Lang Resources & Evaluation
DOI 10.1007/s10579-013-9252-1

ORIGINAL PAPER

Construction of an aligned monolingual treebank for studying semantic similarity

Erwin Marsi · Emiel Krahmer

Abstract Modern paraphrase research would benefit from large corpora with detailed annotations. However, currently these corpora are still thin on the ground. In this paper, we describe the development of such a corpus for Dutch, which takes the form of a parallel monolingual treebank consisting of over 2 million tokens and covering various text genres, including both parallel and comparable text. This publicly available corpus is richly annotated with alignments between syntactic nodes, which are also classified using five different semantic similarity relations. A quarter of the corpus is manually annotated, and this informs the development of an automatic tree aligner used to annotate the remainder of the corpus. We argue that this corpus is the first of this size and kind, and offers great potential for paraphrasing research.

Keywords Monolingual treebank · Semantic similarity · Corpus · Alignment · Semantic relations · Parallel text · Comparable text · Tree alignment · Paraphrase

Preliminary descriptions of the corpus were presented at the 18th Annual Meeting of Computational Linguistics in the Netherlands (CLIN 2007, Nijmegen, The Netherlands) and the 6th International Workshop on Treebanks and Linguistic Theories (TLT 2007, Bergen, Norway).

E. Marsi
Department of Computer and Information Science, Norwegian University of Science and Technology, Sem Sælands vei 7-9, 7491 Trondheim, Norway
e-mail: emarsi@idi.ntnu.no

E. Krahmer (✉)
Tilburg Center for Cognition and Communication (TiCC), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands
e-mail: e.j.krahmer@uvt.nl

Published online: 04 October 2013

1 Introduction

It is a well-known fact that the same meaning can be expressed in words in many different ways. Consider the following pair of Dutch sentences (with English translation):

(1) a. Dagelijks koffie vermindert risico op Alzheimer en dementie.
Daily coffee diminishes risk on Alzheimer and dementia
b. Drie koppen koffie per dag reduceert kans op Parkinson en
Three cups coffee a day reduces chance on Parkinson and
dementie.
dementia

Clearly these sentences are semantically similar, even though they do not express the exact same content.

This kind of semantic similarity is one of the major challenges in building robust natural language processing (NLP) systems. To mention just one example, consider the case of automatic summarisation. Such systems typically rank sentences according to their informativity and then extract the top *n* sentences, depending on the targeted compression rate. Although the sentences are essentially treated as independent of each other, they often are not. Extracted sentences may have substantial semantic overlap, resulting in unintended redundancy in the summaries. This is particularly problematic in the case of multi-document summarisation, where sentences extracted from related documents are very likely to express similar information in different ways (Radev and McKeown 1998).

Imagine a set of documents addressing the link between caffeine intake and neurological disorders. One document may contain sentence (1a), another sentence (1b).[1] A summariser might decide to select both sentences for the summary, but this would be unfortunate. It would be better to merge these two sentences, since they are semantically related. This would essentially allow the summariser to include an additional sentence in its summary, thereby increasing its informativeness. In general, automatic summarisers can be improved with knowledge of how different natural language expressions relate to each other, for instance, in terms of paraphrases or entailments, and the same can be said about many other NLP applications. It is therefore hardly surprising that both paraphrasing and entailment have received considerable attention in recent years (see Madnani and Dorr 2010; Androutsopoulos and Malakasiotis 2010 for recent surveys).

The preferred approach is a data-driven one, in order to avoid the knowledge acquisition bottleneck associated with manual construction of paraphrase resources such as thesauri and paraphrase rules. Generally, this requires a text corpus consisting of pairs of sentences expressing similar information, either monolingual or bilingual. Such paraphrase corpora are used directly to induce algorithms for detecting, extracting or generating paraphrases or indirectly to evaluate performance of such algorithms. Madnani and Dorr (2010:377) stress that "(i)n general, building paraphrase corpora,

[1] Note that the sentences in Example (1) were constructed for illustrative purposes, even though, for example, caffeine intake has been shown to reduce the risk of Parkinson disease (Ross et al. 2000).

whether it is done at the sentence level or at the sub-sentential level, is extremely useful for the fostering of further research and development in the area of paraphrase generation." However, as emphasised in another recent paraphrasing survey by Androutsopoulos and Malakasiotis (2010:154), there is a general lack of such corpora, and the few that exist tend to be small and only lightly annotated.

In this article we present a new, large scale and publicly available paraphrase corpus with detailed annotation of semantic similarity. It contains parallel and comparable Dutch text from several text domains. Similar text spans are aligned at the level of sentences, phrases and words. As this takes the form of pairs of aligned syntactic trees, the corpus can be regarded as a *parallel monolingual treebank*. The alignments not only concern *paraphrases* in the strict sense, i.e., expressions that are semantically equivalent, but extend to expressions that are semantically similar in less strict ways, for instance, where one phrase is either more specific or more general than a related phrase. All alignments are therefore also labeled according to a limited set of semantic similarity relations. Figure 1 contains an example of the aligned and labeled syntactic trees corresponding to Example (1). For expository reasons not all node alignments are shown here. Notice, for instance, that nodes containing unique information, such as *Alzheimer* and *Parkinson*, remain unaligned. The corpus comprises over 2.1M tokens, 678K of which is manually annotated and 1,511K is automatically processed. We argue that this corpus is unique in its size and detailed annotations, and has great potential for paraphrasing research.

The structure of this article is as follows. Section 2 presents our analysis of semantic similarity between sentences by aligning their syntax trees and labelling the alignments with similarity relations. It introduces semantic similarity relations along with alignment examples and provides a formal definition of a labeled tree alignment. Section 3 describes the collection of the text material, linguistic preprocessing and alignment at the sentence level. Section 4 concerns the manual alignment process, followed by a discussion of the characteristics of the resulting corpus. Inter-annotator agreement is evaluated in Sect. 5. Section 6 describes the automatic alignment of the second part of the corpus. In Sect. 7 related work is discussed, focussing on corpora for paraphrasing and the use of semantic relations, as well as a discussion of possible applications of the present corpus. Section 8 concludes the article.

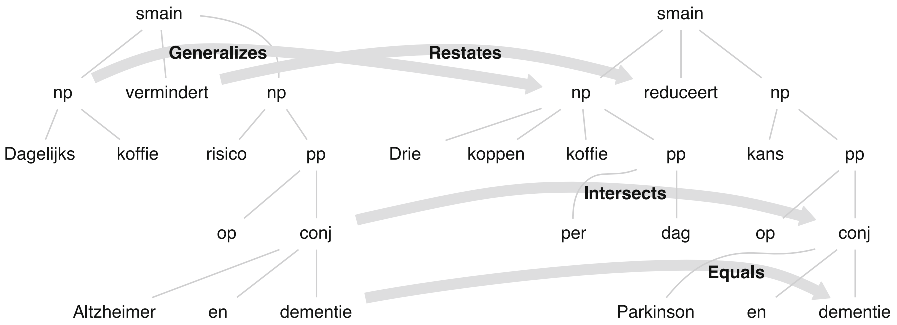


Fig. 1 Example of two aligned and labeled syntactic trees

2 Analysing semantic similarity

Analysis of semantic similarity can be approached from many different angles. A basic approach is to use string similarity measures such as the Jaccard similarity coefficient to measure the amount of overlap in words or the Levenshtein distance for the number of insert/delete/replace operations required to transform one expression into the other (Navarro 2001). Although cheap and fast, such approaches fail to account for less obvious cases such as synonyms or syntactic paraphrasing. At the other extreme we find deep semantic analysis and formal reasoning to infer a meaning relation between linguistic expressions (Bos and Markert 2005). This approach tends to suffer from the lack of coverage and robustness commonly associated with deep linguistic processing. We therefore think that the middle ground between these two extremes has the greatest potential. We propose to analyse semantic similarity between sentences by aligning their syntax trees, where each node is matched to the most similar node in the other tree (if any). In addition, we label these alignments according to the type of similarity relation that holds between the aligned phrases. Tree alignment and semantic similarity relations will be described in detail in the two subsections below.

2.1 Tree alignment

Given a pair of comparable sentences, our objective is to analyse the semantic similarity between these sentences as well as between their parts, down to the level of similar words. In the analysis of similarity, we prefer to stay true to the textual form in the sense that no abstractions such as semantic, logical or conceptual representations intervene. We thus want to directly relate word sequences in one sentence to their most similar word sequence in the other sentence.

However, considering all possible subsequences is impractical for humans and computationally infeasible for machines, particularly for long sentences. Syntactic information is therefore used to constrain the number of possible subsequences to those that form coherent syntactic units. This avoids the need to consider relatively incoherent word sequences such as “in a heavy” or “ therefore it”. We employ *syntactic phrase structure trees* for this purpose, but the same approach will work equally well for *dependency analysis*, or any graph-based syntactic analysis for that matter, as long as each node can be related to a sequence of words. An additional advantage of this approach is that syntactic information such as syntactic categories or dependency relations can be taken into account when relating source to target subsequences.

A potential drawback of syntactic alignment is that a limited number of cases of semantic similarity do not neatly align with syntactic structure. For example, the Dutch verbal expressions “zelfmoord plegen” (committing suicide) and “zichzelf van kant maken” (taking one’s own life) may not be directly aligned, because they do not correspond to separate constituents, at least not in the rather shallow syntactic structures used in the DAESO corpus. However, cases like these turn out to be rare in practice. Moreover, the use of semantic similarity relations—to be described in the next section—usually allows us to align the nodes directly dominating these expressions (typically main clauses) through an appropriate semantic relation, thereby implying the semantic equivalence between the two embedded expressions.

Provided we have syntactic trees corresponding to source and target sentences, our initial objective to analyse semantic similarity between sentences in terms of their most similar word sequences can now be restated in terms of identifying their most similar *syntactic phrases*. This amounts to *aligning* a pair of syntactic trees by pairing those nodes that are most similar. More formally: let t_s and t_t be source and target trees resulting from a syntactic analysis of source sentence s_s and target sentence s_t respectively. Let v_s be a source node from tree t_s and v_t a target node from tree t_t. For a node v, its terminal *yield* $\textsc{str}(v)$ is defined as the sequence of all terminal nodes reachable from v (i.e., a subsequence of words from the sentence). A *node alignment* is a tuple $\langle v_s, v_t \rangle$ meaning that $\textsc{str}(v_s)$ is semantically most similar to $\textsc{str}(v_t)$. This entails that there is no other node v'_s that is more similar to v_t than v_s and conversely that there is no other node v'_t that is more similar to v_s than v_t. A *tree alignment* is a set of node alignments.[2] It is assumed that there exists at least some overlap in meaning between source and target sentences, otherwise their tree alignment will simply be void.

2.2 Semantic similarity relations

Our notion of semantic similarity is not based on any formal definition or axiomatic system. As far as similarity between words is concerned, it involves the same human knowledge encoded in wordnets or ontologies. Our interpretation of semantic similarity is therefore one of common sense rather than strict logic, akin to the definition of *entailment* employed in the recognizing textual entailment (RTE) framework (Dagan et al. 2005). In short, if competent native speakers of the language judge a pair of expressions to share meaning, we call these expressions semantically similar.

However, semantic similarity can take many forms. On one extreme is the trivial case of two identical word sequences. On the other extreme is the case of two expressions which do not have any words in common, yet manage to convey the same meaning through the use of synonyms and paraphrases. Besides paraphrasing, the semantic content of one expression can be contained in that of another or vice versa, or their content can partially overlap. As an example, reconsider the semantic similarity relations in example (1). For a start, some evident similarity between (1a) and (1b) is due to identical words such as “koffie” and “dementie”, which occur in both. Synonym pairs such as “vermindert”/“reduceert” and “risico”/“kans” account for non-literal similarity. Looking at the phrases “dagelijks koffie” and “drie koppen koffie per dag”, it is clear that these share meaning. However, calling these examples *paraphrases* would be stretching the meaning of the concept, as the

[2] Note that this still allows one-to-many alignments if v_t and v'_t are equally similar to v_s. In practice, we found such cases to be rare in our data collection, so we opted for excluding one-to-many node alignments. Formally our tree alignment is therefore a *tree matching*, which is a restricted form of tree alignment in which each node is aligned to at most one other node. Note also that there is no requirement for the alignment to be exhaustive, hence it may also be called a *partial tree alignment*.

second phrase contains extra information regarding the required amount of daily coffee (three cups), which is lacking from the first phrase. Likewise, the phrases "op Alzheimer en dementie" and "op Parkinson en dementie" overlap without being paraphrases. In conclusion, the notion of semantic similarity extends beyond that of synonymy and paraphrasing.

It is for this reason that we distinguish different types of semantic similarity relations. In particular, the following five similarity relations between word sequences are postulated:

1. EQ: v_s **equals** v_t iff lower-cased STR(v_s) and lower-cased STR(v_t) are identical
   Example: "dementia" equals "dementia";
2. RE: v_s **restates** v_t iff STR(v_s) is a proper paraphrase of STR(v_t)
   Example: "diminishes" restates "reduces";
3. GEN: v_s **generalises** v_t iff STR(v_s) is more general than STR(v_t)
   Example: "daily coffee" generalizes "three cups of coffee a day";
4. SPEC: v_s **specifies** v_t iff STR(v_s) is more specific than STR(v_t)
   Example: "three cups of coffee a day" specifies "daily coffee";
5. INT: v_s **intersects** v_t iff STR(v_s) and STR(v_t) share meaning, but each also contains unique information not expressed in the other
   Example: "Alzheimer and dementia" intersects "Parkinson and dementia".

These five relations are mutually exclusive and prioritised: *equals* takes precedence over *restates*, etc. Furthermore, *equals, restates* and *intersects* are symmetrical, whereas *generalizes* is the inverse of *specifies*. The order of the sentences is thus important as well: source and target sentences cannot be swapped because *specifies* and *generalizes* are directional relations. Notice also that these relations may be regarded as the phrasal equivalent of traditional lexical semantic relations where *restates* approximately corresponds to *synonym, generalizes* to *hyperonym, specifies* to *hyponym* and *intersects* to *coordinate terms*.

The type of semantic similarity relation can now be added to a tree alignment as an additional labelling of node alignments. A *labeled node alignment* is a tuple $\langle v_s, v_t, r\rangle$ where v_s is a source node, v_t a target node and r is a label from the set of relations, indicating that relation r holds between the STR(v_s) and STR(v_t). A *labeled tree alignment* is a (possibly empty) set of labeled node alignments. Figure 1 shows an example of labeled tree alignment in this sense.

3 Corpus collection

The parallel monolingual treebank for Dutch described here is called the *DAESO Corpus* with DAESO being the acronym of the *Detecting and Exploiting Semantic Overlap* research project,[3] which gave rise to the corpus. It contains both parallel and comparable Dutch text. A first part of 678K tokens was manually annotated, a second part of 1,511K tokens was automatically aligned. The current and next

[3] http://daeso.uvt.nl.

section pertain to the manually aligned part; automatic alignment is addressed in Sect. 6.

3.1 Text material

The corpus contains Dutch text material of five different types. The composition of the corpus is the result of a trade-off between several constraints. The main objective was to cover the range from true *parallel* text down to loosely associated *comparable text*, preferably across different text genres too. Furthermore, several text types are motivated by potential application of the corpus for automatic summarisation and question answering research. Finally, copyright issues had to be settled in a proper legal manner to ensure that the corpus could be made available to other researchers. This resulted in the following five components.

3.1.1 Part 1: Book translations

Near parallel text was obtained from different Dutch translations of foreign language books. Choices for source material were limited to older books, because alternative translations of recent books are extremely hard to come by. The corpus includes parallel Dutch translations from (parts of) three books: (1) “Le Petit Prince” (de Saint-Exupèry 1960, 2000), (2) “On the Origin of Species” (Darwin 2001, 2002, in the 1st and 6th edition) and (3) “Les Essais” (de Montaigne 2001, 2004). Although the latter two books feature somewhat archaic language, recent translations in modern Dutch were used. Two books could unfortunately not be obtained in electronic format and were (partly) scanned and corrected for OCR errors.

3.1.2 Part 2: Autocue–subtitle pairs

This material comes from the *NOS journaal*, the daily news broadcast by the Dutch public television. It consists of the auto-cue text as read by the news reader and the associated subtitles. It was tokenised and aligned at the sentence level in the ATRANOS project (Daelemans et al. 2004). Because of limited on screen space, subtitles typically present a compressed form of auto-cues, making it an excellent source for research on automatic text compression, one of the subtasks in summarisation.

3.1.3 Part 3: News headlines

Another text type consists of similar headlines from online news articles. These were automatically mined from the Dutch version of Google News. As the news clusters created by Google News are based on the full article content rather than only the headline, there is no guarantee that headlines from the same cluster are indeed similar. Although often very close, completely different headlines are no exception. We therefore performed a manual subclustering in order to eliminate outliers and to extract sets of sufficiently similar sentences.

3.1.4 Part 4: QA-system output

With an eye to application in question-answering (QA), the corpus also contains samples from the QA domain. The Dutch IMIX project developed a multimodal QA system in the medical domain (van den Bosch and Bouma 2011) which answered questions by searching a large collection of text ranging from medical encyclopedia to layman websites. In order to evaluate QA engines, a reference corpus of questions and associated answers as found in the document collection was manually compiled. From this corpus, we extracted all clusters of two or more alternative answers. With about one thousand words, this corpus segment is relatively small.

3.1.5 Part 5: Press releases

Press releases about the same news event make up the main source for comparable text. These were obtained from news feeds of ANP and Novum, the two major Dutch press agencies. This type of material is particularly suited to training automatic multi-document summarisation systems. Similar articles were automatically extracted within a certain time window, relying on simple word overlap measures, and followed by manual correction. The automatic procedure traded a high recall for a low precision, because finding similar articles in two large document collections is much harder for human annotators, than deciding whether a given pair of articles is indeed similar.

3.2 Preprocessing

3.2.1 Conversion to XML

All text material was converted to XML format with UTF-8 character encoding. Book translations were (mostly) automatically converted from their original electronic format (raw text, MS Word, PDF) to XML adhering to the *TEI Lite* standard, the light version of the Text Encoding Initiative markup language (Burnard and Sperberg-McQueen 2006). The original document structure and formatting was preserved as much as possible in the markup; all books at least have markup indicating chapters and sections. In addition, text spans which were not fit for our purposes, e.g., quotes in a foreign language, were manually marked. As TEI Lite is not particularly suited to markup of the other text types, these were converted to custom XML formats. For instance, the XML for headlines stores information regarding timestamps and sources, and indicates clusters and subclusters.

3.2.2 Tokenization

All text material was subsequently processed with a tokenizer for Dutch (Reynaert 2007), with the exception of the autocue–subtitle material, which was already tokenised. Sentence boundaries were marked with sentence tags bearing a unique *id* attribute. We found that tokenization errors were more frequent in the book material, presumably because of the relatively longer and more complex sentences. As errors in end-of-sentence detection are detrimental to the subsequent parsing step—and because tokenization errors are relatively cheap to fix—such errors were manually corrected in all manually aligned book texts. Tokenization errors in the press releases where fixed only in so far as they were noticed by the annotators during the subsequent step of sentence alignment.

3.2.3 Syntactic parsing

Next, the state-of-the-art Alpino parser for Dutch (Bouma et al. 2001; van Noord 2006) was used to parse suitable sentences, excluding chapter headings, footnotes, quotes in a foreign language, etc. The parser provides a relatively theory-neutral syntactic analysis originally developed for the Spoken Dutch Corpus (van der Wouden et al. 2002). It is a blend of phrase structure analysis and dependency analysis, with a backbone of phrasal constituents and edges labeled with syntactic function/dependency labels. The output structures are slightly more powerful than trees because they occasionally contain crossing branches, rendering the yield of a non-terminal node a subsequence rather than a substring (i.e., a non-continuous rather than a continuous part of the sentence). Output is stored as a *treebank*, which is simply a list of parses in the Alpino XML format, with an *id* attribute identical to that of the input sentence in the source document.

Alpino combines knowledge-based techniques (e.g., an HPSG-grammar and lexicon) with corpus-based, statistical techniques (e.g., for part-of-speech tagging and disambiguation). Alpino's performance in terms of concept accuracy approaches the level of agreement among human annotators (see van Noord 2006 for details), but still parsing errors in the form of partly erroneous or fragmented analyses do occur. In addition, a few very long and/or particularly ambiguous sentences failed to pass at all (these are not included in the corpus counts). Due to time and cost constraints such parsing errors were not subject to manual correction. However, we also feel that this is more realistic from the perspective of NLP applications, where perfect parsing information is typically not available. Finally, it is important to stress that even in the case of parser errors, often substantial parts of the sentence are still parsed correctly.

3.3 Sentence alignment

The next stage involved aligning similar sentences (regardless of their syntactic structure). The goal of sentence alignment is to identify pairs of sentences with sufficient semantic similarity, whose syntactic trees can then be aligned at the next stage. This process was carried out in two steps: automatic alignment followed by manual correction. Some text material was already aligned at the sentence level: the autocue–subtitle segment was manually aligned within the ATRANOS project and the alternative answers from the QA reference corpus are implicitly aligned. Sentence alignment was thus only carried out for the book translations, the press releases and the headlines.

3.3.1 Aligning sentences in parallel translations

The book translations form nearly parallel texts. Automatic alignment of sentences in parallel *bilingual* text is a well-studied area for which a number of efficient solutions are available (Gale and Church 1993). For three reasons, however, we opted for a straightforward pragmatic solution. First, even though it is usually assumed that the bulk of the alignments is of the one-to-one variety and that crossing alignments and unaligned sentences are rare, we found that these assumptions are frequently violated in our material.[4] Second, the fact that both texts are in the same language allows for a simpler approach. Third, the initial automatic alignment was subject to manual correction afterwards.

The alignment algorithm 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, where similarity is defined in terms of token n-gram overlap. A relatively low threshold guaranteed a high recall at the expense of precision, which is desirable as in practice manually deleting unintended alignments is an easier task than identifying all correct ones.

Obviously, this approach is sensitive to large gaps due to insertion/deletion of substantial pieces of text. Automatic alignment was therefore carried out in multiple passes, that is, first align chapters, next align sections, then paragraphs and finally sentences. This approach gave very reliable results. For example, automatic sentence alignment on the three chapters from Darwin's "On the Origin of Species", which contained the most dissimilar text in the corpus segment of parallel translations, yielded a precision of 73.6 % and a recall of 96.3 %.

3.3.2 Aligning sentences in comparable news texts

Sentence alignment in comparable texts such as similar press releases is notably different from parallel text alignment: one-to-many alignments—or even many-to-many—are frequent, 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 a clearcut decision whether or not two sentences are translations of the same source sentence, it is often much harder to decide whether two comparable sentences are sufficiently similar to justify alignment.

Our guiding principle for this part of the corpus was that aligned sentences should have at least one *proposition* in common. The term *proposition* is loosely interpreted here as a statement about someone or something. This means that just sharing an entity—such as person, location, etc.—does not suffice, nor does merely sharing a predicate—such as an action, state, etc.—let alone sharing just a function word. Only the combination of sharing both an entity and a predicate in the form of

[4] For instance, the two translations of "On the Origin of Species" are based on different editions and show significant differences, largely due to Darwin's own revisions. These range from long sentences in one translation being split into multiple sentences in the other to substantial pieces of added or removed text (the 6th edition even has a whole new chapter).

a proposition justifies alignment. This heuristic may at times be too tolerant, but those sentence pairs with insufficient similarity are weeded out during the final stage of tree alignment.

3.3.3 Aligning sentences in comparable headlines

As described earlier, headlines were clustered into subclusters of sufficiently similar sentences. Although in principle each headline can be sentence-aligned to every other headline in the same subcluster, some clusters are rather large and would give rise to too many sentence pairs. Instead one headline is chosen as the central one to which all other are aligned. For this the *median string* is calculated, which is defined as the headline which has the overall minimum distance to all other headlines. Distance is measured in terms of the Jaccard coefficient over sets of tokens (excluding punctuation).

3.3.4 Manually correcting sentence alignments

Sentence alignments are stored in a simple XML format which contains two types of information: (1) references to the XML files containing source and target texts; (2) a list of links specifying the tags and id attributes of aligned source and target elements respectively. An XML document containing this information will be called a *parallel text corpus*.

Manual correction of the parallel text corpora was carried out using a newly developed alignment annotation tool, called *Hitaext*, for visualising and editing alignments between textual segments.[5] Hitaext is described in more detail in Marsi and Krahmer (2008). Manual correction turned out to be easy, and resulted in pairs of syntactically parsed sentences containing semantic overlap, which were then subject to manual tree alignment.

4 Manual tree alignment

4.1 Annotation process

4.1.1 Annotation guidelines

Annotators were given an annotation manual containing a detailed description of the tree alignment task, to enforce consistency in alignment and relation labelling. The following general guidelines were supplied to our annotators.

1. Do not align anything if the two sentences do not share any meaning.
2. Create evident alignments only; do not create alignments which rely on elaborate or far fetched reasoning.

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

3. Create a complete alignment of the trees by making sure not to miss any alignments.
4. Always align the top nodes of the trees (except in case 1).

The first case applies to sentence alignment errors where the two sentences do not meet the minimally required semantic overlap of at least one common proposition, which happened rarely. The second and third cases basically state that *all* and *only* valid alignments must be made. The fourth case guarantees that the semantic similarity relation between the two full sentences is always defined, except in the case of unrelated sentences. Although sentence pairs for which the parser failed to deliver a syntactic analysis could in principle be hidden from the annotators, we chose to represent these as a degenerate tree consisting of a single node covering the whole sentence. This not only retains textual coherence, but also permits annotation of the top node relation. In addition to these general guidelines, detailed guidelines and examples per similarity relation were provided, addressing among other things, variations in spelling and punctuation, abbreviations and acronyms, compounding, inflection, word order and pronominal forms.

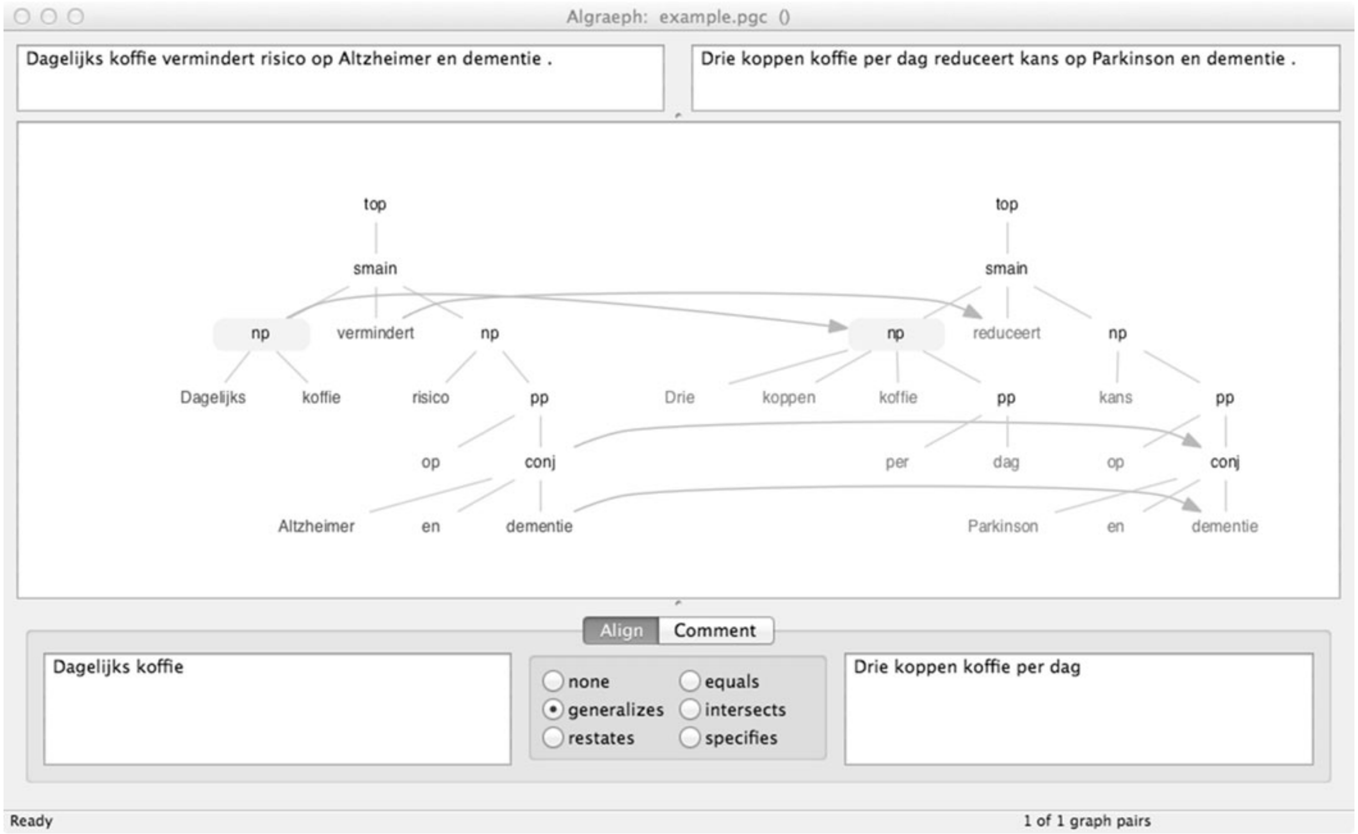
4.1.2 Annotator training

Six different annotators were involved in manual tree alignment. All were university students involved in a humanities program and all were native speakers of Dutch. Each annotator went through a training phase prior to starting annotation for real. Training consisted of aligning parts of a small reference corpus consisting of the first 50 sentences from *Le Petit Prince*. Next an automatically generated list of differences in node alignment and relation labelling was discussed with the authors. This process was repeated until differences were minimal and the authors felt that the annotator had a good grasp of the task at hand, including working with the annotation tool.

4.1.3 Annotation tool

For creating and labelling alignments, a special-purpose graphical annotation tool called *Algraeph* was developed.[6] The screen shot in Fig. 2 shows Algraeph with the sentence pair from example (1). The source and target sentences are shown in the text boxes at the top. The corresponding syntactic trees are shown in the middle, with alignments indicated by coloured lines. The focused nodes and their alignments are shown in yellow. The phrases corresponding to these focused nodes are shown in the text boxes at the bottom, with the alignment relation, which is *generalizes* here, in between. This relation can be changed or removed by clicking any of the radio buttons in bottom panel. Algraeph also allows annotators to add free-form textual remarks to their annotations.

[6] Algraeph is a rewritten and extended version of our earlier tool called Gadget. It is implemented in wxPython, runs on Mac OS X, Linux and Windows, and is available as open source software from http://daeso.uvt.nl/algraeph.



Dagelijks koffie vermindert risico op Altzheimer en dementie .

Drie koppen koffie per dag reduceert kans op Parkinson en dementie .

Dagelijks koffie

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

Drie koppen koffie per dag

Fig. 2 Screen shot of Algraeph, the graph alignment tool used for manual alignment of syntax trees

Large and complex syntax trees corresponding to long sentences pose a challenge to annotators, as these are hard to navigate and require a lot of screen scrolling. Algraeph therefore has a range of options to tweak the rendering of trees and alignments. For instance, there is an option to hide all alignments except those of the focused nodes. Another useful option is to fold or unfold arbitrary parts of the tree.

Attempts were made to speed up the manual annotation process. Alignments of the type *equals* are established automatically whenever this can be done reliably. With regard to the remaining EQ alignment, it suffices to align the largest equal source and target phrases, because EQ alignment of the subphrases down to the level of equal words can then be automatically extrapolated.

Input to Algraeph is a *parallel graph corpus*, a file in a simple custom XML format which contains four types of information. First, it declares the set of semantic similarity relations. Second, it includes one or more references to the treebanks containing the syntactic trees of the aligned sentences. Third, it lists linked trees which are identified by the id's of the source and target trees and the id's of the treebanks they originate from. Fourth, for each pair of source/target trees, it lists aligned nodes in terms of the id's of the aligned source and target nodes as well as the associated alignment relation. In addition, it can hold information such as annotator comments. A parallel graph corpus is automatically derived from the combination of a source and target text document, a parallel text corpus, and the corresponding treebanks.

4.2 Characteristics of the manually aligned corpus

The characteristics of the resulting manually aligned corpus are summarised in Table 1. It contains over 68K aligned trees and over 678K tokens (excluding punctuation). Tokens and trees in unaligned sentences or failed parses are excluded from these counts. The underlying treebanks in fact contain additional trees that were not aligned because no corresponding tree in the parallel/comparable text could be found or because trees were too big/complex to be manually aligned (only in the case of the translations of Montaigne and Darwin). The Autosub (autocue–subtitles) segment is the largest one, followed by Headlines, News and Books. The QA segment is relatively small.

Complexity of the text types is reflected in the average number of tokens per sentence and average number of nodes per tree. Sentences from the Books and QA segments are relatively long and the corresponding trees are relatively big. In contrast, headlines tend to be short and structurally simple.

The figure for uniquely aligned trees expresses the percentage of one-to-one aligned trees. For the Autosub and Books segments, most sentences are indeed uniquely aligned, as is to be expected for parallel text. However, about one out of three trees from the News segment is involved in a one-to-many alignment. Again, this is consistent with expectations regarding comparable text. Note that this measure is not that useful for the Headlines domain, where all headlines are aligned to the median headline, nor for the QA domain, where all answers are exhaustively aligned to each other.

Table 1 Properties of the manually aligned corpus

	Autosub	Books	Head	News	QA	Overall
Trees	18,338	6,362	32,627	11,052	118	68,497
Tokens	217,959	115,893	179,629	162,361	2,230	678,072
Tokens/sent ratio	11.89	18.22	5.51	14.69	18.90	9.90
Nodes	365,157	191,636	318,399	271,192	3,734	1,150,118
Nodes/tree ratio	19.91	30.12	9.76	24.54	31.64	16.79
Uniq. align. trees (%)	92.93	92.49	84.57	63.61	50.00	84.10
Aligned nodes (%)	73.53	66.83	73.58	53.62	38.62	67.62

The percentage of aligned nodes shows how much of the syntax tree is aligned. Interestingly, a substantial number of nodes remains unaligned, even for largely parallel book texts. Still less nodes are aligned for the comparable text types of News and QA.[7]

Figure 3 presents the overall distribution of similarity relations in the manually aligned corpus and the distribution per corpus segment. The distribution is clearly dominated by EQ, which accounts for about 60 % of all relations, and is most frequent in the Autosub segment. Even though SPEC and GEN are each others inverse, the former is overall more frequent than the latter. The reason for this is that SPEC is much more frequent in the Autosub domain, where the auto-cue text is usually more informative than the compressed subtitle text. Another observation is that RE, which is interpreted as paraphrase in the strict sense, occurs most frequently in the Books segment, that is, in parallel text. In contrast, RE occurs below average in the comparable text domains of News and QA. Characteristic for the comparable text domains is the high number of INT relations.

5 Evaluation of inter-annotator agreement

5.1 Evaluation measures

Evaluating inter-annotator agreement requires a measure for comparing different alignments. As described in Sect 2.1, a tree alignment is a set of node alignments $\langle v_s, v_t \rangle$ where v_s and v_t are source and target nodes respectively. As sets can be compared using the well-known *precision* and *recall* measures (van Rijsbergen 1979), these measures can be applied to alignments straight away. Given that A_{true} is a true tree alignment and A_{pred} is a predicted tree alignment, precision and recall are defined as follows:

[7] This may in part be due, however, to the larger number of non-uniquely aligned trees in the comparable text segments, because each count of aligned nodes only concerns one particular pair of source and target trees.

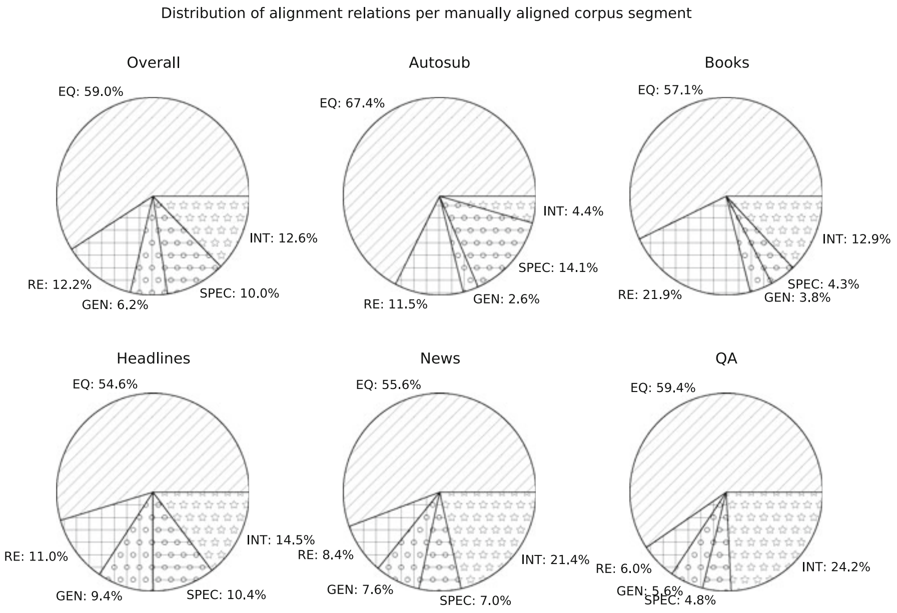


Fig. 3 Overall distribution of similarity relations in manually aligned corpus and distribution per corpus segment

$$precision = \frac{|A_{true} \cap A_{pred}|}{|A_{pred}|} \tag{1}$$

$$recall = \frac{|A_{true} \cap A_{pred}|}{|A_{true}|} \tag{2}$$

Both measures are normalised between zero (worst) and one (best). Precision and recall are combined in the F_1 score, which is defined as their harmonic mean, giving equal weight to both terms.

$$F_1\,score = \frac{2 \times precision \times recall}{precision + recall} \tag{3}$$

The same measures can be used for comparing *labeled* tree alignments in a straightforward way. Recall from Sect. 2.2 that a labeled tree alignment is a set of labeled node alignments $\langle v_s,\, v_t,\, r\rangle$ where v_s is a source node, v_t a target node and r is a label from the set of semantic similarity relations. Let A^{rel} be the subset of all alignments in A with label *rel*.

$$A^{rel} = \{\langle v_s, v_t, r\rangle \in A : r = rel\} \tag{4}$$

This allows us to calculate, for example, precision on relation *equals* as follows.

$$precision^{EQ} = \frac{\left|A^{EQ}_{true} \cap A^{EQ}_{pred}\right|}{\left|A^{EQ}_{pred}\right|} \tag{5}$$

We thus calculate precision as in the unlabelled case, but ignore all alignments—whether true or predicted—labeled with a different relation. Recall and F score on a particular relation can be calculated in a similar fashion.

5.2 Evaluation data

The data used for evaluation consisted of four different samples:

1. the first part of autocue–subtitle text from *NOS Journaal* on March 28 1999
2. a segment from first part of the book *De kleine prins* by Saint-Exupery
3. comparable sentences from 10 similar press releases issued November 2006
4. the first part of Chapter 4, Book 2 of *De essays* by Montaigne

Table 2 shows some properties of these samples. Samples are of roughly equal size in terms of tokens and syntactic nodes, but the Montaigne sample is the most complex in terms of tokens per sentence and syntactic nodes per tree.

5.3 Evaluation procedure

The four samples were independently annotated by six annotators, all of which were involved in the manual annotation of the corpus. Mean agreement scores were calculated using a procedure similar to the Jackknife method. Given the six annotations $A_1, \ldots, A_6$, we repeatedly took one as the A_{true} annotation against which the five other annotations were evaluated. Mean scores and standard deviations over these 30 (=6 × 5) scores were calculated. As a consequence of this procedure, precision and recall are always identical[8] and the corresponding *F* score (i.e., the harmonic average) is consistently a little below the precision/recall, hence only *F* scores are reported here.

5.4 Evaluation results

Figure 4 presents the average *F* scores per relation, as well as regardless of relation (ANY), for each text type. As expected, highest scores are obtained on the EQ relation. At first sight, it may seem that labelling EQ is a trivial and deterministic task, for which the *F* score should always be close to 1.0 rather than around 0.95. However, annotators sometimes fail to identify all identical phrases, especially in long and complex sentences. In addition, the same word may occur multiple times in the source or target sentences, which introduces ambiguity. This frequently occurs with function words such as determiners and prepositions. Finally, choosing among several equally valid EQ alignments may sometimes involve a somewhat arbitrary decision. This situation arises, for instance, when a proper noun is mentioned just once in the source sentence but twice in the target sentence.

[8] Since *precision*(A_n, A_m) equals *recall*(A_m, A_n), the mean over all pairwise precision scores equals the mean over all pairwise recall scores.

Table 2 Properties of the evaluation data (S = sentence)

Text type	S pairs	Tokens	Tokens/S	Nodes	Nodes/S
Autosub	50	1,095	11.53	1,851	19.48
Books: St-Exupery	49	897	10.31	1,459	16.77
News	48	1,089	16.75	1,834	28.22
Books: Montaigne	25	1,280	29.09	2,107	47.89

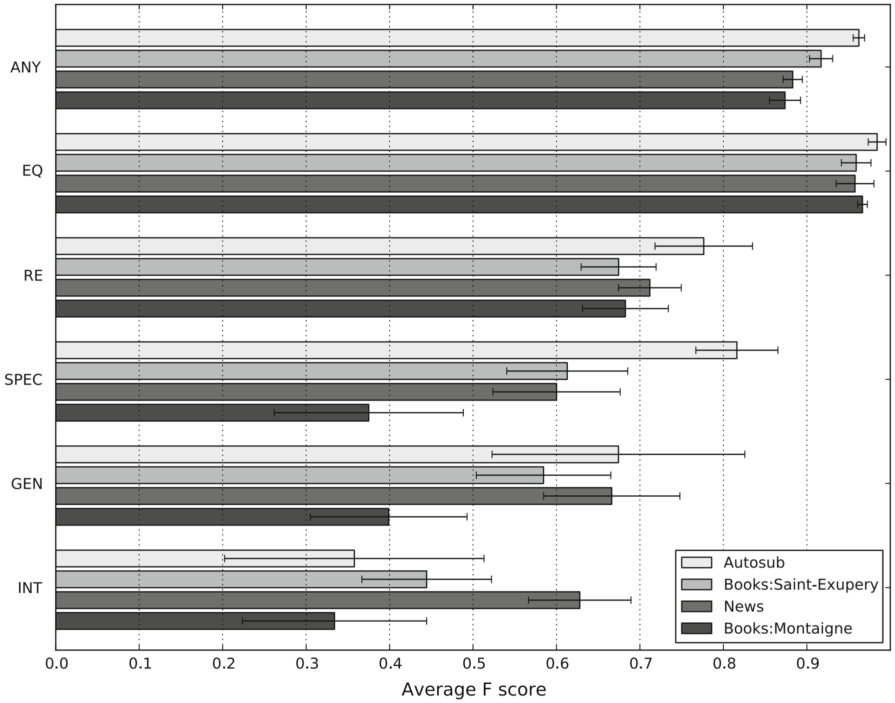


Fig. 4 Inter-annotator agreement per similarity relation in terms of average *F* score on different text types

The scores regardless of relation label (ANY) are slightly lower than those for EQ, but still in the range from 0.87 to 0.96. This is in the same range as reported in Marsi and Krahmer (2005a) on the same task.

Scores on all other relations besides EQ range from best scores on RE, lower but comparable scores on SPEC and GEN, and worst scores on INT. This is in line with our own experience that identifying strict paraphrasing and inclusion relations is more difficult than identifying string identity, whereas identifying partial semantic overlap is most difficult.

With regard to text type, the Autosub sample shows the highest agreement scores. The only outlier—the low score on INT—may be explained by the fact that INT relations are infrequent (4.4 % on average in the whole Autosub corpus) and therefore the Autosub sample may not contain sufficient instances for reliable estimation. Another observation is that the Montaigne sample yields the lowest agreement scores, which is to be expected given that it contains the longest and most complex sentences. There seems to be no clear difference between parallel (Books) and comparable (Autosub and News) text.

In sum, evaluation of inter-annotator agreement indicates that: (1) agreement is higher without relation labelling; (2) agreement is highest on EQ, lowest on INT, and in between on RE, SPEC and GEN; (3) agreement is inversely correlated with the complexity of the text type. Overall the results show substantial agreement among the various annotators.

6 Automatic alignment

6.1 Corpus collection

In addition to the manually aligned part, the DAESO corpus contains a second part of 1,511K tokens that is automatically aligned. It consists of three different segments.

6.1.1 Part 1: Book translations

This segment includes additional chapters from Darwin's "On the Origin of Species" and Montaigne's "Les Essais". OCR errors were manually corrected for one of the translations of "Les Essais"; all others were available in electronic format. Tokenisation errors were not corrected. Sentence alignment was performed fully automatically using the method described in Sect. 3.3.

6.1.2 Part 2: Autocue–subtitle pairs

These autocue–subtitle pairs come from the daily news broadcasts of the Flemish public broadcasting organisation VRT. As mentioned earlier, tokenisation and sentence alignment including manual correction was carried out in the ATRANOS project (Daelemans et al. 2004).

6.1.3 Part 3: News headlines

This part consists of more comparable headlines from online news articles. Recall that for the manually annotated part of the corpus, we performed a manual subclustering in order to eliminate outliers and to extract sets of sufficiently similar sentences. These manual subclusters were used to develop and optimise a system for automatic subclustering of headlines (Wubben et al. 2009), which was subsequently used for preprocessing the automatically annotated part of this corpus segment. The clustering method is based on pair-wise similarity calculations of *tf* × *idf* weighted vector representations of the headlines. It obtains a precision of 0.76 and a recall 0.41. As in the manually annotated data, the median headline is aligned to all other headline in a subcluster.

For reasons of time and resources we decided not to include press releases in this part of the corpus. Additional material in the QA domain was not available.

6.2 Automatic tree alignment

Automatic tree alignment was carried out with newly developed linguistic graph alignment software (Marsi and Krahmer 2010). The core of the system relies on a form of supervised machine learning called *memory-based learning* (Daelemans and van den Bosch 2005), hence the program is called *memory-based graph matcher* (MBGM). It performs the tasks of node alignment and relation labelling simultaneously as a combination of exhaustive pairwise classification using a memory-based learning algorithm, and global optimisation of node alignments using a combinatorial optimisation algorithm. For its input it relies on a combination of morpho-syntactic information from parse trees, lexical resources such as the Dutch wordnet Cornetto (Vossen et al. 2008), and the manually annotated corpus part as training material. Performance of the aligner has been measured on a test set of manually aligned press releases. For the task of node alignment, regardless of relation labelling, performance reached an *F* score of 0.87, which is on par with the 0.88 inter-annotator agreement on this text type reported in Sect. 5. The weighted average *F* score on relation labelling was 0.80, which is again close to the inter-annotator agreement of 0.82. However, even though scores on the EQ and INT relations were good, scores on the other relations were significantly worse than the corresponding human scores (see Marsi and Krahmer 2010 for details).

The properties of the automatically aligned corpus part are summarised in Table 3. By and large the proportional numbers are similar to those for the manually annotated corpus in Table 1.

7 Related work

7.1 Corpora for paraphrasing research

In recent years various techniques for paraphrase recognition and generation have been proposed (Madnani and Dorr 2010; Androutsopoulos and Malakasiotis 2010),

Table 3 Properties of the automatically aligned corpus

	Autosub	Books	Headlines	Overall
Trees	60,182	10,620	88,078	158,880
Tokens	715,970	311,091	483,997	1,511,058
Tokens/sent ratio	11.90	29.29	5.50	9.51
Nodes	1,196,083	514,273	856,197	2,566,553
Nodes/tree ratio	19.87	48.42	9.72	16.15
Uniquely aligned trees (%)	89.54	89.47	87.88	88.61
Aligned nodes (%)	73.35	53.13	69.45	68.00

but the vast majority of these systems is data-driven and crucially rely on text corpora in order to circumvent the knowledge acquisition bottleneck. Different approaches to obtain the necessary data have been explored. One approach is to search for lexical and phrasal paraphrases in large collections of documents (Pasca and Dienes 2005; Lin and Pantel 2001). This usually relies on the assumption of distributional similarity: words and phrases that have a similar distribution, tend to have similar meanings as well. It has been observed that this strategy does not necessarily result in paraphrases, though: antonyms such as *black* and *white* tend to have similar distributions, but are clearly not lexical paraphrases. In general, a limitation of this approach appears to be that there is no guarantee that the pairs that are found are indeed paraphrases.

As an alternative, some researchers have relied on parallel text corpora. Bilingual parallel text corpora are frequently used in data-driven machine translation, where they typically consist of pairs of sentences in different languages which are translations of each other. Since translations of words and phrases from one language to another are not necessarily one-to-one mappings, bilingual parallel corpora can be used to generate paraphrases using a *pivoting* approach. This works by first translating a phrase or sentence from a source to a target language and then back again in order to obtain a source language paraphrase (Quirk et al. 2004; Bannard and Callison-Burch 2005; Zhao et al. 2008).

Another alternative is the use of *monolingual* parallel text corpora, as first suggested by Barzilay and McKeown (2001). Barzilay and McKeown built their corpus using various alternative human-produced translations of literary texts and then applied machine learning or multi-sequence alignment for extracting paraphrases. Similarly Pang et al. (2003) use a corpus of alternative English translations of Chinese news stories in combination with a syntax-based algorithm that automatically builds word lattices encoding paraphrases.

Parallel monolingual texts tend to be difficult to obtain and may not always be representative of the kind of paraphrasing required for applications. Comparable monolingual corpora, containing pairs of sentences with possibly partial semantic overlap, have therefore been investigated as a data source for paraphrase extraction, for example, news reports describing the same event (e.g., Shinyama et al. 2002; Barzilay and Lee 2003; Shen et al. 2006; Wubben et al. 2009).

The corpora discussed so far are all collected automatically and do not rely on human judgments and annotations. The first manually collected paraphrase corpus is the Microsoft Research Paraphrase (MSRP) Corpus (Dolan et al. 2004), consisting of 5,801 sentence pairs, sampled from a larger corpus of news articles. Each pair was judged by human judges who were asked to indicate whether the sentences indeed were paraphrases. This turned out to be a fairly difficult task (judged by inter-annotator agreement), and judges rated 67 % of the pairs as paraphrases. Even though the MSRP corpus is a useful resource, it has a number of limitations (Madnani and Dorr 2010): it is small, limiting its applicability for data-driven paraphrasing techniques, and in addition, one of the constraints used for selecting the sentence pairs is that they must share at least three words, so that pairs that are semantically similar but lexically completely different are never included. Finally, there are no sub-sentential annotations, which would be useful for developing and evaluating paraphrasing applications. A similar corpus is the Parallel Wikipedia Corpus which has over 108K pairs of aligned sentence pairs from Wikipedia and Simple Wikipedia (Zhu et al. 2010). Another recently created paraphrase corpus is the Webis Crowd Paraphrase Corpus 2011 that comprises passage-level paraphrases obtained through Amazons Mechanical Turk for crowdsourcing (Burrows et al. 2012).

The lack of sub-sentential annotations is addressed by Cohn et al. (2008), who develop a parallel monolingual corpus of 900 sentence pairs, including paraphrase pairs from the MSRP Corpus. Each pair is annotated at the word and phrase level, where a pair of phrases is aligned if they can be substituted for each other in the specific context of the sentence. In addition, annotators have to indicate whether substitution is *sure* (substitution is perfectly feasible) or *possible* (substitution may be slightly marked). Inter-annotator agreement scores indicate that this is a feasible albeit fairly difficult task. Madnani and Dorr (2010, p. 377) point out that a paraphrase corpus with detailed alignments is "much more informative than a corpus containing binary judgments at the sentence level".

Compared with the corpora for paraphrasing research discussed so far, the DAESO corpus differs in a number of important ways. For a start, alignment at the phrasal level is linguistically informed in the sense that only syntactic phrases are aligned. Rather than just a corpus of aligned text, it is a parallel monolingual treebank. This tight coupling with syntactic structure allows syntactic information to be used for detecting and extracting semantically similar expressions. Another difference is that all alignments are labeled with a semantic similarity relation. This facilitates aligning a much wider range of semantically similar phrases than just those which are semantically equivalent. In addition, with over 68K manually aligned sentences, the DAESO corpus is substantially larger than any comparable corpus, making it a viable source of training material for bootstrapping data-driven tree alignment and labelling software. In contrast to most paraphrase corpora in existence, the DAESO corpus is also a Dutch language corpus, which supports investigating to what extent computational approaches to paraphrasing developed for English are language-independent. In sum, the DAESO corpus is a rather unique resource that opens up many possibilities for future research.

7.2 Semantic relations

Word alignments in bilingual text corpora are often labeled in terms of *sure* and *possible*, representing an exact versus an approximate translational equivalence respectively (Och and Ney 2003). The same relations have been adopted for phrase alignments in work on aligned bilingual treebanks (Samuelsson and Volk 2006; Zhechev and Way 2008; Tiedemann and Kotzé 2009) as well as in alignment-based paraphrase corpora (Daume and Marcu 2005; Cohn et al. 2008). However, limiting the annotations to translational or semantic equivalence appears to ignore a substantial source of information for learning translation and paraphrase patterns. For example, given the sentences "A was bought by a subdivision of Y" and "Y acquired both A and B", it is a lot easier to hypothesise that "bought" paraphrases "acquired" once we know that "a subdivision of Y" is a specification of "Y" and that "A" and "both A and B" intersect, even though neither of the two pairs involves a strict paraphrase.

The proposed set of semantic similarity relations used in the DAESO corpus is reminiscent to the *natural logic* proposed by MacCartney and Manning (2008, 2009), who argue that having such relations is beneficial for RTEs. Textual entailment in the context of the RTE framework is defined as a directional relation between a pair of texts, called *text* and *hypothesis* respectively, where the former entails the latter under a common sense interpretation (Dagan et al. 2006). Our *specifies* relation may be interpreted as entailment and vice versa, our *generalizes* relation as reversed entailment. Likewise, *restates* may be regarded as mutual entailment. The *intersects* relation, however, cannot be stated in terms of entailment, which makes our relations somewhat more expressive. For instance, it can express the partial similarity in meaning between "John likes milk" and "John likes movies". In a similar way, contradictory statements such as "John likes milk" versus "John hates milk" can not be distinguished from completely unrelated statements such as "John likes milk" and "Ice is cold" in terms of entailment. In contrast, *intersects* is capable of capturing the partial similarity between contradictory statements.

Another difference between RTE's entailment relation and our semantic similarity relations is that the former is primarily a sentence-level relation whereas the latter cover phrase-level and word-level relations. The same goes for the gradient notion of semantic similarity between sentence pairs employed in the Semeval 2012 task 6 on semantic textual similarity (Agirre et al. 2012).

7.3 Applications

There are many opportunities to exploit the DAESO corpus. Paraphrases can be extracted from the corpus in a straightforward way by extracting aligned phrase pairs, possibly filtered by the type of semantic similarity relation. Even though the present corpus is of substantial size, many of these paraphrases, especially those involving larger phrases, will be rather specific. The challenge is therefore to abstract from aligned lexicalised paraphrases to more generic paraphrase patterns and rules. One promising line of research in this respect is the automatic induction of synchronous tree substitution rules from aligned trees, as it not only addresses the extraction of paraphrase rules but also generation of paraphrases by a process of tree transduction (Cohn and Lapata 2009; Cohn et al. 2008). Notice incidentally that sentence compression can be seen as a specific form of paraphrasing, where the result should meet the additional constraint that it is shorter than the original (Knight and Marcu 2002; Filippova and Strube 2008; Cohn and Lapata 2009; Marsi et al. 2010).

As shown earlier, the corpus can also serve to train automatic tree alignment and labelling algorithms. These can in turn be used to bootstrap much larger aligned treebanks. Moreover, automatic tree aligners yield a detailed analysis of the semantic similarity between sentences. This is of great value to many NLP tasks.

Automatic summarisation systems attempt to avoid redundancy by measuring semantic overlap between sentences. This is particularly important in the case of multi-document summarisation, where sentences extracted from related documents are very likely to express similar information in different ways (Barzilay and McKeown 2005). Automatic question–answering systems may benefit from clustering semantically similar candidate answers. The search process of QA may in fact be formulated as detecting semantic overlap between the question and potential answers (Punyakanok et al. 2004; Bouma et al. 2005; Cui et al. 2005). Evaluation of machine translation systems may involve measuring semantic similarity between system output and gold standard translations (Callison-Burch et al. 2006; Zhou et al. 2006; Snover et al. 2009). Automatic duplicate and plagiarism detection beyond obvious string overlap requires detection of semantic similarity (Potthast et al. 2010). Intelligent document merging software, which supports a minimal but lossless merge of several revisions of the same text, must handle cases of paraphrasing, restructuring, compression, etc.

In the area of multi-document summarisation, analysis of semantic similarity between sentences extracted from different documents provides the basis for *sentence fusion*, a process where a new sentence is generated that conveys all common information from both sentences without introducing redundancy (Barzilay and McKeown 2005; Marsi and Krahmer 2005b; Wan et al. 2007; Filippova and Strube 2008). This is an application where semantic similarity relations can provide useful input. Marsi and Krahmer (2005a, b) argue that *generalizes*/*specifies* relations can be used to generate a so-called *intersection fusion*, which is the most general form describing only information shared between two sentences. In contrast, a *union fusion* is the most specific form describing all information from both sentences (but without introducing redundancy). In the context of QA, the same sentence fusion techniques would allow for combining partial answers to obtain full answers, or selecting more specific answers in favour of more general ones, which has been shown to be preferred by users (Krahmer et al. 2008).

8 Conclusion

This article presented a new, large scale (2.1M tokens) and publicly available parallel monolingual treebank with detailed annotations of semantic similarity. Taking its inspiration from earlier work on paraphrasing in English, the corpus contains parallel and comparable Dutch text from several text domains. It includes alignments at the level of sentences, syntactic phrases and words, which are also labeled according to a small number of semantic similarity relations. One part of the corpus (678K tokens) is manually aligned using newly created annotation tools. Inter-annotator agreement was shown to be substantial, although agreement varied per text genre and semantic relation. A second part of the corpus was automatically annotated using a data-driven tree aligner trained on the manually aligned data, resulting in scores which were close to those of the manual part. We argued that this corpus is unique in its size and detailed annotations, and holds great potential for both direct paraphrasing research and for application of indirectly derived resources (e.g., automatic tree aligner) in various NLP tasks.

8.1 Corpus availability

The full version of the corpus can be obtained (free of charge) from the TST-centrale for the Dutch Language Union: http://www.inl.nl/tst-centrale/nl/producten/corpora/daeso-corpus-parallelle-nederlandstalige-monolinguale-treebank/6-69. The Algraeph and Hitaext annotation tools are available as open source software and can be downloaded from http://daeso.uvt.nl/.

Acknowledgments We would like to thank Nienke Eckhardt, Koen van Lierop, Vera Nijveld, Paul van Pelt, Hanneke Schoormans and Jurry de Vos for all their annotation work, Erik Tsjong Kim Sang and colleagues for the autocue–subtitle material from the ATRANOS project, Gosse Bouma and others from the IMIX project for the QA reference corpus, Wauter Bosma for mining the headlines from Google News, and Sander Wubben for automatic subclustering of headlines. We also like to express our gratitude to the publishers and press agencies for providing the raw text material. We thank the anonymous reviewers for their constructive remarks. This work was conducted within the DAESO project (2006–2010) funded by the Stevin program (De Nederlandse Taalunie; *The Dutch Language Union*).

References

Agirre, E., Diab, M., Cer, D., & Gonzalez-Agirre, A. (2012). Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the first joint conference on lexical and computational semantics* (Vol. 1). *Proceedings of the main conference and the shared task* (Vol. 2). *Proceedings of the sixth international workshop on semantic evaluation* (pp. 385–393). Association for Computational Linguistics.

Androutsopoulos, I., & Malakasiotis, P. (2010). A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research, 38*, 135–187.

Bannard, C., & Callison-Burch, C. (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL)* (pp. 597–604), Ann Arbor.

Barzilay, R., & Lee, L. (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of the conference of the North American chapter of the Association for Computational Linguistics on Human Language Technology (NAACL-HLT)* (pp. 16–23), Morristown, NJ, USA.

Barzilay, R., & McKeown, K. (2001). Extracting paraphrases from a parallel corpus. In *Proceedings of the 39th meeting of the Association for Computational Linguistics (ACL)* (pp. 50–57), Toulouse, France.

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

Bos, J., & Markert, K. (2005). Recognising textual entailment with logical inference. In *Proceedings of the conference on human language technology and empirical methods in natural language processing (HLT-EMNLP)* (pp. 628–635).

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

Bouma, G., Mur, J., van Noord, G., van der Plas, L., & Tiedemann, J. (2005). Question answering for Dutch using dependency relations. In *Proceedings of the CLEF 2005 workshop*.

Burnard, L., & Sperberg-McQueen, C. M. (2006). *TEI lite: Encoding for interchange: An introduction to the TEI Revised for TEI P5 release*. Technical report, Text Encoding Initiative.

Burrows, S., Potthast, M., & Stein, B. (2012). *Paraphrase acquisition via crowdsourcing and machine learning*. ACM TIST.

Callison-Burch, C., Koehn, P., & Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the human language technology conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL)* (pp. 17–24), New York City, USA.

Cohn, T., & Lapata, M. (2009). Sentence compression as tree transduction. *Journal of Artificial Intelligence Research, 34*(1), 637–674.

Cohn, T., Callison-Burch, C., & Lapata, M. (2008). Constructing corpora for the development and evaluation of paraphrase systems. *Computational Linguistics, 34*(4), 597–614.

Cui, H., Sun, R., Li, K., Kan, M., & Chua, T. (2005). Question answering passage retrieval using dependency relations. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 400–407).

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge: Cambridge University Press.

Daelemans, W., Höthker, A., & Tjong Kim Sang, E. (2004). Automatic sentence simplification for subtitling in Dutch and English. In *Proceedings of the 4th international conference on language resources and evaluation (LREC)* (pp. 1045–1048).

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL recognising textual entailment challenge. In *Proceedings of the PASCAL challenges workshop on recognising textual entailment*, Southampton, UK.

Dagan, I., Glickman, O., & Magnini, B. (2006). The PASCAL recognising textual entailment challenge. In J. Quiñonero Candela, I. Dagan, B. Magnini & F. d'Alché Buc (Eds.), *Machine learning challenges* (pp. 177–190). Berlin, Heidelberg: Springer.

Darwin, C. R. (2001). *Het ontstaan van de soorten: door natuurlijke selectie ofwel het bewaard blijven van de rassen die in voordeel zijn in de strijd om het bestaan: de definitieve editie* (6th ed.). Amsterdam: Atlas.

Darwin, C. R. (2002). *Over het ontstaan van soorten: Door middel van natuurlijke selectie, of het behoud van bevoordeelde rassen in de strijd om het leven*. Amsterdam: Nieuwezijds.

Daume, H., & Marcu, D. (2005). Induction of word and phrase alignments for automatic document summarization. *Computational Linguistics, 31*(4), 505–530.

de Montaigne, M. (2001). *Essays*. Amsterdam: Boom.

de Montaigne, M. (2004). *De essays*. Amsterdam: Atheneum, Polak and Van Gennip.

de Saint-Exupèry, A. (1960). *De kleine prins*. Rotterdam: Donker.

de Saint-Exupèry, A. (2000). *De kleine prins*. Rotterdam: Donker.

Dolan, B., Quirk, C., & Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th international conference on computational linguistics (COLING)* (pp. 350–356), Morristown, NJ, USA.

Filippova, K., & Strube, M. (2008). Sentence fusion via dependency graph compression. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 177–185), Morristown, NJ, USA.

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

Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: A probabilistic approach to sentence compression. *Artificial Intelligence, 139*(1), 91–107.

Krahmer, E., Marsi, E., & van Pelt, P. (2008). Query-based sentence fusion is better defined and leads to more preferred results than generic sentence fusion. In *Proceedings of the 46th annual meeting of the Association for Computational Linguistics: Human language technologies (ACL)* (pp. 193–196), Columbus, OH, USA.

Lin, D., & Pantel, P. (2001). Discovery of inference rules for question answering. *Natural Language Engineering, 7*(4), 343–360.

MacCartney, B., & Manning, C. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd international conference on computational linguistics* (Vol. 1, pp. 521–528).

MacCartney, B., & Manning, C. (2009). An extended model of natural logic. In *The eighth international conference on computational semantics (IWCS)*, Tilburg, The Netherlands.

Madnani, N., & Dorr, B. J. (2010). Generating phrasal and sentential paraphrases: A survey of data-driven methods. *Computational Linguistics, 36*(3), 341–387.

Marsi, E., & Krahmer, E. (2005a). Classification of semantic relations by humans and machines. In *Proceedings of the ACL 2005 workshop on empirical modeling of semantic equivalence and entailment* (pp. 1–6), Ann Arbor, MI.

Marsi, E., & Krahmer, E. (2005b). Explorations in sentence fusion. In *Proceedings of the 10th European workshop on natural language generation (ENLG)*, Aberdeen, UK.

Marsi, E., & Krahmer, E. (2008). Detecting semantic overlap: A parallel monolingual treebank for Dutch. In S. Verberne, H. van Halteren & P. A. Coppen (Eds.), *Computational linguistics in the Netherlands (CLIN): Selected papers* (pp. 69–84), Rodopi, Amsterdam.

Marsi, E., & Krahmer, E. (2010). Automatic analysis of semantic similarity in comparable text through syntactic tree matching. In *Proceedings of the 23rd international conference on computational linguistics (COLING)* (pp. 752–760), Beijing, China.

Marsi, E., Krahmer, E., Hendrickx, I., & Daelemans, W. (2010). On the limits of sentence compression by deletion. In E. Krahmer & M. Theune (Eds.), *Empirical methods in natural language generation* (pp. 45–66). Berlin, Heidelberg: Springer.

Navarro, G. (2001). A guided tour to approximate string matching. *ACM Computing Surveys, 33*, 31–88.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics, 29*(1), 19–51.

Pang, B., Knight, K., & Marcu, D. (2003). Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proceedings of the conference of the North American Chapter of the Association for Computational Linguistics on human language technology (NAACL-HLT)* (pp. 181–188).

Pasca, M., & Dienes, P. (2005). Aligning needles in a haystack: Paraphrase aquisition across the web. In *Proceedings of the 2nd international joint conference on natural language processing (IJCNLP)* (pp. 119–130), Jeju Island, South Korea.

Potthast, M., Stein, B., Barrón-Cedeño, A., & Rosso, P. (2010). An evaluation framework for plagiarism detection. In *Proceedings of the 23rd international conference on computational linguistics: Posters* (pp. 997–1005). Association for Computational Linguistics.

Punyakanok, V., Roth, D., & Yih, W. (2004). Mapping dependencies trees: An application to question answering. In *Proceedings of the eighth international symposium on artificial intelligence and mathematics*, Fort Lauderdale, FL.

Quirk, C., Brockett, C. C., & Dolan, W. (2004). Monolingual machine translation for paraphrase generation. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)* (pp. 142–149), Barcelona, Spain.

Radev, D., & McKeown, K. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistics, 24*(3), 469–500.

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

Ross, G. W., Abbott, R. D., Petrovitch, H., Morens, D. M., Grandinetti, A., Tung, K. H., et al. (2000). Association of coffee and caffeine intake with the risk of parkinson disease. *The Journal of the American Medical Association (JAMA), 283*, 2674–2679.

Samuelsson, Y., & Volk, M. (2006). Phrase alignment in parallel treebanks. In *Proceedings of 5th workshop on treebanks and linguistic theories*, Prague, Czech Republik.

Shen, S., Radev, D. R., Patel, A., & Erkan, G. (2006). Adding syntax to dynamic programming for aligning comparable texts for the generation of paraphrases. In *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 747–754), Sydney, Australia.

Shinyama, Y., Sekine, S., Sudo, K., & Grishman, R. (2002). Automatic paraphrase acquisition from news articles. In *Proceedings of the human language technology conference (HLT 2002)* (pp. 313–318), San Diego, USA.

Snover, M., Madnani, N., Dorr, B., & Schwartz, R. (2009). Ter-plus: Paraphrase, semantic, and alignment enhancements to translation edit rate. *Machine Translation, 23*(2–3), 117–127. doi: 10.1007/s10590-009-9062-9.

Tiedemann, J., & Kotzé, G. (2009). Building a large machine-aligned parallel treebank. In *Eighth international workshop on treebanks and linguistic theories (TLT)* (pp. 197–208).

van den Bosch, A., & Bouma, G. (2011) *Interactive multi-modal question-answering*. Berlin, Heidelberg: Springer.

van Noord, G. (2006). At last parsing is now operational. In P. Mertens, C. Fairon, A. Dister & P. Watrin (Eds.), *TALN06. Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles* (pp. 20–42).

van Rijsbergen, C. (1979). *Information retrieval* (2nd ed.). London, Boston: Butterworth.

van der Wouden, T., Hoekstra, H., Moortgat, M., Renmans, B., & Schuurman, I. (2002). Syntactic analysis in the spoken dutch corpus. In *Proceedings of the 3rd international conference on language resources and evaluation (LREC)* (pp. 768–773), Las Palmas, Canary Islands, Spain.

Vossen, P., Maks, I., Segers, R., & van der Vliet, H. (2008). Integrating lexical units, synsets and ontology in the Cornetto database. In *Proceedings of the 6th international conference on language resources and evaluation (LREC)*, Marrakech, Morocco.

Wan, S., Dale, R., Dras, M., & Paris, C. (2007). Global revision in summarisation: Generating novel sentences with prim's algorithm. In *Proceedings of the 10th conference of the Pacific Association for Computational Linguistics* (pp. 19–21).

Wubben, S., van den Bosch, A., Krahmer, E., & Marsi, E. (2009). Clustering and matching headlines for automatic paraphrase acquisition. In *The 12th European workshop on natural language generation (ENLG)* (pp. 122–125), Athens.

Zhao, S., Wang, H., Liu, T., & Li, S. (2008). Pivot approach for extracting paraphrase patterns from bilingual corpora. In *Proceedings of the 46th annual meeting of the Association for Computational Linguistics: Human language technologies (ACL-HLT)* (pp. 780–788), Columbus, OH.

Zhechev, V., & Way, A. (2008). Automatic generation of parallel treebanks. In *Proceedings of the 22nd international conference on computational linguistics (COLING)* (pp. 1105–1112).

Zhou, L., Lin, C. Y., & Hovy, E. (2006). Re-evaluating machine translation results with paraphrase support. In *Proceedings of the 2006 conference on empirical methods in natural language processing* (pp. 77–84). Association for Computational Linguistics.

Zhu, Z., Bernhard, D., & Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd international conference on computational linguistics* (pp. 1353–1361). Association for Computational Linguistics.