for these in English WordNet directly. Other subjects that could not be linked often consist of a compound noun. 1168 subject labels were analyzed as a compound by Alpino. As compounding is a productive process, we do not expect all compounds to be present in EWN. Indeed, only 462 compounds (40%) were linked to EWN.

Note that we require that a compound matches exactly with an entry in EWN. Given the fact that Alpino provides a morphological analysis, we could also have linked compound nouns to a more general concept (i.e. the head noun) by means of a hypernym link. For instance, a compound such as *bedrijfspionage* (*industrial espionage*) could be linked to the more general concept *espionage* by means of a hypernym relation. Hollink et al. (2009) observe that compound analysis can be misleading. The compound *aardappel* (*potato, lit. ground-apple*), for instance, should not be linked to *apple*. As the morphological analyzer of Alpino only proposes a compound analysis for nouns not found in its dictionary, such cases are in principle avoided by our approach. On the other hand, the morphological analyzer is not always accurate either. The noun *antilope* (antelope), for instance, is stemmed as *anti_loop*. This noun happens to be absent from EWN. It should not be considered to be a hyponym of *loop* (*walk*), however. As predicting hypernym relations were not part of the task, we have not investigated this issue any further.

Only 5% coverage is lost in the subsequent mapping to English WordNet 2.0. This suggests that the Dutch part of EWN can be considered as almost a proper subset of English WordNet.

Ambiguity is a serious problem for our approach. Ambiguity of the target is not only caused by word sense ambiguity of the Dutch concept labels, but also by the fact that the mapping between synsets in EWN and WN through ILI links is highly ambiguous. Ambiguity of the concept label arises for a concept such as *koninginnen* (queens). This concept (i.e. its singular form *koningin*) has 5 senses in EWN. Two of these are linked unambiguously with a sense in WN (the 'female insect' and 'chess piece' senses). As the links are of type *eq_synonym*, no ambiguity is introduced in the mapping from Dutch EWN to English WN. However, only 631 mappings from EWN to WN are of type *eq_synonym*. The majority of cases involves a *eq_near_synonym* relation. The

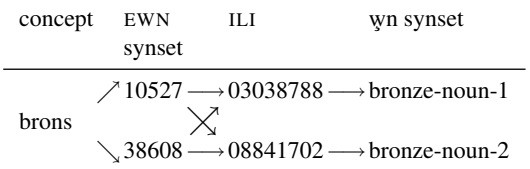


Figure 2: Linking the concept *brons* to two EWN synsets, and two WN synsets.

concept *koningin*, for instance, also has a single 'female ruler' meaning in Dutch EWN, which is mapped by means of an *eq_near_synonym* relation to two meanings in English WN: 'female ruler' and 'wife of a king'.

One might consider reducing the ambiguity by selecting the most appropriate word sense for a given subject label (see Bouma (2009) for some results for Dutch). In the GTAA thesaurus, for instance, *koninginnen* refers to female royalty. However, even if one could develop accurate word sense disambiguation for this particular resource (i.e. concept terms with little context), it will only solve a small part of the problem. The *eq_near_synonym* relation is used much more frequently than the *eq_synonym* relation in mapping EWN to WN. As a consequence, most Dutch synsets are connected to more than one English synset through the near-synonym relation. The situation is illustrated in Figure 2. There are two senses for Dutch *brons*, and two senses for the English word *bronze*. Both Dutch senses are mapped to both English senses. Therefore, even if one resolved the Dutch concept *brons* to the correct EWN synset, it still would be impossible to decide which of the two English WN synsets ought to be chosen. In our results, both targets are given as possible alignment, but lower confidence is given to links involving a near-synonym relation.

The results of our alignment were evaluated by the organizers of the OAEI VLCR task on the basis of a sample of the results (Euzenat et al., 2009). If every link is counted as a potential exact match, we achieve a OAEI precision score of 0.587. If near-synonyms are counted as *closeMatch* relations (thus increasing the possibility that the link is correct, but decreasing the score assigned to the link), a precision of 0.561 results. There was one other group that also took on the challenge of the VLCR task (Nagy et al., 2009). They used a general alignment tool (DSSim) to establish the mapping, and did not use EWN as intermediary. Table 2 shows that the amount of effort we invested in linguistic preprocessing in combination with the fact that we had access to a multilingual resource, helps to improve recall dramatically. The higher precision of the DSSim system is most likely due to the fact that their system requires an exact match of the subject label with a WN sense label. The DSSim system also shows that our decision to ignore names (because they are practically absent in Dutch EWN) is not completely justified, as a substantial number of such names could be aligned with a WN sense directly.

## 3. Aligning GTAA to dbpedia

For linking GTAA entries to dbpedia, we decided to use Dutch Wikipedia as intermediary, and to aim for linking GTAA entries to English Wikipedia pages. Translation of

| | this paper | | | DSSim | | |
|---|---|---|---|---|---|---|
| | # links | Prec | Rec | # links | Prec | Rec |
| Subject | 3,663 | 0.59 | 0.59 | 655 | 0.77 | 0.19 |
| Names | – | – | – | 1,750 | ≈0.50 | – |

Table 2: Aligning Dutch GTAA and WordNet 2.0. Precision and recall are based on OAEI VLCR evaluation figures (based on random samples). *Names* is approximate precision on person names, miscalleneous names, and locations.

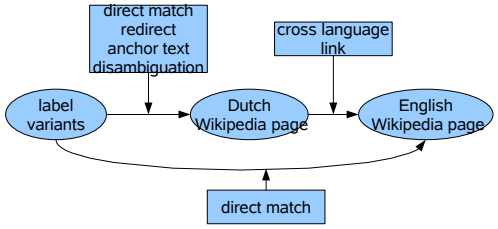


Figure 3: Mapping GTAA to English Wikipedia using Dutch Wikipedia as intermediary

English Wikipedia pages into dbpedia URI's is done by means of a script that adds the correct prefix and deals with special characters. A schematic overview in given in figure 3.

The first step in the alignment is to generate all variants of a label, ensuring that labels start with an upper case letter, are singular, and that person names are of the form *Firstname Lastname*. Alternative labels provided by GTAA are also considered as variants. For instance, for the concept *arenden* (*eagles*), *adelaars* (*eagles*) is given as alternative label.

For all variants of a GTAA concept label, we try to find a matching Dutch Wikipedia page by looking for an exact match with a page title, with a redirect page (in which case the target of the redirect is the desired page), an anchor text (in which case the most frequent target page for that anchor is returned) or with a disambiguation page (in which case all options are returned). Given a suitable Dutch page, we find the English page by following the cross-language link. If a Dutch page (with a corresponding English page) could not be found by means of the techniques above, we tried to find a matching page in English Wikipedia directly, using only page titles. Preference (and a high confidence score) is given to direct matches, followed by redirects, anchors, direct matches in English, and disambiguation pages. If multiple target pages for a given anchor text are found, the most frequent target is selected. If a disambiguation page matches, all its targets are given as matches. The relevant data sets were obtained from an august 2008 dump of Dutch Wikipedia, using the techniques described in Olango et al. (2009). Relevant statistics are given in Table 3.

Examples of the matching process are given in Table 4. The examples illustrate that direct matches and redirects tend to be accurate, but that a match with an anchor text from Wikipedia may have target pages that are more general than the text, or even denote a slightly different concept than the

| GTAA | match | Dutch Wikipedia | English Wikipedia |
|---|---|---|---|
| Aamodt, Kjetil Andre | redirect | Kjetil André Aamodt | Kjetil André Aamodt |
| Abbos, Samira | redirect | Samira Bouchibti | – |
| Abbado, Claudio | direct NL match | Claudio Abbado | Claudio Abbado |
| Abalkin, Leonid | direct EN match | – | Leonid Abalkin |
| Aleksej | anchor text | Aleksej Nikolajevitsj van Rusland | Alexei Nikolaevich, Tsarevich of Russia |
| bedreigde diersoorten | redirect | Bedreigde soort | Endangered species |
| schrijvers (auteurs) (*writers* (*authors*) | direct NL match | Auteur | Author |
| abortusklinieken (*abortion clinics*) | anchor text | Abortus | Abortion |
| computerspelletjes | anchor text | Videospel | Video game |

Table 4: Examples of person names and subjects (concepts) from GTAA for which a match was found in Dutch and/or English Wikipedia, using various matching techniques.

| | |
|---|---|
| page titles | 656K |
| redirects | 198K |
| unique anchors | 1.1M |
| disambiguation pages | 29K |
| cross-language links (NL-EN) | 313K |

Table 3: Statistics for Dutch Wikipedia (aug 2008)

anchor text.

### 3.1. Results

Table 5 gives coverage figures for linking the four different parts of the GTAA thesaurus to English Wikipedia. Coverage is best for subjects and locations. GTAA contains many names of persons and names of other, miscellaneous, entities (organisations, plays, movies, bands) that seem to be absent in both Dutch and English Wikipedia. One might expect coverage of movies to be quite exhaustive in Wikipedia, but of the 204 movie titles in GTAA, 107 could not be linked to a Wikipedia page. Similarly, of the 455 music related names (bands, orchestra's, music awards, etc.), only 253 could be linked to Wikipedia. In the person facet, only 1000 of 2291 actors can be found in Wikipedia. This suggests that, although Wikipedia is often claimed to be very exhaustive especially in area's of popular culture, this is still far from true when compared to a thesaurus dedicated to daily news in the Netherlands.

It should also be noted that coverage of location names is high only because many location names are found in English Wikipedia directly. This holds partly for miscellaneous names as well, but less so for person names. For 6 - 9% of the concepts, a Dutch Wikipedia target could be found, but no corresponding English page. For locations, spelling variation appears to be a source of many missing links. For instance, the place name *Abbasyia* (Lebanon) in GTAA is not found, but search on Google suggests the preferred spelling is *Abbasiya*. Another problem is the presence of additional text in either GTAA or Wikipedia. For instance, for the location *Accomarca* from GTAA there is a matching page *Accomacra (district)* in English Wikipedia, but the *district* classifier makes it hard to find. Similarly *Albertville (Belgisch Kongo)* in GTAA can only be linked through a disambiguation page (*Albertville*) to the correct page *Kalemie* (formerly known as *Albertville*).

Detailed precision scores, based on the OAEI VLCR evaluation (Euzenat et al., 2009), are given in Table 6. Precision varies sharply for different parts of the thesaurus, and also depends on the question whether all links are evaluated as exact match. If only direct matches as seen as exact matches, and all other cases as *closeMatch*, precision drops substantially for parts involving proper names. This is probably due to the fact that *closeMatch* relations generally do not exist for names. Precision for miscellaneous names is considerably lower than for other names. This is caused among others by the fact that this facet contains names for ships, pop-bands, and other entities, that are frequently linked to their name-givers in Wikipedia (i.e. the ship *Carole Lombard* is linked to the English page for the actress *Carole Lombard*).

| | close | exact |
|---|---|---|
| subject-dbp | 0.854 | 0.860 |
| person-dbp | 0.684 | 0.905 |
| misc-name-dbp | 0.527 | 0.627 |
| location-dbp | 0.800 | 0.941 |

Table 6: Precision of the mapping from GTAA to English Wikipedia (dbpedia) for 4 facets of the thesaurus as provided by the OAEI VLCR organizers (Euzenat et al., 2009)

A comparison between our system and the *DDSim* system Nagy et al. (2009), following the results in Euzenat et al. (2009), is given in Table 7.

Again, it can be seen that the fact that we used a Dutch-English resource as intermediary (i.e. the cross-language links in Dutch Wikipedia) helps to improve recall substantially (this is clear for subject terms, but for names we also find more links with equal or better precision, which implies that recall must be higher in these cases as well).

## 4. Discussion

A mapping between two ontologies in different languages can be achieved using appropriate multilingual resources. The mapping to WordNet owes much to the existence of

| | subject | | misc name | | location | | person | |
|---|---|---|---|---|---|---|---|---|
| link type | links | % | links | % | links | % | links | % |
| nlpage | 2,027 | 52.3 | 3,128 | 11.5 | 5,135 | 36.7 | 7,311 | 7.5 |
| redirect | 423 | 10.9 | 984 | 3.6 | 400 | 2.9 | 762 | 0.8 |
| anchor | 621 | 16.0 | 616 | 2.3 | 357 | 2.6 | 176 | 0.2 |
| enpage | 260 | 6.7 | 4,085 | 15.1 | 3705 | 26.5 | 9,246 | 9.5 |
| linked | 3,127 | **80.6** | 8,830 | **32.6** | 9,602 | **68.6** | 17,521 | **17.9** |
| no-english | 357 | 9.2 | 2,197 | 8.1 | 878 | 6.3 | 5,721 | 5.9 |
| no-link | 394 | 10.2 | 16,077 | 59.3 | 3,512 | 25.1 | 74,375 | 76.2 |
| total | 3,878 | 100.0 | 27,104 | 100.0 | 13,992 | 100.0 | 97,617 | 100.0 |

Table 5: Coverage of the mapping from GTAA to Dutch and English Wikipedia (dbpedia)

| | this paper | | | DSSim | | |
|---|---|---|---|---|---|---|
| | # links | Prec | Rec | # links | Prec | Rec |
| subject-dbp | 3,381 | 0.86 | 0.62 | 1,363 | 0.70 | 0.30 |
| person-dbp | 17,516 | 0.91 | – | 2,238 | 0.79 | – |
| misc-name-dbp | 9,023 | 0.63 | – | 3,989 | 0.64 | – |
| location-dbp | 9,527 | 0.94 | – | 5,566 | 0.80 | – |

Table 7: Aligning Dutch GTAA and English Wikipedia (dbpedia). Precision and recall (for subject terms only) are based on OAEI VLCR evaluation figures (based on random samples).

EuroWordNet, which solves the multilingual aspect of the task to a large extent. On the other hand, EuroWordNet does not help much in deciding which synset for a given English term is the appropriate one.

Our results for Wikipedia linking could still be improved in a number of ways. We hardly employed hierarchical and categorical constraints. The GTAA thesaurus comes in four parts. Each part is a different category. This information could be used to block the link from *A4* in the locations file to *A4 (paper format)* in Wikipedia. Word overlap could also be used to select the correct target page (i.e. to prefer *highway A4 in the Netherlands* over that in *Austria*)). Alternatively, one might use the information that concepts with the same *scopeNote* are likely to be linked to Wikipedia pages with identical or closely related Wikipedia categories to detect outliers. The name *Carole Lombard* has a *scopeNote ship*, for instance, that could be used to rule out the link with the *actress Carole Lombard*, for instance. Note, however, that this requires a mapping between labels used as *scopeNote* and Wikipedia categories, along with a notion of incompatibility of categories, something that might be a challenge in itself.

For selecting the most promising target, we experimented with a simple preference scheme (which always prefers the link given by the most reliable relation), and a simple weighting scheme (which adds scores when multiple links to the same target are found). Weighting was used for the final results. No doubt, more subtle schemes could be developed. For instance, at the moment we only take into account the most frequent target of an anchor text. Alternatively, one might consider all targets pointed to by anchor text as potential targets, and use the frequency of these links as a weight.

Somewhat surprisingly, we discovered that cross-language links are not reversible. Initially, we used cross-language links harvested from English Wikipedia, as this is the larger resource, and we expected that this might also be more thorough in providing cross-language links. However, since English Wikipedia has more pages than Dutch Wikipedia, several English pages may be linked to the same Dutch page (i.e. *Bowling* and *Ten pin Bowling* both point to the Dutch page *Bowling*). If one works with cross-language links harvested from Dutch Wikipedia, this situation does occur less frequently, although similar problems can occur here as well (i.e. in the versions of Wikipedia we used, the Dutch *A4 highway* was linked to an English page which redirected to a general page on Dutch highways).

## 5. References


Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific American*.

C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. DBpedia-A crystallization point for the Web of Data. *Web Semantics: Science, Services and Agents on the World Wide Web*.

Gosse Bouma. 2009. Linking Dutch Wikipedia Categories to EuroWordNet. In *Proceedings of the 19th Computational Linguistics in the Netherlands meeting (CLIN 19)*. Groningen, the Netherlands.

J. Daude, L. Padro, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 504–511. Association for Computational Linguistics Morristown, NJ, USA.

J. Euzenat, A. Ferrara, L. Hollink, A. Isaac, C. Joslyn, V. Malaisé, C. Meilicke, A. Nikolov, J. Pane, M. Sabou, F. Scharffe, P. Shvaiko, V. Spiliopoulos, H. Stuckenschmidt, O. Ŝváb-Zamazal, C. Trojahn dos Santos V. Svátek, G. Vouros, and S. Wang. 2009. Results of the ontology alignment evaluation initiative 2009. In *Fourth International Workshop on Ontology Matching*, Washington, DC.

Laura Hollink, Véronique Malaisé, and Guus Schreiber. 2009. Thesaurus enrichment for query expansion in audiovisual archives. *Multimedia Tools and Applications*.

V. Malaisé, A. Isaac, L. Gazendam, and H. Brugman. 2007.

Anchoring Dutch cultural heritage thesauri to Wordnet: two case studies. In *Proceedings of the ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, pages 57–64, Prague.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics*, 33(4):553–590.

Miklos Nagy, Maria Vargas-Vera, and Piotr Stolarski. 2009. DSSim results for OAEI 2009. In *Fourth International Workshop on Ontology Matching*, Washington, DC.

Proscovia Olango, Gerwin Kramer, and Gosse Bouma. 2009. Termpedia for interactive document enrichment. In *Computational Linguistics Applications (CLA) workshop at IMCSIT*, Mragowo, Poland.

P. Shvaiko and J. Euzenat. 2005. A survey of schema-based matching approaches. *Lecture notes in computer science*, 3730:146–171.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA. ACM Press.

Gertjan van Noord. 2006. At last parsing is now operational. In Piet Mertens, Cedrick Fairon, Anne Dister, and Patrick Watrin, editors, *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42.

P. Vossen. 1998. Eurowordnet a multilingual database with lexical semantic networks.