FP7-ICT-2007-3-231161

Deliverable D4.2.1

Vocabulary alignment as an interactive and replicable workflow

M. Hildebrand (VUA), J. van Ossenbruggen (VUA),
V. de Boer (VUA)

18/02/2011

Document Administrative Table

Document Identifier	PP_WP4_D4.2.1_VocabularyAlignmentMethodology Release	0
Filename	PP_WP4_D4.2.1_VocabularyAlignmentMethodology_v1.0.pdf	
Workpackage and	WP4 - Long-term audiovisual access and use in changing contexts	
Task(s)	WP4T2 - Vocabulary alignment and annotation	
Authors (company)	Michiel Hildebrand (VUA), Jacco van Ossenbruggen (VUA) Victor de Boer (VUA)	
Contributors (company)	Antoine Isaac (VUA)	
Internal Reviewers (company)	Werner Bailer (JRS), Johan Oomen (B&G)	
Date	18/02/2011	
Status	Release	
Version	1.0	
Type	Deliverable	
Deliverable Nature	Report	
Dissemination Level	Public	
Planned Deliv. Date	M24 - 31/12/2010	
This IsPartOf		
This HasPart		
Abstract	Interactive alignment methodology for SKOS vocabularies	

DOCUMENT HISTORY

Version	Date	Reason of change	Status	Distribution
v0.01	2010-10-14	Introduction and requirements for methodology	Outline	Confidential
v0.02	2011-01-26	Modify requirements	Working Draft	Confidential
v0.03	2011-02-01	GTAA use case	Working Draft	Confidential
v0.04	2011-02-08	WordNet Use case	Working Draft	Confidential
v0.05	2011-02-10	GEMET Use case	Working Draft	Confidential
v0.05	2011-02-11	Description of interactive approach	Working Draft	Confidential
v0.04	2011-02-14	Discussion	Working Draft	Confidential
v0.05	2011-02-14	First Final Draft version	Final Draft	Confidential
v1.00	2011-02-18	Release corrected according to feedback from internal reviewers	Release	Public

Table of contents

1 Introduction

In the library, archive, museum and many other domains, objects are routinely described using terms from predefined vocabularies. When object collections need to be merged or linked, a typical question that needs to be answered is how those vocabularies relate. More specifically, one would like to know which concepts from different vocabularies correspond to one another. We will call a set of such correspondences an *alignment*.

There is an active research field that studies methods and techniques to generate alignments automatically, and the tools produced by this field are evaluated yearly in the context of the Ontology Alignment Evaluation Initiative (OAEI)[1]. A key insight from this field is that two concepts can be similar or dissimilar along many different dimensions [1]. Automatically finding similar concepts typically requires some hybrid approach that combines different techniques, each addressing a part of the total set of potentially interesting dimensions. Another important insight is that the application context in which the alignment will be deployed often influences what constitutes a "good" alignment [8]: two concepts might be regarded as sufficiently similar in one context, but not in another. Not surprisingly, the main approach in vocabulary alignment is to develop tools that (a) apply some smart combination of available techniques to generate an alignment, and (b) allow the developer to tune the tool so that the alignment fits a specific application context.

While the approach sketched above is well established, our work in the cultural heritage field has proved that it also has some major limitations. Currently, alignment tools are evaluated according to a reference alignment. In practice, it is, however, unclear for data providers how well a tool would perform for their specific vocabulary. In the cultural heritage domain this creates a bottleneck to align vocabularies, as data providers want to have tight control over the quality of their data. In the remainder of this paper, we will discuss the key limitations of current tools in more detail and propose an alternative approach. We will show how this approach has been used in three alignment use cases, and demonstrate how it is currently supported by our "Amalgame" alignment platform [2].

[1] http://oaei.ontologymatching.org/
[2] http://semanticweb.cs.vu.nl/amalgame/

2 Requirement analysis

The requirements in this chapter are partly based on previously published work [4, 5, 6], and partly based on feedback received from domain experts during our work in the MultimediaN E-Culture[3], Europeana(Connect)[4] and PrestoPrime projects[5].

First, domain experts find it hard to determine how well a tool would perform for their alignment task. From the alignment research literature, it is clear how each tool performs on the data used in the evaluation experiments. However, it is unclear why the tools perform well and if the same performance is achieved for a different vocabulary. Therefore, it is hard for experts to predict which tool would perform well on their own data set.

Second, experts perceive the current tools to not support the large and shallow vocabularies that are typical for their domain. Most alignment tools target complex vocabularies with different ontological relations, but only several 100s or 1000s of classes. In the cultural heritage domain the vocabularies typically contain only a few thesaurus relations, but may contain 10,000s or 100,000s of concepts.

Third, when an alignment run finishes, it typically produces a result set with a large number (e.g. over 100k) of correspondences, but provide little support to assess the quality of these results. Furthermore, the quality of the correspondences might not be homogeneously distributed across the alignment result set. Different subsets of alignments might have different features that determine the quality of the end result. Transparent and interactive assessment is crucial to be able to decide whether the result is of sufficient quality.

Fourth, when the results are not sufficient it is unclear how the tool should be (re-)configured to improve the results. Experts need to be able to understand why the tool found erroneous correspondences and how to get rid of them in a next step. When the tool failed to find correct correspondences, the experts need to know how to find those in a next step. This requires insight in how the alignment algorithms work, and how to configure them to adjust them to the specific needs of vocabularies at hand.

We conclude that to effectively support the alignment of real vocabularies in the cultural heritage domain users need: (i) the ability to quickly run different matching algorithms in an interactive environment, and (ii) the ability to analyze large sets of correspondences to determine the effect of the algorithms. Such an interactive environment needs to fulfill the following specific requirements that would distinguish it from most current alignment toolkits:

Speed

To allow interactive scenarios, it is important that the matchers are sufficiently fast to be run in an interactive setting. Therefore, when aligning large vocabularies in an interactive environment, it is better to use simple matching strategies that are computationally cheap, instead of the computationally expensive ones used by most alignment tools today.

Transparency

In an interactive setting, it is also important that users understand each intermediate result to be able to judge how to improve it in a next step. They would favour simple matching algorithms in which the pros and cons are easy to understand over complex ones for which domain experts cannot explain the results. Domain experts often have a deep understanding about the characteristics of their vocabularies. If a tool also allows them to understand the matching techniques used, they should have sufficient knowledge to make informed choices about the design of an alignment strategy that is targeted to the specific needs of

[3] http://e-culture.multimedian.nl/
[4] http://www.europeanaconnect.eu/
[5] http://www.prestoprime.org/

their own data.

Configurable components and overall workflow

Assuming that the user indeed has sufficient knowledge about the data and the available matching algorithms to design a targeted alignment strategy, we need a method in which the user can tune the parameters of each matcher to the specific needs of her data set. In addition, the user also needs to be able to configure which matchers to run for which part of the data and in which order.

Vocabulary and result analysis

Assuming that the above requirements are met, users can quickly configure and run different matchers on even large data sets. The most expensive step in such an interactive environment would then no longer be finding the correspondences, but the analysis of the large result sets (and optionally an re-analysis of the vocabularies) to decide on which step to take next. Tool support for the analysis of the large sets of correspondences that are typical for this domain is thus also crucial.

Provenance

An additional requirement follows from the potential for the use of the resulting alignments in other application scenarios. Within the context of the Semantic Web and Linked Open Data, more and more alignments are being published on the Web and reused in a widely different set of contexts. For many of these alignments, it remains unclear how they have been generated, and how the same or a similar data set could be reproduced. For example, as new versions of the underlying vocabularies become available, one might want to update the associated alignments by rerunning the same alignment technique on the new versions. When alignments are the result of a scientific experiment described in a research paper, it also desirable to be able to replicate the experiment and have the results confirmed by others. However, most alignments currently published have insufficient metadata to allow alignments to be reproduced. To address this, we require that an interactive alignment tool records sufficient information about each individual step, and the order in which the steps are executed, that the result set can be fully reproduced later.

In the next section, we sketch an alignment approach that is based on these requirements. We then show the feasibility of our approach by discussing three use cases of vocabulary alignments in which we have used this approach.

3 The amalgame approach to vocabulary alignment

To address the requirements above, we developed an alignment approach that improves the speed and transparency of the alignment process by drastically reducing the complexity of the technology, allowing the user to combine a limited number of basic building blocks into a alignment workflow targeted to the data set at hand. Each building block should be sufficiently simple to produce an understandable result. Which blocks to use and in what order or combination is fully controlled by the user. Furthermore, produced alignments —both intermediate and end results— can be easily and quickly evaluated to give insight in their quality.

We have built a prototype alignment service that has been designed with this approach in mind, and used the prototype to create alignments in three different use cases, that will be discussed in the next section. Here we sketch an high level overview of the amalgame alignment methodology and will flesh out some interesting details in the context of the use case descriptions.

3.1 Vocabulary analysis

An assumption of the interactive approach is that the user has knowledge of the vocabularies being aligned. Here, we focus on vocabularies that can be represented by SKOS [3]. For such SKOS-like vocabularies we identify two types of characteristics.

First, the user needs to know the number of concepts that the source and target vocabularies contain. For example, if vocabulary A is ten times the size of B, it is unlikely to find equivalence relations for more that 10% of A's concepts, and looping over all concepts in B to find correspondences in A will be more efficient than *vice versa*. Furthermore, large vocabularies are often heterogeneous, in the sense that they can contain different types of concepts that may require different matching techniques. This makes it important that the user knows the different types of concepts the vocabularies contain. By matching only similar types of source and target concepts, the workload is reduced and the precision can be increased. There is, for example, no need to align concepts representing persons with those representing locations.

Second, the user has to identify the concepts' properties that can be used in the matching process. In SKOS-like vocabularies preferred and alternative labels are likely candidates for simple string matching techniques. Using alternative labels typically improves recall over using only preferred labels, but may also reduce precision. A vocabulary owner should be able to make her own trade off what to use in which case. Knowing which correspondences result from preferred label matches only and which not may also help in designing a more targeted evaluation strategy (e.g. by deciding to evaluate a relative smaller sample of preferred label matches) or by deciding what next steps to take (e.g. that the preferred label matches are of sufficient precision but that the alternative label matches need additional filtering to remove false positives). In addition, there may be labels in different languages. Typically, the user wants to avoid matching of labels that are in different languages to prevent false matches, but sometimes multi-linguality can also be used as an advantage, for example when syntactic label match found in multiple languages can be interpreted as extra evidence for a given correspondence.

Besides the labels, the vocabularies may provide other properties that can be matched. Typically, longer textual descriptions found in SKOS definitions or scope notes may provide another source of textual information. Hierarchical or associative relations provide structural information. These properties can be used to create new sets of alignments (boosting recall) but also to improve already existing alignments, for (boosting precision). For example, in geographical thesauri the hierarchical containment is often sufficient to distinguish between ambiguous label matches.

Properties such as SKOS editorial notes are typically not useful for matching, but there are exceptions in individual cases. For example, we have came across several cases where editorial notes associated with many concepts in one vocabulary contained the unique identifier of the concept of another vocabulary from

which the concept had once been copied. Such information obviously simplifies alignment drastically, but requires that the user knows these notes are present and that this knowledge can subsequently be used as input in the alignment workflow.

3.2 Workflow components

Our approach is to have the user interactively construct an alignment workflow. The individual building blocks of this workflow consist of: (i) selectors to define which concepts to use from the source and target vocabularies, (ii) matchers to find correspondences between the source and target concepts, (iii) partitioners to split sets of correspondences and mergers to create the union of subsets, (iv) analyse tools to investigate the mappings, and (v) filters to select specific correspondences and discard others.

Selectors

The user starts by defining the source and target vocabularies and possible selectors, based on the analysis of the vocabularies. For example, the user can select a specific concept scheme or specific types of concepts. Once a set of alignments is created the user can also select the set of concepts that are not yet aligned.

Matchers

The definitions of the source and target provide the input for a matching technique. Which technique is most suited for the first step depends on the source and target vocabularies, and might vary for different types of concepts or properties within these vocabularies. Our method is independent of the specific matching technique used. We assume each technique takes a specification for the source and target concepts as input and produces a set of correspondences. In addition, each technique can have parameters to tune it for the data at hand.

We distinguish matching techniques that use textual properties from techniques that use structural information. The literature provides a wide range of similarity metrics to match textual properties. Our tool provides a number of these metrics out-of-the-box including exact label matching, prefix matching and jaccard matching. Structural matching uses the position of a concept in the graph or tree structure of the vocabulary. Typical structural information in SKOS like vocabularies is provided by the hierarchical and associative relations, where the number of steps that are considered in the matching are typical parameters. Again, our tools provides reasonable defaults but leaves tuning to the user.

Partitioners and mergers

Alignment in our approach is inherently an iterative process, the user needs to determine — after each matching technique that has been applied — what to do next. This decision depends on the correspondences that are generated. Amalgame supports the user in this interactive process by allowing the user to partition the mappings in specific subsets and analysis tools to investigate these subsets. A typical partitioners divides the mappings into a set of 1-1 and 1-n correspondences. The 1-1 subset contains all correspondence for which each source only has one target, whereas the 1-n subset contains all correspondences where a source has multiple targets. Vica-verse the users might also want to combine the mappings from the different matchers. For this case, Amalgame provide a simpler merger to create the union of multiple sets of mappings.

Analysis tools

Set (or subsets) of mappings can be analysed as a whole, or as individual correspondences. Currently, Amalgame supports the analysis of a set of mappings by presenting a number of statistics. Figure 1 shows a screenshot of such statistics. These include for each matching technique the number of the source and target concepts that are matched and the total number of alignments. In addition, Amalgame contains an evaluation tool that enables the user to inspect individual correspondences. When the user is confident that a set of matches is correct, for example the 1-1 mappings, she can limit investigation to a small sample. Figure 2 shows a screenshot of the evaluator. On the left it shows the source concept and on the right two matching target concepts. For each concept additional information, such as alternative labels, descriptions, related concepts and the location in the hierarchy are shown. For the target concepts the user can manually indicate wether it is a good match. The details of this evaluation process are outside the scope of this paper.

Filters

Based on the analysis of the results the user may decide that the set of mappings should be filtered, for example, to distinguish true from false correspondences. We can distinguish two cases. In the first case the result set already contains sufficient evidence to make a distinction. For example, Amalgame provides several filters to determine the best target concept from a 1-n correspondence, such as selecting the target that was found by the most number of matching labels or choosing a target found by a preferred label over a target found by an alternative label. In case the current result set does not provide insufficient information to make such a selection, additional matching techniques can be run to find extra evidence to discriminate true from false correspondences. Such filters can, for example, use structural information, such as comparing the similarity in the hierarchical structure. In fact, in Amalgame any matcher can also be used as a filter.

3.3 Interactive alignment

Alignment within Amalgame is a process where the user iteratively applies matchers, partitions the result set, and applies new matchers or a filter. After each step the user typically analysis the results to determine the next step. We identify five typical scenarios, depending on the outcome of the analysis.

- The first scenario is that a user decides the results are no good at all, in which case all results are simply discarded after analysis. Assuming the technique used is sufficiently simple, the user will understand from the analysis what caused the failure and will be able to try another matching run, using another technique or a better configuration of the technique used in the previous run.
- The second scenario is that the results are good, but that recall is low. To improve recall, the user can proceed by matching only the concepts that have not yet been aligned. Note that this result set is typically a smaller set, so the user may decide to deploy computationally more expensive matching techniques to improve recall in subsequent runs.
- The third scenario is that the results are good, but that precision is low. To improve precision, users need to find filters that allow them to distinguish true from false correspondences. Again, more expensive techniques can be used to boost precision for smaller subsets.
- The fourth scenario is that a user decides that the results are of sufficient quality, after which she exports them to the desired format and we consider the alignment task to be successfully finished.
- The fifth scenario is that the user finds the results of insufficient quality, but is out of options and does not know how they can be further improved, in which case we consider the alignment task to be failed.

In practice, we found the first scenario useful to quickly try some alternative matchers, and to compare, analyse and discard the results, just to develop some intuition before the real alignment task starts. Many

alignment tasks, including the first two use cases discussed below, are based on an iteration of the second and third scenario. Ideally, with each iteration the set of concepts that have to still be aligned (to improve recall) and the set of correspondences that still have to be filtered (to improve precision) decreases, or, if not, the user gains some knowledge to achieve this in the next step.

4 Use cases

In practice, we found that the in-house vocabularies from different institutes are sometimes directly aligned with each other, but typically they are indirectly related because by aligning them to the same external vocabulary. As the first use case we explore such an alignment of an in-house vocabulary to an external vocabulary. We consider the alignment of the thesaurus of the Netherlands Institute for Sound and Vision, GTAA, with a general linguistic vocabulary of Dutch, Cornetto. A benefit of alignment with such an external vocabulary is that this also makes the alignments of this vocabulary available for the in-house vocabulary. For example, Cornetto already contains links the English WordNet. A different example where alignment is required, is when a new version of a vocabulary is released, and no direct links between the two are maintained. In the second use case we consider the mapping of two different versions of WordNet. As a third use case we include the alignment of multilingual vocabularies, to demonstrate how the overlap in different languages can be incorporated.

4.1 In-house to general: GTAA to Cornetto

The Netherlands Institute for Sound and Vision uses an in-house thesaurus for the documentation of audiovisual content. This so-called GTAA thesaurus (Dutch acronym for Common Thesaurus Audiovisual Archives) contains approximately 160,000 terms in six facets: subjects, locations, person names, organization names, maker names and genres. In this use case we focus on the terms in the subjects facet.

Cornetto is a lexical semantic database of Dutch that contains 70,000 synsets [9]. Compared to the GTAA subject terms, the synsets provide a large number of additional synonyms and an extended description. The synsets are linked by 59 different types of semantic relations, including a fine-grained hierarchical structure.

An alignment from GTAA to a Cornetto would provide additional labels (e.g. synonyms) and semantic relations to GTAA's subject terms, increasing the ways to access the audiovisual collection. In addition, the existing alignment between Cornetto and WordNet could also provide an English access point to the archive.

For this use case we map the GTAA subject terms to Cornetto synsets. As Cornetto contains the same words in different synsets (e.g. homonyms) , we can expect string matching techniques will find multiple synsets for many GTAA subject terms. Our focus is choose the right target synset(s) for each source. Typically, this will be one synset per GTAA subject term, but there might be cases where multiple synsets are good candidates. In this case, the aim is to get all correct alignments between the source and different targets.

4.1.1 Vocabulary analysis

We start the alignment process with an exploration of the GTAA subject terms. In total there are 3,932 subject terms. All terms have at least one preferred label, often an alternative label and one or more related terms, and some have a description. In addition, the subjects are organized in an hierarchical structure. We observe that the majority of the terms are nouns. In Cornetto this part of speech distinction is explicit, as each synset is of word type: noun (52,845), verb (9,017), adjective or adverb. Ideally, we would like to map the nouns in GTAA to the nouns in Cornetto. However, there is no explicit information in GTAA to automatically distinguish the nouns from the verbs. We choose the next best solution and start with the alignment of all GTAA subject terms to the nouns in Cornetto. We assume that there will be no or little verbs from GTAA that will be mapped to the nouns in Cornetto.

We also observe that the most labels of the GTAA subject terms are in plural form, whereas the labels in Cornetto are in singular form. When matching the labels we should account for this difference. Finally, we

Sel	Abr	Source	# mapped	Target	# mapped	Format	# maps	Named Graph URI
☐	A	gtaa:GTAA	2493	Cornetto Dutch Lexical Database	3533	edoal	3667	<stemming_pref>
☐	D	gtaa:GTAA	1785	Cornetto Dutch Lexical Database	1776	edoal	1785	<stemming_pref_1n>
☐	E	gtaa:GTAA	708	Cornetto Dutch Lexical Database	1817	edoal	1882	<stemming_pref_11>
☐	F	gtaa:GTAA	2725	Cornetto Dutch Lexical Database	4403	edoal	4715	<stemming_pref_alt>
☐	G	gtaa:GTAA	1655	Cornetto Dutch Lexical Database	1646	edoal	1655	<stemming_pref_alt_1n>
☐	K	gtaa:GTAA	1070	Cornetto Dutch Lexical Database	2832	edoal	3060	<stemming_pref_alt_11>
☐	M	gtaa:GTAA	1190	Cornetto Dutch Lexical Database	1637	edoal	1670	<exact_pref>
☐	N	gtaa:GTAA	880	Cornetto Dutch Lexical Database	878	edoal	880	<exact_pref_1n>
☐	O	gtaa:GTAA	310	Cornetto Dutch Lexical Database	780	edoal	790	<exact_pref_11>
☐	P	gtaa:GTAA	1319	Cornetto Dutch Lexical Database	1980	edoal	2050	<exact_pref_alt>
☐	Q	gtaa:GTAA	829	Cornetto Dutch Lexical Database	821	edoal	829	<exact_pref_alt_1n>
☐	S	gtaa:GTAA	490	Cornetto Dutch Lexical Database	1177	edoal	1221	<exact_pref_alt_11>
							24204	*Total (double counting)*

Figure 1: Screenshot of the mapping statistics in Amalgame. For different type of matching techniques it shows the number of source and target concepts that are matched and the total number of matches found.

	Preferred labels			Preferred + alternative labels		
	total	1-1	1-n	total	1-1	1-n
exact	1,190 (30%)	880	310	1,319 (33%)	829	490
stem	2,493 (63%)	1785	708	2,725 (69%)	1655	1070

Table 1: Number of alignments between GTAA and Cornetto. Horizontally, the labels used: preferred labels only and including alternative labels. Vertically, the label similarity metric: exact matching or matching after stemming.

observe that where GTAA discriminates between preferred and alternative labels, Cornetto only has one type of label, which has been mapped to `skos:altLabel`.

4.1.2 Interactive alignment

Given the discussions above, it is not *a priori* clear which string matching strategy to use. We expect the use of GTAA alternative labels will increase recall, but are unsure at what expense it terms of precision. Similarly, we expect that stemming might deal with the the plural nouns in GTAA and singular nouns of Cornetto, but stemming might also introduce new problems. We decide to explore different options and try matching including and excluding GTAA alternative labels. We also run different matchers: using exact label matching and matching after stemming.

Figure 1 shows the statistics generated by Amalgame for for all four combinations. The key statistics are summarized in Table 1. From the column labeled *total*, we observe that there is indeed a large increase when stemming is used. We can also observe that by including the alternative labels more alignments are found. Based on these observation we might opt for the approach that gives us the most alignment, matching the stems of both the preferred and alternative labels. Before we make this decision there is, however, an important characteristic of the results that we should consider. How many target concepts are found for each source concept? And in case multiple targets are found, is this caused by ambiguity of the source concept or are all targets valid alternatives?

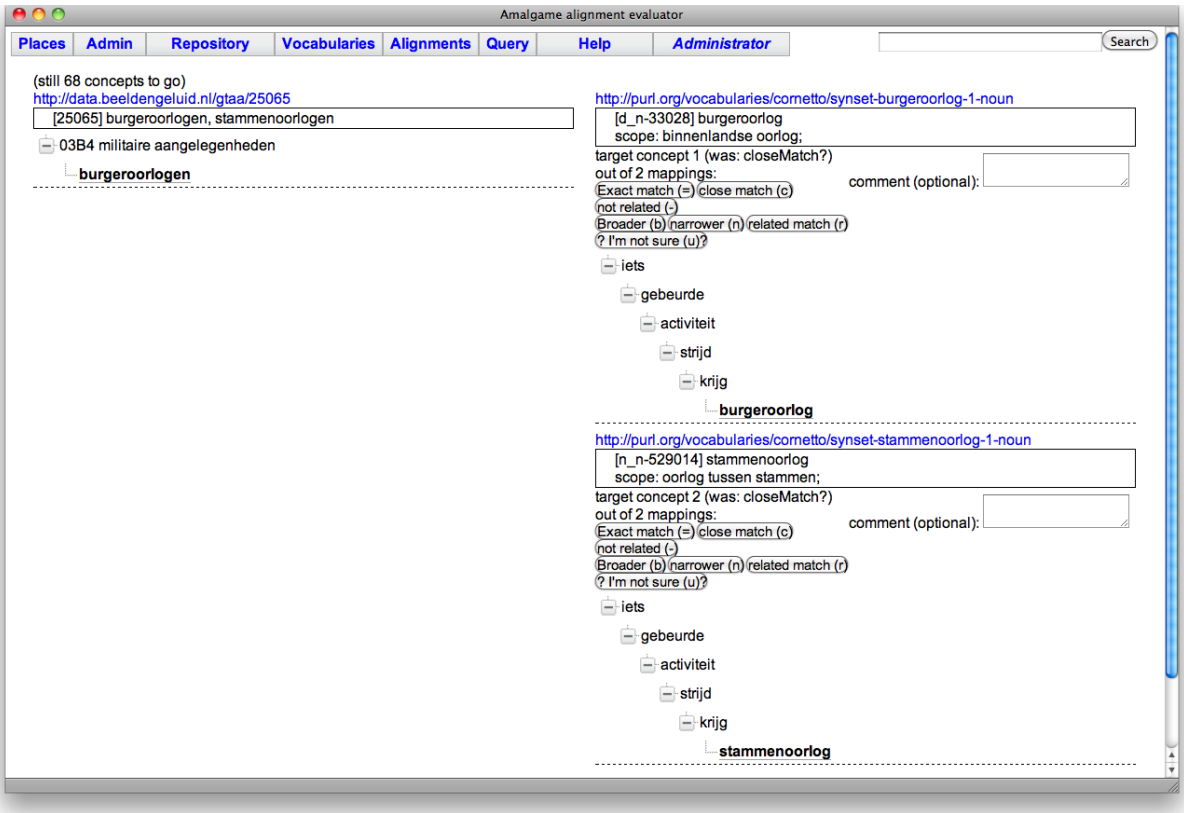

Figure 2: Screenshot of the Amalgame evaluation prototype. On the left the source concept from GTAA and on the right two target concepts from Cornetto. The GTAA concept “burgeroorlog” (dutch for civil war) with alternative label “stammenoorlog” (dutch from tribal war) is mapped to two different targets. The target concepts in Cornetto, civil war and tribal war, are siblings as they are two specific types of war.

To investigate different types of alignments we use Amalgame to partition the set of alignments. We partition them in a set of alignments where the source concepts have only 1 target, and another set where the source concepts have multiple targets. Table 1 lists the number of sources that are mapped to only 1 target, *1-1*, and mapped to multiple targets, *1-n*. We observe that the number of *1-1* alignments is larger when only the preferred labels are included. In other words, the alternative labels primarily introduce extra targets for sources that were already mapped. Do these alignments introduce unnecessary ambiguity, or are the additional targets valid alternatives?

To analyze the results in more in detail we use Amalgame to visualize alignments including the relevant information of the source and target concepts. In this case, we are interested in the *1-n* mappings introduced by the alternative labels. We produce this set by subtracting the 708 *1-n* mappings found by preferred labels from the larger set of 1070 *1-n* mappings found by preferred and alternative labels. From the resulting set of alignments we take a random sample of 25 alignments to investigate in detail. Figure 2 shows a screen shot of this investigation. For a single source concept it lists the multiple target concepts. In addition, all alternative labels, descriptions and related terms are shown. Going through the sample set we found four different types of 1-n mappings:

1. One of the targets is more generic then the other. Cornetto is more fine-grained then GTAA. A single concept in GTAA containing multiple labels, e.g. “poison, pesticide”, is mapped to different targets in Cornetto, where “poison” is more generic than “pesticide”. In this case we want to select the most generic term. Optionally, we could create narrower matches between the other targets, but this is

1-n	most generic	siblings	hierarchy similarity	other
1070	91	72	397	510

Table 2: Number of alignments between GTAA and Cornetto after different disambiguation strategies.

outside the scope of this paper.

2. The targets are siblings of each other. Again the granularity difference between the vocabularies often causes a single concept in GTAA, e.g. civil war and tribal war, to be mapped to different targets in Cornetto, where civil war and tribal war are siblings as they are more specific types of war (shown in Figure 2). In this case all targets are valid alternatives and we want to keep all alignments to the siblings.

3. The targets are about the same topic. Some concepts in GTAA contain labels for different types of things, but related in topic e.g. bee-keeping and honey edge. In Cornetto these are different terms in completely different parts of the hierarchy. We choose here to keep both matches. If we would have the rights to modify GTAA, we could also decide to split the source concepts into two separate concepts.

4. The targets are different senses of the source concept. A GTAA concept is matched to one concept from Cornetto by its preferred label and to another by its alternative label. For example, by the preferred label “capitulate” a single concept from Cornetto is found. By the alternative label “surrender” it finds the same concept, but also the concept that refers to “surrender of attention”. In this case the source concept is ambiguous and only one target should be selected.

We conclude that by using only the preferred labels valid alternatives are excluded. Therefore, we choose to include alternative labels and match them after stemming. Alignments are likely to be correct as we used a simple matching algorithm that fits well with the labels in our vocabularies. Evaluation of a random sample of 25 alignments confirms this assumption, as all alignments are indeed correct. At the other hand we have a large set of 1-n mappings. The analysis of these alignments provides us with a number of different cases. How can we use this knowledge of the 1-n mappings to find the valid 1-n mappings and, in case of ambiguity, select the best candidate?

To automatically detect different types of alignments and select the best target candidates Amalgame provides a number of strategies. We configure these strategies for the different types of *1-n* alignments. We start with the set of *1-n* alignments and try to identify the alignments for each case. Table 2 shows the total number of source concepts, 1070, in the *1-n* alignments, and the number of source concepts we detected for each case. For 91 source concepts we can find a target that is more generic than the other targets. These concepts are found by configuring the Amalgame partitioning component to check for hierarchical relations between the targets. From the remaining alignments 72 sources are identified as having a set of targets that are siblings.

For the remaining alignments we try to automatically detect the most suited candidate. We observed that the wrong targets can occur in different sub trees of Cornetto. Therefore, we can identify the best target by the hierarchical similarity to the source target. For each ambiguous alignment we check if the source and target have similar ancestors or descendants. To test for similarity between the terms in the hierarchy we use as a base set the 1-1 alignments. When the hierarchy of one target has more alignments to the hierarchy of the source it is a better candidate. As this method adds new 1-1 alignments, it extends the base set, possibly relevant for further disambiguation. Therefore, we repeat this procedure until no more additional matches are found. In total, for 397 source concepts we manage to find a distinguishing target.

Finally, we decide to align all remaining GTAA subjects to the verbs in Cornetto. Analysis of the vocabularies also makes clear that the labels of the verbs, in both vocabularies, are in infinitive form. Therefore, we

choose to align them using exact string matching. For 115 source concepts we find alignments, 78 of these are *1-1* mappings, while 37 are *1-n* mappings. As the set of *1-n* mappings is very small, we can manually evaluate it. Within 14 minutes we manually disambiguated 19 sources, and accepted multiple alternatives for two sources. For the remaining 14 source concepts we decided they were falsely mapped. All were nouns that were not mapped due to limitations of the stemming algorithm. We expect the same stemming problem causes errors in the set of *1-1* alignments, and also manually evaluate these. Within only 5 minutes we found the 13 source concepts were it went wrong.

4.1.3 Results

Figure 4 (in Appendix) shows the final workflow for the mapping between GTAA and Cornetto. In total we found matches for 2840 (72%) concepts from the GTAA subjects facet. From these the large majority 2725 (69%) were matched to Cornetto nouns. For 42% of the GTAA subjects we found a mapping to only one target. As we used a simple matching technique, we expected high precision for this subset. In an evaluation of a small sample of this set all mappings were judged to be correct. In the remaining set we identified four ways in which multiple targets were found. We configured the filter components to identify these cases. For more than half of the 1-n matches we managed to either select the best target or confirm that all targets are valid alternatives. To judge the other half of the matches manual evaluation is required. In future work we would like to perform such an evaluation with the users of GTAA. Finally, only 115 GTAA subject terms were mapped to Cornetto verbs. This small set we manually evaluated in only a few minutes.

4.2 Versioning: WordNet 3.0 to WordNet 2.0

WordNet is a large lexical database of English published by Princeton University. It groups nouns, verbs, adjectives and adverbs into sets of cognitive synonyms (synsets), each expressing a distinct concept [2]. W3C released an RDF version of Princeton's WordNet 2.0 in June 2006 [7]. In August 2010 we released an RDF conversion of Princeton's WordNet 3.0 as Linked Open Data. Until now, there is no reliable data set that specifies which synset in version 3.0 correspond to which synset in version 2.0.

4.2.1 Vocabulary analysis

To be able to treat WordNet as a SKOS vocabulary, we use a simple schema mapping: WordNet synsets are mapped to SKOS concepts, WordNet sense labels to SKOS altLabels and WordNet glosses to SKOS definitions. WordNet 2.0 consists of 115,424 concepts with a total of 203147 labels. WordNet 3.0 has been extended, counting 117,657 concepts with 206,976 labels.

Based on the similarity of these numbers, and the fact that WordNet maintenance is largely a manual effort, we expect to be able to automatically map a large set of concepts relatively easily. Concepts that we will choose to leave unmapped are those 2.0 synsets that have been dropped in the new version, without having a counterpart in the new version and the 3.0 synsets that are newly added without having a counterpart in the old version. Concepts that we would like to map but could be hard to do automatically include concepts that have splitted or merged between versions, and concepts of which so many properties have changed that it is hard to tell if we are dealing with the "same" concepts or not.

Both vocabularies are splitted into nouns (70%), verbs (12%), adjectives (15%) and adverbs (3%). We assume that by mapping only nouns to nouns, verbs to verbs, etc. we can both reduce the search space and avoid many erroneous mappings between homonyms in different parts of speech. This approach risks missing concepts that moved to another part of speech category, but we assume this to occur very infrequently or not at all.

4.2.2 Interactive alignment

When aligning WordNet 3.0 to 2.0 we would like to explicitly use our knowledge the fact that we are aligning two versions of the same vocabulary. For example, given the large amount of homonymy, we expect a simple label match to produce many mappings, most of which will be wrong. In contrast, we expect the definitions to be unique for most concepts, and since manually updating many definitions is hard manual work, we expect the majority of the concepts to have the same definition in both versions.

So as a first step, we try a quick case insensitive match on skos:definition. Selecting only the 1 to 1 mappings results leaves us with 103.531 mappings (set 1a), covering already 89.7% of all 2.0 synsets. Of the n to m mappings, 922 can be reduced to a 1-1 mapping (set 1b) by simply matching also the labels. We quickly evaluate the remaining 24 mappings (set 1c) manually, and conclude these are all cases where two synsets from the old version have been merged into one synset in the new versions, so all the remaining mappings turn out the be correct too. After this simple first step, we need to align less than 10% of the original number of concepts, and can afford more expensive techniques in the following steps.

As a second step, we run a quick case insensitive label match on the remaining concepts. This yields another 4095 1-1 mappings (set 2a), which we assume to be mostly correct. As expected, it also results in a relatively large number (6006) of n-m mappings, of which we expect many will be wrong homonym matches. To filter out those, we first run a cheap disambiguation technique: just selecting the alternative with the most matching labels yields 2451 1-1 mappings (set 2b). We disambiguate the remaining 3446 n-m mappings by running a more expensive string distance matcher on the definitions, this yields another 2557 1-1 mappings (set 2c). The remaining 211 we subject to structural matching techniques. By looking if broader, narrower or related terms have already been matched, we are able to disambiguate 21 more concepts (set 2d). We manually evaluated a sample of the remaining 121 n-m mappings and concluded that the overwhelming majority is wrong, as most of them are cases of concepts that are new in WordNet 3.0 that have been wrongly mapped to WordNet 2.0 homonyms.

4.2.3 Results

We have created three distinct subsets of mappings in the first step and four subsets in the second step. Together, these seven sets consist of 113,599 mappings for an equal number of WordNet 2.0 concepts, covering 98.42% of all 2.0 synsets. For each subset, we can easily describe how the mappings have been created, and why we would or would not trust the mappings they contain. A more thorough manual evaluation could take this into account, by taking strategic samples from each subset. The coverage can be further increased by trying to map concepts for which (all) the labels have been changed between versions, as happens when spelling errors are detected or new spelling conventions are applied, but this is out of scope for this paper.

4.3 Multilingual: GEMET to AGROVOC

The use case described in this section exploits the multilinguality of the thesauri that are to be aligned. The evidence from multiple language-specific matching techniques can be combined to achieve a higher quality mapping between these thesauri. For this use case, we only used simple label matching algorithm and did not consider any other properties or structural information.

In this use case, the source vocabulary is the GEneral Multilingual Environmental Thesaurus (GEMET), a thesaurus comprising environmental information. It has been developed as an indexing, retrieval and control tool for the European Topic Centre on Catalogue of Data Sources (ETC/CDS) and the European Environment Agency (EEA), Copenhagen.

Our source vocabulary is the AGROVOC thesaurus. AGROVOC is a multilingual thesaurus describing

concepts from the fields of agriculture, forestry, fisheries, food and related domains (e.g. environment). It was created and is maintained by the Food and Agriculture Organization of the United Nations (FAO).

4.3.1 Vocabulary analysis

We used SKOS versions of both the GEMET and the AGROVOC thesaurus for our experiment. GEMET is considerably smaller than AGROVOC, with respectively 5,244 and 28,953 concepts. These concepts are described by labels in multiple languages. A difference between the two thesauri is that GEMET does not make use of alternative labels, where AGROVOC does use both preferred and alternative labels.

GEMET's concepts have a combined total of 109,612 labels which are spread over 21 languages[6]. Most concepts have labels in either 20 or all of the languages.

AGROVOC has 28,953 concepts with in total 329,932 preferred and 146,791 alternative labels. The version we used has labels in 16 different languages[7] but not all concepts have labels in each of languages. A concept has on average about 11.4 different languages for preferred and 5 different languages for alternative labels.

4.3.2 Interactive alignment

In this use case, we show how by combining evidence from multiple matchers, a higher mapping accuracy can be achieved. We exploit the fact that the source and target thesuari have labels in multiple common languages. In this case, GEMET and AGROVOC have 10 languages in common. Therefore, in the first step, we run ten language-specific matchers on the labels (for AGROVOC we use both alternative and preferred labels). This results in ten different mappings, one containing pairs of concepts whose German labels match, one for concepts whose Hungarian labels match etcetera.

We are not specifically interested in language-specific matches but rather in how multiple sources of evidence result in higher quality mappings. We therefore merge the mappings, combining the evidence in for the language-specific matches. In total, the merged mapping consisted of matches for 2323 source concepts.

The merge graph was then split into strata. A Stratum X is defined as the set of matches for which we have evidence from exactly X languages (or methods). For example, Stratum 2 contains a match between concepts for which only the German and Spanish label matches as well as concepts for which the French and English label matches.

The result is ten strata, for which we show the number of concepts matched is shown in Table 3.

Results

To analyze the quality of the different strata, we performed an evaluation of a sample for each of the ten strata. The samples were evaluated by a domain expert. Each of the strata was sampled. For stratum 1 we randomly selected 40 random matches, for stratum 2-9 we extracted 20 random matches and for stratum 10 we used all 8 matches. The evaluation was performed using Amalgame's built in evaluation tool. For each match, the evaluator was asked to indicate whether the concepts should be considered SKOS exactMatch, closeMatch, broaderMatch, narrowerMatch or related. The concept pair could also be evaluated as being

[6]Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Basque, Finnish, French, Hungarian, Italian, Dutch, Norwegian, Polish, Portuguese, Russian, Slovak, Slovenian and Swedish

[7]Arabic, Czech, German, English, Spanish, Persian, Hindi, Hungarian, Italian, Japanese, Laothian, Polish, Portuguese, Slovak, Thai and Chinese

Stratum	Source concepts matched
1	725
2	341
3	304
4	295
5	240
6	196
7	185
8	193
9	135
10	8

Table 3: Alignment result, listed per stratum

unrelated. Finally an option “unsure” was available. After the evaluation experiment we found that four matches that were left unevaluated. These were then evaluated by a second evaluator.

The results of the evaluation can be found in Figure 3. Here the percentages are shown of the match type indicated by the evaluator. The figure shows that the quality of the matches increases for higher strata. For Stratum 1, seven out of 40 matches are evaluated as being an exact match(17.5%), whereas for stratum 2, this is already 12 out of 20 (60%) and for higher strata, this is even higher. If we consider both exact and close matches to be “correct”, the precision for Stratum 1, 2 and 3 is 27.5%, 75% and 90% percent respectively. For higher strata, this precision stays above 90%. The drop in precision for stratum 10 in this respect can only be explained by the fact that the sample size is too small.

To verify the quality of the evaluation, we had a second person evaluate six strata (1-5 and 9). The evaluators agree on 68% of the matches. For exact matches only, the inter-annotator agreement is 81%.

If we select only mappings from Stratum 2 or higher, we end up with 1897 matches and an projected precision of 91% (when we consider both exact and close matches as “correct”). If we only use matches from Stratum 3 or higher, these numbers rise to 1556 and 95%.

The analysis and evaluation of the strata shows that combining evidence from multiple sources has a huge positive effect on the precision of the produced mapping. We here use different language matchers as sources of evidence, but using the Amalgame tool, we can combine label match techniques with structure-based approaches, matches on (scope)notes etcetera.

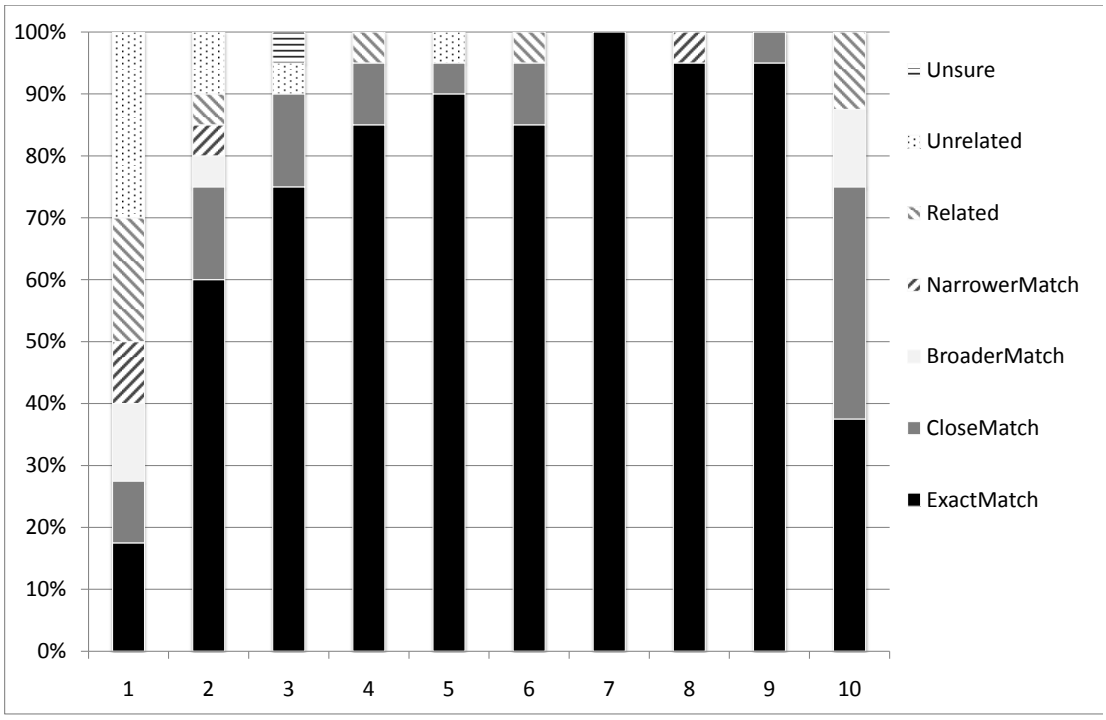

Figure 3: Percentages of evaluated match types for each of the ten strata in the GEMET-AGROVOC alignment task. The sample size for Stratum 1 is 40, for strata 2-9 is 20 and for Stratum 9 is 8 alignments.

5 Discussion

We conclude it is feasible to construct an alignment workflow for relatively large SKOS-like vocabularies by combining simple techniques. With the prior knowledge of the vocabularies and analysis of the correspondences we iteratively increased recall and precision. The resulting mappings are comprised of multiple homogeneous subsets of correspondences. This allows for targeted evaluation per subset. In addition, this allows to combine evidence from multiple subsets to increase precision, or strategically select multiple subsets to increase recall.

A potential drawback of our approach is that the selection, configuration and combination of components is the responsibility of the user. This makes the approach less attractive were fully automatic approaches produce results of sufficient quality. A potential risk is that we assume a finite and relatively small set of basic components. Amalgame currently provides a number of such components, some of these were used across use cases. During the specific use cases, however, we also found a need for additional components. Creating these components was straightforward. New use cases might require new components as well.

The workflows for the use case presented in this paper were created by the authors, using an experimental interface. Our longer term goal is to support vocabulary owners to create their own alignments. This requires a user interface to iteratively construct alignment workflows. Currently we are developing such a user interface. The interface will combine the construction of a workflow, with the analysis of mappings. In such an interface the user will construct, for example, the workflow shown in Figure 4 as a step-by-step process. Thus, each time extending a single node and using the analysis tools to investigate intermediate results. In future work we will evaluate such an interface with the vocabulary owners.

All mappings produced in the use are available[8], along with all information required to replicate the construction of these mappings. All software is available as open source, and the vocabularies are publicly available.

[8] http://semanticweb.cs.vu.nl/lod/prestoprimeD421/

Acknowledgements

We thank Carmen Reverté Reverté for her assistance and evaluation in the GEMET-AGROVOC use case. Mark van Assem produced the RDF conversions for WordNet 2.0 and 3.0.

References

[1] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg (DE), 2007.

[2] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. Language, Speech, and Communication Series. MIT Press, 1998.

[3] A. Miles and S. Bechhofer. Skos simple knowledge organization system reference. W3C Recommendation, August 18 2009.

[4] A. Tordai, J. van Ossenbruggen, and G. Schreiber. Combining vocabulary alignment techniques. In *K-CAP '09: Proceedings of the fifth international conference on Knowledge capture*, pages 25–32, New York, NY, USA, 2009. ACM.

[5] A. Tordai, J. R. van Ossenbruggen, A. Ghazvinian, M. A. Musen, and N. F. Noy. Lost In Translation? Empirical Analysis Of Mapping Compositions For Large Ontologies. In *Proceedings of International Workshop on Ontology Matching 2010 (5)*. CEUR-WS, November 2010.

[6] A. Tordai, J. R. van Ossenbruggen, G. Schreiber, and B. Wielinga. Aligning Large SKOS-Like Vocabularies. In *Proceedings of European Semantic Web Conference 2010 (7)*, Lecture Notes in Computer Science, pages 198 – 212. Springer, May 2010.

[7] M. van Assem, A. Gangemi, and G. Schreiber. Conversion of WordNet to a standard RDF/OWL representation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006.

[8] W. R. van Hage, A. Isaac, and Z. Aleksovski. Sample evaluation of ontology-matching systems. In R. Garcia-Castro, D. Vrandecic, A. Gómez-Pérez, Y. Sure, and Z. Huang, editors, *EON*, volume 329 of *CEUR Workshop Proceedings*, pages 41–50. CEUR-WS.org, 2007.

[9] P. Vossen, I. Maks, R. Segers, and H. van der Vliet. Integrating lexical units, synsets and ontology in the Cornetto database. In E. L. R. A. (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, 2008.

Appendix: Workflow GTAA-Cornetto

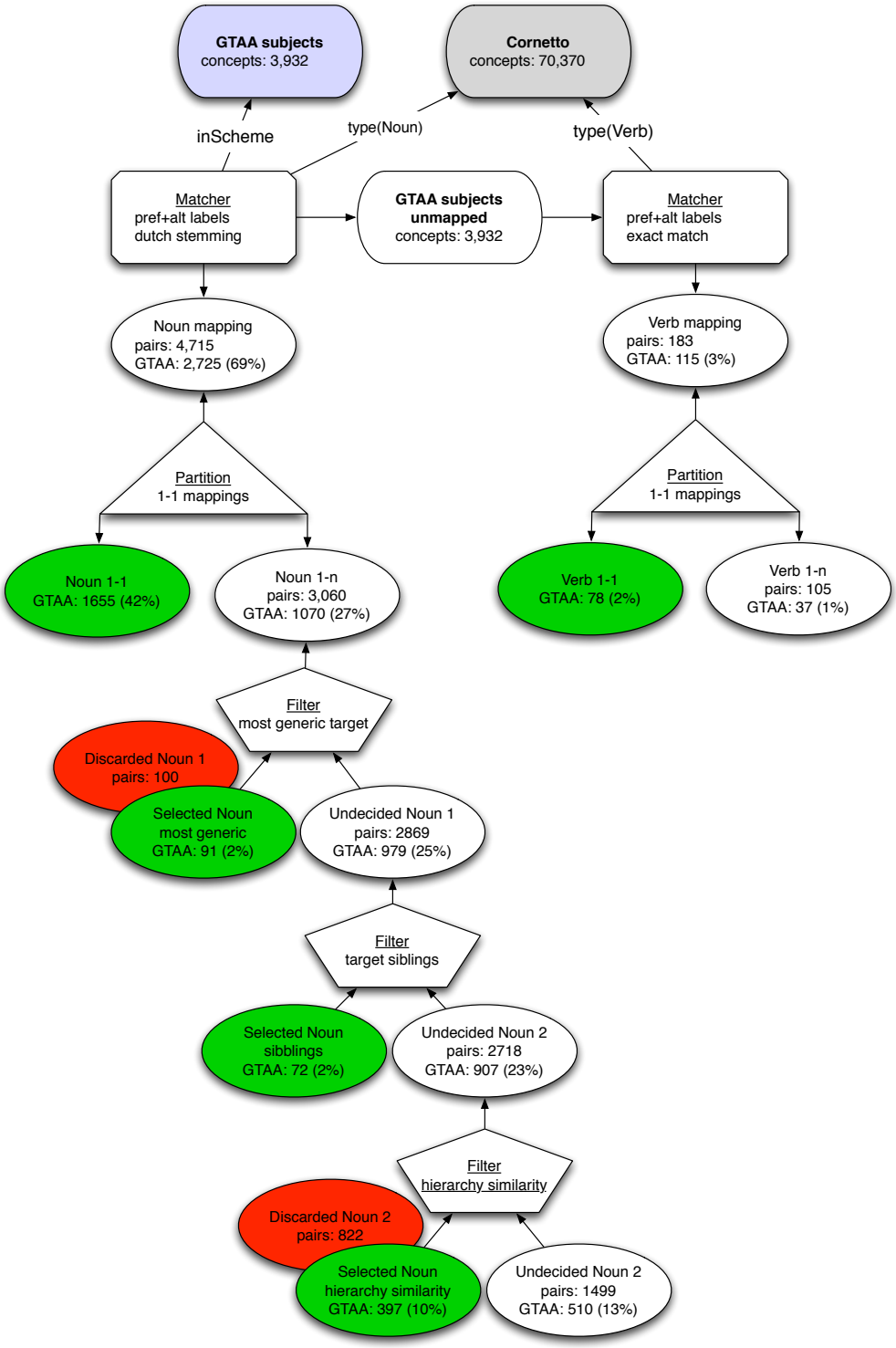

Figure 4: Workflow of the GTAA Cornetto alignment.