Annotating a parallel monolingual treebank with semantic similarity relations

Erwin Marsi and Emiel Krahmer
Tilburg University
Dept. of Communication and Information

Abstract

We describe an ongoing effort to build a large-scale parallel/comparable monolingual treebank for Dutch of 1 million words, where nodes of dependency trees are aligned and labeled according to a limited set of semantic similarity relations. We address alignment of sentences and dependency trees, both manual and automatic. We introduce new annotation tools, present results from pilot experiments, and discuss complications. We discuss applications in multi-document summarization, question-answering and paraphrase extraction.

1 Introduction

Treebanks of syntactically annotated sentences have become a core part of computational linguistics and many related areas. Not only for developing and systematically validating computational models of syntax, but also for data-driven development of natural language processing tools such as part-of-speech taggers, chunkers and parsers. In a similar vein, large *parallel corpora* of *bilingual* text have become an essential ingredient of statistical and example-based machine translation. Typically, the text material in a bilingual parallel corpus is aligned at the level of sentences, words or arbitrary substrings. Convinced of the need for more syntactic structure, several researchers have explored *parallel treebanks* with aligned phrase-structure trees or dependency structures (see e.g., Gildea 2003; Samuelsson and Volk 2006).

A similar type of corpora, parallel corpora of *monolingual* text, have proved to be useful for automatic extraction of synonyms and paraphrases, which in turn have a wide range of applications from machine translation to information retrieval. This has also inspired work on *comparable corpora* of loosely associated text, e.g., entries from different encyclopedia on the same topic (Barzilay and Elhadad 2003).

A logical combination of these two trends – parallel bilingual treebanks on the one hand and monolingual parallel/comparable text corpora on the

De Smedt, K., Hajič, J. and Kübler, S. (Eds.)
Proceedings of the Sixth International Workshop
on Treebanks and Linguistic Theories (2007)

other – gives rise to the notion of *a parallel monolingual treebank*, which we define as a corpus of parallel/comparable text in the same language with aligned parse trees. It seems that so far no published research has addressed this idea (although (Ibrahim, Katz, and Lin 2003) comes close). In our opinion parallel monolingual treebanks hold great potential, not only for paraphrasing, but also in general for studying the mapping from meaning to alternative surface realizations, and in many NLP applications such as multi-document summarization, question answering and recognizing textual entailment. We will elaborate on applications in Section 5.

A second idea we would like to introduce here is that of *typed alignment relations*. The notion of a parallel treebank implies alignments between structural units like words, phrases or tree nodes. In fact, this makes the alignments to some extend similar to the dependencies among words in a syntactic dependency structure. However, in contrast to dependencies, which are normally typed in terms of a particular set of dependency relations, alignments are unlabeled. We propose an extension to alignments typed in terms of a limited set of *semantic similarity relations* such as "X specifies Y" or "X generalizes Y". We think that such an extension has many interesting theoretical and practical implications, some of which are discussed in Section 5.

In this paper we describe an ongoing effort within the context of the DAESO project[1] to build a large-scale parallel/comparable monolingual treebank for Dutch of 1 million words, where nodes of dependency trees are aligned and labeled according to a limited set of semantic similarity relations. We will first describe the text material and syntactic annotation. Next, we will discuss alignment at the sentence level, both automatic and manual, followed by alignment of dependency trees. We will introduce some newly developed annotation tools, review some of the results from pilot experiments on annotation, and share our experiences so far in building a large-scale corpus. We finish describing some tools and applications we intend to address in future work.

2 Corpus material and annotation

2.1 Text material

The corpus contains written Dutch text from five different sources, ranging from true parallel text to loosely associated comparable text. The target size is 1 million words, half of which will be processed with partly manual annotation and correction, whereas the other half will be processed fully automatically. We limit ourselves to the first half here. The composition of the corpus results from a trade-off between several constraints: (1) intended

[1] Detecting And Exploiting Semantic Overlap – see http://daeso.uvt.nl

coverage of different text styles; (2) availability in electronic format; (3) targeted applications; (4) strong requirements regarding copyright of the texts imposed by the research funder.

Book translations Parallel text comes from alternative translations of the same book (125k words). The corpus includes two Dutch translations from (parts of) each three books: (1) “Le Petit Prince” by Antoine de Saint-Exupéry, (2) “On the Origin of Species” by Charles Darwin in the 1st and 6th edition and (3) “Les Essais” by Michel de Montaigne. Although the original books are quite old, we use modern translations.

Autocue-subtitle pairs This material comes from the *NOS journaal*, the daily news broadcast by the Dutch public broadcasting channel, and consists of the autocue text as read by the news reader and the associated subtitles (125k words). It was collected, tokenized and aligned at the sentence level in the ATRANOS project (Daelemans, Höthker, and Sang 2004). Because of space constraints, the subtitles typically present a compressed form of the autocue.

News headlines Our third text source consists of headlines of clustered news articles which were automatically mined from the Dutch version of Google News (25k words). As the clustering is based on the full article rather than only the head, we found substantial differences between headlines, so manual subclustering was required in order to get parallel sentences.

QA-system output For future work aimed at Question-Answering (QA - see Section 5), the corpus also contains samples from the QA domain. The IMIX project has developed a multimodal question-answering system in the medical domain (Theune, van Schooten, op den Akker, Bosma, Hofs, A.Nijholt, Krahmer, van Hooijdonk, and Marsi 2007). Questions are answered by searching a large collection of text ranging from medical encyclopedia to layman websites. In order to evaluate the QA engines, a reference corpus of questions and associated answers as encountered in the available texts was manually compiled. From this corpus, we extracted clusters of two or more alternative answers. With about 1k words, this segment is relatively small.

Press releases The final source delivers comparable text in the form of press releases about the same news topic obtained from ANP and Novum, two Dutch press agencies (225k words). The selection of comparable press releases was initially automatic, aiming at a high recall at the expense of precision, and later on manually corrected.

2.2 Tokenization and syntactic parsing

All text material was tokenized using the Dutch tokenizer developed within the D-COI project (Reynaert 2007). We found that tokenization errors were more frequent in the book material, probably because of long and complex sentences. Since especially sentence-splitting errors will be fatal in the subsequent parsing step, and because tokenization errors are relatively cheap to fix, we undertook manually correction in this corpus segment.

Next, the Alpino parser for Dutch (Bouma, van Noord, and Malouf 2001) was used to parse all sentences. It aims at providing a relatively theory-neutral syntactic analysis as originally developed in the context of the Spoken Dutch Corpus (van der Wouden, Hoekstra, Moortgat, Renmans, and Schuurman 2002). It assigns dependency links to pairs of tokens, labeling them with dependency relations such as *head/subject*, *head/modifier* and *coordination/conjunction*.

Due to time constraints, parsing errors are not subject to manual correction. Evidently this will have a negative effect on the final step of dependency tree alignment, but to what extent remains to be seen.

3 Sentence alignment

Part of the material was already aligned at the sentence level: the autocue-subtitle segment was aligned within the ATRANOS project; the alternative answers from the QA reference corpus are implicitly aligned, and the same goes for all sentences in a subcluster of news headlines. Alignment of sentences was thus required for the book translations and for the press releases. This process took place in two steps: automatic alignment and subsequent manual correction.

3.1 Automatic alignment of parallel translations

Automatic alignment of sentences from parallel translations is a well-studied area for which a number of standard solutions are available, e.g., (Gale and Church 1993). It is usually assumed that the majority of the alignments is of the 1-to-1 type, and that crossing alignments and unaligned sentences are rare. We found these assumptions are frequently violated, for example, in the two translations of “On the Origin of Species”, where there are many differences due to Darwin’s own revisions. These range from added or removed text segments (the 6th edition even has a whole new chapter) to long sentences in one translation being split in multiple sentences in the other.

As the automatic alignment is manually corrected anyway, we opted for a fairly straightforward pragmatic approach that is nevertheless robust to above problems. It takes a sentence from the first translation and checks for all sentences in a sliding window over the second translation at approximately the same position whether two sentences are sufficiently similar to justify alignment. Similarity is defined in terms of n-gram overlap. We use a relatively low threshold to get a high recall at the expense of precision, because in practice manually deleting incorrect alignments takes less time than finding all correct ones.

Obviously, this approach is sensitive to gaps due to insertion/deletion of large text segments. We therefore found it beneficial to carry out alignment in multiple passes. That is, first align chapters, next align sections, then paragraphs and finally sentences.

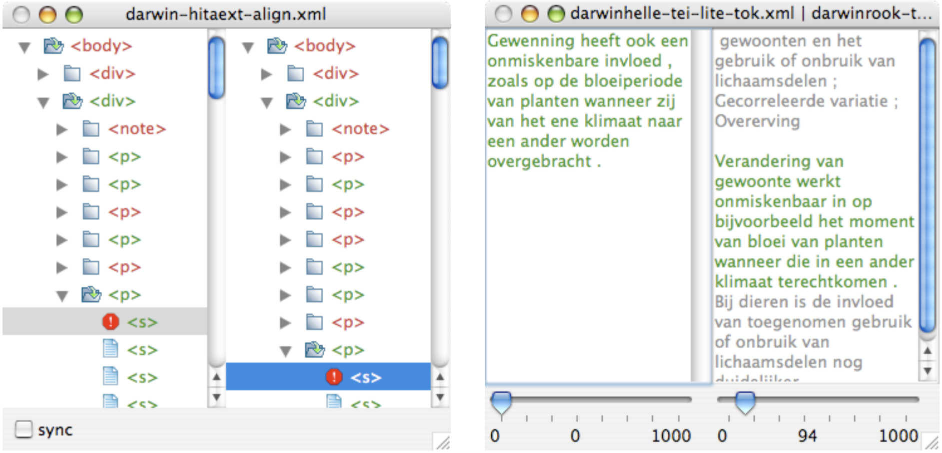
3.2 Manual correction of sentence alignments

We developed a special alignment annotation tool, called *Hitaext*, for visualizing and editing sentence alignments. In fact, Hitaext is a general graphical tool for aligning text elements in pairs of arbitrary XML documents. Distinguishing features of Hitaext in comparison with other text alignment tools are its lack of a predefined input format and its support for simultaneous alignment at arbitrary annotation levels (e.g. words, sentences, paragraphs, or chapters). It takes as input a pair of marked-up texts in the form of XML documents. A third XML file contains the alignments (possibly none yet) as well as a simple style sheet for rendering the text. Hitaext provides two different views on the input documents: the tree view and the text view.

The *tree window* – see the left side of Figure 1 – visualizes the hierarchical structure of the XML elements in the form of two parallel tree controls (or tree widgets). These allow a user to walk through the XML elements using mouse and/or keyboard. In our case, the documents are typically TEI XML documents, and the elements correspond to chapters, section, paragraphs and sentences. Large documents remain manageable because a user can expand or collapse arbitrary parts of the tree. In addition, irrelevant elements can be hidden by configuring the style sheet.

The *text window* – see the right side of Figure 1 – shows the two pieces of text corresponding to the two elements currently focused (or selected) in the tree window. These are typically the texts of sentences, paragraphs, chapters or even whole documents (when the focus in on the root node). The sliders at the bottom of the text window allow one to reveal a variable amount of the surrounding text.

If an element is aligned, its tag is shown in green in the tree window and its corresponding text is shown in green in the text window. Conversely, unaligned tags and texts are shown in red. The current focus is always an element in the tree window, either in the left or right tree. If this focused element has alignments, the aligned elements in the other tree are marked by means of an exclamation mark icon. This way Hitaext facilitates 1-to-1, 1-to-n and n-to-m alignments. By selecting two elements from either tree



Gewenning heeft ook een onmiskenbare invloed , zoals op de bloeiperiode van planten wanneer zij van het ene klimaat naar een ander worden overgebracht .

gewoonten en het gebruik of onbruik van lichaamsdelen ; Gecorreleerde variatie ; Overerving

Verandering van gewoonte werkt onmiskenbaar in op bijvoorbeeld het moment van bloei van planten wanneer die in een ander klimaat terechtkomen . Bij dieren is de invloed van toegenomen gebruik of onbruik van lichaamsdelen nog duidelijker

Figure 1: Screen shot of Hitaext, the tool used for aligning text segments

and subsequently hitting the space bar, a user can toggle (switch on/off) the alignment between them.[2].

3.3 Alignment of comparable text

The assumptions about parallel text clearly do not hold in the case of comparable text: 1-to-many alignments – or even many-to-many – are to be expected, just as crossing alignments and large portions of unaligned material. Moreover, similarity between sentences, and therefore the decision to align or not, is much more gradient. Whereas with parallel translations it is virtually always evident whether or not two sentences are translations of the same source sentence, it turns out to be much harder to decide whether two comparable sentences are sufficiently similar to justify alignment. We have started manual alignment of comparable text and are in the process of developing annotation guidelines. Below we share some of our considerations.

There is no need for aligned sentences to be paraphrases of each other. One sentence may contain additional information which is not present in the other. Likewise, information in one sentence may be more specific/general than in the other. However, aligned sentences should have at least one proposition in common. We interpret this notion loosely as a statement about someone or something. Examples of (partial) sentences including the same proposition:

- Balkenende is the minister-president of the Netherlands
- Balkenende, who is the minister-president of the Netherlands, ...

[2] Hitaext is implemented in wxPython, runs on Mac OS X, Linux and Windows, and is released as open source software from http://daeso.uvt.nl/hitaext

- Balkenende, the minister-president of the Netherlands ...
- Balkenende as the minister-president of the Netherlands ...
- Balkenende being minister-president of the Netherlands ...

However, the following examples do not (although they may share some phrases):

- Balkenende is a wine expert
- Balkenende likes to barbecue
- Bush likes to barbecue
- the minister-president of the Netherlands
- the capital of the Netherlands

We do not attempt to align each sentence with *the* most similar sentence (one-to-one alignment). Instead, we align a sentence to every other sentence with which it has at least one proposition in common, effectively creating one-to-many alignment.

Finally, we allow use of common sense. Consider the following pair:

- Keith Urban left US rehabilitation clinic
- Keith Urban cured from addiction

In the strict logical sense, these statements differ: in theory, one may leave the clinic without being cured, or one may be cured but remain in the clinic. However, in the context of two texts on the same topic, we prefer to view them as identical for all practical purposes. This is in the same spirit as *natural entailment* is defined in the Recognizing Textual Entailment task (Dagan, Glickman, and Magnini 2005). Other examples include approximately identical locations, quantities, times, etc. In a similar vein, referring expressions like pronouns or generic definite descriptions may be interpreted in the context.

However, we refrain from alignment in cases where inferring similarity requires elaborate reasoning and background knowledge, as in:

- The Radicals now hold 80 of the 250 seats in parliament
- The SRS is currently the biggest party in Serbia

Notice that this would requires one to know that 80 out of 250 is a majority because all other political parties are smaller.

We carried out a pilot experiment with two annotators who each aligned the same 10 pairs of comparable press releases, varying in length from 4 tot 33 sentences. The total number of possible one-to-one alignments to consider was 1492. Both annotators agreed on 44 but disagreed on 32 alignments. While discussing the differences, we encountered some difficult cases which gave rise to revision of the annotation guidelines. However, it turned out that the majority of the disagreements were due to the fact that an annotator overlooked a particular alignment.

After revision of the guidelines, we repeated the experiment with a different set of 10 comparable press releases, this time with 1337 possible alignments to consider. Agreement was 51, while one annotator made 15 unique alignments and the other 39. This time by far the most disagreements were caused by missed alignments. This confirms our impression that the task of identifying *all* pairs of similar sentences is harder than deciding on similarity of a given sentence pair, causing the precision to be substantially better than the recall.

4 Dependency tree alignment

4.1 Semantic similarity relations

Dependency tree alignment can be described informally as: given two dependency analyses, align those nodes that are semantically related. More precisely: for each node v in the dependency structure for a sentence S, we define $\textsc{str}(v)$ as the substring of all tokens under v (i.e., the composition of the tokens of all nodes reachable from v). For example, the string associated with node *persoon* in the left dependency structure in Figure 2 is *heel veel serieuze personen* ('very many serious persons'). An alignment between sentences S and S' pairs nodes from the dependency trees for both sentences. Aligning node v from the dependency tree D of sentence S with node v' from the tree D' of S' indicates that there is a semantic similarity between $\textsc{str}(v)$ and $\textsc{str}(v')$.

We distinguish five potential, mutually exclusive, similarity relations between nodes, with illustrative examples from "Le Petit Prince":

1. v **equals** v' iff $\textsc{str}(v)$ and $\textsc{str}(v')$ are literally identical (abstracting from case). Example: "a small and a large boa-constrictor" equals "a large and a small boa-constrictor";
2. v **restates** v' iff $\textsc{str}(v)$ is a paraphrase of $\textsc{str}(v')$ (same information content but different wording). Example: "a drawing of a boa-constrictor snake" restates "a drawing of a boa-constrictor";
3. v **specifies** v' iff $\textsc{str}(v)$ is more specific than $\textsc{str}(v')$. Example: "the planet B 612" specifies "the planet";
4. v **generalizes** v' iff $\textsc{str}(v')$ is more specific than $\textsc{str}(v)$. Example: "the planet" generalizes "the planet B 612";
5. v **intersects** v' iff $\textsc{str}(v)$ and $\textsc{str}(v')$ share some informational content, but also each express some piece of information not expressed in the other. Example: "Jupiter and Mars' intersects "Mars and Venus"

In interpreting these relations we adhere to the same principles we discussed in the previous Section on alignment of comparable sentences, for instance, use of common sense is allowed.

4.2 Manual alignment of dependency nodes

For creating manual alignments, we developed a special-purpose annotation tool called *Gadget* ('Graphical Aligner of Dependency Graphs and Equivalent Tokens'). It shows, side by side, two sentences, as well as their respective dependency graphs. When the user clicks on a node v in the graph, the corresponding string ($\textsc{str}(v)$) is shown at the bottom. Alignment takes place by selecting two nodes, followed by selection of the appropriate alignment relation. The tool offers additional support for folding parts of the graphs, highlighting unaligned nodes and hiding part-of-speech or dependency relation labels.[3] We recently discovered the Stocholm Tree Aligner (Volk, Gustafson-Capkova, Lundborg, Marek, Samuelsson, and Tidstrom 2006) which is a similar tool intended for aligning bilingual treebanks in Tiger XML format using two types of alignments ("good" and "fuzzy").

In (Marsi and Krahmer 2005a) we reported on a pilot experiment which involved aligning dependency trees using the first five chapters from "Le Petit Prince". Results indicated that humans can perform this task well, with an F-score of .98 on creating alignments and an F-score of .95 on assigning semantic similarity relations. We also presented results on automatic annotation, which achieved an F-score on alignment of .85 and an F-score of .80 on semantic relation classification (assuming some prior knowledge). Our corpus allows us to repeat these experiments on larger scale and with more challenging text material.

5 Tools and Applications

While we are currently still building the corpus, this section outlines planned future work.

Tools for automatic alignment Apart from the two graphical alignment tools we will develop software to automatically align sentences from parallel and comparable text sources. Likewise, we will continue work on automatic alignment of dependency trees. Some initial work and results are described in (Marsi and Krahmer 2005b). Evidently, the corpus is an excellent resource for this.

[3] Gadget is implemented in wxPython, runs on Mac OS X, Linux and Windows, and will be released as open source software from from http://daeso.uvt.nl

Corpus expansion Once reliable tools for alignment are in place, we intend to double the size of our corpus to 1 million words by automatically aligning more book translations, news headlines, press releases and other sources.

Sentence fusion in multi-document summarization We intend to evaluate the tools for automatic alignment in the context of a number of NLP applications, starting with multi-document summarization. Given a set of similar documents, a multi-document summarization system must first identify the most important sentences for inclusion in the summary. To avoid redundancy, the system must detect similar sentences, which amounts to the task of sentence alignment in comparable texts. Summarizers which attempt to produce real summaries – instead of merely abstracts – must also revise sentences, thereby removing irrelevant information and merging similar sentences. One can envision this as aligning and merging dependency trees, and subsequently generating revised sentences using techniques from Natural Language Generation – an approach called *sentence fusion* by Barzilay and McKeown (2005). Some of our initial work in this area is described in (Marsi and Krahmer 2005a). We intend to continue this in the context of multi-document summarization of press releases and news articles. In particular, we intend to take advantage of the semantic labeling of the alignments, which allows us to generate fused sentences which are more specific, equivalent or more general than the original ones.

Clustering answers in Question-Answering Question-Answering (QA) systems typically analyze a question, search for potential answers in a large body of text material, produce a list of potential answers ranked in order of decreasing likelihood, and show only the topmost answer to the user. For questions of the "open" type, like "What are the risks of overweight?", the topmost answer is unlikely the be optimal. On the one hand, it may be incomplete in the sense that it does not exhaustively list all the risks of overweight encountered in the full text collection. On the other hand, as it is a piece of text extracted from a particular context, it may contain additional information which is irrelevant to the question. We think that detecting and merging similar answers will lead to answers which are both more comprehensive and more to the point. The corpus segment containing QA answers is intended to facilitate initial work in this area.

Acknowledgments We would like to thank Jurry de Vos en Paul van Pelt for all the annotation work, Erik Tsjong Kim Sang and colleagues for the autocue-subtitle material from the ATRANOS project, Gosse Bouma for the QA reference corpus from the IMIX project, and Wauter Bosma for mining the headlines from Google News. This work was conducted within the DAESO project funded by the Stevin program (De Nederlandse Taalunie).

References

Barzilay, R. and N. Elhadad (2003). Sentence alignment for monolingual comparable corpora. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 25–32.

Barzilay, R. and K. R. McKeown (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics 31*(3), 297–328.

Bouma, G., G. van Noord, and R. Malouf (2001). Alpino: Wide-coverage computational analysis of Dutch. In W. Daelemans, K. Sima'an, J. Veenstra, and J. Zavre (Eds.), *Computational Linguistics in the Netherlands 2000. Selected Papers from the Eleventh CLIN Meeting*, pp. 45–59. Amsterdam, New York: Rodopi.

Daelemans, W., A. Höthker, and E. T. K. Sang (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pp. 1045–1048.

Dagan, I., O. Glickman, and B. Magnini (2005). The PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, Southampton, U.K.

Gale, W. A. and K. W. Church (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics 19*(1), 75–102.

Gildea, D. (2003). Loosely tree-based alignment for machine translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan, pp. 80–87.

Ibrahim, A., B. Katz, and J. Lin (2003). Extracting structural paraphrases from aligned monolingual corpora. In *Proceedings of the second international workshop on Paraphrasing*, Volume 16, Sapporo, Japan, pp. 57–64. ACL.

Marsi, E. and E. Krahmer (2005a, 8-10 August). Explorations in sentence fusion. In *Proceedings of the 10th European Workshop on Natural Language Generation*, Aberdeen, GB.

Marsi, E. and E. Krahmer (2005b). Semantic classification by humans and machines. In *ACL 2005 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan.

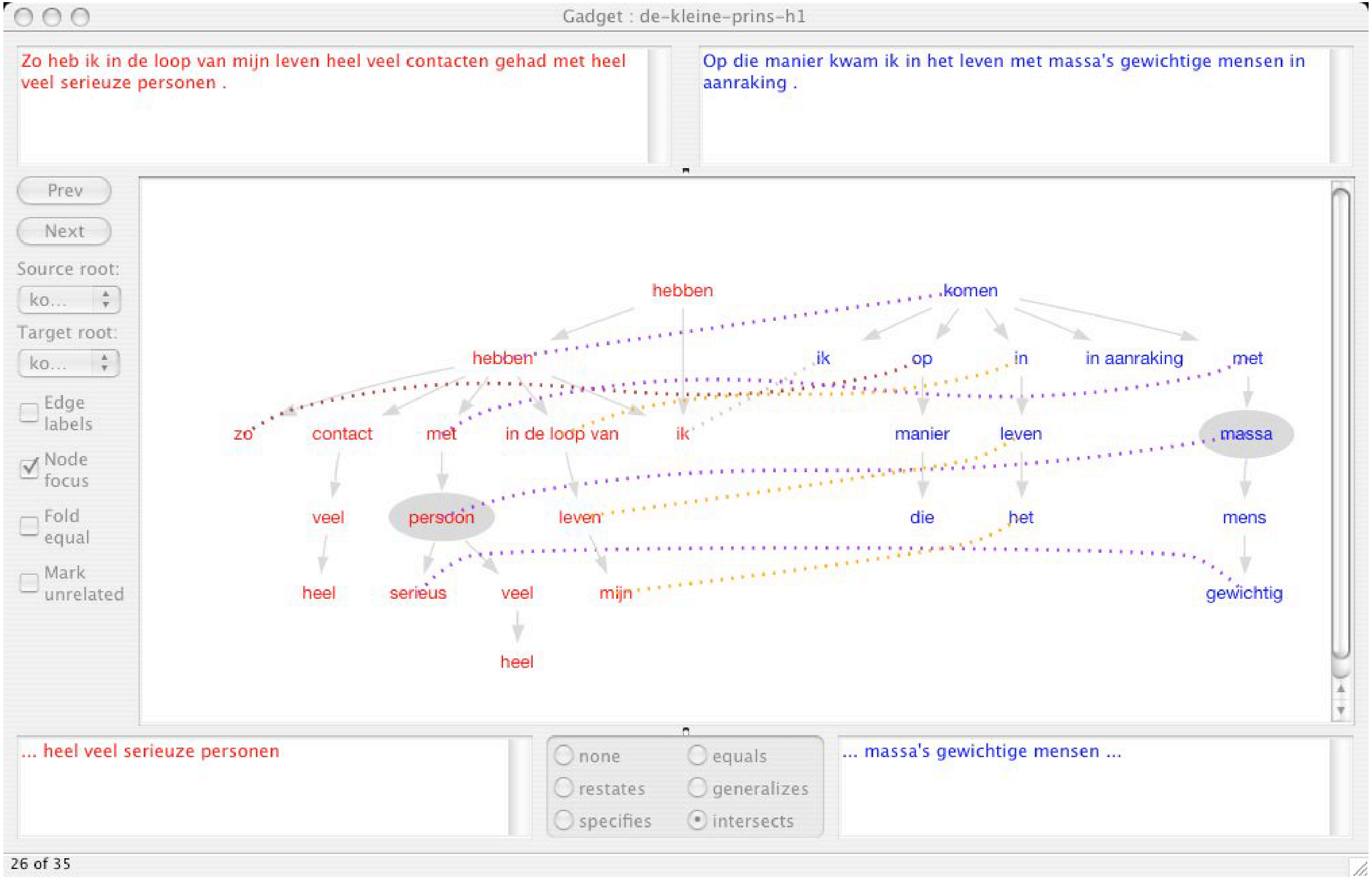
Reynaert, M. (2007). Sentence-splitting and tokenization in d-coi. Technical Report 07-07, ILK Research Group.

Samuelsson, Y. and M. Volk (2006). Phrase alignment in parallel treebanks. In *Proceedings of 5th Workshop on Treebanks and Linguistic Theories*, Prague, Czech Republik.

Theune, M., B. van Schooten, R. op den Akker, W. Bosma, D. Hofs, A.Nijholt, E. Krahmer, C. van Hooijdonk, and E. Marsi (2007). Questions, pictures, answers: Introducing pictures in question-answering systems. In L. R. Miyarez, A. M. Alvarado, and C. A. Moreno (Eds.), *ACTAS-1 of X Symposio Internacional de Comunicacion Social*, Santiago de Cuba, pp. 450–463.

van der Wouden, T., H. Hoekstra, M. Moortgat, B. Renmans, and I. Schuurman (2002). Syntactic analysis in the spoken dutch corpus. In *Proceedings of the third International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, pp. 768–773.

Volk, M., S. Gustafson-Capkova, J. Lundborg, T. Marek, Y. Samuelsson, and F. Tidstrom (2006). XML-based Phrase Alignment in Parallel Treebanks. In *Proceedings of EACL Workshop on Multidimensional Markup in Natural Language Processing*, pp. 93–96.



Gadget : de-kleine-prins-h1

Zo heb ik in de loop van mijn leven heel veel contacten gehad met heel veel serieuze personen .

Op die manier kwam ik in het leven met massa's gewichtige mensen in aanraking .

Source root: ko...

Target root: ko...

□ Edge labels

☑ Node focus

□ Fold equal

□ Mark unrelated

... heel veel serieuze personen

○ none ○ equals
○ restates ○ generalizes
○ specifies ● intersects

... massa's gewichtige mensen ...

26 of 35

Figure 2: Screen shot of Gadget, the tool used for aligning dependency structures of sentences. The two dependency structures shown are for the sentences *Zo heb ik in de loop van mijn leven heel veel contacten gehad met heel veel serieuze personen.* (lit. 'Thus have I in the course of my life very many contacts had with very many serious persons') and *Op die manier kwam ik in het leven met massa's gewichtige mensen in aanraking.*.