Morphological knowledge and alignment of English-German parallel corpora

Patrick Tschorn,
Institute of Cognitive Science
University of Osnabrück
ptschorn@uos.de

Anke Lüdeling,
Institute of German Language and Linguistics
Humboldt-University, Berlin
anke.luedeling@rz-hu-berlin.de

1 Introduction

Alignment is an important step to linguistically exploit parallel corpora. In this paper we introduce a morphological component that improves the alignment of German-English parallel texts and helps find correspondences between morphological elements on the sub-word level. This paper deals with a small aspect of an alignment system, namely the improvement of a dictionary-based distance measure through a morphological analyser.

What is alignment? For the purposes of this paper we define a bilingual parallel text as a text (L1) and its translation (L2). A sentence level alignment then maps groups of L1-sentences to corresponding groups of L2-sentences. These groups are often called "beads". An alignment can be viewed as a sequence of beads that covers the entire parallel text. While most beads usually express the correspondence between a single L1-sentence and a single L2-sentence, other types of beads arise when sentences are split, merged, deleted, added or changed in order by the translator. Each sentence belongs to exactly one bead.

To illustrate some of the difficulties, consider the following excerpt from the very beginning of 'The War of the Worlds' parallel text[1]:

0: The Eve of the War	0: 1. Am Vorabend des Krieges
1: But who shall dwell in these worlds if they be inhabited ? ...	
2: Are we or they Lords of the World ? ...	
3: And how are all things made for man ?	
4: - - Kepler (quoted in The Anatomy of Melancholy)	
5: No one would have believed in the last years of the nineteenth century that this world was being watched keenly and closely by intelligences greater than man 's and yet as mortal as his own ; that as men busied themselves about their various concerns they were scrutinised and studied , perhaps almost as narrowly as a man with a microscope might scrutinise the transient creatures that swarm and multiply in a drop of water .	1: Niemand hätte in den letzten Jahren des 19 . Jahrhunderts daran gedacht , daß unsere Welt beobachtet würde ; daß andere intelligente Wesen , größer als die menschlichen und doch ebenso sterblich , uns bei unserem täglichen Tun fast ebenso intensiv belauschen und erforschen könnten , wie jemand mit dem Mikroskop jene kurzlebigen Lebewesen erforscht , die in einem Wassertropfen ihr Wesen treiben und sich darin vermehren .
6: With infinite complacency men went to and fro over this globe about their little affairs , serene in their assurance of their empire over matter .	2: Mit einem unendlichen Behagen schlenderte die Menschheit mit ihren kleinen Sorgen kreuz und quer auf dem Erdball umher , in gelassenem Vertrauen auf ihre Herrschaft über die Materie .
7: It is possible that the infusoria under the microscope do the same .	3: Es ist möglich , daß die mikroskopischen Lebewesen unter dem Brennglas dasselbe tun .
8: No one gave a thought to the older worlds of space as sources of human danger , or thought of them only to dismiss the idea of life upon them as impossible or improbable .	4: Niemand dachte daran , daß von anderen Planeten Gefahren für die Menschheit herrühren könnten . 5: Jede Vorstellung , daß sie bewohnt sein könnten , wurde als unwahrscheinlich oder unmöglich

[1] In this paper, we draw all our examples from 'The War of the Worlds' parallel text (Wells 1898). We also studied European Union documents (RAPID 1995-2002) which contain many productively formed compounds and thus benefit from morphological analysis.

	abgetan .
9: It is curious to recall some of the mental habits of those departed days.	6: Es ist seltsam , sich heute der geistigen Verfassung jener vergangenen Tage zu entsinnen .
10: At most terrestrial men fancied there might be other men upon Mars , perhaps inferior to themselves and ready to welcome a missionary enterprise .	7: Es kam höchstens vor , daß Erdbewohner sich einbildeten , es könnten Wesen auf dem Mars leben , minderwertige allenfalls , zumindest aber solche , die eine irdische Forschungsreise freudig begrüßen würden .
11: Yet across the gulf of space , minds that are to our minds as ours are to those of the beasts that perish , intellects vast and cool and unsympathetic, regarded this earth with envious eyes , and slowly and surely drew their plans against us .	8: Aber jenseits des gähnend leeren Weltraums blickten Geister , uns so überlegen wie wir den Tieren , ungeheure , kalte und unheimliche Geister , mit neidischen Augen auf unsere Erde . 9: Bedächtig und sicher schmiedeten sie ihre Pläne gegen uns .

Table 1: Alignment of the first sentences of 'The War-of-the-Worlds' parallel text.

Computing the alignment is made difficult by the fact that the English version contains a quotation at the very beginning that is missing in the German version and by the two instances where more than one German sentence corresponds to one English sentence (so-called 1:2-beads; 8:4,5 and 11:8,9).
Different approaches to automatically compute sentence level alignments have been proposed since the mid nineteen eighties. A short history can be found in (Veronis, 2000).
Conceptually, an alignment program needs two main components:

- a distance function: a function to judge the degree of correspondence between two given segments of a parallel text
- an optimizer: a means of finding the sequence of beads that maximizes the overall degree of correspondence (composed of the assessments of the individual beads)

A popular choice is to solve the optimization task by dynamic programming, but a range of other possibilities exists. This aspect is of no importance to this paper.
Distance functions numerically express the relatedness of two segments of text. A variety of different distance functions have been proposed. The kind of knowledge employed to judge the relatedness of segments is of importance:

- *Length-based distance functions* are based on the observation that between many languages long sentences tend to be translated into long sentences, and short sentences into short sentences. The involved knowledge boils down to a probability distribution that predicts how likely it is that two segments, given their lengths, are translations of each other. Lengths can for example be measured in characters or words. This approach can easily be adapted to many language pairs. Despite its simplicity it works surprisingly well on clean parallel texts such as the Canadian Hansards. (Gale & Church 1991, Canadian Hansards) However, they do not work so well on texts with a slightly freer translation (e.g. literature). In many texts – as illustrated above - we find elisions or insertions. Because the length-based distance function does not assess the content of a sentence, missing sentences cannot reliably be detected. This causes problems in sequences of sentences with similar length.
- *Lexical distance functions*, on the other hand, try to find corresponding words in the two halves of the parallel text. The more correspondences between L1-words and L2-words found in a segment, the higher the degree of relatedness is deemed to be. The simplest and cheapest way of finding correspondences is by computing string similarity (cognates). Even though many sentence pairs contain cognates, they alone are not sufficient for a reliable distance function. Therefore it is desirable to employ a bilingual dictionary. There are in principle two possibilities to obtain a dictionary: either an ad hoc dictionary is automatically extracted from a parallel text or a high quality machine readable dictionary is taken.
    - *Cognates* are strings that are orthographically (or otherwise) similar across languages, as for example *observatory* - *Observatorium*. Proper names, foreign words and numbers are often suitable cognates. String similarity can be judged by comparing the number of n-grams shared by two strings to the total number of n-grams. Other popular methods are related to the Levenshtein Edit Distance (see e.g. Steven 1994). Cognates are easy to detect and are often used to determine anchor-points for length-based alignment programs. Cognates can also be used to complement other lexical resources.

- Methods of *automatically extracting an ad-hoc dictionary* from aligned and unaligned parallel texts (see e.g. Fung 2000) or during the alignment (e.g. Kay & Röscheisen 1993) have been proposed. These approaches usually assume that translation relations hold between single words only and are based on some notion of frequency of co-occurence. It is, however, well-known that words in texts follow a Zipfian distribution, that is, texts consist of a small number of highly frequent words (such as articles and prepositions) and of a large number of very rare words (Baayen 2001). Highly frequent words occur in nearly every sentence and thus say virtually nothing about the relatedness of two segments of a parallel text. Rare words are a much better indicator for relatedness but are very hard to capture by co-occurence counting methods.
- The third possibility mentioned above are *dictionary-based distance functions* that rely on an existing machine-readable dictionary. Dictionary-based distance functions perform better than distance functions that employ less knowledge because one word can have many translations and multi-word units can be taken into account. There are, however, a number of disadvantages and problems: first, dictionary-based functions are language dependent and therefore have to be obtained for each new language pair. Dictionary resources can sometimes be difficult to obtain. Another fundamental problem is the necessary incompleteness of any dictionary resource (see the following section).

In this paper we take a dictionary-based distance function as a starting point and show how we can deal with the problem of incompleteness by using a morphological analyser. The distance function and the morphological analyser are part of an experimental alignment system (Tschorn 2002). The dictionary contains approx. 120,000 lemmatised words. The texts are also lemmatised (Schmid 1994). A list of stop words (frequent words that do not provide information on relatedness) is excluded.

The remainder of the paper is organized as follows: first we establish the need for a morphological analysis by showing how incomplete lexicons are and giving some examples which illustrate where morphological analysis would help find more correspondences (Section 2). Then we introduce the morphological analyser (Section 3) and show how it finds correspondences in the bitext (Section 5). In Section 6 we evaluate the morphological component.

2 Lexical knowledge and the incompleteness problem

Dictionary-based approaches are, of course, dependent on the quality and the scope of the employed dictionary. But no matter, how good and large the dictionary is, it will always be incomplete because of morphological productivity which can produce an "in principle unlimited number of new formations" (Schultink 1961, 113). This is especially obvious in languages like German which has excessive compounding in addition to derivation but it is just as true in languages like English or even Romance languages if the notion word does not refer to 'graphemic word' but rather to something like 'semantic unit' (as it should in translation). That means that there will often be unknown/unmatched words (or sequences) in the parallel text. Matching (some of) these words can significantly improve the distance function. The remainder of this paper shows how a morphological analysis of complex words can help in this task. The basic idea is the more matches the better – even if this results in partial matches and some morphological principles are violated along the way.

In this section we describe the phenomena that can be tackled by a morphological analyser – the analyser itself is described in Section 4.

For a first example, consider Table 2. Some unknown/unmatched words are marked in boldface. Consider first the case of the noun-noun compound *Wassertropfen* which consists of *Wasser* and *Tropfen.* Although this compound is not listed in the dictionary and can therefore not be matched, its parts can straightforwardly be matched to words (*drop* and *water*) in the English half of the parallel text. In this case it would be sufficient to analyse the German word. The case of the adjective *menschlich* 'human' which consist of the noun *Mensch* and the adjective suffix *-lich* is more complicated. Here we see that the corresponding word *man* belongs to a different part-of-speech – this happens very often in translation, and such cases can hardly ever be found in a dictionary. Here we see that complex derivations also have to be analysed. Only part of the derivation is used to improve the alignment (*-lich* remains unmatched). Sometimes complex English words have to be analysed as well, cf. Table 3.

5: No one would have believed in the last years of the nineteenth century that this world was being watched keenly and closely by intelligences greater than **man** 's and yet as mortal as his own ; that as men busied themselves about their various concerns they were scrutinised and studied , perhaps almost as narrowly as a man with a microscope might scrutinise the transient creatures that swarm and multiply in a **drop of water** .	1: Niemand hätte in den letzten Jahren des 19 . Jahrhunderts daran gedacht , daß unsere Welt beobachtet würde ; daß andere intelligente Wesen , größer als die **menschlichen** und doch ebenso sterblich , uns bei unserem täglichen Tun fast ebenso intensiv belauschen und erforschen könnten , wie jemand mit dem Mikroskop jene kurzlebigen Lebewesen erforscht , die in einem **Wassertropfen** ihr Wesen treiben und sich darin vermehren .	1. kurzlebigen :: transient 2. beobachtet :: watched 3. Welt :: world 4. größer :: greater 5. Mikroskop :: microscope 6. Jahren :: years 7. Jahrhunderts :: century 8. fast :: almost 9. Wesen :: being 10. letzten :: last 11. Lebewesen :: creatures 12. doch :: yet 13. sterblich :: mortal 14. Tun :: concerns 15. intelligente :: intelligences 16. 19 :: nineteenth 17. andere :: themselves 18. Wesen :: men 19. erforschen :: scrutinised 20. erforscht :: scrutinise 21. gedacht :: studied 22. vermehren :: multiply 23. intensiv :: keenly **24. wasser|tropfen:: drop of water** 25. **mensch|(lich) :: man**

Table 2: A bead from 'The War-of-the-Worlds' parallel text (Wells, H G 1898). Dictionary correspondences are listed in the rightmost column under 1 – 23. 24 and 25 are correspondences found by the morphological analyser.

24: … at its nearest distance only 35,000,000 of miles **sunward** of them , a **morning star …**	24: ... in nächster Entfernung , nur 35.000.000 Meilen **sonnenwärts** , einen **Morgenstern ...**	**…** **morgen|stern :: morning star** **sonne|(n)|(wärts) :: sun|ward** **…**

Table 3: A bead from 'The War-of-the-Worlds' parallel text: Correspondences found by our tool are listed in the rightmost column; morphological elements are delimited by '|', bound morphological elements are bracketed

To summarize the desiderata. In order to find more matches – and thus to improve the distance function

- complex words should be analysed into their morphological elements in both languages
- as many elements as possible – but not necessarily all elements – should be matched with corresponding elements in the other language

In the following section we describe the morphological assumptions and the functioning of the word formation component.

3 The word-formation component

One basic assumption that lies behind our word formation component is that many of the unknown/unmatched complex words are productively formed words and as such are formed by regular morphological procedures. This means that chances are high that we will find translations for the morphological elements that are the result of our analysis. There are two general problems in the automatic analysis of complex words: (1) finding the morphological elements and (2) choosing the correct analysis from a number of possible analyses. Both problems are notoriously difficult (see ten Hacken & Lüdeling 2002 for an overview of automatic word formation systems).

Problem (1) – finding the morphological elements of a complex word – is difficult in concatenating languages like German (English or Romance languages, for example, have much more graphemic cues - spaces or hyphens that show morpheme boundaries). This is made even more difficult by stem changes such as linking elements (1a), elisions (1b) or umlauts (rounding of front vowels, 1c) in word

formation (morpheme boundaries are marked by '·'). The same element may have different forms in derivation and compounding (or even in different compounds). Stem changes are lexical and thus cannot be dealt with by rules (Fuhrhop 1998).

(1a) Glocke 'bell' – Glocken·läuten 'bell ringing';
Erlösung 'deliverance' – Erlösungs·botschaft 'message of deliverance'
(1b) Sprache 'language' – sprach·lich 'linguistic – Sprach·kurs 'language class'
(1c) Stadt 'city' - städt·isch 'urban' - Städte·bau 'urban development'

We deal with stem changes in a heuristic way. If a form is unknown we try to 'remove' all possible linkers and undo umlauting. This leads to many unwanted ambiguities, of course. This is a problem that could be dealt with by providing lexical word formation stems as suggested in (Lüdeling & Fitschen 2002).
Another factor that leads to ambiguities are bound elements. As described above, we need to analyse derivations as well as compounds.[2] This can only be done if bound elements (affixes) are listed in the lexicon along with the free elements. Therefore we added 192 bound elements for German and 56 bound elements for English.[3] These bound elements tend to be short and therefore also lead to many ambiguities. Many of these – such as (s)|(o)|(n)|(n)|(en)|(wärts) for *sonnenwärts* – 'sunwards' – are, of course, morphological nonsense but only a full-blown morphological system could resolve these, and such a system is currently not available for German (see Schmid et al. 2001).
This brings us to problem (2) : sometimes there are many possible analyses for one complex word. In our case – unlike in many other NLP applications – we do not have to choose a 'best' or 'correct' analysis between many possible analyses because we have as an added cue the corresponding words in the other language. As long as the correct analysis sonne|(n)|wärts is available the analysis (s)|(o)|(n)|(n)|(en)|(wärts), for example would not be chosen because in one case both elements can be matched while in the other case only wärts can be matched. We can therefore live with a crude and heuristic morphology system (although sometimes, of course, incorrect analyses are chosen, but in most cases our approach leads to more matches between two sentences and therefore to a better distance function).
Our algorithm first matches those word pairs that can be found in the dictionary and then processes unknown complex words: they are first analysed into their morphological parts – for all possible analyses, all the elements are then matched against the left over words in the other language. Disambiguation of multiple analyses is thus made possible by the other language and new correspondence pairs can be added to the dictionary. Note that not every morphological element needs to have a corresponding element in the other language – even partial matches improve the alignment quality. The next section describes how complex words are broken down into possible sequences of elements. This is followed by the presentation of two different strategies for matching the elements of a complex word with their correspondences in the other language.
A complex word is analysed by recognizing its elements from right to left.[4] Usually, there are a number of strings (known free and bound morphemes) that are suitable as elements, which leads to a variety of possible analyses. All arising analysis alternatives are collected and returned.
An agenda-based search is conducted to construct alternative analyses. The agenda contains analysis states – i.e. snapshots of analyses in progress. A analysis state comprises of the string remaining to be decomposed and a sequence of previously recognized elements. An element can be a known free or bound morpheme.The agenda is initialized with an analysis state consisting of the complex word and an empty list of recognized elements.
While the agenda is non-empty

- the first analysis state is removed and examined.
    - Should the analysis state be complete, it is collected in the result set. An analysis state is complete if its remaining string is empty, i.e. nothing is left to be analysed. The list of recognized elements then constitutes a legal analysis.

[2] Conversion is not a problem in our approach We cannot deal with other types of word formation here.
[3] Up to now the bound elements in our lexicon have no translations. It is possible to find regular translation correspondences for affixes that can then be added to the lexicon and used to establish more correspondences. There are obvious cases such as the German negative prefix *un-* that very often corresponds to *in-* or *un-* in English (*unaufgeklärt – unexplained, unsichtbar – invisible*). There are, however, interesting less obvious cases such as the German suffix *-artig* which often corresponds to *-like* (*tischartig – tablelike, spinnenartig – spiderlike*) but which is often *-ular* in neoclassical words (*fühlerartig – tentacular*) or *-y* in native English words (*milchartig – milky, tintenartig – inky*). We are still in the process of evaluating our data.
[4] This is due to the fact that German and English are morphologically right-headed. However, since we take all possible analyses into account the analysis direction does not affect the result.

- Otherwise the analysis state is incomplete. In this case its successors are computed and appended to the agenda.

When the agenda has become empty, the result set is returned.
Successors can be computed for each incomplete analysis state as follows:

- Orthographic variations of the remaining string are obtained. For example umlauting in German is reversed.
- For each variation (including the original string)
  - a set of free and bound morphological elements (stems, affixes, linkers) that match the right end of the remaining string is constructed using a dictionary.
  - for each of the applicable morphological elements a new analysis state is created whose remaining string is shortened by the element which is added to the list of recognized elements. If no morpological element can be cut off the end of the remaining string, no successors are generated.

The computed analyses are ranked using a simple quality measure (quality := number of free morphemes / total number of morphological elements). The closer to 1 the better.

4 Finding correspondences of complex words in parallel text segments

As noted above, many complex words have compositional translations. The distance function component described in this paper is aimed at these. Non-compositional or non-literal translations cannot be found by the component. Each element of an analysis can have a variety of translations. Thus there is usually a large number of possible literal compositional translations for one complex word.
Two matching strategies are described in the following sections: the first strategy (the simple matching strategy) tries to match elements with unanalysed words while the second strategy (the complex matching strategy) tries to match elements with elements.
Two constraints apply: only 1:1 correspondences are possible and the involved elements must not cross sentence boundaries. Both strategies allow partial matches.

4.1 Simple matching strategy

In a given parallel text segment, all unmatched words of one language are analysed. Words that do not yield possible analyses are ignored. For each analysis an attempt is made to match its elements with unanalysed words of the other language half of the parallel text segment. A match between an L1-element and a L2-word is possible if the two are translations of each other according to the employed dictionary. All targeted words must be in the same sentence. Figure 1 illustrates this strategy.

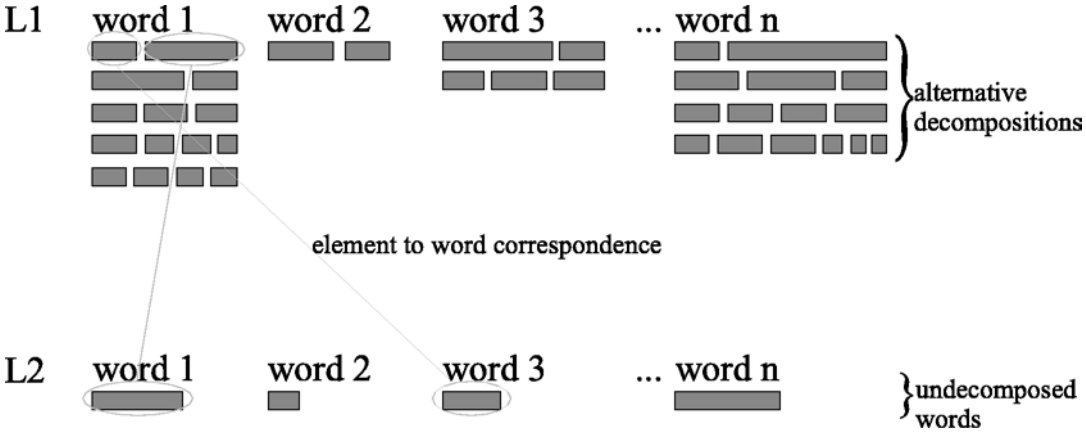

Figure 1: Illustration of the simple matching strategy

If an element does not have a correspondence in the second language half of the segment, it is ignored, thus making partial matches possible, as illustrated above in Table 2 or the correspondence between neighbour and Nachbarplanet (literally 'neighbour planet') in Table 4.)

19: The secular cooling that must someday overtake our planet has already gone far indeed with our **neighbour** .	18: Die allmähliche Abkühlung , die auch unserem Planeten bevorsteht , ist bei unserem **Nachbarplaneten** schon weiter fortgeschritten .	**...** **nachbar|planet :: neighbour** **...**

Table 4: Partial correspondence between *neighbour* and *Nachbarplanet*. Example taken from 'The War-of-the-Worlds'-parallel text.

The translation candidates are sorted according to their "length".

length := 1 + number of matched L2-words

Candidates with a length < 2 are rejected since they do not have any matching elements.

While the sorted list of translation candidates is not empty, the system

- commits itself to the longest candidate
- removes all further candidates that share words with the longest candidate to prevent the same words being part of more than one translation relation – an illegal situation

This strategy matches one complex L1-word with one or more unanalysed L2-words and prefers completely matchable analyses over partially matchable analyses. In this sense, the best decomposition of a complex word is picked from the set of available alternatives – disambiguation is achieved. The chosen analysis is not necessarily the correct one.

The procedure described above constitutes a simple but effective strategy for finding correspondences for complex words not included in the dictionary.

The elements of many English (and some German) compounds are joined together by hyphens. Separating tokens that contain hyphens in a preprocessing stage makes the simple matching strategy more successful.

67: That night another invisible missile started on its way to the earth from Mars , just a second or so under **twenty - four** hours after the first one .	73: In dieser Nacht nahm ein zweites unsichtbares Geschoß seinen Weg vom Mars hin zur Erde , bis auf ein oder zwei Sekunden genau **vierundzwanzig** Stunden nach dem ersten .	**…** **vier|und|zwanzig :: twenty four** **…**

Table 5: A bead from 'The War-of-the-Worlds' parallel text. Separating hyphenated words during preprocessing improves the simple matching strategy. Twenty-four as one graphemic word can only be treated by the complex matching strategy.

4.2 Complex strategy: sub-word level matching

Whereas the simple strategy matches elements with words, the complex strategy matches elements with elements. In a given parallel text segment, all unmatched words of both language halves are analysed. All words are also made available as unanalysed matching targets. A great number of sub-word level matches are possible since there are several analysis alternatives whose elements possibly match a number of different elements of other analysis alternatives. Consider Figure 2.

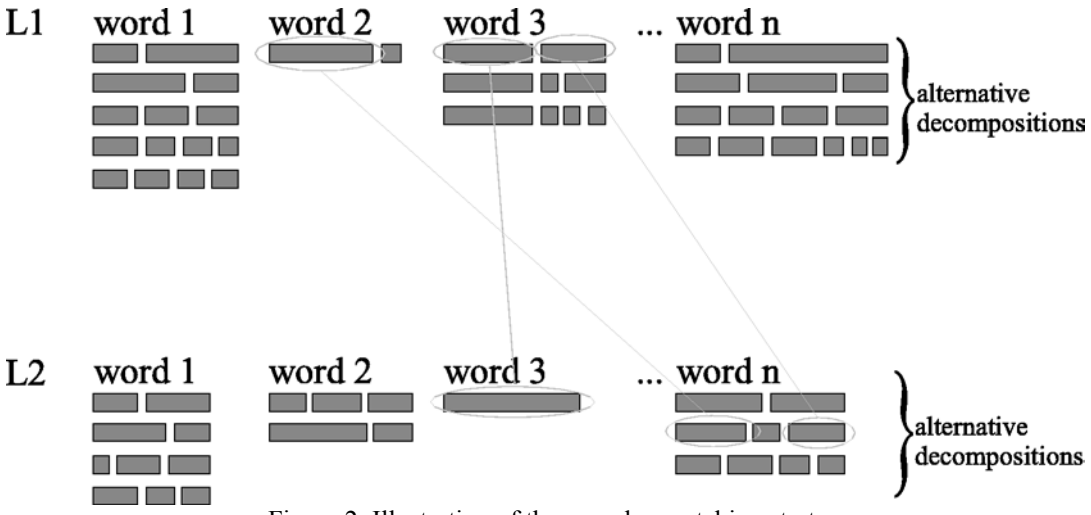

Figure 2: Illustration of the complex matching strategy

The complex strategy aims towards selecting the most promising sub-word level match. All analyses of all thus processed words are sorted according to their "quality" as defined above and put into an agenda. As long as the agenda contains analyses, the best candidate is removed and an attempt to construct a sub-word level matching for the candidate is made. If the attempt is successful, all words involved in the matching are marked as unavailable and all remaining analysis candidates that depend on these words (or their elements) are removed from the agenda. The successful matching is stored in a result set which is returned when the agenda becomes empty.
A best-first search through the space of possible sub-word level matches is conducted. This space consists of states that represent such matchings in progress. Such a state consists of a list of elements belonging to an analysis alternative that still have to be matched (todo-list) and a list of those elements that have already been matched or deleted (done-list). A state with an empty todo-list is a final state that represents a complete matching. The search begins in an initial state that contains in its todo-list all the elements of the analysis chosen from the agenda. The successor states are explored until suitable complete states have been found.
To generate successor states

- the first element is removed from the todo-list
- for each element of the other language that is a correspondence according to the dictionary, a new state is created whith the following properties:
    - the new done-list is the parent's done-list plus the first element and its corresponding element
    - the new todo-list is the parent's todo-list without the first element and its correspondence plus all the unmatched elements that are also part of the analysis to whose element the first element corresponds.
- a further state with the first element moved to the done-list is created. This state represents a deletion of the first element which is useful if the element cannot be matched.

It is important that targeted elements of a word must belong to the same analysis alternative.
To guide the best-first search, the different sub-word level matching states are assessed using the following formula:

value := 1.0 - (completion / numwords)
completion := number of matched free elements / total number of elements
numwords := number of actual words involved

The closer to 0 the value is, the better the sub-word level matching state is deemed to be. The assessment aims to favour analyses with few elements (most of which are free) that involve a small number of words.

5 Evaluation

Two things have to be assessed. On the one hand we would like to know how well the morphological analyser performs its task and what kinds of mistakes we find. On the other hand it is important to consider the effect the analyser has on the alignment quality.

5.1 Performance of the morphological analyser

In order to judge the quality of the morphological analyser we qualitatively evaluated the longest chapter from 'The War-of-the-Worlds' bitext (Chapter 2-07). We looked at the complex words in the German half of the text and their processing by the simple and complex analysis strategy. Some numbers first: the simple strategy processed 118 words making 13 errors. The complex strategy processed 38 words making 7 errors – it can be seen that the simple strategy seems to perform better (see below). In many cases correspondences could be found for parts of the complex words only. We find two types of errors: (1) analysis errors and (2) matching errors. Analysis errors are most frequent. The are usually due to one or two letter elements. Examples for a German analysis error and an English analysis error are given in Table 6.
Some analysis errors can be excluded by simple heuristics (in the style of: a complex word in German can never begin with a one-letter element except a) but in the long run we would welcome a fully-fledged morphological analysis.
Matching errors are errors where the correct analysis is chosen but the elements are not matched to the corresponding words. Sometimes it happens that there are two (almost) identical possible matches, as shown by the example *rötlich* (which should have been matched with *redly* but instead was matched

with *red*) in Table 7 while sometimes a different reading is chosen, as illustrated by the example scheinen which in this case is matched with *flash* instead of *glow* (*scheinen* can mean both).

419: After an interminable string of games , we supped , and the artilleryman **finished** the champagne .	432: Nach einer endlosen Reihe von Spielen nahmen wir unser **Abendessen** ein , und der Artillerist trank den Champagner aus .	**…** **ab|ende|(s)|(s)|(e):: finished** **…**
406: " We can dig better on this **Thames** - side burgundy , " said I.	419: " Es ist vielleicht besser , wenn wir bei unserem Burgunder weitergraben " , sagte **ich** .	**…** **ich :: (th)|(a)|me|(s)** **…**

Table 6: Two beads from 'The War-of-the-Worlds' parallel text. Analysis errors. *Abendessen* 'supper' in the first row is incorrectly analysed to contain *ende* 'finish'. In the second row the German *ich* 'I' is matched to the *me* in Thames.

426: The northern hills were shrouded in darkness ; the fires near Kensington glowed redly , and now and then an orange - **red** tongue of flame **flashed** up and vanished in the deep blue night .	439: Die nördlichen Hügel waren in tiefes Dunkel gehüllt , die Feuer in der Nähe von Kensington **schienen rötlich** herüber , und hier und da zuckte eine orangefarbene Feuerzunge auf , um in der tiefblauen Nacht gleich wieder zu verschwinden .	**...** **schein|(en) :: flashed** **rot|(lich) :: red** **…**

Table 7: A bead from 'The War-of-the-Worlds' parallel text: Matching errors

5.2 Improvement of the distance function

The War of the Worlds parallel text was aligned using four different distance functions: the base distance function consists of bilingual dictionary lookup plus cognate recognition. The other three functions are composed of the base function plus the simple and/or complex matching strategy. The thus obtained alignments were compared to a manually aligned reference. Table 8 shows the resulting recall (recall := number of correct beads / number of reference beads).

	base	base + simple	base + simple + complex	base + complex
War of the Worlds	96.51 %	97.21 %	97.16 %	96.37 %

Table 8: recall on 'The War of the Worlds' parallel text using different distance functions

The simple matching strategy improves the overall alignment quality (recall) by 0.7%. While this might seem to be only a small improvement, the difference on individual chapters was as large as 7.6% which we consider to be very successful.
Contrary, it seems that the present implementation of the complex matching strategy leads to a decrease in performance. We assume that this can mainly be attributed to awkward analyses, some of which could be filtered out using simple heuristics such as introducing a minimum word length. The impact of incorrect matches seems to outweigh that of correct matches. Thus the complex matching strategy (in its present form) should not be used in the distance function. However, when applied to the beads of a previously computed alignment, the complex matching strategy can be used to extract interesting matches that can be reviewed by a human expert.

6 Discussion

We have shown that morphological analysis can improve a dictionary-based distance function. This is true for a literary text such as 'The War of the Worlds' as well as for the rather technical EU debates. There are still some problems, however: the rough analysis of complex words, sketched here, often fails – in the long run a linguistically more adequate morphological analyser is desirable. Our approach can also help to improve the dictionary since correspondences can be added to it.

Acknowledgements

We would like to thank Petra Prochazkova for her help in the evaluation and Kim Wallum for critical comments.

References

Canadian Hansards:
aligned versions available from http://www.isi.edu/natural-language/download/hansard/

Fuhrhop, N 1998 *Grenzfälle morphologischer Einheiten*. Tübingen, Stauffenburg

Gale, W A & Church, K W 1991 A Program for Aligning Sentences in Bilingual Corpora. In: *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, CA, pp 177 –184

Fung, P 2000 *A statistical view on bilingual lexicon extraction.* In: Veronis, J (ed) 2000, pp. 219-243

ten Hacken, P & Lüdeling, A 2002 Word Formation in Computational Linguistics. Tutorial. In: *Proceedings of TALN 2002*, Nancy, pp. 61-87

Kay, M & Röscheisen, M 1993 Text-Translation Alignment. *Computational Linguistics 19*, pp. 121-142

Lüdeling, A & Fitschen, A 2002 An integrated lexicon für the automatic analysis of complex words. In: *Proceedings of the Tenth International EURALEX Congress*, Copenhagen, vol 1, pp. 145 – 152

RAPID, The Press and Communication Service of the European Commission 1995-2002, RAPID is a database giving a daily view of the activities of the European Union, http://europa.eu.int/rapid/start/welcome.htm

Schmid, T, Lüdeling, A, Säuberlich, B, Heid, U, Möbius, B 2001 DeKo: Ein System zur Analyse komplexer Wörter. In *GLDV - Jahrestagung 2001* pp. 49-57.

Schmid, G 1994 TreeTagger - a language independent part-of-speech tagger. *Manuscript*. Available as http://www.ims.uni-stuttgart.de/Tools/DecisionTreeTagger.html.

Steven, G A 1994 *String Searching Algorithms*. Singapore, World Scientific Publishing

Tschorn, P 2002 *Automatically aligning English-German parallel texts at sentence level using linguistic knowledge*. Unpublished Master's thesis, University of Osnabrück

Veronis, J (ed) 2000 *Parallel Text Processing – Alignment and use of Translation Corpora*. Kluwer, Dordrecht

Wells, H G 1898 *The War of the Worlds*.
English version available from: www.fourmilab.ch/etexts/www/warworlds/
German version available from: www.geocities.com/Area51/Corridor/8282/