Exploiting Linguistic and Statistical Knowledge in a Text Alignment System

Dissertation

zur Erlangung des Grades eines Doktors der Kognitionswissenschaft

eingereicht am Fachbereich Humanwissenschaften

der Universität Osnabrück

vorgelegt von

Bettina Schrader

Osnabrück, März 2007

Exploiting linguistic and statistical knowledge 0.0 Bettina Schrader 2 Dissertation

Überarbeitete Fassung

beruhend auf der begutachteten Fassung mit dem Titel
Exploiting Linguistic and Statistical Knowledge in a Text Alignment System

Dissertation zur Erlangung des Grades eines Doktors der Kognitionswissenschaft,
eingereicht am Fachbereich Humanwissenschaften
der Universität Osnabrück

Acknowledgements

It is a laborious way from a first, sketchy idea for a thesis to its final version. In my case, the idea was born when I was writing the final section of my Master's thesis at the *Institut für Maschinelle Sprachverarbeitung* in Stuttgart. Now, four years later, I've completed my dissertation at the *Institut für Kognitionswissenschaft* in Osnabrück, a vital, lively place where everyone is open to new ideas, and concepts, sometimes radically so.

And so I am deeply indebted to my supervisors Prof. Dr. Peter Bosch, Dr. habil Helmar Gust and Prof. Dr. Stefan Evert for encouraging, and supporting me, to elaborate my sketchy idea to a fully grown dissertation – a text alignment system that differs radically from any I've ever heard of before. Additionally, many thanks to Martin Volk for co-refereeing this thesis.

Special thanks also go to Dr. Carla Umbach, the coordinator of the Cognitive Science Graduate Program, and the members of the Doctorate Colloquium, for helpful hints and discussions.

I deeply appreciate the discussion with and help of Philip Reuter, Klaus Dalinghaus, Graham Katz, Angela Grimm, Peter Geibel, Joscha Bach, Gemma Boleda (for help with Spanish), Martin Volk (who also provided morphological annotations), Judith Degen (especially for her word alignment annotations), and the participants of the Stockholm Symposium on Parallel Treebanks which took place in September 2006. I also thank my proof-readers Arne Fitschen, Julia Ritz, and Mark Hopkins. Many thanks also to the members of my dance club, most importantly to Britta Deuwerth, Thomas Vinke, Constanze Weth and Jürgen Volbers for taking my mind off the thesis every now and then.

Last, but definitely not least, I thank my parents and my brother: I would not have embarked on the mission to write a PhD thesis, much less completed, without their unwavering support.

Abstract

Within machine translation, the alignment of corpora has evolved into a mature research area, aimed at providing training data for statistical or example-based machine translation systems. Moreover, the alignment information can be used for a variety of other purposes, including lexicography and the induction of tools for natural language processing. The alignment techniques used for these purposes fall roughly in two separate classes: sentence alignment approaches that often combine statistical and linguistic information, and word alignment models that are dominated by the statistical machine translation paradigm. Alignment approaches that use linguistic knowledge provided by corpus annotation are rare, as are as non-statistical word alignment strategies. Furthermore, parallel corpora are typically not aligned at all text levels simultaneously. Rather, a corpus is first sentence aligned, and in a subsequent step, the alignment information is refined to go below the sentence level.

In this thesis, the distinction between the two alignment classes is withdrawn. Rather, a system is introduced that can simultaneously align at the paragraph, sentence, word, and phrase level. Furthermore, linguistic as well as statistical information can be combined. This combination of alignment cues from different knowledge sources, as well as the combination of the sentence and word alignment tasks, is made possible by the development of a modular alignment platform. Its main features are that it supports different kinds of linguistic corpus annotation, and furthermore aligns a corpus hierarchically, such that sentence and word alignments are cohesive. Alignment cues are not used within a global alignment model. Rather, different sub-models can be implemented and allowed to interact. Most of the alignment modules of the system have been implemented using empirical corpus studies, aimed at showing how the most common types of corpus annotation can be exploited for the alignment task.

Contents

Chapter 1

The Global Picture

> Voilà, une fois de plus!
> Les dictionnaires n'expliquent bien que les mots qu'on connaît déjà.
> (There again!
> Dictionaries don't explain anything but what we already know.)
>
> (Schmitt 2003)

Although young Momo, protagonist in "Monsieur Ibrahim et les fleurs du Coran" (Schmitt 2003), talks about an encyclopaedia rather than a lexicon, his statement is true about bilingual dictionaries, too: They contain information on words that we already know, but when looking up a word that we aren't familiar with, we might encounter difficulties in the amount of information that is necessary to use the translation appropriately. Augmenting dictionaries with information on what we *don't* know, therefore, is a necessity. On the other hand, many researchers have noted that "existing translations contain more solutions to more translation problems than any other available resource" (Isabelle et al. 1993, p. 205). Translations contain all those expressions that one might search for hopelessly in existing dictionaries, as well as example contexts showing when to use an expression appropriately. If there is a means to tap this translation knowledge, it will be possible to augment standard dictionaries, improve machine translation systems, or extend one's own foreign language skills.

Fortunately, this translation knowledge can be accessed by using *parallel corpora*: texts that are available in the language they have been written in (the *source language*) and one or more translations in other languages (the *target language(s)).* A technique called *text alignment* has been developed, originally for purposes in machine translation, that automatically determines which specific segment of the source language text is translated where, and how, in the target language. The segments of text alignment are usually either sentences or words, and aligning them means setting them into correspondence such that they *mean the same*, i.e. that they are *translations of each other*. If a parallel corpus is aligned, this alignment information can be used for a variety of purposes: one might look up an expression in German, and use it to retrieve its translation in English. This procedure is exactly the same lookup strategy that one might use with a dictionary. But instead of an ordered list of words and their translations, it is a coherent *text* that is used for the lookup, and unlike the dictionary, this texts also contains context information, namely at least one example sentence.

However, a parallel corpus seems to be an artificial construction, and the idea of aligning its words and sentences seems likewise artificial. Parallel texts seem to be something that do not exist in ordinary life, nor does one usually engage in aligning them. Quite the opposite is true: parallel corpora lead a very ordinary existence, and many people have aligned them every now and then: each and every technical equipment for sale in the EU comes with a parallel text, namely a manual, and sometimes it is easier to understand a manual by comparing the instructions given in different languages. Furthermore, tourists and business travellers alike encounter parallel texts on signs at airports, international railway stations, immigration offices etc. Finally, students learning a foreign language can use parallel texts when starting to read novels in that foreign language. In this case, the novel is viewed in parallel, the native language on one, the foreign one on the other page, so that the student may refer to the text in the native language to resolve those comprehension problems he or she encounters in the foreign language.

In order to make the most of such a “parallel” novel, the student will automatically align at least some parts of the two texts, i.e. he or she will scan it in both languages, and determine which expression of the one language is translated how in the other. In order to be fast, the student will probably stop short of understanding the texts, but will use heuristics to rapidly narrow down the search space to the places he or she is interested in. This process is imitated in text alignment systems within natural language processing (NLP), and once the information on translation pairs has been extracted from the corpus, it is used within machine translation and lexicography, in order to train machine translation systems and extend bilingual dictionary information. Furthermore, parallel corpora and alignment information have been, and are being, used in other research areas within and outside NLP.

Several methods have been developed to align parallel texts on the sentence and on the word level. While sentence alignment works comparatively well, the quality of the word alignment information is relatively low, as word alignment systems depend heavily on statistical models that do not capture linguistic structure adequately: sentences e.g. are often wrongly taken as internally unstructured *bags of words*. Humans, on the contrary, are sensitive towards linguistic structure and would rarely assume that a sentence in a foreign language was an unordered sequence of words. Accordingly, to improve word alignment quality, it seems plausible to direct attention more towards linguistic structures in texts, and towards finding strategies that use linguistic information for word alignment. Just like humans are aware of linguistic features and linguistic structures, whether phonological, morphological, syntactical, or semantic, it should be possible to make an alignment system aware of this information, and to enable the system to use this knowledge for the alignment task.

The aim of my PhD thesis has been to develop such an alignment system: a system that can exploit structured information in that it uses diverse levels of linguistic descriptions for aligning sub-sentence elements like phrases and words. These levels include lemma and word category information, or information on syntactic constituency. Additionally, this system has been designed to compute alignments driven by both linguistically motivated rules and statistically derived similarities between words. The resulting system architecture can integrate different alignment strategies, and it is capable of processing hierarchically structured input text as well as hierarchically align texts. This system can simultaneously align a parallel, bilingual corpus both at the sentence, word and at any other level, e.g. at the phrase level. All alignment information that is generated during the processing of the corpus is available afterwards, both as a bilingual dictionary and an aligned corpus.

To my knowledge, this is the first time an alignment system has been developed that is not restricted to aligning either sentence or word alignment, hence its name is program:

ATLAS – an **a**lternative **t**ext a**l**ignment **s**ystem.

The one thing that ATLAS does not do, and does not even aspire to, is to align texts in a human like fashion. It does use linguistic information, but there is no reason to suppose that its strategies are the same that humans use, because (a) no one knows how humans align, (b) secondly, because it would have been beyond the scope of this thesis to investigate human language capacity and simultaneously develop a system that uses these strategies.

However, research on alignment techniques helps one to learn about human language skills. Firstly, designing a system that computes alignment information with reasonable quality requires us to formalize at least certain aspects of these human linguistic skills. Secondly, the system's output, the aligned corpus, can be used to extract further linguistic knowledge by means of contrastive linguistic research, etc.

1.1 A Closer Look

Up to now, the discussion of the alignment task and its applications has been rather informal. Subsequently, the task will be more formally defined, including a more formal definition of a parallel corpus. Some examples of parallel corpora and treebanks will also be given, and I list research areas that have been profiting from alignment information. Finally, I describe the requirements on a hierarchical sentence and word alignment system, especially if it has to align linguistically annotated text.

1.1.1 Definition of the Task

Parts of the terminology have been used ambiguously: the term *alignment* is often both used and understood as describing i) the task of finding and annotating the correspondences between sentences, words, etc. in a parallel corpus, ii) the set of all correspondence pairs within an aligned corpus, and iii) a specific instance contained in a global alignment path, e.g. a specific sentence or word pairing. Furthermore, specific sentence or word pairings are often called *beads* (cf. Chen 1993; Tschorn 2004; Moore 2002). But there is also a second, fully synonymous term for bead, namely the term *link* (cf. Tiedemann 1999; Melamed 1997a; Melamed 2000; Cherry and Lin 2003). Finally, parallel texts are also often referred to as *bitexts*, a concept first introduced by Harris (1988): A bitext is a source language text being simultaneously present with its translation in another language.

Subsequently, this concept has been adopted within research on machine translation and text alignment, and given a geometrical interpretation (cf. Melamed 2001; Simard and Plamondon 1998): In this perspective, the source and language texts of a parallel corpus are the axes of a rectangular search space, with the bits or characters of each texts constituting the points on these axes. In this theoretical framework, the alignment task consists of generating a *bitext map* that gives information on which coordinates of the bitext are *points of correspondence*, i.e. positions where information in the source language and target language texts are equivalent.

Fortunately, the two theoretical frameworks are compatible: bitext maps can be converted into sentence and word alignments, but with some loss of information: a bitext map typically records the correspondences for each and every coordinate in the bitext. Converting this information to sentence or word alignment means discarding all but the "best" correspondence points.

Alignment Definitions used here

In this thesis, the distinction is drawn between a global alignment, which holds between two sequences, and the correspondences that hold between the elements of these sequences. These correspondences are based, on translational equivalence, i.e. two aligned elements convey the same meaning, although they do so using different languages. The correspondences are called *links*, and it is possible to specify whether the segments in question are sentences (a sentence link), words (a word link), or any other textual unit.

Formally, the terms *sequence*, *links*, and *alignment* are defined thus:

1. A sequence is an ordered list of elements.

2. Given two sequences S_1 and S_2, an ordered pair consisting of n elements of S_1 and of m elements of S_2 is a *link* $(n, m \in \mathbb{N}_0, n+m > 0)$[1].

3. An alignment of two sequences S_1 and S_2 is a list of all links formed by the elements of S_1 and S_2, where each element of S_1 and each element of S_2 must belong to exactly one link.

4. If two sequences consist of elements of type *b*, the links of those elements are called *b links*, and the resulting alignment is a *b alignment.* If a sequence S_1 consists of elements of type *b*, and the second sequence S_2 consists of elements of type *c*, I call the links of those elements *b-c links*, and the resulting alignment is a *b-c alignment.*

An alignment system or alignment method is used to automatically determine an alignment via the links between the elements of the two input sequences. If a bilingual corpus is aligned, the two halves of the corpus, one a text in the source language, the other in the target language, are the sequences to be aligned, and the elements of these sequences can be paragraphs, sentences, sentence-internal phrases, or words.

Bilingual and Multilingual Corpora

In principle, it is possible to compute an alignment between any number of texts at once, whether these are written in two, three, or, in the case of European Union texts published before the last accession of countries, in 11 languages. However, aligning more than two languages is prohibitively expensive in terms of computational complexity: when aligning the sentences within two texts, one containing *n*, the other containing *m* sentences, there are *n* times *m* possibilities for 1:1 links. Furthermore, as soon as other link types like insertions and deletions (n:0 or 0:m links) are allowed, the number of possible link types increases further. If a third language is added to the alignment task, then the number of possible link types increases even more. Generally, the complexity of the alignment task is $O(n^l)$, n being the length of the corpus, and l the number of languages involved.

Some researchers have worked on multilingual alignment, arguing that *three languages are better than two.* Simard (1999b) and Simard (1999a) e.g. demonstrates that multilingual alignment is feasible, at least on the sentence level, and moreover can be used to improve bilingual alignment information. One strategy to keep the computational complexity low is to first compute bilingual alignments for each language pairing in the corpus, to determine which language pairs are more similar, and hence more reliably alignable. The ”best” language pairing, i.e. the one with the best overall alignment score, is used as starting point to extend the alignment information to more and more languages: each alignment link is aligned to a segment in a third language. Thus, the task of multilingual alignment is broken down into stepwise bilingual alignment.

A second strategy to arrive at a multilingual alignment is to compute bilingual alignments, and combine them by computing their transitive closure: given e.g. the two links (a↔b) and (b↔c), a, b, and c being segments of three different languages, the transitive closure would yield the additional linking (a↔c). This transitive closure strategy has been used by Borin (2000) to improve word alignment quality in a trilingual corpus.

[1] Thus deletions and insertion (0:n links) are allowed as well as any other link type, e.g. 1:1 links or 1:7 links. The *empty link* (0:0 link) is excluded.

Apart from improving bilingual alignment information however, there seems to be no real purpose in computing truly multilingual alignments. Furthermore, the techniques used by Simard (1999a), Simard (1999b) and Borin (2000) seem to have exhausted the need to do research on multilingual alignment: as it is based on bilingual alignment, it is not necessary to pursue this topic on its own. Accordingly, I do not work on multilingual alignment but restrict research on how to link correspondences to the bilingual case.

Parallel and Comparable Corpora

Another precondition of the alignment task is that the texts are *parallel*, i.e. they convey exactly the same meaning, usually because they are a text and its translation in another. Of course, it is also possible that both texts are translations from a third source, or that the original language is unknown. In both cases, each segment of a parallel corpus can be taken to have a translation in the other language(s) of the corpus, and the alignment task consists in identifying which segments are translations of each other.

Parallel corpora have to be seen in contrast to *comparable* ones, i.e. corpora available in two or more languages that cover roughly the same topic or genre, but that are *not* translations of each other, nor need be created from a common source. Aligning comparable corpora thus presents the additional problem of finding out which segments of the source and target language texts actually are translations of each other, or which segments are so close in meaning that they can be taken to be translationally equivalent. It is only these segments that should be aligned.

When aligning parallel corpora, the difficulty of distinguishing between translated and not translated segments is avoided. My PhD is only concerned with aligning parallel corpora with the idea that once text alignment on parallel corpora is reasonably well understood, and can be done with a sufficiently high confidence, it becomes feasible to adapt research results and techniques to the alignment and exploration of comparable corpora.

1.1.2 Existing Parallel Corpora and Treebanks

Since the first alignment programs have become available, more and more parallel corpora have been set up, for more and more languages. Most often, the English-French Canadian Hansards (cf. Simard et al. 1992; Brown et al. 1993) and the multilingual Europarl (cf. Och and Ney 2004; Och et al. 1999; Vogel et al. 2000; Köhn et al. 2003) are used for developing alignment techniques. Furthermore, *parallel treebanks* are currently being created, i.e. parallel corpora where both source and target language texts are syntactically annotated, and where alignment information is available on the sentence and below.

It is nowadays virtually impossible to keep track with the development of parallel corpora – too many have been created and are currently being created. Any overview on parallel corpora, including treebanks, will hence be incomplete. Accordingly, the following is only an overview of those corpora that have played, or are still playing, an important role for the development and testing of alignment techniques.

Additionally, three parallel treebanks are included that are currently being set up, as I expect them to play an important role for further research: due to their rich annotation, they can be used to develop and train alignment systems. Furthermore, parallel treebanks provide information that is reliable enough to be used as evaluation data, whether for quantitatively assessing alignment systems, or for qualitatively analysing alignment errors.

Parallel Corpora

Most parallel corpora have been preprocessed scarcely: they have been tokenized and sentence segmented, and the sentences have been aligned. Higher level annotations like information on parts of speech (POS) are absent. As such, the corpora are well-suited for unsupervised training of alignment and programs for statistical machine translation (SMT), but they cannot efficiently be used for in e.g. contrastive linguistics, or for developing an alignment system that makes use of e.g. POS information.

Canadian Hansards One of the most prominent parallel corpora are the *Canadian Hansards*, the complete collection of debates of the 36th Canadian Parliament. These texts are available in the two official languages of Canada, English and French, and were among the first texts that were used for developing alignment techniques.

The complete corpus consists of roughly 1.4 million sentence pairs, having been sentence-aligned using geometric sentence alignment (GSA) (Melamed 2001), with 22.190,000 English and 13.771,000 French tokens. The release described by Germann (2001) is slightly smaller, containing only 1:1 sentence links, every other link type having been filtered out. However, it still consists of roughly 89% of the complete corpus, in 1,278,000 sentence links, containing 19,830,000 English and 21,242,000 French tokens.

BAF Another very well-known English–French corpus, the BAF, has been developed in the context of the ARCADE alignment evaluation project (Simard 1998). Accordingly, it has been *manually* annotated with sentence alignment information. The corpus consists of texts from various genres, among them texts from the Canadian Hansards, the Canadian Supreme Court, UN documents, four scientific articles, a technical manual and Jules Verne's novel "De la terre à la lune". The corpus consists of roughly 400,000 tokens per language, i.e. it is considerably smaller than the Canadian Hansards.

HKUST English–Chinese Parallel Bilingual Corpus Unlike many parallel corpora, the parallel corpus that has been compiled by the *Hong Kong University of Science and Technology* contains text from a non-Indoeuropean language, Chinese. Similar to the Canadian Hansards, it consists of parliamentary proceedings of Hong Kong. After preprocessing and automatic sentence alignment, the corpus consists of 3.3 million English and 3.2 million Chinese tokens (Wu and Xia 1995).

Verbmobil This corpus was originally created as training data for the *Verbmobil* machine translation system, aimed at the online translation of spontaneous speech during travel or business negotiations. The languages of the corpus are English, Japanese and German, but for the purposes of word alignment research, the language pair English–German is primarily used. This data set contains roughly 34,500 sentences in English and German, with roughly 150,000 tokens per language. A subset of 354 sentence links has since been manually aligned to serve as gold standard for word alignment evaluation (Och and Ney 2003).

Europarl Nowadays, the multilingual EUROPARL corpus (Koehn 2005; Köhn 2003) is widely used. It is a large corpus of parliamentary debates having taken place between 1996 and 2001. All of the texts are available in the eleven official EU languages of that time, thus the corpus can be used for research on many different language pairs. The EUROPARL corpus has been sentence aligned using the length-based approach by Gale and Church (1991b)[2] and contains roughly 30

[2] A discussion of this approach can be found in section 2.1.1.

million tokens per language. It has lately been added to the OPUS corpus collection (see below), and linguistic annotation like POS information has been added for some languages.

Acquis Communautaire A very recently released corpus, the Acquis communautaire, even features 20 EU languages plus Romanian (Steinberger et al. 2006). The texts are legal documents of the European Union, some of them dating back to the 1950s. They are manually classified according to subject domain, and automatically aligned at the paragraph level. Furthermore, the texts have been POS-tagged and lemmatized (Erjavec et al. 2005). On average, the texts consist of 8.825,544 tokens per language.

Parallel Treebanks

Unlike the parallel corpora described above, parallel treebanks are annotated syntactically for all languages. Furthermore, they are typically not only aligned at the sentence level, but also at the phrase or word level.

Prague Czech–English Dependency Treebank The *Prague Czech–English Dependency Treebank* has been built in order to facilitate research in MT (Čmejrek et al. 2004). It contains roughly 20,000 sentences taken from the English Penn treebank (Marcus et al. 1993) along with their Czech translations. The translations have been created by human annotators, with the guiding principle to translate as literally and as closely as possible . Most of the Czech translations have been automatically tokenized, POS-tagged, lemmatized and parsed. The English originals, annotated with phrase structure information, have been converted automatically into the dependency structures used in the Prague Dependency Treebank (Böhmová et al. 2001). A small subset of the parallel corpus has been annotated manually, containing 1,257 English and 515 Czech sentences.

Stockholm Trilingual Treebank A trilingual English–Swedish–German parallel treebank has been created in Stockholm, consisting of the first two chapters of Jostein Gaarder's novel *Sophie's World* and economy texts (Volk et al. 2006). This treebank is relatively small with roughly 1000 sentences per language, but unlike the Prague dependency parallel treebank, the sentences are taken from different genres, 500 sentences having been taken from the novel, the others being economy reports. The monolingual annotations have been added semi-automatically, using a procedure similar to that of the TIGER treebank (Brants et al. 2002), with an additional node insertion step for the German and Swedish texts (Samuelsson and Volk 2004). The phrase and word alignment information, finally, has been annotated manually.

The CroCo Corpus As of June 2005, an English–German treebank is finally being built, in Saarbrücken in the CROCO project[3]. This treebank is being created for research in translation studies, and it will be balanced with respect to translation direction and genre: each of the 8 genres is to be represented by parallel samples of 2,000 words, in both translation directions English→German and German→English (Neumann and Hansen-Schirra 2005; Hansen-Schirra et al. 2006). The annotation is done mostly automatically, using a variety of publicly available NLP-tools, and includes POS and morphological information, as well as shallow syntactic analyses and information on grammatical functions. Furthermore, words and syntactic constituents are automatically aligned.

[3] http://fr46.uni-saarland.de/croco/

Other Sources for Parallel Texts

An important resource for parallel corpora is the OPUS corpus collection (Tiedemann and Nygaard 2003; Tiedemann and Nygaard 2004), consisting mainly of technical manuals from open source software. In 2004, the collection consisted of roughly 30 million words of technical manuals in 60 languages. Afterwards, the EUROPARL corpus mentioned above has been added to the collection. During corpus preprocessing, sentence alignment information is computed using the length-based approach by Gale and Church (1991b), and linguistic annotation is provided for those languages with publicly availably NLP-tools, hence the extent of the annotation information varies from language to language.

Another possibility to collect parallel corpus material is to use the internet, e.g. with the technique suggested by Resnik (1999). There have also been experiments on automatically creating parallel texts using machine translations from natural languages (Callison-Burch and Osborne 2003): existing parallel texts are used to train statistical MT models; in a second step, these SMT models are used to translate monolingual texts, thus creating additional parallel MT training data where an original text is paired up with MT output.

1.1.3 Uses of Aligned Corpora

Aligned corpora and alignment techniques have become increasingly popular over the past 15 years for a variety of applications and purposes. Sometimes, the corpora are used for machine translation (cf. Brown et al. 1990; Brown et al. 1993; Tiedemann 2001; Imamura 2001), or to extract bilingual dictionaries or term glossaries (cf. Klavans and Tzoukermann 1996; Smadja et al. 1996). Additionally, aligned corpora can be used to *bootstrap* or *induce* linguistic tools or resources (cf. Yarowsky et al. 2001; Kuhn 2004b; Mitkov and Barbu 2004; Gale and Church 1991a; Bentivogli and Pianta 2004).

Bilingual Applications: Machine Translation, Cross-Language Information Retrieval, and Lexicography

The prime purpose behind aligning corpora and developing better alignment techniques has always been machine translation (MT): First and foremost, because aligned corpora can be used to train translation models for statistical machine translation (SMT). Foster et al. (2003) e.g. use GIZA++, an implementation of the well-known IBM translation models (Brown et al. 1990; Brown et al. 1993) and the equally famous HMM model by Vogel et al. (1996)[4] in order to train an SMT system for translation from Chinese to English. Fox (2005) even extends the alignment information in a sentence aligned corpus to word and dependency structure alignments in order to train an SMT system that can translate between Czech and English.

Aligned corpora are also useful for extending and revising the dictionaries that MT systems use (Klavans and Tzoukermann 1996; Smadja et al. 1996), and for creating the knowledge bases used within example-based machine translation and translation memories (Tiedemann 2001; Samiotou et al. 2004; Rahman and Azih 2004; Brown et al. 2005; McTait 2001; Imamura 2001). As Thurmair (2006) e.g. notices, even MT dictionaries with several 100,000 entries still have gaps that can be decreased using this corpus information. Furthermore, corpus information is helpful in order to choose a correct translation among several available possibilities.

Parallel corpora aligned at the word level are also important for Cross-language Information Retrieval (CLIR): In these applications, a search engine uses a query in a source language to retrieve documents in a target language (Hiemstra 1998; Hiemstra et al. 1997; Hiemstra 1996). Hence it needs a bilingual term dictionary to translate a query.

[4] See section 2.2 for a description of the SMT models

Finally, word-aligned corpora can be exploited for lexicographic tasks like dictionary creation and term extraction. Sahlgren and Karlgren (2004) use random indexing on parallel corpora to arrive at bilingual dictionaries for the word pairs Swedish–Spanish and German–English, whereas Conley (2002) extends word_align, a language-independent word alignment method (Dagan et al. 1993) to align multiword sequences and construct a bilingual dictionary. Martin et al. (2003) align an English–Inuktitut parallel corpus in order to automatically create a term glossary for that language pair. Hull (2001) also uses parallel corpora in order to generate bilingual terminology dictionaries for use in translation memories and CLIR, and Moore (2003) uses aligned parallel corpora in order to translated named entities. Finally, parallel corpora are useful for extracting *poor man's synonyms* (Tschorn 2004) and paraphrases (Bannard and Callison-Burch 2005).

Monolingual Applications: Improvement and Development of NLP-tools

An application where it is crucial that the alignment algorithms used are language independent is a bootstrapping approach where alignment information is used to speed up the development of monolingual NLP-tools. In this approach, a language for which NLP-tools such as POS-taggers, lemmatizers or parsers are available, is paired up with a language for which these tools do not exist. The text of the resource-rich language is annotated, and aligned to the resource-poor language. Via the alignment links, the annotation is then projected from the one language to the other. As a result, an annotated corpus has been created for the resource-poor language, which then can be exploited to train NLP-tools.

Yarowsky et al. (2001) showed how to create a POS-tagger and several other NLP-tools using this bootstrapping approach: the researchers annotated the English texts of two different parallel corpora with POS-information and aligned them to the other languages using an off-the-shelf word alignment tool. Secondly, the POS-tags were transferred from the English texts in the corpora onto those of the target languages, in this case, French and Chinese. As a result, the French and Chinese texts were annotated with POS-information. The tagger achieved an accuracy of 97% using a core tagset, i.e. the set contained tags for word classes such as *noun, verb, etc.*, but without more fine-grained morphological distinctions. Even using a more fine-grained tagset, the induced POS-tagger still achieved an accuracy of 93%, thus indicating that NLP-tools can be developed for a language without manually creating training data first, but using a parallel corpus instead.

Less shallow NLP-tools, like probabilistic grammars can also be induced using a parallel, word aligned corpus (Kuhn 2004b). Furthermore, experiments have been carried out how to learn probabilistic context-free grammars for *both* languages of a parallel corpus, and with synchronous parsing of word aligned texts (Kuhn 2004a; Kuhn 2005). Lately, Hwa et al. (2004) have used parallel corpora to also induce dependency parsers.

Aligned corpora can also be used to improve the performance of monolingual NLP-tools. Mitkov and Barbu (2004) use a parallel corpus, aligned at the word level, to improve the performance of two pronoun resolution systems: In a two-pass process, anaphora resolvers for the two languages are run on a French–English corpus. In the second pass, the resolution decisions by the two programs are examined and revised based on information from the translation: the morphological gender markings in French are used to improve English anaphora resolution, while English syntactic patterns help to recognize and improve errors of the French anaphora resolver.

Research has also been carried out on how to automatically learn semantic relations between words from parallel data (Dyvik 2004): translation pairs from a parallel corpus are compared based on the assumption that the different senses of a polysemous word will receive different translations while synonyms will share translations. Huang et al. (2002), on the other hand, explore the possibility of translating English wordnet information into Chinese. Gale et al. (1992) use a parallel corpus to extract training and test data for a word sense disambiguation systems.

In this research area, the focus is not only on inducing NLP-tools. Rather, the annotation projection also serves at speeding up the creation of monolingual annotated training data like corpora for training word sense disambiguation tools (Gale and Church 1991a; Bentivogli and Pianta 2004)

1.2 Requirements

To sum up, parallel corpora are used in a variety of domains and for a variety of purposes, and in all of these approaches, it is the alignment information that makes the parallel corpora useful. In order to achieve a maximum utility of the corpora, it is then necessary to achieve high quality alignments both on the sentence and on the word level.

One possibility to achieve this goal is to design an alignment system that can exploit different kinds of corpus annotation during the computation. Such a system can never be called truly language independent, as it has to rely on the existence of NLP-tools such as POS-taggers, lemmatizers, and the like. Or, it has to include strategies to lemmatize, POS-tag, etc. a parallel corpus while aligning it. Further, the alignment system should be modular in order to facilitate the accommodation of new language pairs and kinds of corpus annotations. In other words, it should not be language-pair specific, i.e. tailored down to a specific language pair, or to a specific type of corpus annotation. Its output should at least be corpus alignment information, and ideally, it would also compute a probabilistic bilingual dictionary based on its alignment information. With respect to the system's performance, alignment quality should be preferred over robustness and speed.

The design requirements that were finally adopted for the development of ATLAS were that the system should

- align *bilingual parallel corpora*,
- be *language-pair independent*: it should be neither specifically applicable to a single language pair like German–English, nor should it be so language-independent that only statistical cues are used.
- use *linguistic corpus annotation*: The kinds of supported corpus annotation should at least include information on word category membership, supplied by a part-of-speech tagger (POS-tagger), lemmas, provided by a lemmatizer or stemmer, and syntactic constituency supplied by a parser or chunker[5].
- align using a *variety of alignment strategies*, which may be statistical, heuristic, or rule-based in nature,
- be *modular* to allow extensions to other kinds of language pairs and annotations
- align *hierarchically* and be structure-sensitive, such that word, phrase, and sentence alignment information can be generated in parallel,
- produce corpus alignment information as well as a bilingual dictionary,
- prefer *high-precision* alignment information, at the sacrifice of speed and robustness.

ATLAS has been designed to meet all of these requirements, as will be described in subsequent chapters.

[5] These types of information should be supported because automatic means to supply them are available for a wide range of languages, and they tend to be the first being developed for languages with scarce resources.

1.3 Overview

First, I discuss the various approaches towards sentence, word and phrase alignment (chapter 2), including their merits and disadvantages.

Then, I describe the alternative text alignment system in more detail: in chapter 3, I give an overview on the system's architecture, including a description of the development corpora that I used. Afterwards, I describe the alignment strategies that are currently part of the alignment system, and experiments that I have carried out on the development corpora, concerning the usefulness of specific kinds of linguistic corpus annotation or statistics (chapter 4). I also discuss which evaluation methods have been used to assess the performance of sentence and word alignment systems, and do a thorough evaluation of the text alignment system ATLAS on a specifically designed gold standard (chapter 5).

Last, but not least, I wrap up the most important results and insights gained during the work on ATLAS, and summarize directions for further work (chapter 6). Technical details can be found in the appendix (appendix 6.2).

Chapter 2

Previous Approaches to Bilingual Text Alignment

> Do we really know how we translate or what we translate? What is the ’interlingua’? Are we to accept ’naked ideas’ as the means of crossing from one language to another? [...] Translators know they cross over but do not know by what sort of bridge. They often re-cross by a different bridge to check up again. Sometimes they fall over the parapet into limbo.
>
> (Firth 1957, p. 197)

Many alignment methods have been suggested for aligning a text at the sentence, phrase, or word level. *Sentence alignment* is concerned with setting the sentences of a bilingual corpus into correspondence with each other. *Paragraph alignment* is usually considered a part of the sentence alignment task: usually, the paragraphs of a parallel corpus are aligned prior to, but using the same techniques as, the sentence alignment. Hence there is no need to describe paragraph alignment strategies in detail.

Word alignment is concerned with aligning all words of a parallel corpus so that each word link is a translation pair. An alternative to word alignment is the alignment of *phrases*, where sequences of words are set into correspondence with each other. These sequences can be syntactic constituents or any other arbitrary sequences. Some approaches even compute phrase alignments via word links. In this case, a parallel corpus is first aligned at the word level, and then the word links are examined to generate and link phrases, or if phrases are annotated, they *inherit* the alignment information from the words that they contain.

Usually, the different alignment methods are applied iteratively, i.e. phrase or word-level alignment depends on a prior sentence (and paragraph) alignment. However, information on sub-sentence links – such as word links – may be a by-product of a sentence alignment method. In these cases, the word links are usually discarded after the sentence alignment task has been completed, i.e. they are not re-used for the subsequent word alignment.

Firstly, I discuss the most important sentence alignment strategies, which basically fall into the classes of length-based (section 2.1.1) and anchor point-based strategies (section 2.1.2), but some approaches combine the two different strategies (section 2.1.3). After summarising the different approaches, I turn to the different word alignment strategies (section 2.2). Finally, I also highlight the most important phrase alignment approaches (section 2.3).

2.1 Sentence Alignment

Many methods have been proposed and implemented for aligning sentences. They can be divided into two major groups: length-based versus anchor point-based approaches. Of course, both of them can be and are combined to *hybrid* sentence alignment approaches. Thus, it is sometimes hard to draw the line between length- and anchor-based approaches.

2.1.1 Methods using Sentence Length

Length-based methods compute sentence alignment by using the respective sentence lengths as alignment cues. The main idea is that

> longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences.
>
> (Gale and Church 1991b, p. 78)

The method was first introduced by Gale and Church (1991b). It uses the number of characters in the sentences of a corpus as alignment cues: two sentences are aligned if their character numbers indicate that they are of similar length.

Gale and Church (1991b) base their method on empirical findings on the UBS corpus containing German bank reports that have been translated into English and French. The corpus consists of roughly 14,700 tokens (725 sentences and 188 paragraphs) per language. According to the analysis by Gale and Church (1991b), the *paragraph lengths* of a parallel corpus are highly correlated, i.e. there is reason to assume that long paragraphs are translated by long paragraphs, and vice versa short ones are translated by short ones. As a result, Gale and Church (1991b) define a similarity measure δ,

$$\delta = \frac{length_{L2} - length_{L1} \cdot c}{\sqrt{length_{L1} \cdot s^2}} \tag{2.1}$$

that uses the difference between the lengths $length_{L1}$ and $length_{L2}$ of two sentences in order to determine whether they could be translations of each other. This difference is further modified using the mean c of the sentence length ratios and the variance s^2.

The probabilistic model computes the probability $P(\text{sentence link}|\delta)$ of each sentence link with the similarity measure δ using Bayes Theorem as

$$P(\text{sentence link}|\delta) = \frac{P(\delta|\text{sentence link}) \cdot P(\text{sentence link})}{P(|\delta|)} \tag{2.2}$$

where the probability $P(\delta|\text{sentence link})$ is defined as

$$P(\delta|\text{sentence link}) = 2 \cdot (1 - P(|\delta|)) \tag{2.3}$$

and where the prior probability P(|sentence link|) is estimated by counting how many 1:1, 1:0, etc. sentence links occur in the development corpus. According to Gale and Church (1991b), 1:1 links occur most often with a probability of 0.89 (see table 2.1).

The distance measure is used in combination with a Viterbi search to compute the paragraph alignment of a text and then the alignment of the sentences within the paragraph links.

For the evaluation, the development corpus UBS was aligned manually by one annotator, with the aid of two additional annotators for difficult passages. Gale and Church (1991b) report that their program correctly aligns all paragraphs of the trilingual corpus after excluding one file from

Link type	Frequency	P(sentence link)
1:1	1167	0.89
1:0 or 0:1	13	0.0099
2:1 or 1:2	117	0.089
2:2	15	0.011

Table 2.1: Probabilities for n:m links, estimated by Gale and Church (1991)

the corpus that proved very difficult to align due to its low translation quality. Furthermore, they report that their sentence aligner aligned correctly in most of the cases: deletions, insertions, and larger n:m links receive error rates up to 100%, while 1:1 links seem comparatively easy to align. Overall, the English–French texts were aligned with 94.35% accuracy (585 links of 620), i.e. the error rate is 5.65%. The results for the English–German texts were better: here, the sentence aligner correctly aligned in 97.26% of all cases, i.e. the error rate is 2.74%.

Additionally, Gale and Church (1991b) report on various experiments carried out to further tune the aligner. Most importantly among them is the discussion that prior paragraph alignment has a huge effect on the quality of the subsequent sentence alignment. They also report that the paragraph links can serve as hard delimiters to the sentence alignment in that no sentence pair can be linked across paragraph boundaries.

Despite the good evaluation results, there are three reasons for criticism. First of all, the method is based on a very small corpus, containing no more than 725 sentences per language. Furthermore, the data analysis has not been carried out on the sentences that Gale and Church (1991b) want to align, but on the much smaller set of 188 paragraphs per language. Accordingly, it is hard to predict whether the analysis results carry over to much larger corpora, for instance the EUROPARL corpus with roughly 340,000 paragraphs and more than a million sentences per language. Additionally, the question remains whether estimates on *paragraph lengths* can sensibly be used for aligning *sentences*. It would be interesting to repeat the analysis of Gale and Church (1991b) on another, possibly larger parallel corpus, and to compare the results. Furthermore, the data analysis is described *informally*, with plots showing the correlations between paragraph lengths. Information such as average paragraph lengths for both languages, standard deviation and variance is not given. Judging from the plots, most paragraphs seem to consist of roughly 500 characters on average. Depending on the standard deviation, the similar length heuristic can cease to be meaningful: if the standard deviation is small, i.e. if most of the paragraphs are of roughly equal lengths anyway, then the number of suitably sized target language paragraphs that have to be aligned with a source language sentence is too high for the method to reliably distinguish between good link pairs and bad ones. Third, the evaluation of the length-based approach is worth discussing: it is carried out on the development corpus, and although Gale and Church (1991b) give a thorough discussion of their evaluation results, they do not give precision and recall values. Instead, they give error rates calculated for the different link types, and leave it to the reader to derive the error rate per language pair.

Brown et al. (1991) follow a strategy very similar to the one suggested by Gale and Church (1991b). However, there are differences: first, Brown et al. (1991) define sentence length as the number of tokens in a sentence. Second, they pre-align their corpus using anchor points.

Brown et al. (1991) begin by defining so-called *anchor points*, i.e. elements in a corpus that can be aligned very reliably and that can simultaneously serve to segment a parallel corpus into smaller sections that are translations of each other. In the case of the *Canadian Hansards*, the corpus that Brown et al. (1991) use, things like comments describing the speaker of a passage, exclamations from the auditorium and time stamps serve as anchor points. Moreover, Brown et al.

(1991) distinguish between *major* and *minor* anchor points, the latter being speaker comments. Prior to the sentence alignment, the anchor points in the corpus are aligned in a two pass process: first, major anchor points are aligned using minimum edit distance, then they are accepted or rejected given the distribution of minor anchor points between two aligned anchor points. After the anchor point alignment has been completed, the sentences and paragraphs within the sections are aligned based on their respective lengths.

When doing an analysis on the Canadian Hansards, the authors observe a correlation between the lengths of the English and French sentences, as well as a predominance of 1:1 translations, i.e. translations where one sentence in the source language corresponds to exactly one sentence in the target language. Accordingly, the authors define a Hidden Markov Model and train it "on a large sample of text" (Brown et al. 1991, p. 175) using the EM-algorithm. Link types are not modelled using estimates as done in the approach described previously. Instead, the sentence lengths themselves are used as cues to determine during the Markov computations if sentences have to be merged or deletions or insertions have to be assumed.

For their evaluation, Brown et al. (1991) aligned the Canadian Hansards and sample a total of 1000 links for manual inspection. Within these 1000 links, only 6 errors are found, i.e. the error rate of the strategy is 0.6%. These evaluation results are very good. However, the evaluation was not carried out on unseen data, i.e. the evaluation and development corpus are identical.

Another problem is the vague description of the data analysis. Similarly to the description given by Gale and Church (1991b), it is informal: the authors do state the average lengths of English and French sentences, respectively. However, they do not give information about the standard deviation or variance, hence it is difficult to judge whether sentence length based on the number of tokens is a good cue for the alignment task. However, the modelling of insertions, deletions, and other types of links is more straightforward than the approach taken by Gale and Church (1991b). Brown et al. (1991) artificially create link types such as 2:2 links and calculate their cumulative sentence lengths: If two sentences a and b of the source language are merged, e.g. the length of this unit is given by the sum of the lengths of the two sentences a and b. This unit length can be directly compared to the length of a second unit in the target language, whether the target language is a sentence, two merged sentences, or of any other length and type. Accordingly, it is not necessary to estimate the probabilities of the various link types from a development corpus. As a result, this strategy might carry over well to corpora of other genres and sizes with possibly other distributions of link types.

2.1.2 Methods using Anchor Points

Another major approach to sentence alignment involves the detection and alignment of so-called *anchor points*: tokens occurring in the corpus that can be aligned very reliably. Brown et al. (1991) use them to pre-segment a parallel corpus before applying their length-based algorithm. The general strategy is simple: pre-defined anchor points are searched for, aligned, and subsequently used to direct the alignment of further segments of the parallel corpus.

In principle, virtually anything can serve as an anchor point, from text markup (such as chapter or speaker tags) to word pairs from a bilingual dictionary to cognates. Cognates are word pairs that show obvious phonetic or orthographic similarities and which are hence taken to be close in meaning as e.g. the word pair

(1) error ↔ erreur

which shows striking orthographic similarities and is a correct English–French translation pair. However, cognates may also be false friends, as an often cited word pair shows:

(2) library ↔ librarie
(bibliothèque ↔ book shop)

though orthographically highly similar, library and librarie are not translationally equivalent. Rather, as indicated by the glosses, the French word for "library" is bibliothèque, while French "librarie" means "book shop" in English.

Methods using a Bilingual Lexicon

The general idea behind using bilingual lexica to compute sentence alignment is based on the

> observation that a pair of sentences containing an aligned pair of words must themselves be aligned.
>
> (Kay and Röscheisen 1993, p. 122)

Hence, it is possible to derive sentence alignment information from a partially word aligned corpus. The word alignment information itself can be taken from an additional, external knowledge base (cf. Tschorn 2004), or it can be induced from the corpus itself (cf. Kay and Röscheisen 1993, Fung and Church 1994, Fung and McKeown 1994).

Kay and Röscheisen (1993) suggest an algorithm that induces a lexicon from a parallel corpus, and use this lexicon information for computing a sentence alignment. In their algorithm, a first, rough sentence linking is initially performed in order to compute the co-occurrences of the words within the aligned segments. Words with similar frequencies and occurrence vectors are subsequently linked. As a similarity measure, the Dice-coefficient is used:

$$\mathrm{sim}(\mathrm{word}_{L1}, \mathrm{word}_{L2}) = \frac{2 \cdot c}{\mathrm{freq}(\mathrm{word}_{L1}) + \mathrm{freq}(\mathrm{word}_{L2})} \tag{2.4}$$

where c is the number of times the two words word_{L1} and word_{L2} co-occur, and $\mathrm{freq}(\mathrm{word}_{L1})$ and $\mathrm{freq}(\mathrm{word}_{L2})$ are the frequencies of the two words in the corpus. Word pairs that achieve a similarity value below 0.5 are discarded. The remaining word pairs are used to link the sentences of the corpus they occur in. The whole alignment process is iterated such that in each iteration, only the best word pairs are used to link sentences, and thus only the best sentence links are kept after each iteration has been completed. The best word pairs are defined as those with the highest similarity values, and word pairs with higher frequencies are preferred. The value of the sentence links depends on the number of word pairs that they contain. The more and better word pairs are contained within a sentence pair candidate, the more likely it is that the sentence link is correct.

Further requirements are that crossing links, i.e. cases where the order of sentences within a parallel corpus is reversed, are avoided, and that only those sentences are linked for which no alignment information from previous iterations is available.

The algorithm has been developed and evaluated using two articles from the journal *Scientific American* and their German translations in the journal *Spektrum der Wissenschaft*. These two articles consist of only 469 and 462 sentences, respectively. The first of these articles, containing 214 English and 162 German sentences has been used for developing the alignment strategy, while the latter serves as evaluation corpus. On this article, the sentence alignment strategy achieved a *correctness* of 99.7% after the fourth pass, correctness indicating the percentage of links correctly computed by the algorithm.

No information is given about whether the bilingual lexicon that has been induced during the sentence alignment computations is discarded at the end of the process, or not. Furthermore, no information is given on how good the induced lexicon is. Only some examples show that the algorithm is able to determine correct translation pairs, but also that it may generate word pairs that are somehow related in meaning, though not identical. One example for this relatedness is the word pair

(3) primary ↔ sekundären
(primary ↔ secondary)

where the English *primary* is wrongly linked to German *sekundären* (secondary), which does have a different meaning. However, their meanings can be taken as related.

The algorithm of Haruno and Yamazaki (1996) is a refinement of the strategy described above in that it produces a rough sentence alignment on Japanese and English texts, using Mutual Information as well as t-score to induce a bilingual statistical lexicon[1]. This lexicon information, along with information extracted from online dictionaries, is used to determine anchor points: a sentence pair is accepted as aligned if it contains word pairs taken from the dictionary or word pairs that are highly similar according to mutual information and t-score, provided i) they occur in isolation, i.e. the word pair does not occur in neighbouring sentences, and ii) the resulting sentence link does not introduce crossing links into the alignment information. During the alignment process, the statistically computed word pairs are used in the order of their reliability, i.e. word pairs receiving high Mutual Information and t-score values are preferred over word pairs with lower scores. A threshold is also used to discard word pairs that are likely to be wrong.

The algorithm is evaluated on a small, manually aligned Japanese–English corpus containing articles from various genres. Overall, the corpus consists of only 421 Japanese and 407 English sentences. The algorithm is tested with different parameter settings. First, it is tested as described above. Second, Haruno and Yamazaki (1996) also test the algorithm when used with i) information from existing dictionaries, and ii) induced and pre-existing dictionary information, combined. The best results by far are achieved for the algorithm if both pre-existing and induced dictionary information is used, with a precision value of approximately 94.94% and a recall value of 95.08%[2].

A different approach to inducing a bilingual lexicon and using it for the sentence alignment task is suggested by Chen (1993). The author defines a statistical model that computes the probability of a sentence link by using both sentence length and word translation probabilities. The basic translation model computes the probability of a specific sentence link, P(sentence link)

$$P(\text{sentence link}) = \sum_{B} \frac{P(\text{l}_B)}{\text{N}_{l_B} \cdot \text{length}_{L1}! \cdot \text{length}_{L2}!} \prod_{i=1}^{l_B} P(\text{word link}_i) \tag{2.5}$$

using a set of word links B that are consistent with the sentence link, the lengths of the two respective L1 and L2 sentences, $length_{L1}$ and $length_{L2}$, and a normalization factor N_{l_B} ($P(l_B)$ is the probability distribution of the word links in B).

This basic model is further modified to account for the various link types. For word links, only 1:0, 0:1 and 1:1 links are assumed, while sentence links may be 0:1, 0:1, 1:1, 2:1 and 1:2. Furthermore, the word frequencies that serve as estimates to word probabilities are modified such that cognate word links automatically receive higher probabilities than non-cognate word links.

[1] While all words occurring in the English part of the corpus are used to induce the lexicon, only nouns, verbs, and adjectives of the Japanese translation are used for the lexicon induction. This modification is due to the vast difference between the two languages.

[2] The authors give separate precision and recall values for each of the four articles in their gold standard. The precision and recall values I am giving here are the average of these values.

However, Chen (1993) very narrowly defines cognates as only those words that are identically spelled. The translation model is trained using the EM-algorithm, on a very small training corpus containing not more than 100 sentence links.

For his evaluation, Chen (1993) uses the algorithm to align a French–English corpus consisting of 6,000,000 sentences. Most of the text is taken from the Canadian Hansards, a third of the sentences stemming from proceedings of the European Economic Community. He also analyses roughly 500 links where the alignment information differs from the one suggested by the Brown et al. (1991) algorithm. This difference analysis is used to estimate the error rate of the approach suggested by Chen (1993). Given that the error rate of the Brown et al. (1991) algorithm is 0.6%, the author estimates that his algorithm achieves comparable results with an error rate of 0.4%. Moreover, Chen (1993) states that his algorithm even aligned the "hard" data set that had been discarded during the evaluation by Brown et al. (1991), a success that makes his approach better than the one suggested by Brown et al. (1991).

However, while the estimated error rate sounds impressive, it is hard to follow Chen (1993)'s reasoning: the exact number of analyzed links remains unknown, as well as information on the relative alignment difficulty in those places – do all examined links occur at places where computing the correct alignment is difficult or easy? How many times was the difference observed by Chen (1993) due to an error of his aligner, and how often was the error found in the other program's output? Finally, if parts of the data were excluded, i.e. not aligned by the Brown et al. (1991) algorithm, how can this non-existing alignment information be compared to the alignments produced by the Chen (1993) aligner?

A different way to use dictionary information in order to align sentences has been presented by Tschorn (2004): he uses an existing bilingual dictionary in order to align English–German parallel texts, and augments his algorithm with several lexicon-based alignment strategies. Furthermore, Tschorn (2004) explicitly uses linguistic knowledge, relying on tagged and lemmatized corpora, and his algorithm is the only one I am aware of that does not restrict link types, i.e. apart from n:m links where n,m ≤ 2, larger link types like 7:1 may be used. Finally, the algorithm is restricted to aligning German and English texts. However, the author notes that his algorithm may be ported to other language pairs, if POS-taggers and lemmatizers are available for these languages.

In detail, the algorithm by Tschorn (2004) is a cascade of six different alignment strategies, where each strategy is only responsible for aligning those words that have not been aligned by previous strategies. All word alignment results are used to compute which sentences of the source language should be aligned to which sentences of the target language: The similarity between two sentences of source and target language

$$\text{sim}(\text{sent}_{L1}, \text{sent}_{L2}) = \frac{\text{\#matched words}_{L1} + \text{\# matched words}_{L2}}{\text{words}_{L1} + \text{words}_{L2}} \tag{2.6}$$

corresponds simply to the number of aligned words divided by the number of words contained in the two sentences $sent_{L1}$ and $sent_{L2}$. Stop words, i.e. highly frequent words, are ignored, and the final computation of the sentence alignment path is done using an A* search.

First, a dictionary lookup in a pre-existing bilingual dictionary, containing roughly 120,000 entries, is performed. Words not included in the dictionary, and hence left unaligned, are subsequently processed by a morphological strategy that decomposes German compounds and tries to translate their component words. These component translations are used to compute English candidate translations for the complete German compound. If such a candidate translation is found in the text, it is aligned to the German compound.

Next, a finite-state transducer converts numbers to words, translating e.g. "2" into German "zwei" (two). Afterwards, the converted numbers are aligned using dictionary-lookup. This procedure links numbers written either as words or not, as in the example (42 ↔ forty-two).

Then, two cognate-based approaches are used to align words with high orthographic similarity. The first employs a simple 3-gram strategy, matching words depending on the number of shared trigrams. The second uses approximate string matching, computing the longest common substring between two candidate cognates. Word pairs are taken to be translationally equivalent if they achieve a similarity above the empirically determined threshold 0.78. Cognates with a very high similarity (>0.99) are added to the bilingual dictionary.

Another, very interesting dictionary-based strategy extends the bilingual dictionary by detecting synonym translations. This strategy is based on the idea that two source language words are synonyms if they can be translated by the same target-language word.

Two further strategies have been implemented by Tschorn (2004), but as they did not improve the performance of the sentence aligner, they have been discarded. The first of these aims at detecting word class transformations, to e.g. link deverbal nouns like German *Lösung* (solution)[3] to their English equivalent expressions (in this case, the English verb *solve*). Unfortunately, while the strategy does not decrease alignment quality, it does not increase it either, and hence is not used by the algorithm by default.

The second strategy that is not used by default is the only strategy of the algorithm that is does not use dictionary information, namely the well-known length-based approach by Gale and Church (1991b). As Tschorn (2004) reports, this strategy decreased the accuracy of the algorithm and hence sentence alignment is computed without it.

On an evaluation on the development corpus, the sentence alignment algorithm achieves an accuracy of 97.23%, punishing partial matches between the automatic and the gold alignment severely. With respect to the word alignment, Tschorn (2004) reports that 65% of the tokens were left unaligned because they are stop words. The remaining non-stop words are primarily aligned based on the bilingual dictionary. Only 7% of the non-stop words are aligned based on the automatically generated synonym lexicon, and further 14% are aligned because of morphological decomposition or because the words are cognates.

Another sentence aligner that uses pre-existing dictionary information (Ma 2006) also attributes weights to word pairs, assuming that rare word pairs are more reliable than frequent ones. Furthermore, it does not exclusively handle deletions and insertions via the probabilities of these link types. Rather, it proposes sentence links only if lexical cues are present, and treats sentences as deleted/inserted in all other cases. It has been developed for the alignment of noisy English–Chinese texts, and also allows link types not covered by the standard length- or anchor-point-based approaches (these are the link types 1:3, 3:1, 1:4 and 4:1).

As a similarity measure, Ma (2006) adapts *term frequency* (tf) and *inverse document frequency* (idf), well-known from Information Retrieval, to the alignment task. Analogous to tf, Ma (2006) defines *segment-wide term frequency* (*stf*)

$$stf = \text{freq(word) in sentence} \tag{2.7}$$

as the frequency of a term within a sentence, here called a segment. Instead of idf, Ma (2006) uses *inverse document-wide term frequency* (*idtf*),

$$idtf = \frac{\text{\#all terms in document}}{\text{\#occurrences of specific term}} \tag{2.8}$$

where the term *document* is used to refer to a specific sentence rather than a complete text. *Idtf* and *stf* are then used in a similarity measure $sim(sent_{L1}, sent_{L2})$, computed between a source language and a target language sentence[4]

[3] German *Lösung* is derived from the verb *lösen*, to solve.

[4] Unfortunately, a closing parenthesis has been lost in the original formula.

Link type	Probability (Gale and Church 1991b)	Probability (Ma 2006)
1:1	0.89	0.894
1:0 or 0:1	0.0099	0.064
2:1 or 1:2	0.089	0.041
others	0.011	0.001

Table 2.2: Probabilities for n:m links

$$\text{sim}(\text{sent}_{L1}, \text{sent}_{L2}) = \sum_{i=1}^{k} \log(\text{stf}(\text{word}_{L1}, \text{word}_{L2}) \cdot \text{idtf}(\text{word}_{L1}) \cdot \alpha + \text{penalty}(\text{sent}_{L1}, \text{sent}_{L2}) \tag{2.9}$$

such that *stf* is computed for each word pair $(word_{L1}, word_{L2})$ within the two sentences, multiplied by the *idtf* for the relevant source language word $(word_{L1})$, a factor α depending on the type of linking, and finally a length penalty. α is set to 1 for 1:1 links, and to some value below 1 for all other link types. Unfortunately, the specific value for α for these other link type is not mentioned, nor how it was determined. Additionally, Ma (2006) seems to have abandoned the idea to estimate the probabilities for the different link types, as the different values of α cannot add up to 1. So the author seems to use a non-probabilistic model, after all. Finally, the length penalty, described simply as “a function of the length of the source segment and the length of the target segment” is added to the similarity value.

Ma (2006) evaluates the sentence aligner on a manually aligned English–Chinese corpus containing 3788 English and 3866 Chinese sentences[5]. However, and very interestingly, Ma (2006) reports that the frequencies of the different link types, and hence their probability estimates, differ considerably from the ones reported in Gale and Church (1991b), especially with respect to non-1:1 links (see table 2.2).

For the translation lexicon, Ma (2006) harvested and merged dictionary information from various Internet resources, resulting in a bilingual English–Chinese dictionary containing 58,000 head words. The size of this translation lexicon was varied during the evaluation: the best results were achieved when the lexicon contained at least the 4000 most frequent head words in the two languages, along with all their translations. Using this lexicon size, the aligner achieved precision and recall around 96.4%. A further small increase in precision and recall was achieved using the full set of 58,000 head words. Regarding the reliability with which the aligner suggested 1:1 etc. links, Ma (2006) reports a high reliability for 1:1 links, but also very low precision and recall values for the other link types. In the worst case, precision dropped to 35.3% for 2:2 links (with 60% recall) and recall was lowest for insertions and deletions with a value of 45.3% (the precision for this link type was 54.6%).

Lexicon Induction

Obviously, a subtask of the lexicon-based approach to sentence alignment is the computation of a bilingual lexicon: although external resources can be used, it is also possible to estimate which co-occurring words in a parallel corpus are translation pairs. These translation pairs can subsequently be aligned and used to compute sentence alignment information.

These techniques work for estimating a bilingual lexicon on a corpus that has not been sentence segmented, but they can also be adapted to parallel corpora where information on sentence

[5] Token numbers to further characterize the corpus size are omitted.

boundaries is given, or where this information is not reliable. K-vec, the algorithm presented by Fung and Church (1994), and its refinement DK-vec by Fung and McKeown (1994), both can induce a bilingual dictionary without using sentence boundaries.

In K-vec, a parallel corpus is split up into k segments that are linearly aligned. These aligned segments are then used to compute word vectors, which are subsequently compared to determine which words are translation pairs. In more detail, after each side of the parallel corpus has been split into k segments, these segments are taken to be linearly aligned, and used to estimate the joint probability P(word$_{L1}$,word$_{L2}$) of two words word$_{L1}$ and word$_{L2}$. Additionally, the frequencies of the two words are used to derive the probabilities P(word$_{L1}$) and P(word$_{L2}$) of the two words. Using Mutual Information (MI),

$$\log_2 \frac{\mathrm{P}(\mathrm{word}_{L1}, \mathrm{word}_{L2})}{\mathrm{P}(\mathrm{word}_{L1}) \cdot \mathrm{P}(\mathrm{word}_{L2})} \tag{2.10}$$

K-vec computes the similarity between the two words word$_{L1}$ and word$_{L2}$. As the authors state that MI gives unreliable results for small frequency counts, the results are further filtered: only word pairs that have significant MI values according to the t-score (p>0.95) are added to the bilingual lexicon.

Apart from an excerpt of an automatically generated lexicon, induced from the English–French Canadian Hansards, and several dotplots that visualize the alignment computed for the corpus, no evaluation information is given.

The algorithm DK-vec by Fung and McKeown (1994) is a refinement of K-vec, where the initial segmentation step into k parallel segments has been abandoned. Furthermore, MI has been replaced by so-called recency vectors that give information on the distances between the occurrences of a word.

In DK-vec, word vectors are computed based on their offsets. Each word vector is then used to compute a recency vector for the word which encodes for each two “neighbouring” occurrences of a word the distance between their offset positions. Secondly, all source language words are linked to all target language words, and their recency vectors are compared: word links are abandoned if the difference in their vector lengths is too great, and if the first occurrences of the two words are too far apart. All word pairs are aligned using dynamic time warping, and a distance function that compares the occurrences of the two words.

Fung and McKeown (1994) informally evaluate their algorithm on Chinese-English data: out of the best-scoring 42 word pairs, 32 were indeed translations of each other. However, according to the authors, the overall quality of the induced lexicon is not high, and they used their algorithm as a preprocessing and lexicon-induction step for further alignment.

Methods using Cognates

The first researchers that used cognates for computing sentence alignments were Simard et al. (1992), based on the assumption that there should be

> a significantly higher number of cognates between pairs of sentences which are mutual translations than between random pairs of sentences.
>
> (Simard et al. 1992)

Technically, the authors define cognates as either numbers, punctuation symbols, or strings that begin with an identical sequence of four letters, and their algorithm is basically the same as in the approach by Gale and Church (1991b), with the difference that the similarity measure is a function that computes the degree of *cognateness* between two sentences.

Simard et al. (1992) test their hypothesis by manually aligning a very small subset of 102 English and 94 French sentences taken from the Canadian Hansards. On this hand-aligned data set, they compute the average degree of cognateness, i.e. the degree to which translation pairs contain cognates. Additionally, they create a random alignment of the same data set, compute the average degree of cognateness between these random links, and compare the obtained value to the value computed for the hand-aligned data. According to Simard et al. (1992), the average cognateness between correct translation pairs is significantly higher than the average cognateness of the random alignment. The analysis further reveals that the number of shared cognates in the sentences pairs roughly follows a binomial distribution.

Accordingly, Simard et al. (1992) define the similarity measure

$$\mathrm{sim}(\alpha,c,n) = -\log\left[\frac{P(c|n,t)}{P(c|n)} \cdot P(\alpha)\right] \tag{2.11}$$

where α is a particular link type like 1:1, c describes the number of cognates shared in a sentence pair, and n is the average size of the sentences. For α, the authors use the link type probabilities estimated by Gale and Church (1991b).

This similarity measure is combined with the full length-based approach by Gale and Church (1991b) in a two-pass process: in the first pass, a set of best alignment paths is computed using the length-based similarity measure. In the second pass, cognateness is used to pick a unique best alignment from the set of alternatives.

Simard et al. (1992) evaluate their algorithm on two different data sets: the so-called *easy* set contains 2775 paragraph links of the English–French Canadian Hansards, corresponding to 7123 sentence links. The second, *hard* data set is much smaller with only 790 paragraph links, and 2693 sentence links. The paragraph links in the hard data set are assumed to be difficult to align as they are highly *asymmetric*, i.e. the paragraphs in a link do not contain the same number of sentences. Both data sets were manually aligned by a group of eight annotators. Unfortunately, details about whether each link was annotated once or by at least two different annotators, have been omitted.

After the annotation, the corpus was used to evaluate four different alignment strategies on the *easy* data set. In the first case, only the length-based similarity measure was used to compute the sentence alignment, and achieved an error rate of 1.8%. Then the purely cognate-based approach was evaluated, and it was found to compute more errors, with an error rate of 2.6%. The combined, two-pass strategy described above achieved better results than even the length-based approach, with an error rate of 1.6%. Finally, an alignment computed only on the basis of the linear ordering of the sentences did worse, but not as bad as expected, with an error rate of 9.6%.

Unsurprisingly, the results are worse on the *hard* data set. In an additional error analysis, Simard et al. (1992) discover that a large number of the errors is due to 'unorthodox' translation patterns, i.e. n:m links where either of the two values is larger than 2. Furthermore, wrong sentence segmentation cause alignment errors, and the algorithms also fails for some of the more ‘orthodox’ link types.

In subsequent years, the definition of *cognates* has been re-defined by several researchers. Melamed (1995) defines cognateness in terms of the *longest common subsequence ratio* (LCSR), i.e. the notion of cognateness between strings does not depend on the identity of the first four characters, but instead on the length and number of shared subsequences. The more and longer the shared subsequences of two words, the higher is the probability that those two words are cognates.

Ribeiro et al. (2000) replace the traditional heuristics on filtering noise out of the anchor-point alignment with two filters based on linear regression and confidence bands. The first filter is used to discard all those word pairs that occur too far away from the linear regression line to be reliable. The second computes the confidence band for each source language position, and discards all those anchor points that are situated outside the confidence band.

The same authors also give a definition of cognates, where cognates do not need to share continuous subsequences, but may also be “typical non-contiguous sequences of characters” (Ribeiro et al. 2001): In a preprocessing step, the authors determine the set of character n-grams that both languages of a parallel corpus share using Mutual Expectation. As all possible n-grams of the input corpus are computed before the algorithm determines which of them are likely to be cognates, it is computationally expensive.

Another cognate-inspired approach uses 4-grams of characters for aligning noisy corpora (Church 1993). However, it neither computes word nor sentence alignment. Instead, it can be used to determine which sequences of characters are probably “translations” of each other and derive sentence alignment information via this character linking.

The concept of extracting anchor points and using them for aligning a corpus is also considered in a quite different approach using *bitext maps* (Melamed 1996; Melamed 1997b; Melamed 2001; Chang and Chen 1997)[6].

In these approaches, a bilingual parallel corpus is considered a bitext that defines a rectangular bitext space. In this bitext space, an algorithm like *Smooth Injective Map Recogniser* (SIMR) detects *true points of correspondence* (TPCs) using *matching predicates*, and determines the best path from the origin of the bitext to its terminus along the TPCs (Melamed 1996; Melamed 1997b; Melamed 2001). This best path is converted into alignment information using a second algorithm, called *Geometric Sentence Alignment* (GSA).

The true points of correspondence are equivalent to anchor points, and a matching predicate is a similarity measure that relates the points of the bitext map to each other. For his English–French development texts, Melamed (1996) uses both cognateness, computed using LCSR, and a pre-existing bilingual dictionary as matching predicates.

SIMR produces a bitext map in a two-pass process. First, candidate points of correspondence are detected based on the matching predicates. Second, these candidate points are submitted to several filters. Based on the assumption that correct points of correspondence are linearly ordered, the dispersal of the correspondence points within a chain is inspected. If the points are dispersed too much, the whole chain is rejected. Also, the slope of a chain of true points of correspondence is assumed to be similar to that of the diagonal from the origin to terminus of the bitext. If the slope of a chain of correspondence points differs too widely from that of the diagonal, the chain is rejected. Additionally, true points of correspondence never overlap, or are identical, i.e. no point on the x- or y-axis of the bitext map can participate in more than one true point of correspondence. Finally, the size of a chain of correspondence points is limited: it may only contain 6 to 11 correspondence points. These thresholds and parameters are learnt on training data using simulated annealing.

[6] Melamed (2001) summarizes and collects work published previously. Hence I will often refer to the previously published articles (Melamed 1996) and (Melamed 1997b), instead of mentioning Melamed (2001).

To optimize the bitext map generation, the algorithm never uses the full bitext to search for candidate points of correspondence. Rather, it generates small, rectangular search regions within the bitext, and searches for candidate points only within these search areas. If no candidate points are found or if they cannot be chained together, the search region is enlarged and the search is iterated. On the other hand, if more than one chain of correspondence points is found, then the chain with the least dispersed points is chosen, and its alternatives are discarded. Most interestingly, the search regions are created adjacently to the chain of correspondence points that has been accepted most recently such that it is directed from the origin of the bitext to the terminus. In other words, the bitext map is scanned along the diagonal between the origin and the terminus of the bitext.

The bitext map that SIMR produces is subsequently converted into sentence alignment information using the algorithm called *Geometric Sentence Alignment* (GSA): given a bitext map and information on sentence boundaries, it computes the transitive closure over the input correspondence relations, and furthermore forces all aligned segments to be contiguous. If the resulting segments happen to contain more than one sentence on both sides, the whole linking is re-evaluated using the length-based similarity measure of Gale and Church (1991b). If this similarity measure suggests a more fine-grained linking than proposed by GSA, the GSA-proposal is discarded and replaced by the more fine-grained information. The same length-based similarity measure is used to align sentences for which bitext map information is missing.

Melamed (1996) evaluates both his algorithms on the hard and easy data sets of the English–French Canadian Hansards. When generating the bitext map, SIMR achieves a root mean squared error of 13 on the easy Hansard data, and 9.8 to 8.6 on the hard data set. Also, the sentence alignment information resulting from applying GSA to the SIMR bitext map is evaluated: in this task, the error rate is between 1.1 and 1.6 on the easy Hansards data set, and between 1.7 and 2.3 on the hard data set, depending on the availability of paragraph alignment information.

2.1.3 Hybrid Approaches

The first hybrid approach, i.e. an approach that combines purely statistical and lexicon-based cues to align sentences was suggested by Debili and Sammouda (1992). In their approach, a symmetric sentence alignment is computed using cues on sentence length, sentence position, and word pairs from an automatically induced bilingual lexicon.

However, Debili and Sammouda (1992) do not use the probabilistic model for comparing sentence lengths as suggested by Gale and Church (1991b). Rather, their similarity measure computes the difference β between the lengths of two sentences, β being a value between zero and 1. Similarly, Debili and Sammouda (1992) define a parameter α designed to capture whether two sentences appear at similar positions in the corpus, based on the assumption that the order of sentences is retained during translation. This linearity parameter exploits the *relative positions* of the two sentences, thus adjusting for insertions, deletions and other non-1:1 links. Finally, similarly to Kay and Röscheisen (1993), sentences are assumed to be translations if they contain words that are translations of each other, the bilingual word pairs being computed on the basis of their cognateness and simple co-occurrence counts.

As evaluation results, Debili and Sammouda (1992) report that their approach gives “mediocre up to excellent” results on their French–English development corpus, consisting of 339 French and 350 English sentences. This rather informal assessment is supported by an error analysis describing the percentages of correctly aligned, incorrectly aligned, and at least partially correctly aligned sentences. Unfortunately, these values are hard to translate into the evaluation metrics used in other approaches to sentence alignment. Furthermore, the numbers have been computed on the basis of a small French-English development corpus. No information is given about how the algorithm performed on the French-Arabic texts that allegedly have been used. Additionally, the full similarity measure using both parameter α and β plus the bilingual lexicon is not given.

Wu (1994) extends the length-based approach by Gale and Church (1991b) by incorporating lexical information, and reports improvements in alignment quality on the HKUST English–Chinese corpus. Using only sentence length alignment cues, only 86.6% of all sentence links are correct, while after incorporating lexical information, the accuracy increases to 92.1%.

A very interesting hybrid approach to sentence alignment was suggested by Simard and Plamondon (1998): it is a two-pass process where in the first pass, an approximate alignment is computed on the basis of *isolated cognates*, a concept that has also been used by Haruno and Yamazaki (1996). In the second pass, the approximate alignment information is used to learn a statistical translation model in order to link sentences and derive a final sentence alignment.

An *isolated cognate* is a source language token that does not show any similarity to its surrounding tokens. However, it is highly similar to a target language token that is equally isolated: the target language token also does not have anything in common with its neighbours. The neighbourhood of the cognates is given by a text window of varying size: at the beginning, it may be the whole parallel corpus but its size decreases during the alignment process.

In the first pass of the alignment strategy, a bitext map is created, and within this bitext map, isolated cognates are determined. These cognates are accepted as anchor points if they occur roughly on a diagonal between the source and end points of the bitext-map. The process is iterated: after the first anchor points have been determined, they are used to pre-segment the bitext and to determine further anchor points within the segments. The iteration stops when no new anchor points can be found.

In an intermediate step, the bitext map is translated into a rough sentence alignment. This sentence alignment is in turn used in the second phase of the process to learn a statistical translation model that computes the translation probabilities for sequences of words (i.e. sentences) rather than translation probabilities for single words[7]. The trained translation model finally refines the sentence alignment information.

Simard and Plamondon (1998) carry out an extensive evaluation on the English–French BAF corpus, also comparing their approach to various others. As evaluation metrics, they use precision and recall defined along the lines of the ARCADE project (see section 5.2.1). According to Simard and Plamondon (1998), their two-pass sentence aligner scores 98.46% precision in the best case, on the UN reports included in BAF. The worst precision value is obtained on the literature part of the corpus; on this text, only a precision of 54.49% are achieved. The recall values are considerably better: on the Canadian Hansards, the program achieves a recall of 99.06%, and even its worst recall amounts to 85.25%, on scientific articles. According to the error analysis, the approach by Simard and Plamondon (1998) encounters difficulties in dealing with deletions and insertions. Furthermore, wrong sentence segmentations cause further errors.

A very similar hybrid approach is presented by Moore (2002): firstly, an initial, rough sentence alignment is computed using the length-based approach by Brown et al. (1991). Secondly, the most reliable 1:1 links are extracted from the alignment and used to train the IBM-1 translation model (section 2.2). In the final pass of the program, the word translation model is used to refine the sentence alignment, i.e. the word translation model is used in combination with the length-based sentence alignment strategy.

[7] In fact, the authors note that the program is capable of aligning sentences as well as smaller sequences of words. Furthermore, it tends to compute links for sequences of one to three words.

For his evaluation, Moore (2002) uses the manually aligned 1:1 links contained in two English-Spanish technical manuals, and computes the error rates for several parameter settings of his hybrid aligner. With the best parameter setting, the aligner achieves an error rate of 0.006 to 0.03%. However, as these numbers are calculated only for the 1:1 links, but not for all links contained in the technical manuals, these error rates do not give insights into the overall alignment quality.

Among the hybrid approaches using both sentence characteristics and anchor points is also the alignment strategy suggested by Ceauşu et al. (2006). Their approach is to use a Support Vector Machine trained on a very small sample of gold-standard links using various length-based and anchor point-based features. The gold standards that Ceauşu et al. (2006) use are 1000 sentence links taken from the *acquis communautaire* corpus (section 1.1.2), in the language pairs English–French, English–Italian, and English–Romanian. Each sample was aligned manually, and the features for training the SVM were added. Examples of wrong sentence links were automatically created by systematically distorting the gold standard data.

As features, Ceauşu et al. (2006) use translation equivalence, estimated using the IBM-1 word alignment model (section 2.2), plus the features used by Tufiş et al. (2005), namely i) word sentence length correlation, i.e. the correlation of the lengths of two aligned sentences, counted as numbers of words, ii) character sentence length correlation, which is the same correlation as before, but sentence lengths are defined as numbers of characters instead of numbers of words, iii) word rank correlation, and finally iv) non-word sentence length correlation. The word rank correlation is a simple lexicon induction strategy based on the assumption that words with similar frequencies are translation pairs. The non-word sentence length correlation is another sentence length correlation differing from the ones described before by only counting punctuation marks and other tokens that are, strictly speaking, not words.

The alignment process itself consists of two stages: first, only the length-based features are used to compute a preliminary sentence alignment. Then, the best sentence links are used to induce a bilingual lexicon for the refinement and correction of the sentence alignment.

Finally, Ceauşu et al. (2006) evaluate their aligner on additional gold standard samples, containing roughly 1000 sentence links per language pair. According to the authors, the aligner achieves precision results between 98.99 and 99.60%, depending on the language pair, and equally high recall values between 98.96 and 99.53%.

2.1.4 Discussion of the Sentence Alignment Approaches

Most sentence alignment approaches are concerned with finding algorithms that are language-independent, i.e. not tied to a specific language pair. However, the majority of them have been applied to and evaluated on English and French texts. As alignment cues, sentence length and anchor points are used, and the anchor points are either cognates or word pairs contained in a bilingual lexicon. The lexica used to generate sentence alignment information are either induced from a corpus, or they come from independent sources like online-dictionaries. After the sentence alignment is computed, the induced lexicon information is usually discarded, along with potential word alignment information. Often, no precise evaluation is carried out to give insights into the quality of the corpus-induced lexica.

With respect to the usefulness of the alignment cues, obviously the use of dictionary-based approaches depends on whether this information is available prior to the alignment process. Several methods have been suggested to induce a lexicon automatically from a corpus for those cases in which online dictionary and the like do not exist. Another relevant issue in these approaches is the size and quality of the used bilingual dictionaries; surprisingly, only one author reports results showing which lexicon sizes lead to reliable and good alignment results: high precision and recall values are already achieved with a lexicon containing 4000 highly frequent head words (Ma 2006).

Reference	Technique	Language Pairs	Corpus	Alignment Quality
Gale and Church (1991b)	length	en–de, en–fr	UBS corpus	2-4% error rate
Brown et al. (1991)	length	en–fr	Canadian Hansards	0.9% error rate
Kay and Röscheisen (1993)	dictionary	en–de	Scientific American	0.3% error rate
Haruno and Yamazaki (1996)	dictionary	en–jap	various genres	∼95% recall / precision
Chen (1993)	dictionary	en–fr	various	0.4% estimate error rate
Tschorn (2004)	dictionary	en–de	War of the Worlds	97.23% accuracy
Ma (2006)	dictionary	en–chin	various genres	∼96% recall / precision
Simard et al. (1992)	cognates	en–fr	Canadian Hansards	1.6% error rate
Melamed (1995)	bitext map	en–fr	Canadian Hansards	1.1% error rate
Debili and Sammouda (1992)	hybrid	en–fr	various genres	"mediocre to excellent"
Simard and Plamondon (1998)	hybrid	en–fr	BAF	∼ 99% precision / recall
Ceauşu et al. (2006)	hybrid	en to fr, it, rom	Acq. Communautaire	∼99% recall / precision

Table 2.3: Short Description of Sentence Alignment Methods

Concerning cognates, most researchers note that this option may only be available for closely related language pairs such as English and French, where, due to the close relationship between the two languages, the amount of cognates will be high. However, Melamed (1997b) suggested the use of *phonetic* cognates, i.e. words that show phonetic similarities instead of orthographical ones, may be an option.

Fortunately, the dictionary-based approaches have the asset of being less sensitive to deletions and insertions in a parallel corpus. This puts them in opposition to length-based approaches that show bad performance if the parallel corpus they have to align contains many or large deletions and insertions. Apart from this disadvantage, sentence length is a popular cue and is often combined with dictionary information in order to arrive at a hybrid sentence alignment algorithm: An algorithm that is both robust with respect to deletions and insertions, and depends less on the size and quality of a bilingual lexicon than pure dictionary-based approaches.

Another interesting issue in the design of a sentence alignment program is that virtually all researchers use the "main diagonal" of a parallel text as an alignment cue, either relatively directly as a filter or similarity measure (cf. Debili and Sammouda 1992; Melamed 1995; Simard and Plamondon 1998) or indirectly, assuming a directed search path through the bitext (cf. Gale and Church 1991b; Simard et al. 1992; Melamed 1995). Sometimes, such a "dummy alignment" is even used as a baseline during evaluation. However, I do not know of any study that investigates more thoroughly to which extent the linear ordering of translations can directly be exploited for the sentence alignment task, and whether the good alignment results of "purely" length-based approaches is influenced by the directionality inherent in the approaches.

In order to measure the success of the strategies, all approaches to sentence alignment have been evaluated quantitatively, and often also with a qualitative assessment of alignment errors. As evaluation metrics, the error rate has most often been used, but some methods have also been evaluated using precision and recall. Unfortunately, no numerical evaluation results were given for one approach, and in a second case, the error rate was estimated rather than computed on the basis of a gold standard.

In any case, the evaluation results achieved so far suggest that sentence alignment is a comparatively easy task that can be performed automatically with high precision and recall where the best values of these metrics approach 100%.

Finally, the best-performing sentence alignment approaches almost invariably use some kind of linguistically-informed strategy[8], whether by using direct dictionary lookup or by using lexical cues like cognates. Moreover, these systems tend to be hybrids, i.e. to use alignment cues

[8] i.e. with the exception of the purely length-based approach by (Brown et al. 1991)

coming from a variety of sources. However, the good success of using dictionary information for the sentence alignment task constitutes a dilemma: it seems that near-perfect sentence alignment is hard to achieve without aligning words. But if sentence alignment is done prior to word alignment, then word alignment information may not be available, or the available word alignment information may be partial and hence may not suffice to guide the sentence alignment process. A first solution to this dilemma can only be to either use a dictionary (but this would make most of the subsequent word alignment task obsolete) or to first align on the word level (but this requires techniques that do not operate on sentence aligned corpora).

2.2 Word Alignment

While the strategies that have been suggested for sentence alignment use different kinds of information, namely statistics, dictionary information or quasi-linguistic notions like cognates, word alignment nearly exclusively relies on statistical translation models. In these models, a word L1 of the source language is taken to be aligned to a word L2 of the target language if the source language word can be translated by the target language word[9]. Thus, statistical translation models and alignment models are equivalent.

Most often, statistical translation models are directional, i.e. they are used to compute how a source language word is translated in a target language, but they do not give any information on how to translate a target language word into the source language. Further restrictions concern e.g. the types of links computed, i.e. if a SMT-model can produce 1:1 links only, or allows other types of n:m links.

2.2.1 The Basic Idea: A Statistical Translation Model

The first and best-known of these translation and alignment systems is a cascade of five statistical translation models by Brown et al. (1990) and Brown et al. (1993)[10], also called the IBM-models. In this approach to machine translation, any string of a source language can in principle be translated by any other string of the target language, and the task is to compute which translation pair of the set of all possible translation pairs has the highest probability.

Under the *noisy channel model*, the basic SMT model computes the probability P(L1|L2) of a source language expression L1, given a target language expression L2

$$P(L1|L2) = \frac{P(L1) \cdot P(L2|L1)}{P(L2)} \tag{2.12}$$

using language models for the languages L1 and L2, and a translation model $P(L2|L1)$. The target language translation model is usually disregarded, thus leaving

$$P(L1|L2) = P(L1) \cdot P(L2|L1) \tag{2.13}$$

as the core of the statistical translation model. This probability $P(L1|L2)$ of a source language utterance, given its translation, will be maximal for true translation pairs, and considerably lower for incorrect translation pairs. The (ngram-) language model is generally given by

$$P(w_1 w_2 ... w_n) = P(w_1) \cdot P(w_2|w_1)...P(w_n|w_1 w_2 ... w_{n-1}) \tag{2.14}$$

i.e. the probability of each token w_j depends on the preceding words $w_1 ... w_{j-1}$.

[9] The alignment path of a parallel corpus thus is a sequence of word links, each standing in a translation relation.

[10] In the following, I will only cite Brown et al. (1993) because there the authors provide the most detailed description of their models.

The translation probability $P(L2|L1)$

$$P(L2|L1) = \sum_a P(L2,a|L1) \tag{2.15}$$

of two sequences L1 and L2 can be defined as the sum of the joint probability of each L2 sequence and alignment link *a*, given the source language sequence L1.
Unfortunately, a source language word need not correspond to exactly one target language word. In order to cover this phenomenon, Brown et al. (1993) use the notion of *fertility* $P(n|w)$: it describes the number *n* of target language words that are equivalent to a source language word *w*. Secondly, Brown et al. (1993) define *distortion* $P(i|j,l)$ to model the effect of word order changes: the position *i* of a target language word thus depends on the position *j* of its source language equivalent and on the length *l* of the target language sentence.

Such an alignment model has to be trained on large amounts of text. Whereas only monolingual text is needed to train the language model, a parallel sentence-aligned corpus is needed for training the translation model, the sentence alignment information being necessary to restrict the parameter space.

The First Translation Model (IBM-model 1) In the first model, neither fertility nor distortion are used. Rather, a sentence is taken to be an unordered *bag of words*. This simplification leads to a significant reduction in the parameter space: the translation model only depends on the two expressions L1 and L2 and on their lengths *m* and *l*, the former being the length of the target language, the latter being the length of the source language expression[11].

$$P(L2|L1) = \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(L2_j|L1_{a_j}) \tag{2.16}$$

The model is trained using the EM-algorithm, just like the subsequent four models. Unlike the other models however, it has only one local maximum and thus the parameter setting obtained after EM-training does not depend on the initial parameters.

The Second Model (IBM-Model 2) The assumption that a sentence is an unordered bag of words is abandoned for model 2, thus adding distortion to the translation model. This is achieved by taking each link *a* to be dependent on *l*, too. As a result, $P(L2|L1)$ is

$$P(L2|L1) = \varepsilon \sum_{a_1=0}^{l} \dots \sum_{a_m=0}^{l} \prod_{j=1}^{m} t(L2_j|L1_{a_j}) a(a_j|j,m,l). \tag{2.17}$$

The initial parameter estimates of this model are those of the model 1.

The Models 3 to 5 (IBM-Models 3 to 5) Starting with model 3 however, the parameters need to be determined by approximate iterations of the EM-algorithm. In the third model, fertility is used as an additional parameter. Unfortunately, in this model, the probabilities used for model 3 might not add up to one, hence this model is deficient. Model 4 is lexicalized in addition to using fertility and distortion, but also deficient. Model 5 is a non-deficient version of model 4.

Brown et al. (1993) train their translation models of roughly 1.800,000 translations extracted from the Canadian Hansards. Hapax legomena, i.e. words occuring only once, are discarded from the English and French vocabularies of the corpus, in order to "eliminate some of the typographical

[11] t($L2_j|L1_{a_j}$) is the translation probability of $L2_j$ given $L1_{a_j}$, whereas ε is the probability of the length m of the L2 expression, given the L1 expression.

errors that abound in the text"(Brown et al. 1993). After twelve iterations, the perplexity of the translation model cascade has dropped to 30.65, and on average, each English word is linked to 39 different French words.

These statistical translation models were implemented during a six-week MT workshop, and they were made available to the research community as the *EGYPT MT toolkit* (Al-Onaizan et al. 1999). The toolkit has since been replaced by GIZA++ (Och 2000).

2.2.2 Competing Alignment Models

Following the seminal work by Brown et al. (1993), more SMT and alignment models have been suggested, with the aim of improving MT quality and solving remaining MT problems. However, the basis of the models, the assumption that co-occurring words are probable translations of each other, has not changed.

The HMM-model to Alignment

Vogel et al. (1996) developed a first-order HMM-model for statistical machine translation that is very similar to IBM-1 *and* IBM-2: sentences are considered to be bags of words, but words are assumed to form clusters. Thus, local word order variations are accounted for. Hence, translation probabilities are computed by a statistical model weighted by the local context of each word pair.

Vogel et al. (1996) assume that Indo-European languages show local word order similarities, i.e. although the word orders in two languages may be different, there may as well be local contexts where the word orderings are either highly similar, or where word order changes appear in a very limited context window. These word order changes, the authors argue, often occur within a context window of only three words. As an example of such a localisation effect, an English–German sentence pair

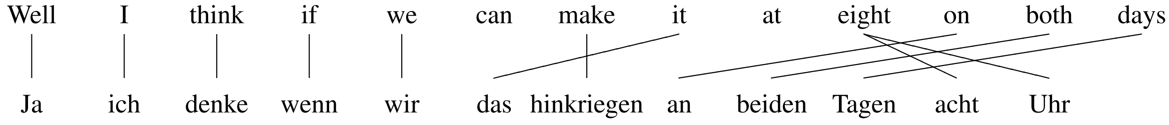


Figure 2.1: Example taken from Vogel et al. (1999), p. 838

is given where only local word order changes affect the word pairs (make ↔ hinkriegen) and (it ↔ das), but there are no longer range word order changes (situations where the differences between word positions are bigger than 3). The HMM-model is accordingly based on a statistical translation model where such local reorderings are allowed. It is defined as

$$P(L2_1^J|L1_1^I) = \sum_{a_1^J} \prod_{j=1}^{J} [p(a_j|a_{j-1},I) \cdot p(L2_j|L1_{a_j})] \tag{2.18}$$

the variables J and I describing the respective sentence lengths, and i, j describing specific word positions[12]. The probability of a specific link a_j depends, according to the model, on its leftmost neighbour a_{j-1}. Thus, the model corresponds roughly to the IBM-2 model, the only difference being that word order changes are much more restricted in the HMM-model. Just like IBM-2, however, the HMM-model has been designed to align only 1:1 links, thus the example above showing a 1:2 link (eight ↔ acht Uhr) is slightly misleading.

[12] The notation $L1_1^I$ is used to indicate that the token sequence ranges from positions 1 to I in the whole L1 expression. Similarly, $L2_1^J$ describes the token sequence in the L2 expression, ranging from position 1 through J.

The HMM-model has been tested using three different corpora, the first containing avalanche bulletins issued by the relevant Swiss authority. It contains 62,849 tokens of French and 44,805 tokens of German text. The second corpus, containing tourism-related texts, is considerably smaller with only 14-15,000 tokens per language, the languages in question being Spanish and English. Finally, Vogel et al. (1996) also use the German and English texts from the trilingual Verbmobil corpus. This corpus is the largest of the three, containing roughly 150,000 tokens per language. According to Vogel et al. (1996), the HMM-model achieves a perplexity of 20.18. However, the authors report that the HMM-model should be extended to account for larger "jumps", i.e. for word order differences that cannot be modelled locally. Vogel et al. (1996) further notice that the constraint to allow only 1:1 links results in a very poor alignment quality on the Verbmobil corpus, and that hence the HMM-model should be extended to allow further link types.

A joint-probability Model for CLIR

Within the Twenty-One project, two statistical translation models have been developed that are symmetric, i.e. they can be used to translate in any direction between two languages (Hiemstra 1998; Hiemstra et al. 1997; Hiemstra 1996). Secondly, they are not aimed at a use within Machine Translation. Rather, the statistical translation models have been designed to create bilingual dictionaries to be used for CLIR. Furthermore, one of them has specifically been designed to correctly compute 1:m and n:1 links, a feature that is necessary when aligning from languages that make considerable use of compounding (like Dutch or German) to languages that prefer multiword expressions (like English and French).

The translation models both assume that a sentence is a *bag of words*, i.e. word order is not taken into account. The first model,

$$P_{L1\to L2}(L1,L2) = P_{L2\to L1}(L1,L2) = \prod_{i=I}^{l} P(L1_i, L2_i) \tag{2.19}$$

assumes 1:1 links between the words of a parallel corpus, while the other allows for n:m links. Both translation models assume a multinomial distribution and estimate the probability of each word pair using maximum likelihood estimation. Just like the models by Brown et al. (1993), the two symmetric translation models are trained using the EM-algorithm.

On the Dutch-English development corpus consisting of roughly 5800 sentences and 150,000 words per language, the models achieve a precision of 89% to 97% and a recall of 69% to 74%. These values have been calculated on the basis of 20 manually aligned sentence pairs, seen during training, and only those links are part of the evaluation that have received a probability of 0.5 or more. Thus, only the best links are evaluated, i.e. no overall performance of the alignment models is assessed. Furthermore, the 20 evaluation sentence pairs constitute a very small gold standard. Surprisingly, the second model, allowing n:m links, performs worse than the simple 1:1 model. Linguistic preprocessing mostly improves these alignment results slightly: each word string is extended to include information like POS-tags or lemmas. However, contracting tokens to noun phrase strings decreases alignment quality.

Although the evaluation results of the alignment models are very good, there are some caveats. First of all, the test sentences were part of the training data, and moreover a test set of 20 sentence links is very small. Then, only the best of the automatically aligned links were evaluated, i.e. the precision and recall values do not reflect the overall performances of the two alignment models. Furthermore, no information is given about what types of links usually achieved probabilities above the arbitrarily chosen threshold.

The experimental results using linguistic preprocessing are very interesting, first of all because they seem to show how easily one can incorporate linguistic information into the alignment computation, and secondly, because the evaluation results suggest that no big improvements can be gained from the preprocessing. Still, the linguistic preprocessing is used in an arbitrary way – the information is tightly connected to the word forms, i.e. the co-occurrence statistics are still computed over word forms albeit those that explicitly incorporate certain kinds of linguistic knowledge. Hence, although equivalence classes of the tokens are computed, the impact of the linguistic knowledge is impeded. This effect is obvious when looking at the different vocabulary sizes of the experiments: using noun phrases increases the vocabulary sizes of the development corpus drastically. It is not surprising then that the evaluation results of this experiment are considerably lower than those achieved in the others – as indicated by the larger vocabulary size, the experiment suffers from a data sparseness problem.

Word Association Alignment

Apart from the statistical translation models, a few methods have been suggested that apply word association tests to the word alignment task. The most prominent one uses the ϕ^2 association measure, a χ^2-like statistic test, to compute which word pairs may be translationally equivalent (Gale and Church 1991a). Only word pairs are considered as good that exceed the threshold *t* that depends on the variance of ϕ^2. The authors report that their approach succeeds in linking 61% of English words to *some* French word in their sample of 800 sentences, with most of the suggested links (95% of the 61%) being correct. In other words, their approach achieves a coverage of 61%, a precision of 95% and a recall of 55%.

A rather different approach to word alignment is the *cue alignment* method presented by Tiedemann (2003): an iterative algorithm is used that detects translation pairs within a parallel corpus, determines their reliability and aligns them in the order of their reliability, i.e. translation pairs that are most likely to be correct are linked first.

Translation pairs are generated based on the notion of *alignment cue*. Such a cue is basically a confidence value of a linking between two specific expressions that indicates the reliability of a link. Alignment cues are typically weighted, i.e. they are first generated based on a specific knowledge source, and then modified depending on the general reliability of the knowledge source. The weights can, of course, be set manually, but experiments with genetic algorithms have been done in order to learn them (Tiedemann 2004). For generating alignment cues, everything from a statistical association score up to a machine-readable dictionary may be used. In fact, the cue sources used by e.g. Tiedemann (1999) are cognateness and the association measures Dice-coefficient, Mutual Information and t-score. Corpus annotation can be used to derive additional cues from previously seen links, sequences of word pairs can e.g. be used to infer POS-patterns and how to link them.

The alignment process includes a sophisticated segmentation procedure in which the corpus is tokenized such that multiword expressions are recognized as single units rather than token sequences. This way, it is possible to analyze and align them in 1:1 links to their translations in the other language. Unfortunately, only multiword units occurring more than three times are used.

In an evaluation on English-Swedish data from the PLUG corpus, the aligner achieves a precision of 74.75% if all implemented cues are used, and a corresponding recall of 63.73%. Unfortunately, the aligner is evaluated in a translation spotting task, i.e. no full word alignment information is compared to a gold standard (see section 5.2.4).

Word Alignment without Sentence Links

Another word alignment system that corresponds to IBM-2, is *word_align* (Dagan et al. 1993), an extension of the *char_align* program by Church (1993). *Word_align* takes the output of *char_align* as input and computes word link probabilities weighted by offset probabilities, i.e. by probabilities that a target language word occurs at a certain distance k from the source language word.

As neither *char_align* nor *word_align* uses the notion of sentence alignment or sentence boundaries, the words, or indeed any kind of corpus item, need not be aligned within the boundaries of aligned sentences. Rather, a context window of size 40 is used to restrict the number of possible links. Furthermore, the offset probability P(k) is used to compute the expected distance between the corpus positions of a translation pair.

Dagan et al. (1993) restrict their model to align neither high-frequency nor words occuring less than three times. As a result, the word alignment will be partial. However, the authors argue, this partial information may already be sufficient for lexicographic and term extraction purposes, which are the applications that Dagan et al. (1993) have in mind.

Word_align has been tested on a 160,000 token large sample from the Canadian Hansards. According to this evaluation, 55% of the links suggested by the aligner are correct. Faulty links were usually close to the correct solutions in terms of corpus positions. Similar evaluation results are reported for a second test corpus, containing noisy parallel data from technical manuals.

While the word alignment technique that Dagan et al. (1993) use is very interesting, their evaluation results are relatively bad: although they were at least not far off the mark, 45% of the links computed by *word_align* are completely wrong, thus using a method like *word_align* does not seem very fruitful. Further, no information is given with respect to gaps in the alignment information, and the way of sampling alignment errors is rather dubious.

Discriminative Word Alignment Approaches

Two discriminative word alignment approaches have recently been suggested: in the first, the similarity score of a word link is given by the multiplication of its feature scores (Moore 2005). Features that can be used by the algorithm may be log-likelihood scores, positional heuristics favouring word links that occur at similar corpus positions, heuristics for n:m links, and the like. In an evaluation on the Canadian Hansards, the method achieved an alignment error rate of 7.5%, corresponding to 89.8% recall and 94.7% precision.

The second discriminative approach to word alignment is very similar to the one presented by Moore (2005), but is more restricted in that it can only link words in a 1:1, 1:0 or 0:1 fashion. On the same evaluation data, this approach achieves an alignment error rate of 5.4% using a variety of features, including the probabilities of the IBM-model 4 (Taskar et al. 2005). These results, in contrast to those by Moore (2005), seem to indicate that the correct alignment of n:m links, i.e. multiword units, is vital to achieve good alignment quality.

Matrix and Vector Space Approaches

Another approach uses a matrix operation similar to singular value decomposition to align words within a parallel corpus (Goutte et al. 2004). In this framework, alignment is seen as a linking of each word of the parallel corpus to a so-called cept, with the numbers of cepts per sentence being estimated by a probabilistic model. When computing an alignment, a translation matrix is filled with values from some knowledge source, e.g. from pre-computed GIZA++ alignments. Afterwards, the dimensions of the matrix are reduced to the most probable number of cepts.

The method was tested on the trial and test data of the 2003 word alignment evaluation campaign (Mihalcea and Pedersen 2003), where it achieved an alignment error rate of 10.81%, corresponding to 86.56% precision and 34.30% recall.

A word alignment approach that does not use direct co-occurrence counts has lately been developed by Sahlgren and Karlgren (2004). In this approach, words and their contexts are represented in an n-dimensional vector space. Due to random indexing, the dimensionality of the vector space is constant and need not be reduced arbitrarily using singular value decomposition or similar methods. Comparisons of the word vectors reveal whether two words are translationally equivalent.

In their evaluation, Sahlgren and Karlgren (2004) compute bilingual dictionaries from Swedish–Spanish and German–English texts, and compare them to existing, freely available dictionaries. The authors report an overlap between their automatically created and the existing dictionaries of up to 100% for the German–English dictionary. Unfortunately, the evaluation methodology only assesses the amount of true positives within the dictionaries, and no indication is given how many errors are also included in the automatic dictionaries. As a good dictionary will contain as many correct entries as possible, and as few incorrect ones as possible, the number of erroneous dictionary entries would also allow insights into the dictionary quality.

2.2.3 Improvements of Word Alignment Approaches

Since the introduction of statistical machine translation models, several problems have been identified that hinder the models from showing good performance both in computing alignments and translation. One is the correct computation of n:m links involving more than one word in either language. Rare words also present a problem to most statistical approaches, basically because they occur so rarely that statistical models and co-occurence tests tend to yield unreliable results. The most common approach for dealing with rare words is simply to exclude them entirely.

Another problem for statistical word alignment approaches is word order differences and structural divergences between source and target languages as in these cases, translationally equivalent words do not occur in equivalent corpus positions, if at all.

There have been many attempts at improving alignment quality and at finding solutions to these and other problems. These improvements are realized by modifying the alignment process or the parallel data during a specific stage: the corpus may be ”normalized” in order to make the statistical computations more reliable or the parameters of the statistical models may be optimized. The training scheme has also been found to influence alignment quality, and finally the process-final alignment information may be filtered and manipulated in order to repair errors. In most of these cases however, linguistic information is only used during pre- or postprocessing, while the underlying assumption that alignment has to proceed based on statistical computations is rarely modified.

Preprocessing of the Corpora

Linguistic knowledge is almost exclusively used for preprocessing corpora such that the data better fits the statistical alignment models: to arrive at a better frequency distribution of word types, the corpora are lemmatized. Or, in order to overcome structural divergences, the source or target language input is transformed to more closely resemble the input in the other language.

Increasing Frequency Counts for Specific Word Pairs A very simple means to improve word alignment quality is done by feeding extra lexical knowledge into the training data before the word alignment is computed. This extra lexical knowledge might take the form of a word pair list: a bilingual lexicon is fed into the alignment process as additional training data (Dejean et al. 2003). In this approach, a corpus is aligned twice. The first alignment process aims at generating a bilingual dictionary that can be transformed into a list of word pairs. The corpus and the list of word pairs together constitute the training data for the second, final alignment process. Dejean et al. (2003) report a decrease in alignment error rate using this procedure. However, it is comparatively small, dropping from 28.3% to 28.1% in the best case. This small improvement is an artifact of the training scheme – the word pairs constitute additional statistical evidence for the co-occurrence between the words. This does not mean, however, that the additional information is sufficient to influence the alignment computation.

Vogel et al. (2000) and Och and Ney (2000a) also experiment with using dictionary information and automatically induced word classes for improving alignment quality: the bilingual dictionary information is also simply added to the training data, and trained alongside the parallel texts. Secondly, the authors induce word classes using the strategy by Och (1999). Vogel et al. (2000) and Och and Ney (2000a) report that both strategies lead to a decrease of alignment error rate, at best to 6.0 using the IBM-model 4.

Kondrak et al. (2003), on the other hand, first extract lists of cognates from the corpus and then append this list to the training data. This enhancement of the training data led to an improvement of alignment error rate from 17.6% to 15.8% on the Canadian Hansards.

Using Lemmatization A very obvious possibility to decrease the vocabularies of a parallel corpus is to lemmatize it and carry out the word alignment on the lemmas (Dejean et al. 2003; Schrader 2002). However, both Dejean et al. (2003) and Schrader (2002) report that lemmatization may not be desirable as different usage patterns, usually indicated by different word forms, may be conflated to an extent that the remaining information is useless.

Furthermore, full lemmatizing decreases alignment quality, a phenomenon observed by both Schrader (2002) and (Dejean et al. 2003). The latter also experiment with partial lemmatization, e.g. with lemmatizing only rare words up to a certain frequency. In these cases, small improvements are observable. However, no explanation for these improvements is offered, nor are the authors able to argue for or against a certain frequency threshold for partial lemmatization.

Using Parts-of-Speech information The systematic use of POS-tags is also suggested by Tufiş (2002) and Tufiş and Barbu (2002): In their iterative algorithm, a translation pair has to have the same word category membership, or at least follow regular alternations such as *a gerund is being translated as a noun*. Secondly, Tufiş (2002) and Tufiş and Barbu (2002) use log-likelihood estimates, string similarity and a distance function to determine the translation pairs. Word pairs occuring less than three times are excluded from the base algorithm. In a lexicon extraction test on the Slovene and English texts of the MULTEXT-EAST corpus, the algorithm scored 98.7% precision and 22.7% recall, recall being computed using *all* lemmas of the corpus, disregarding whether they occurred often enough to pass the frequency threshold.

Ahrenberg et al. (1998) assume that lexical expressions are translated by lexical expressions, whereas functional expressions are translated by functional expressions, i.e. they do not assume that a lexical expression may be linked to a functional one. However, they do not back up their assumption with empirical data, nor do they state how their algorithm distinguishes between lexical and functional words.

Detection of Multiword Units Depending on the languages involved, compounds may be split up into their components during the preprocessing stage. Alternatively, the preprocessing can be used to recognize and *contract* multiword expressions, usually so that a sequence of words is glued together to become a single token.

Detecting multiword sequences and glueing them together to become single tokens is a very dangerous means to improve alignment quality: it may increase the data sparseness already inherent in text corpora, i.e. a “glued” phrase may be rarer than its single-token components. Thus, computing a correct linking for a multiword sequence may still be impossible after the “gluing” operation, and in addition, the correct linking of its components may be impeded, simply because their frequencies are affected by the process.

Melamed (1997a) suggests a neat means to re-segment a corpus such that contiguous multiword sequences are recognized as single units: after the induction of a base translation model, all bigrams occurring at least four times in the corpus are contracted to become a single unit. After this resegmentation, a second translation model called trial model is trained on this modified data, and the two models are compared using a predictive value function. The best-scoring model according to this function is retained, and the algorithm is iterated on the resegmented data. Using the Canadian Hansards, Melamed (1997a) shows that this resegmentation method can improve statistical translation models. However, the success is rather small, yielding an increase in F-measure from roughly 54% to 55%.

While the method itself is a very elegant way of recognizing multiword expressions and using this information to optimize a statistical translation model, it has several drawbacks: firstly, it can recognize only those multiwords that occur frequently enough. Furthermore, the handling of non-contiguous multiwords is awkward at best.

Moore (2001) uses parse trees to detect and align contiguous and non-contiguous multiword units. In a first step, the lemmas of content words are extracted from the corpus and their association scores are computed. Then, the association scores are used to detect multiword units which need not be contiguous. The algorithm combines all those links to multiword units that are asymmetric, i.e. while a word s_i of the source language is most strongly linked to a word t_k of the target language, this target language word t_k itself is most strongly linked to a second source language word s_j. After the creation of n:m links, the association scores for multiword and single word units are recomputed, and this information is finally used to align the corpus. This method was tested on English–French technical texts and manually evaluated. According to this evaluation, the method achieved an accuracy of 55-59.5%, and a coverage of 63.2%.

As the contraction of multiwords to single tokens is dangerous, due to its tendency to increase data sparseness, the opposite strategy – to split compound expressions into their components – seems viable: the components of a compound may be linked to their corresponding expressions in the other language in 1:1 links, and so the problem of computing n:1 or m:1 links can be avoided. The problem here is how to recognize and split compound expressions reliably, especially if there are decomposition alternatives, as in the German example

(4) Staubecken → Staub|ecken (dusty corners; literally: dust corners)
or
Staubecken → Stau|becken (reservoir; literally; dam basin)

However, even wrong decompositions might help, as Köhn and Knight (2003) argue: they suggest a strategy that considers all decompositions of a compound, as long as the compound components are legitimate words of the corpus, i.e. as long as the components occur elsewhere in the corpus as single tokens. The disambiguation between different possible structures is done via a heuristic: the alternative with the highest geometric mean of its component frequencies is taken to be the correct one. Secondly, a compound is not decomposed if it occurs more often than its components. After the decomposition, the parallel data may be aligned as usual using a statistical translation model. After the alignment, the decomposed compounds can be recovered in order to arrive again at 1:m or n:1 links.

Köhn and Knight (2003) evaluate this method on a set of 1000 nominal and prepositional phrases, consisting of roughly 3500 tokens. The best results are achieved if, in addition to the parallel corpus, POS-categories are used to filter the results. In this case, the method achieves a precision of 93.1% and a recall of 90.1%, calculated on all compounds within the data set. Thus, non-compounds are excluded from the evaluation. Unfortunately, while these numbers sound impressive, few of the 3500 tokens are compounds, i.e. the evaluation sample is rather small. Finally, the authors report that there was not enough training data to correctly split and align “obscure”, i.e. rare compounds like “Passagieraufkommen” (amounts of passengers). In the second part of the evaluation, Köhn and Knight (2003) test the impact of the compound decomposition strategy on SMT quality, and report an increase in quality of roughly 10%.

Another study systematically tests which kinds of linguistic knowledge can be used for corpus preprocessing in order to improve word alignment quality, measured by the quality of an extracted bilingual dictionary (Schrader 2004). In several experiments, lemma information, POS-tags, and chunk information are used to change the input data, e.g. by replacing all word forms by their lemmas, or by deleting all words except those contained in nominal and prepositional chunks. Afterwards, the corpora are aligned using a symmetric statistical alignment model (Hiemstra 1996), and the dictionary that is computed by the aligner is inspected manually: precision is virtually always affected negatively by the preprocessing.

Syntactic Reordering Syntactic annotation can also be used to adapt the corpus data more to the requirements of the statistical models: Drábek and Yarowsky (2004) extract information on syntactical divergences between the source and target language manually from existing grammars, and use this information in order to transform the source language sentences such that they appear more similar to the target language. The alignment proceeds then on the normalized data, and the transformation of the source language are reversed after the alignment is complete. On the Romanian–English evaluation data of the 2003 word alignment evaluation task (Mihalcea and Pedersen 2003), the system achieves only very small changes in F-measure: using the smallest set of training data (1000 sentence links), e.g. the alignment achieves an F-measure of 35.6%, and after reordering, the F-measure increases slightly to 35.9%.

Changes of the Alignment Model

Melamed (2000) suggests ways to condition the statistical translation models towards specific assumption, e.g. towards assuming that a words is linked to at most one other word and that many words are not translated at all. Secondly, Melamed (2000) explains how to condition a translation model towards taking word categories into account. His evaluations are done on 250 manually aligned verses from English–French Bible texts; according to these evaluations, precision and recall values up to 40% were achieved.

The heuristic that most words have to be linked in a 1:1 fashion is frequently used (Vogel et al. 2000; Tufiş 2002; Tufiş and Barbu 2002; Chang and Chen 1997). However, it is not all too clear to which degree this heuristic holds: lexical words are mainly assumed to correspond in a 1:1 fashion, while function words are ignored. However, lexical and even functional words may occur within multiword expressions, and thus, the efficiency of this heuristic is doubtful.
Nießen and Ney (001b) e.g. change the statistical translation model slightly to

$$p_i(L2|L1) = \sum_{[t_0^i]} p(t_0^i|L1) \cdot p(L2|t_0^i, L1) \tag{2.20}$$

where t_0^i is the linguistic annotation of a word form up to the i^{th} tag. As an effect, morphological information, i.e. information on the lemma and inflectional features of a word can be taken into account when computing an alignment or training an SMT model. However, the model allows only to access the lemma of a word *plus* morphological feature, i.e. it is not possible to simply ignore a word's lemma and use any given subset of the word's morphological features. Apart from these features, no other linguistic information can be used. Furthermore, Nießen and Ney (2001) do not give information on the effect of their model on the alignment quality. They only report that the word alignment quality on their corpus, the VERBMOBIL corpus, was such that using different features did not have much effect.

In another approach, the HMM-model by Vogel et al. (1996) is modified to take part-of-speech information into account (Toutanova et al. 2002). The authors change the underlying statistical model further to allow for n:m links and to account for nonlocal word order changes. The probabilities for null-links are also re-estimated from the training corpus. With these changes, the word alignment quality can be improved up to an alignment error rate of 10.69%.

Cherry and Lin (2003) present a statistical alignment model that takes as input a sentence-aligned corpus where the source language information includes a dependency tree. Word links are generated based on the assumptions that a word can only be linked once, that word order is only changed locally, and that the word links must be cohesive: The source language dependency tree is projected onto the target language such that there are no crossing dependencies within the target language dependency tree. In other words, the dependency structure restricts the possible word links towards structure isomorphism of the source and target dependency structures. On the Canadian Hansards, this structural alignment achieves an alignment error rate of 8.7% which is 3% smaller than the competing IBM-model 4. As the strategy was first restricted to 1:1 word links, Lin and Cherry (2003) introduce a link expansion method to create contiguous n:m links. Adding this method to the alignment model yields a further small decrease in alignment error rate to 8.3%.

Another syntax-informed means to compute word alignment is by annotating both source and target language with dependency information, and stepwise transformation of the source language dependency tree to resemble both the target language surface form and its dependency tree (Gildea 2003). As source and target language need not show structural isomorphism, nodes can be copied elsewhere in the dependency tree, and thus allow for structural divergences. This alignment method achieves an alignment error rate of 36% on Korean–English test data. A further refinement is achieved by disregarding the dependency node labels during the alignment process (Ding et al. 2003). Unfortunately, this refinement is not tested on the same Korean–English evaluation data, nor using the same evaluation metric, so a direct comparison between the original and its refinement is not possible. Instead, the alignment model is tested on a 500 sentence link sample from a Chinese–English corpus, and achieves an F-measure of 56.32%.

A modification of the basic alignment method that comes very close to suggesting a completely new alignment model is alignment via syntactic propagation (Ozdowska 2005; Ozdowska 2004). Here, both sides of a parallel corpus are dependency parsed and the annotations are used to guide the word alignment process. Anchor points are determined using similarity measures like the Jaccard association score and string similarity. Taking these word links as starting points, the dependency graphs are traversed to compute additional word links: if two words $word_{L1}$ and $word_{L2}$ stand in identical dependency relations to the words of the anchor link, then $word_{L1}$ and $word_{L2}$ are aligned, as well. The procedure includes sophisticated alignment rules that allow traversing the dependency graph in either direction, i.e. from governor to dependent and vice versa, and to disambiguate between alignment alternatives if necessary. Unfortunately, the success of the strategy depends on the reliability of the dependency annotations, and whether the dependency relations are encoded in similar ways for both languages. Furthermore, the alignment model needs rules or patterns to disambiguate between alignment alternatives when traversing the dependency graphs from governors to dependent. How these disambiguation rules come about is unclear. So far, it seems that they are defined manually.

In her evaluation, Ozdowska (2005) notices that precision and recall are both lower than those achieved by a baseline statistical alignment model: the syntax-based alignment scores 83% precision and 58% recall on the evaluation data of the shared task in 2003 (Mihalcea and Pedersen 2003), whereas GIZA++ achieves a precision of 95% and a recall of 85% on the same data set. This effect may be due to several factors: there may be simply too few anchor links in order to compute a full-fledged alignment, and the coverage of the disambiguation patterns is very restricted: only isomorphisms, i.e. cases where linked words share exactly the same word category and dependency relation, are covered.

Training of the Statistical Models

The statistical models can be optimized during training by influencing the different parameter settings. Moore (2004) achieves an improvement of the IBM-model 1 by smoothing the probabilities for rare words and by giving extra weight to the probabilities for null-links. On the Canadian Hansards, the optimization of parameters done by Moore (2004) leads to an improvement of alignment error rate from 29.8% to 27.1%.

Considerable success is also achieved if the corpus is enriched with the link information itself, i.e. if the corpus is taken to be semi-aligned. As Callison-Burch et al. (2004) have found out, a pre-alignment of at least some parts of the corpus leads to a considerable decrease of the alignment error rate. On 16.000 sentence links of the German–English VERBMOBIL corpus, their best-scoring alignment model achieves an alignment error rate of 12.04% on the raw data. The use of a dictionary leads to a decrease of the alignment error rate down to 10.13%, and adding word links yields a drop of the alignment error rate down to 8.80%. Fraser and Marcu (2006) experiment along similar lines with changing their model training to accommodate pre-existing word links. The major insight of both works is that increasing the influence of pre-existing, or high-confidence word links, increases the alignment quality.

The training of the alignment models itself is significantly changed by Talbot (2004): During training, the probabilities of ”unwanted” word links are set to zero, thereby strongly biasing the alignment model. Talbot (2004) shows that this training scheme leads to an improvement, as alignment error rates drop from 22% to a value below 20% on the German–English part of EUROPARL.

Postprocessing of the Alignment

Another possibility to improve word alignment quality is to align a corpus in both translation directions using a directional alignment model and to combine the two alignments paths (Och et al. 1999; Vogel et al. 2000; Och and Ney 2004). The two alignments may be intersected: links occuring in both alignments are taken to be correct. Gaps in the intersected alignment $A_{intersected}$ are filled using a heuristic: a link occuring in only one of the two directional alignments can be added to $A_{intersected}$ if it has a vertical or horizontal neighbour that is already part of $A_{intersected}$, and if the union of the link and $A_{intersected}$ does not contain any (other) link that has both a horizontal and a vertical neighbour. As an effect, n:m links are disallowed, and the combined alignment may be incomplete. Thus the precision of the alignment is increased with respect to 1:1 links, but n:m links are either discarded or mutilated to incomplete linkings.

An alternative is to compute the union of two alignments, whether they are produced by the same aligner, using two different translation directions (Och et al. 1999; Vogel et al. 2000; Och and Ney 2004), or by using alignments generated by two different aligners (**?**)Tufis:2005). In the latter case, the combined aligner scores 70,84% precision and 76.67% recall in the shared task 2005 (Martin et al. 2005), but by refining the combination procedure, precision and F-measure increase to 87.17% and 77.8%, respectively. Recall decreases to 70.25%, and the alignment error rate of this alignment combination is 22.2%. Och et al. (1999) and Vogel et al. (2000) also point out that this method can be used to arrive at 1:m and n:1 links. However, as the alignment models are incapable of producing n:m links where both n and m are greater than one, no n:m links are generated. Still, Och et al. (1999) report an increase of precision from 83.3% and 81.8%, respectively, to 88.4% on the VERBMOBIL corpus. No recall value is given, hence it is unknown whether the quality can be increased for a substantial amount of the alignment data or not. Secondly, no information is given on the quality of the 1:m and n:1 links computed with this strategy, i.e. no information is given whether the principal goal of the strategy, to generate good 1:m and n:1 links, is achieved.

Finally, when combining two alignments, one for either translation direction, the links not in the intersection of the two alignments can be exploited to recover n:m links (Lambert and Castell 2004). If a 1:1 link of the source-to-target alignment is included within a m:1 link of the opposite direction, then this can be taken to indicate that the m:1 link is correct. Likewise, overlapping 1:m and n:1 links of the two alignments can be combined to form n:m links. Using these strategies, Lambert and Castell (2004) decrease alignment error rate on the VERBMOBIL corpus from 18.57% to 17.72%, and on the HANSARDS from 9.13% to 7.37%.

Another possibility to improve alignment quality is to constrain the alignment by partitioning the input sequences further, and by disallowing links crossing these sequence boundaries (Simard and Langlais 2003). In this approach, the sentences in a sentence link are arbitrarily partitioned into subsequences, and for each subsequence, an alignment is computed. Afterwards, the different subsequence alignments are concatenated to arrive at a global alignment path. In this way, word alignment information can be constrained locally simply because links across partitions are disallowed. In the two extreme cases, a sentence may be partitioned into as many subsequences as it contains words, or, it will not be partitioned at all. How to choose which subsequence alignments to keep for a final alignment path remains unclear. Furthermore, the strategy only works for n:1 and 1:m links, but not for full-fledged n:m links, and the partitioning concerns contiguous sequences, only. Simard and Langlais (2003) report that their strategy can be used to improve alignment quality on the English-French gold standard of the 2003 shared task (Mihalcea and Pedersen 2003), from 62.6% precision and 32.12% recall in the worst “pure” alignment up to 77.56% precision and 36.81% recall in the best case when partitioning the input sentence links and imposing what they call “compositionality constraints”. Still, the question is whether locally constraining a statistical word alignment works well for language pairs with more differences in word order than English and French. Indeed, the approach of Simard and Langlais (2003) yields considerably worse results on the Romanian-English gold standard of the same shared task.

Postprocessing of the Bilingual Dictionary In Melamed (1995), several filters are cascaded in order to remove incorrect translation pairs from an automatically induced bilingual lexicon. The lexicon, having been computed using word co-occurrence statistics, is basically a cross-product of the words found in the parallel corpus, i.e. all words have received many different translations, most of them being incorrect and improbable. This lexicon is then passed through a cascade of four filters: a POS-filter, an existing dictionary, a cognates heuristic and a word-aligned corpus.

According to the POS-filter, word category changes are forbidden, i.e. a source language word is translated by a target language word of the same category. The category comparisons are based on an automatically simplified tagset for both languages. Secondly, translation pairs are judged good if they are also contained in a dictionary, following the heuristic “don’t guess if you know”. Thirdly, word pairs are considered good based on their degree of orthographic similarity (their cognateness) exceeding an empirically set threshold. Finally, the lexicon is compared to a word aligned corpus: if a translation pair has been used to link words in an actual corpus, this is seen as evidence that the translation pair is actually good.

These four filters can be cascaded in any order, but the best results are achieved using only two filters (Melamed 1995). Additionally, the filters seem to augment each other, neither judging all good translation pairs correct, but the union of the filter results being relatively clean. The findings are based on an evaluation of an automatically induced lexicon, generated out of 100,000 sentence links taken from the Canadian Hansards, and tested against further 15,000 sentence links from the same corpus.

Still, these filters and experiment results have to be treated with caution. The language-specific tagsets have been mapped onto a more general tagset based on machine learning, and the mappings may not be linguistically well-founded. Using an existing dictionary as filter may lead to unwanted side effects. The intersection of the two dictionaries contains what has already been known, and hence cannot be used to extend a new dictionary. Furthermore, correct translation pairs may be discarded just because, by chance or because its size is too small, do not occur in the dictionary, and the same is true if an aligned corpus is used as a filter. The corpus-as-filter method may also be inefficient if the alignment quality of the corpus is low. Finally, the amounts of cognates found between two languages varies, and hence the applicability of a cognate-filter.

2.2.4 Discussion

Summed up, a variety of word alignment methods and improvements have been introduced but all of them rely heavily on co-occurrence statistics (table 2.4). Hence all models cannot reliably link rare words. Moreover, most models cannot, or can badly, align multiword sequences. Word order differences, although they can be accounted for by using syntactic information, are still a problem.

Moreover, linguistic knowledge is predominantly used only during corpus preprocessing if at all. The core of the alignment strategies still is a statistical model of word co-occurrences without more refined or linguistically motivated parameters. This may be due to the sheer number of parameters that would be added if one began to include information on POS, lemmas, etc. Including linguistic cues into the statistical models would most probably lead to a data sparseness problem.

Reference	Technique	Language Pairs	Alignment Quality
(Brown et al. 1993)	asymmetric SMT	English–French	30.65 perplexity
(Vogel et al. 1996)	HMM model	English–German	20.18 perplexity
(Hiemstra 1996)	symmetric SMT	English–Dutch	89-97& prec., 69-74& rec.
(Gale and Church 1991a)	*phi*^{2} association score	English–French	55% correctness
(Tiedemann 1999)	various statistical cues	English–Swedish	74.75% prec., 63.73% rec.
(Dagan et al. 1993)	positional	English–French	55% estimated correctness
(Moore 2005)	discriminative	English–French	94.7% prec., 89.8% rec.
(Taskar et al. 2005)	discriminative	English–French	5.4% alignment error rate
(Goutte et al. 2004)	matrix factorisation	English-French	86.56% prec., 34.30% rec
(Sahlgren and Karlgren 2004)	random indexing	Swe.–Span./Engl.–Ger.	up to 100% overlap

Table 2.4: Short Description of Word Alignment Methods

Concerning the quality of the word alignments, it must be said that the gold standard data, evaluation metrics and results vary considerably. Some baseline experiments on the Romanian–English and French–English data of the 2003 shared task, however, suggest that very simply strategies, e.g. aligning words along the diagonal, or alignment using a nearest neighbour classifier can result in F-measures of up to 46.35% (Henderson 2003). Usually, however, the relationship between precision and recall is unbalanced: a nearest neighbour classifier may be able to link words with up to 86.99% precision. However, the accompanying recall may drop as low as 10.12%.

Current statistical word alignment methods achieve up to 28.86% alignment error rate on the Romanian–English test data, and the best result for the French–English data is an alignment error rate of 5.71%. More recent results indicate that language-independent methods score worse than those using resources like parsers and lexicons when confronted with few training data (Martin et al. 2005): on the largest data set, achieving an alignment error rate as low as 9.46% is feasible, while on the smallest data, even the best-scoring system achieve only an alignment error rate of 32,12%. Moreover, recent developments in word alignment have not even achieved much improvement on known data sets: while the best-scoring system achieved an alignment error rate of 28.86% on the Romanian–English data set, two years later the alignment error rate decreases to 26.10%, which is still a relatively high value[13].

So overall, word alignment quality can still be improved, and the alignment qualities measured differ considerably. Moreover, effort should still be put into designing alignment techniques that can correctly link multiword sequences and rare words, especially. One direction for research that has already proven fruitful for sentence alignment is i) to explicitly use linguistic information and ii) to design hybrid word alignment systems. Such a hybrid might, e.g. *not* use a "one fits all" statistical model. Rather, it could combine different strategies, all defined for specific subtasks of word alignment. One such strategy might be used to exclusively detect and align multiword sequences, while another aligns lexical words, etc.

2.3 Phrase Alignment

Phrase alignment has been done using two complementary approaches. The first is to find correspondences between phrases, and subsequently align the words within these phrases. The second does the opposite: a parallel corpus is first aligned at the word level and then these word links are used to derive phrase links. In most of these approaches, the term *phrase* applies to contiguous sequences of words that appear with a certain frequency in the parallel corpus. Thus, there is a clear difference between these word sequences and linguistic phrases, i.e. well-formed syntactic constituents.

The need to define phrases as word sequences that occur at least twice in a parallel corpus is thoroughly discussed by Köhn et al. (2003): they argue that linguistic constituents do not cover recurring word sequences like "there is" and that *not* having phrase links for these constructions leads to a severe lack in coverage. In their experiments, they show that a phrase-based machine translation system achieves a better BLEU-score (Papineni et al. 2002) if the model is not restricted to syntactic phrases. However, the phrase alignment quality itself is not assessed, so it is not clear whether the low coverage of syntax-based phrase alignment is at least correlated with high precision or not.

2.3.1 Alignment of Phrases based on their Similarities

An first approach to phrase alignment is to annotate the corpus with phrase or at least chunk information, and then to link these phrases (van der Eijk 1993; Wu 1995; Wu 1997; Wu 1999; Conley 2002; Marcu and Wong 2002; de Gispert and Marino 2005).

van der Eijk (1993) chunks a Dutch–English corpus, and the chunks are linked based on their co-occurrences. This approach leads to a recall of 64% on a manually linked data set of 1100 noun phrase pairs and a precision of 68%. A main error source of the approach is that Dutch noun chunks only partially translate to English noun-noun compounds, i.e. the Dutch chunking information is insufficient for finding phrase links between translation pairs.

A second prominent approach is to use *stochastic inversion transduction grammars* (SITGs) to bilingually parse and align a parallel corpus (Wu 1995; Wu 1997; Wu 1999). In this approach, transduction grammars are extended to allow for the inversion of the right-hand-side of the grammar rules and the grammar rules are augmented with probabilities. Thus, even a very small SITG can parse a bilingual corpus, and the resulting parallel bracketing of the corpus is used to derive the phrase alignment. On a test set of 2000 sentence links, the method achieves a bracketing precision

[13] In 2003, this best-scoring alignment method was language independent, while the "winner" of 2005 was using additional linguistic resources. The best-scoring aligner of 2005 that was language-independent achieved an alignment error rate of 26.55%.

of 80.4% and a recall of 78.4%. A random sample extracted of the phrase alignment suggests an alignment precision of 81.5%. Zhang and Gildea (2005) further refine the approach by lexicalizing the SITGs. With this approach, they achieve a precision of 69% and a recall of 51% on a small Chinese–English sample of 47 sentence links.

Another phrase alignment approach called seq_align (Conley 2002) is an extension of the language-independent word_align system (Dagan et al. 1993). It aligns words and word sequences based on their statistical co-occurrence, and the size of the word sequences is not restricted. The computational load in aligning every possible word sequences of the source language with every other possible word sequence of the target language is reduced using several heuristics: those sequences occuring only once or containing only stop words are excluded from the computation. Moreover, a windowing filter is used to align only those sequences that occur at roughly the same corpus positions. Unfortunately, the approach does not yield any improvement over the performance of the original word_align algorithm: On JOC, one of the ARCADE corpora[14], the seq_align algorithm achieves an overall precision of 52%, which is slightly lower than the precision of word_align on the same corpus (55%). Recall, on the other hand, is slightly higher than the value of its competitor (53%) with 55%, and the F-measure is the same for both systems.

Marcu and Wong (2002) defined a joint-probability model for phrase-based machine translation along similar lines. This model aligns contiguous n-grams of up to six words if they appear at least five times in the corpus. Thus the model is much more restricted than the sequence alignment algorithm of Conley (2002).

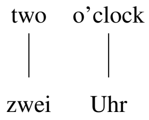
A combined word and phrase alignment approach is finally also presented by de Gispert and Marino (2005): In their approach, a very specific set of phrases is selected for alignment and the most similar phrases are linked based on the ϕ^2 association measure (Gale and Church 1991a). Subsequently, the remaining unlinked words and phrases of the corpus are aligned using two additional constraints: the linked items must occur at roughly the same corpus positions, and they must obey a *cohesion* constraint: they must not cross phrase boundaries. For phrases, this approach achieves a precision of 99.03% and a recall of 19-93% on 400 manually aligned sentence pairs in an English–Spanish corpus. The overall precision of the approach, counting both phrase and word alignments, is 96.37% with a recall of 80.75%

2.3.2 Alignment of Phrases based on Word Links

The basic idea in these approaches is to word align a corpus and then examine all link sequences with respect to the link behaviour of their words: if all source language words of the linking are only aligned to the target language words of the link sequence, and vice versa, then the link sequence is taken to constitute a bilingual phrase pair, and the source and target language word sequences are taken to be phrases (Och and Ney 2004; Och et al. 1999; Vogel et al. 2000).

Others (Och et al. 1999; Vogel et al. 2000; Och and Ney 2004) also suggest an *alignment template approach*: after computing an alignment for both translation direction and combining the two alignments, word equivalence classes are determined using bilingual word clustering (Och 1999). Then, alignment patterns are detected in the alignment path: an alignment pattern is an n:m link where the source language words are either fixed or belong to a fixed word equivalence class, the same being true for the target language words. Thus, an alignment pattern may be a linking between two time expressions, *two o'clock* ↔ *zwei Uhr* where some parts may be fixed, like *o'clock* and *Uhr*, while others have to belong to a specific equivalence class like *numbers*.

[14] See section 5.2.1



Unfortunately, no information is given how the alignment template approach influences the alignment quality (Och et al. 1999; Vogel et al. 2000; Och and Ney 2004). Instead, they use it for setting up a machine translation system. It would be interesting to see how a trained alignment template system performs when aligning new text of the same language pair. Furthermore, as the approach depends on a prior word alignment which may be erroneous, there must be a certain amount of error propagation and hence a close inspection of the data would be advantageous to better understand how best use and train the alignment template approach.

An early approach to aligning phrase structures by exploiting word alignment information is presented by Imamura (2001): Here, both texts of a Japanese–English corpus are tagged, parsed, and word links are computed. Subsequently, the word links are exploited to compute the phrase alignment: two phrases are linked if they dominate the same set of word links, and if these phrases have the same syntactic type, e.g. if they are both noun phrases. Additional heuristics are used for the disambiguation of problematic cases. In order to evaluate the phrase alignment method, the author conducts a series of experiments on 300 semi-manually annotated sentence links. In most of these experiments, the accuracy is roughly 86%.

A phrase alignment construction algorithm that uses dependency structures has been developed by Fox (2005) in order to train a syntax-based Czech–English SMT system. The training data is prepared by annotating the data with POS-tags and dependency structures in both languages. Secondly, the IBM-model 4 (Brown et al. 1993) is used to word-align the training corpus. In the third step, the word links are used to derive phrase links. The algorithm allows for direct inference, i.e. whenever the dependency structures of the two languages are sufficiently similar, then the different nodes can inherit the alignment information directly. If structural changes occur, then the phrase alignment is achieved via first erasing superfluous nodes.

However, Fox (2005) describes work in progress, and so it is not fully clear how the construction of phrase alignment information works, especially concerning structural changes: when does the algorithm allow structural changes? Moreover, as the algorithm both removes superfluous nodes from the source language structures and inserts them, under which circumstances does the algorithm remove or insert nodes? Finally, of course, it would be interesting to know which phrase alignment quality can be achieved with this procedure.

One relatively resource-poor approach is to word align a bilingual corpus, and to subsequently chunk the corpus in both languages simultaneously (Wang and Zhou 2002). Firstly, the word alignment is done using a standard off-the-shelf tool, and the bilingual chunkers are trained on existing treebanks. Then, the chunk are aligned using a so-called crossing constraint: the chunks simply inherit the linking information from the words contained in them. Wang and Zhou (2002) report that this strategy leads to 85.31% precision and 81.07% recall on 1000 sentence links of their English–Chinese corpus.

Recently, Brown et al. (2005) have suggested a phrase alignment scheme that can align n-grams of arbitrary sizes without using word links or co-occurrence statistics. In this approach, all possible word n-grams of a source language sentence are compared to those word n-grams of the corresponding target language sentence and only the best-scoring n-gram pairs are linked. The scoring function favours n-grams that have roughly the same size, occur at roughly the same corpus positions, or are neighbours to already established word links. No co-occurrence statistics are used, but a probabilistic dictionary can guide the phrase alignment. In experiments on the English–Romanian development data of the 2003 shared task, this phrase alignment strategy achieves an alignment error rate of 36.44, with a precision of roughly 64.47% and approximately 63% recall.

2.3.3 Summary

Algorithms that construct phrase alignment information via word links are always dependent on the prior word alignment quality, and it would be interesting to know the effect of bad word links on the phrase alignment quality. Secondly, the question remains why one should adopt a strategy that cumbersomely reconstructs phrase alignment via word links, instead of computing the required information directly.

On the other hand, approaches that compute phrase alignments directly, without using prior word alignment information, have to rely on the precision and robustness of the syntactic parsing, *as well* as on the similarity measure or probability model that links corresponding phrases. So it might seem that using existing parsers or chunkers for phrase annotations before computing alignments between them is much more error-prone and fragile than relying on word alignment information and deriving phrase alignment via heuristics. But as current chunkers and probabilistic parsers are developed to be robust, this error source for direct phrase alignment seems negligible.

Another option might be a hybrid approach to phrase alignment: phrases could be annotated in a parallel corpus and aligned based on structural similarities, thereby allowing to align constituents that are too rare to be linked in approaches that reconstruct phrase links via word link information. Word link information, on the other hand, could be used in a bottom-up fashion to disambiguate between competing phrase links.

2.4 Discussion

Summed up, the different alignment strategies differ considerably with respect to the achieved alignment quality and with respect to the specific techniques.

Sentence alignment information can be computed based on sentence length, anchor points like cognates, and based on lexical information. Furthermore, experiments have been carried out to induce bilingual lexica from an unaligned parallel corpus thereby allowing sentence alignment based on lexical cues irrespective of whether lexicographic resources exist for a specific language or not. Moreover, these alignment cues have been combined in order to improve sentence alignment quality even further. A minor drawback is that even those approaches that use lexical cues for the alignment task typically discard this information once the sentence alignment is completed.

The situation is different for word alignment: Statistical translation models and co-occurrence statistics are almost exclusively used, but hybrid approaches are rare. Moreover, several properties of the statistical models have been identified and addressed by many researchers that impede word alignment quality. Rare words cannot be accounted for by these models. The models are ill-suited to detect and align multiword sequences, or behave badly if the word orders of the source and target language vary considerably. Linguistic knowledge such as information on syntactic constituency serves basically as filter in order to reduce the combinatorics of the alignment procedure. Those improvements suggested in the research community tackle specific alignment problems and modify the standard word alignment procedures. However, real hybrids that use radically different alignment cues, as have been suggested for sentence alignment seem not to exist.

Despite these problems, word alignment seems to work reasonably well, with reported precision and recall values up to 95%, or alignment error rates as low as 5%. However, there is reason to suppose that the evaluation metrics and gold standards used do not give adequate insights into the alignment quality, a topic that is pursued more thoroughly in chapter 5.

Phrase alignment is done either using statistically computed word alignment information or vice versa but hybrids that use word alignment to derive or verify phrase alignments, and that can also work top down to generate or verify word links using phrase alignment information, seem not to exist.

Major characteristics of the different alignment approaches are that

- alignment is typically done only for one level, i.e. a corpus is either aligned at the sentence, or at the word, or at the phrase level. A text alignment system for all text levels, however, still has to be developed.
- whereas best sentence alignment quality is achieved using a mixture of techniques, word and phrase alignment approaches typically fall within the statistical paradigm. Hybrids that use a variety of alignment cues and techniques seem not to exist.
- alignment approaches have to be language-independent in that the algorithms are developed in order to be used for any language pair, and additionally using as few linguistic cues as possible.

However, the research results achieved so far indicate that language independence and statistics alone are not sufficient if a decent word alignment quality has to be achieved. Rather, statistical approaches should be at least linguistically-informed. Furthermore, the distinction that is drawn between sentence, phrase, and word alignment approaches seems artificial: high-quality sentence alignment requires word alignment information, and aligning a parallel corpus at the sentence level using lexical cues but discarding this lexical information in order to laboriously compute word alignments again with a second tool seems like an unnecessary repetition of alignment efforts.

If a new word alignment approach is to be developed, then it should be a hybrid, i.e. it should be able to use information from a variety of sources, possibly using a variety of techniques, or different models for different subtasks. The approach should decidedly not be restricted to word statistics. Moreover, if it is possible to compute phrase alignment information using word links, and if phrase alignment information can affect word alignment computations, then it should be possible to design an alignment approach where there is interaction between all textual levels. In this approach, sentence links could be derived from dictionary information, but the same sentence links could also be used to influence phrase alignment computations, etc.

Of course, the design of such an alignment approach must also include the design of a procedure that ensures alignment cohesion: the sentence alignment must be such that all words that occur within a specific sentence link are linked to each other, but not to words outside the sentence link, and the same degree of cohesion must hold between all textual levels.

The motivation behind the development of ATLAS is thus two-fold: on the one hand, it has to combine different, statistical and linguistic, information sources and alignment techniques in order to compute word alignment and solve problems of the standard statistical approaches. Secondly, it has to allow interactions between the different text levels, and compute a cohesive text alignment for paragraphs, sentences, words, and phrases.

Chapter 3

The Text Alignment System ATLAS

> Word alignment is a real-life problem: We are looking for links in the complex world of parallel corpora and we need good clues in order to find them.
>
> (Tiedemann 2003, p. 346)

3.1 Design Principles

The alternative text alignment system ATLAS has been designed to address problems of current statistical word alignment systems. Basically, these statistical approaches are restricted in that they cannot easily incorporate linguistic information, that they align rare words unreliably, and typically are not suited to align multiword sequences. Word order variations between the two languages in question are usually not well-modelled, too. Furthermore, standard aligners compute alignment on either a sentence or paragraph level, or at the word or phrase level, but none of them operates on all alignment levels simultaneously.

In contrast to the standard approaches to text alignment, the development of ATLAS was guided by the requirements to

- be *language-pair independent*, i.e. its applicability is not restricted to a specific language pair; instead, it may be used for any combination of languages,
- align *bilingual, parallel corpora*,
- use *linguistic corpus annotation*, i.e. the parallel corpora may be POS-tagged etc.,
- *use a variety of alignment clues and strategies*, i.e. linguistically motivated rules, statistical information or heuristics,
- be *modular* to allow for an easy integration of new languages, annotation types, and alignment strategies,
- align *hierarchically, and simultaneously* on the sentence, word, and other text levels.
- produce *high-quality alignments*, thus favouring precision over recall and speed,

These design decisions lead to a few implications: if the aligner has to be language-pair independent, it must be developed using more than one language combination. Using only one language pair would easily lead to the development of specific alignment strategies that work well for that particular language pair, but these strategies would also be difficult to re-use for other language pairings. So, the more language pairs are used, the better[1].

The range of supported languages should be restricted to those that share the same, or roughly the same, alphabet, or that belong to roughly the same language group, thereby allowing to ignore problems connected to different character sets. Furthermore, this approach allows the definition of alignment clues that will be highly relevant for a specific language group, irrespective of what works best for a specific language pair. Finally, the development of an aligner that makes use of linguistic corpus annotation cannot achieve much progress if developed for languages with scarce resources, i.e. for languages without the most basic NLP-tools like POS-taggers and stemmers.

As has been described in the introduction (section 1.1.1), it is not necessary to compute multilingual alignments at once. Rather, if multilingual alignment information is needed, multiple bilingual alignments can be computed and combined. When aligning parallel corpora, it is not necessary to distinguish between translated, non-translated, and paraphrased text which makes the alignment task easier than when using comparable corpora. So the text aligner should be enabled to align parallel texts.

Another design decision concerns the size of the development corpus: it should be large enough to allow a variety of alignment strategies, including statistics. Finally, the genre needs to be decided on: Different text domains present different difficulties to the alignment methods. technical text is usually translated so that it is still close to the original, and resembles it as much as possible in terms of choice of words. Literary translations will be more flexible, allowing for a higher degree of synonymy or a larger vocabulary. Political and legal texts, finally, have been used widely for the development of alignment systems, mainly because enough parallel data is available for this genre. Using this type of text allows to compare a system's performance to that of others. However, it makes sense not to develop an alignment system based on a single genre. Rather, corpora taken from different genres can be used to test and improve the system on different levels of difficulty.

As one design decision is to exploit linguistic corpus annotation for text alignment, it is necessary to define what types of corpus annotations are to be supported, and where the corpus annotation come from. One possibility is to include other NLP-tools like stemmers into the alignment system. However, it is beyond the scope of this thesis to develop a text alignment system along with robust, reliable NLP-components to add the required kinds of corpus annotation. Moreover, it is unnecessary since for most languages, POS-taggers and other NLP-tools are already available or have to be developed as independent tools in any case.

Depending on the types of corpus annotation, it should be possible to define various alignment strategies: an alignment strategy might use only a single type of corpus information, such as word frequency statistics, or it might use several, e.g. POS-tags and word lengths. Furthermore, the alignment clues may be reliable enough to allow for alignment *rules* defining that if a certain condition holds, then two corpus segments *are* translations of each other . Or, the strategies may be statistical or heuristic in nature, constraining that if a certain condition holds, it *is usually the case* that the two corpus segments are translations of each other. As the alignment strategies may work independently of each other, it is necessary to compare and merge their alignment information. This is done via treating all links suggested by an alignment strategy as hypotheses, which may interact with, verify, or contradict each other.

[1] However, the number of language pairs should be restricted to those that the system developer can use to manually assess the quality of the alignment output.

The three requirements

- to be able to extend ATLAS to more and other language pairs,
- to allow different kinds of corpus annotations, and
- to allow different alignment strategies

are the reasons why the system is required to be modular: Otherwise, it would simply be impossible to add more languages, different types of corpus annotations, or alignment strategies without changing the whole structure of the alignment system.

As ATLAS has been developed to simultaneously align on the sentence and word level as well as on other levels, it has to align hierarchically in that the corpus structure – a paragraph contains sentences which contain phrases which, finally, contain words – has to be taken into account. Furthermore, textual cohesion must be ensured: the process-final alignment information should not contain overlapping or contradicting links. This condition has to be met both within the same link types (if two word links overlap, then one of them is surely wrong), and across link types. If the first paragraphs of the source and target language texts are translations of each other, and if they are linked, then a link between the first word of the source language and the penultimate word of the target language should not occur. This alignment disambiguation should not just ensure cohesion, but also discard erroneous links (but retain correct ones).

With respect to the system's performance, precision should be preferred over recall and speed for a simple reason: optimization can still take place after the system has been enabled to produce high-quality alignments. Furthermore, a parallel corpus need be aligned only once. Once the corpus has been aligned properly, there is no need to align it a second time.

In the following, I first describe for which languages ATLAS has been developed (section3.2) and give an overview on the corpora used for the development (section 3.3). Then, I list the requirements and possibilities for corpus annotation (section 3.4). In the following section (section 3.5), I describe the system architecture of ATLAS in more detail, starting with the *task manager*, i.e. that part of the program that is responsible for core functionalities such as data base management and alignment disambiguation (section 3.5.1). Then, I give a rough overview how alignment modules are integrated into the system (section 3.5.2; more information on the implemented alignment strategies can be found in chapter 4). Afterwards, I explain the alignment disambiguation in more detail (section 3.5.3), and describe an example alignment process (section 3.7).

3.2 Supported Language Pairs

I have developed the aligner using the language pair German–English, extending the tests and development of alignment strategies to other language pairs, specifically German–French and German-Swedish, wherever possible. Care was taken to design the alignment strategies such that any arbitrary language pairing of the supported languages is possible. Thus ATLAS can also align the language pair English–French. Apart from these four languages, the aligner supports those additional Indo-European languages that use the Latin alphabet. In particular, additional basic support for the languages Spanish and Italian is provided.

The reasons for intensively testing the performance of ATLAS with the language pairs English–German and French–German were entirely practical in that enough parallel data is available to thoroughly test and experiment with ATLAS, and that I have enough language and linguistic skills to analyze and check the alignment results. Those experiments involving Swedish were made possible through the help of Judith Degen and Martin Volk, the former providing annotation and language skills, the latter NLP-tools for Swedish.

3.3 The Development Corpora

When choosing which corpora to use for the development of ATLAS, the main objectives were to find corpora that are

1. sufficiently large to make statistical analysis feasible,
2. available in English and German, and possibly in French and Swedish, to test how well the system performs for different language pairs, and to compare the test results,
3. from different domains or genres.

Accordingly, the development corpora include political texts as well as literary ones, the former in order to compare ATLAS to other alignment systems. The literature corpus has been included as a challenge: as literary texts usually contain more metaphors, idioms, puns, directed speech etc. than other genres, they cannot be translated easily, least be aligned. Trying to align them accordingly helps to detect and tackle alignment problems. Finally, if an alignment system is capable of aligning such difficult text, then it should not break down if confronted with the reputedly easier technical or news documents. The third development corpus contains news bulletins, again because this text type has been used in previous alignment approaches. During the development of ATLAS, these texts were found to be very cleanly translated. Thus, they serve as an ideal starting point for the development of an alignment system.

In detail, three different development corpora were used: The first containing political texts is the multilingual EUROPARL corpus that has been used in previous alignment approaches (section 3.3.1). The second is a corpus containing six literary texts in English and German (section 3.3.2). The last and smallest corpus contains news bulletins of the European Investment Bank in the three languages English, German, and French (section 3.3.3)[2] .

3.3.1 The EUROPARL Corpus

This corpus consists of verbatim protocols of the European Parliament from April 1996 until September 2003 (Koehn 2005) and is available at the OPUS open corpus website (Tiedemann and Nygaard 2004). I have used the German, English, French and Swedish monolingual corpus files along with the sentence alignment information for the language pairs German–English, German–French, and German–Swedish.

Overall, the corpus consists of roughly 490 documents and around 30 million tokens per language (The size details are given in table 3.1). One protocol file, containing roughly 100,000 tokens per language, was designated as evaluation corpus, and hence not used during the development of ATLAS. It will be described in more detail in chapter 5.

	files	tokens	types	sentences	paragraphs
English	488	28.842,367	130,935	1.064,462	340,297
French	492	33.238,913	153,728	1.089,670	346,817
German	492	27.759,028	373,994	1.123,309	345,854
Swedish	442	23.73,4858	264,404	1.057,410	315,121

Table 3.1: Europarl: English, French, German, Swedish

[2] All three corpora have been added to the corpus collection of the Institute of Cognitive Science in Osnabrück.

3.3.2 The LITERATURE Corpus

The literature corpus consists of five short stories, harvested from the German *Gutenberg project*[3], and the novel *Madame Bovary* by Gustave Flaubert. While the five short stories are only available in English and German, *Madame Bovary* is also available electronically in the original language (French) and translations into Italian, Spanish and Catalan. For the development of ATLAS, however, only the German and English translations of the novel were used. After preprocessing, the corpus contains roughly 260,000 tokens and 20,000 types per language (table 3.2).

	files	tokens	types	sentences	paragraphs
English	6	261,719	16,731	14,453	4,493
German	6	264,152	26,929	15,906	5,401

Table 3.2: Literature: English, German

3.3.3 The EUNEWS Corpus

This small, trilingual news corpus was harvested manually from the website of the European Investment Bank (Investment Bank (BEI) 2004). In sum, the corpus consists of 15 news bulletins of the bank, published between October and December 2003. All of the bulletins are available in English, German, and French. After the preprocessing, this corpus consists of 10.307 tokens of English text and their translations into French and German (table 3.3).

	files	tokens	types	sentences	paragraphs
English	15	10.306	1.856	582	263
French	15	13.471	2.062	651	276
German	15	10.925	2.141	634	267

Table 3.3: EU news: English, French, German

3.4 The Currently Supported Types of Corpus Annotation

As has been mentioned above, ATLAS has been designed to use corpus annotations as linguistic alignment clues. Furthermore, the system does not arrive at these corpus annotation due to system-internal resources. It depends rather on the existence of external NLP-tools.

ATLAS needs a parallel corpus to be at least tokenized and sentence segmented. Apart from this very basic information, the system relies on the presence of information on lemmas and word category membership. Information on word category membership is provided via POS-tags, and additionally parametrized such that language-specific POS-tags are translated into more general word category classes like *noun*, *verb*, *preposition*, etc. Furthermore, ATLAS can use information on syntactic constituency, whether this information has been generated by a parser or chunker (dependency graphs or functional annotation of constituents cannot be used so far). Finally, the corpus annotation may include morphological information.

It is also possible to add paragraph or sentence alignment information to the corpus annotation. In these cases, ATLAS will only align phrases and words, but it will not recompute the paragraph or sentence alignment. A pre-existing dictionary can also be made available to the system.

[3] http://www.gutenberg2000.de

The Annotations of the Development Corpora

Most texts in the collection had to be preprocessed: For EUROPARL, a simple format conversion from the XCES format into the internal format used by ATLAS has been done. During this conversion, meta information on languages or identities of the speakers of a parliament debate have been removed. Despite the corpus being annotated using XCES, the automatic format conversion has been impeded: not every token has been marked up properly in the original XML files (figure 3.1).

```
<CHAPTER ID="5">
  <P id="363">VOTE</P>&lt; SPEAKER ID = 74 NAME = " " &gt; " Area of freedom ,
   security  and justice " Motion for a resolution ( B5-0095 / 2000 )
 <P id="364">
```

Figure 3.1: EUROPARL: noisy format

Secondly, although the English texts of EUROPARL have been lemmatized, this lemma information is not available for each and every token of the corpus. Rather, it seems that whenever a lemma was unknown to the preprocessing tool, no information status on the lemma was given.

In contrast to EUROPARL, more preprocessing efforts have been necessary for the other two development corpora: After the format conversions from HTML or DOC to ASCII, they have been tokenized, POS-tagged and lemmatized using the IMS tree-tagger (Schmid 1994). Afterwards, the texts have been sentence-segmented using a simple perl-script. Additionally, the LITERATURE corpus and parts of EUROPARL have been chunked.

The Preprocessing Tool

The tree-tagger is publicly available for evaluation, research and teaching purposes and has been developed for tagging and lemmatizing English and German texts. Lately, it has been augmented by parameter files for POS-tagging and lemmatization of French, Spanish and Italian texts, as well[4]. It has a reported accuracy of 96.36% on the Penn Treebank, and 97.5% accuracy on a German newspaper corpus (Schmid 1994; Schmid 1995).

The tree-tagger uses the tagset *Stuttgart-Tübingen-Tagset* (STTS) for German (Thielen et al. 1999), and the *Penn Treebank* tagset for English (Santorini 1990). The tagsets for French and Italian have been developed by Achim Stein[5] and are available via the tree-tagger homepage[6]. The Spanish tagset is described on the homepage, as well (The supported tagsets are described in appendix D). The Swedish tagset SUC, finally, has been developed for the *Stockholm-Umeå-Corpus* (Ejerhed et al. 1992).

Additionally, the tree-tagger supports chunking of German, French, and English texts. As it hence provides a uniform annotation style for the three languages, it has been used to chunk the English and German LITERATURE corpus as well as the evaluation corpus. Thus, any kind of parametrization of the syntactic annotation has been unnecessary.

[4] Parameter files for other languages are also available at the tree-tagger homepage.

[5] Professor Dr. Achim Stein, Institute for Linguistics / Romance Languages, University of Stuttgart, Germany

[6] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html

3.5 The System's Architecture

Within the text alignment system, the task of hypothesis generation has been separated from that of generating a final alignment path for the whole corpus. Secondly, there is a clear division between the hypothesis generation on the one hand and core functionalities like data base management on the other (see figure 3.2).

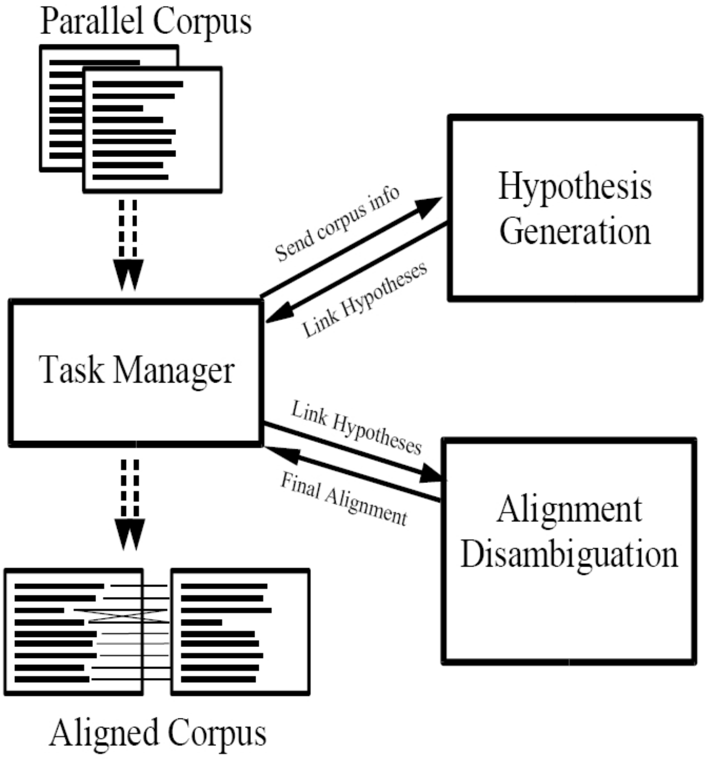


Figure 3.2: The architecture of ATLAS

These core functionalities, like the storing and updating of corpus annotations (including alignment hypotheses), are managed by the so-called task manager. This task manager also directs the alignment process itself and manages the dynamic bilingual dictionary that is created during the alignment process.

The hypothesis generation is triggered by the task manager submitting corpus segments to the different alignment modules. Each alignment module need not compute a full text alignment. Rather, it computes a set of alignment hypotheses e.g. on which words may be translation pairs.

After alignment hypotheses have been generated, a single, unambiguous text alignment of the whole corpus is computed. This step is implemented as a constrained best-first search, referred to as *alignment disambiguation*. In short, it generates a cohesive, hierarchical alignment on the basis of the generated link hypotheses, discarding hypotheses that are probably incorrect, and retaining hypotheses with a high degree of reliability.

3.5.1 The ATLAS Task Manager

The *task manager* of the ATLAS alignment system has been devised as a central platform that separates the generation of probable alignment links, here called alignment *hypotheses*, from the management of all knowledge sources like dictionaries and corpus annotation, as well as from the process-final choice of an alignment path from the first units of the parallel corpus to its last ones. The generation of alignment hypotheses itself is transferred to so-called ALIGNMENT MODULES that are implementations of different alignment strategies[7].

As there is a division between a central platform and the various alignment strategies that can be used there is no need to restrict the process to only compute sentence or word alignment. Instead, the different modules may take different textual units as input and align within them, with the task manager being responsible for initialising the correct alignment modules.

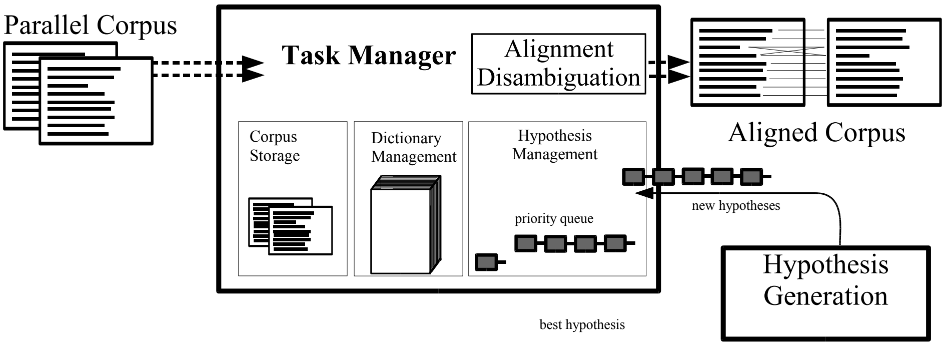


Figure 3.3: The architecture of ATLAS

The task manager thus controls and guides the text alignment process, managing

1. the reading in of all input data,
2. corpus storage and corpus access,
3. dictionary generation and management,
4. initiation of the hypothesis generation by the various alignment strategies,
5. execution of the alignment disambiguation, and finally
6. the output of the alignment results.

[7] As each alignment module implements a specific alignment strategy, the two terms are used interchangeably.

Reading of Input Data

In the first step, the unaligned, annotated corpus is read in, each source language file simultaneously with its target language translation, until all corpus data has been transferred to the system's data bases. In addition to the corpus files, dictionary information as well as pre-existing alignment information is processed.

During the reading process, the corpus information is indexed such that each index gives information on the type of the corpus information, i.e. whether it is a file, paragraph, sentence, word, noun phrase, or any other kind of phrase. Additionally, the index records information on the *parent* of the corpus element, and on its corpus position. As an example, the index of the first word of a corpus would end with W:1, thereby indicating that it is of type word (abbreviated W) and at position 1 in the linear ordering of word tokens in the corpus. Further, the full index F:1-S:1-W:1 gives information on the parent of the word w:1, which is the first sentence (S:1) contained in the first file F:1. In other words, each index is a path from the root of the text to the indexed element. Phrase types are also recorded within the index, i.e. the first noun phrase of the corpus would have an index F:1-S:1-NC:1.

Corpus Storage and Corpus Access

All corpus information is recorded in a data base where the data base entries for word tokens contain their word forms, a lexicon index that allows lookup in the bilingual, system-internal dictionary, and additional information like its lemma, and which syntactic category it belongs to. For larger units, i.e. sentences, paragraphs, or phrases, only information on their children is recorded, i.e. a data base entry for a sentence is a set of word tokens and phrases, represented by their indices.

Dictionary Management and Generation

Additionally, ATLAS stores lexicon information in a system-internal bilingual dictionary. This dictionary is dynamically populated using the word alignment hypotheses generated during the alignment process. Each entry in the lexicon data base records information on its syntactic category, its translations, and whether it is part of a multiword unit. Additionally, if a pre-existing, additional dictionary is made available to the system, its information is automatically added to the system-internal information, i.e. the two dictionaries are merged. Thus, pre-compiled dictionary information can be used during the alignment process.

The information that is recorded in the corpus and lexicon data bases is available to all alignment modules, but they do not directly update or change this information. Instead, the data bases are updated by the task manager. This encapsulation of the knowledge bases allows the use of central management routines that ensure that there are not multiple copies of the same hypothesis within the lexicon and that information from various alignment strategies is merged efficiently.

Hypothesis Management

The task manager is in full control of the alignment process, and it starts the alignment process by generating process-initial *alignment hypotheses*. These hypotheses concern, at the beginning, all file pairs of the corpus and suggest that two files A and B, containing texts in language one and two, are translations of each other, with a certain confidence. If alignment information is already available process-initially, it is transformed into process-initial hypotheses, as well.

More generally, an alignment hypothesis is about two corpus units A and B being translation pairs irrespective of whether they are files, paragraphs, sentences, or any other textual units, and the hypothesis is augmented by a confidence value and information on its type.

The confidence value indicates the reliability of the hypothesis, i.e. how correct the hypothesis probably is, with high confidence values indicating high reliability. It does *not* describe the probability of an alignment. Rather, each alignment hypothesis is created with a certain probability or similarity value, depending on the hypothesis generating module. In order to allow for a straightforward interpretation of these values, they are always restricted to a range from 0 to 1. A value close or equal to 1 will indicate that the hypothesis has a high degree of correctness, corresponding to its high probability (computed by a probabilistic alignment module) or that the corpus segments linked by the hypothesis show a high degree of similarity (as defined within a heuristic module). A value close to zero, on the other hand, will indicate the opposite.

The probability or similarity value of a hypothesis is secondly multiplied with the confidence of the parent hypothesis, based on the assumption that no child hypothesis can be more reliable than its parent. Thirdly, the reliability of the alignment module is taken into account: confidence values computed by unreliable modules are reduced, while those of highly reliable modules are increased. So far, the reliability of an alignment module is determined by the system's developer, based on intuition and test runs of the alignment modules. Currently, typical values are 1 for reliable modules (like the two cognate-based modules described in chapter 4) and 0.01 or 0.001 for unreliable ones (the reliability of the length-based sentence alignment module is for example judged to be this low). In the future, machine learning approaches will be used to determine the optimal reliability values for the alignment modules.

The type of the hypothesis indicates to which alignment module the hypothesis should be dispatched for further refinement. If a sentence hypothesis e.g. is to be processed, the task manager will hand it over to word and phrase alignment strategies and these will subsequently compute alignment hypotheses for the elements, i.e. words, contained in the sentences. The general rule is that sentence and paragraph hypotheses are dispatched to modules that will align within these sections, i.e. a paragraph hypothesis is used as starting point to align the sentences within the hypothesized paragraph link. Sentence hypotheses are used to generate phrase and word link hypotheses within the sentence pair, etc. Word link hypotheses, on the other hand, are used to populate the system-internal dictionary, and thus are cues to align phrases, sentences, and paragraphs.

The process-initial alignment hypotheses are ordered according to their confidence values and used to initiate a priority queue. The hypothesis with the highest confidence value is subsequently removed from the queue and handed over to the appropriate alignment modules by the task manager (see also figure 3.4).

The alignment results, i.e. new alignment hypotheses, are handed back to the task manager in order to update the data base. The task manager also adds new alignment hypotheses to the priority queue, after which point the alignment cycle starts over again with the task manager removing the hypothesis with the highest confidence from the queue, and transferring it to the appropriate alignment modules (the descriptions of the different modules can be found in chapter 4).

New hypotheses are submitted to the task manager, which checks whether there is another hypothesis that covers the same corpus items. If a hypothesis is completely new, it is submitted to a data base that records which and how many hypotheses have been generated. If it has been generated before by a different module, the hypothesis with the highest confidence is used, and the other is discarded. It would have been possible to compute an average of the two confidence values. However, in case the hypothesis has received a low confidence by one tool, and a high confidence by another, averaging out would decrease confidences of good hypotheses drastically. Thus, correct hypotheses may be ignored in the alignment disambiguation because of their (averaged) confidence values. In the worst case, such ignorance will lead to an overall poor performance of the aligner. Another possibility would be to compute the combined confidence values by weighting those of the contributing hypotheses.

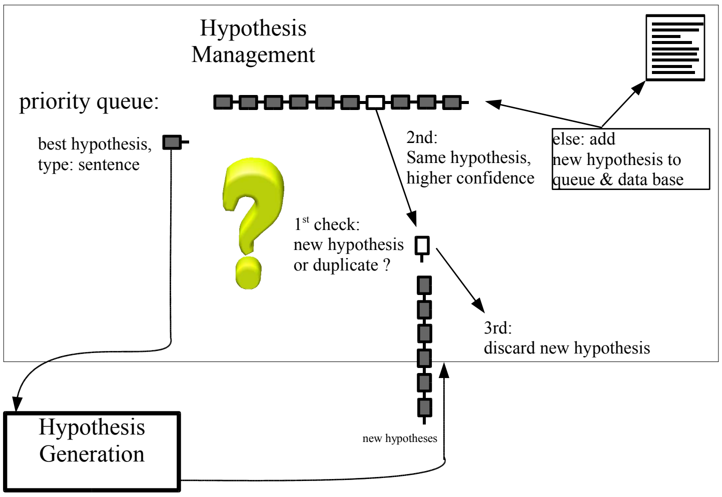

Hypothesis Management
priority queue:
best hypothesis, type: sentence
2nd: Same hypothesis, higher confidence
else: add new hypothesis to queue & data base
1st check: new hypothesis or duplicate ?
3rd: discard new hypothesis
new hypotheses
Hypothesis Generation

Figure 3.4: The architecture of ATLAS

A combined confidence value might also be computed using 80% of the confidence of a reliable hypothesis, and the remaining 20% are contributed by several low-reliability modules. However, the weights of the different modules are hard to determine manually.

Additionally, its confidence value is modified using a reliability factor that depends on which module the hypothesis has been generated with. This factor may be (close to) zero, indicating that a module is unreliable, or it may be higher[8]. Finally, the hypothesis is submitted to the priority queue.

Execution of the Alignment Disambiguation

The interaction between hypothesis generation, data base update, and removal of hypotheses off the priority queue is repeated until no new hypotheses have been generated and the priority queue has been emptied. Afterwards, the task manager starts the *alignment disambiguation* that will be explained in more detail in section 3.5.3.

The alignment disambiguation is implemented as a constrained best-first search that chooses an optimal, unambiguous set of alignment hypotheses as output. The decisions of this step are based on two types of information: firstly, hypotheses with higher confidence values are preferred, and secondly, hypotheses are chosen that do not contradict each other.

As the hypothesis generation and the alignment disambiguation are kept separate, it is possible that the process-final alignment contains gaps in those cases where the alignment disambiguation had to delete erroneous links from the data base. Hence it is vital for having a complete alignment path that the alignment modules generate many, and preferably all the correct, alignment hypotheses. As it is not possible to generate alignment hypotheses after the alignment disambiguation is complete, the alignment modules must systematically overgenerate hypotheses. As a result, the computational cost for the whole alignment process is high. How many hypotheses are generated, on the other hand, depends on the quality of the similarity measure used: if the similarity measures generate highly reliable hypotheses, i.e. hypotheses that are very likely to be correct, then it

[8] The reliability factors are currently determined by manually tuning the system, see section 4.12 for an example.

is possible to keep the number of alignment hypotheses low. Each module may also be tuned to a sensible trade-off between hypothesis overgeneration and computing only single-best hypotheses: This can be achieved by requiring each module to generate only the n best hypotheses per corpus segment, n e.g. setting to 3. This way, overgeneration is limited, and the chances that the final alignment will contain gaps is reduced.

The Output of the Alignment Results

After the alignment disambiguation, the disambiguated information is returned in the output format specified in the options file. The output will contain *all* available alignment information, i.e. alignment information available prior the alignment process as well as newly-computed alignment information are merged.

3.5.2 The Alignment Strategies

Currently, thirteen alignment modules have been implemented and tested for ATLAS. Five modules generate sentence or paragraph alignment hypotheses, five modules are used to compute how sub-sentence units should be aligned, and three further modules can be used to align at every level. During alignment, each strategy or module will receive a parent hypothesis, and will generate new alignment hypotheses based on the parent. A length-based approach to paragraph alignment will be used to align the paragraphs within a corpus. Subsequently, each paragraph hypothesis will be dispatched to a cognate-based alignment module. This, in turn, will link those cognates of the paragraph pair, i.e. it will generate word hypotheses. These word hypotheses will be added to the bilingual dictionary, and they will also be used to align sentences (figure 3.5).

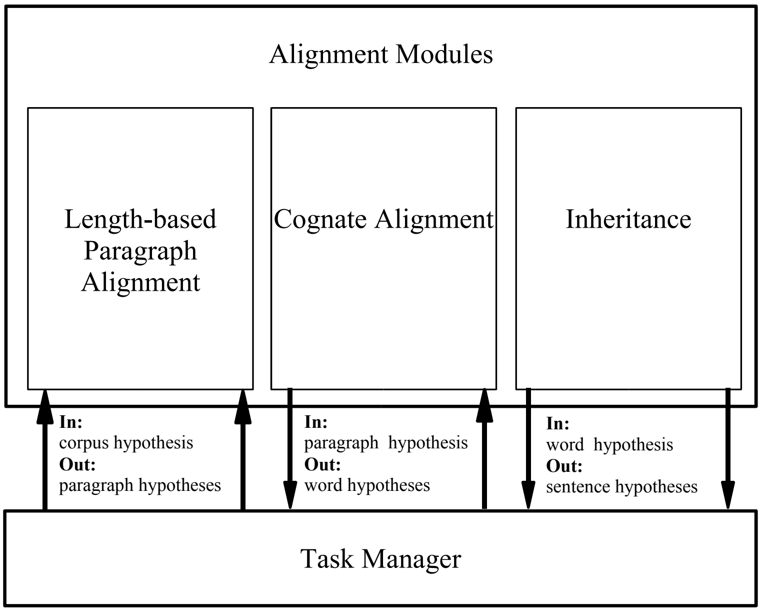


Figure 3.5: The architecture of ATLAS

With respect to paragraph and sentence alignment, some strategies have been implemented that have already been introduced in the literature, namely the length-based method of Gale and Church (1991b), two variants of the cognate-based approach developed by Simard et al. (1992), and a dictionary-based strategy for both sentence and word alignment. However, the strategies were modified to fit well into the ATLAS alignment procedure: the computation of a text alignment involves firstly a similarity comparison between corpus items to determine which of them are likely to be translation pairs, and secondly a search through the similarity matrix to derive a global, best alignment path through the bilingual corpus. The former part of each well-established strategy is implemented in the alignment modules, but not the latter as ATLAS computes the global, best alignment during the disambiguation step described in section 3.5.3.

The modifications of the well-established strategies have been made possible by ATLAS' ability to use corpus annotation: The classic cognate-based approach (Simard et al. 1992) is modified such that cognates must be members of the lexical classes nouns, proper names, adjectives, and verbs. In the dictionary-based alignment strategy, word category information is used to restrict the search space such that if a word like *fish* is marked as a verb in the corpus, then the dictionary lookup is performed only for the verb *fish* and its translation, *not* for the noun. ATLAS is able to use a variety of well-established or new clues to arrive at word and phrase hypotheses.

A module that computes word links based on statistical co-occurrence, however, has not been added to the system, as the focus of the system development was on finding additional, linguistically oriented alignment cues. A second reason for doing without a statistical alignment module is conceptual in nature. A typical scenario for which ATLAS has been developed is that sentence and word alignment are computed simultaneously, i.e. prior to word alignment, sentence alignment information is not available. In this scenario, statistical word alignment strategies that heavily rely on sentence alignment information are not applicable, hence implementing them would be futile. Secondly, although sentence alignment information is generated by ATLAS, this process is incremental, i.e. at any given point during the alignment process, the sentence alignment information is partial and hence not yet sufficient for statistically computing word link hypotheses. The only remaining possibility to allow ATLAS to link words based on statistical co-occurrence is then to use this strategy at the latest possible point during the alignment process, i.e. directly before the alignment disambiguation, when a sufficient number of sentence alignment hypotheses have been generated. Due to the systematic overgeneration of hypotheses, however, a statistical word alignment module will have to compute word hypotheses on the basis of a large number of noise and overlapping sentence hypotheses. Thus, it is not too likely that adding a statistical word alignment strategy is worthwhile[9]. All alignment strategies are described in the next chapter (chapter 4).

The sentence alignment modules have been tested using a semi-automatically constructed test set: the texts of the corpora EUNEWS and LITERATURE have been aligned semi-automatically and iteratively: first, ATLAS computes the paragraph alignment for the German and English texts. After a manually inspection and correction of the data, ATLAS re-processes the corpus to compute the sentence alignment, based on the corrected paragraph alignment. The sentence alignment has been corrected manually, too. The German EUNEWS texts have also been aligned to the French ones. Thus alignment strategies can be tested on relatively error-free data for the two language pairs German–English and German-French.

The performance of the word alignment modules has been monitored indirectly: the bilingual dictionaries that ATLAS generates based on its word alignment have been examined: errors in the bilingual dictionaries are indications of wrong word alignments and hence sufficient clues to determine the strengths or weaknesses of a specific module. As examining the bilingual dictionaries does not reveal the coverage of the system, i.e. because non-aligned words will not be listed in the dictionary, a thorough evaluation will be conducted in chapter 5.

[9] As ATLAS is modular, however, such a module can be added any time.

3.5.3 The Alignment Disambiguation

As the alignment modules may generate many alternative hypotheses for each single sentence, word, or other corpus unit, it is necessary to *disambiguate* in order to arrive at a single cohesive corpus alignment. Moreover, it is necessary to impose hierarchy or cohesion constraints on the alignment. The process-final text alignment may not contain links that contradict each other. As has already been explained, this condition has to be met both within the same link types (if two word links e.g. overlap, then one of them is surely wrong) and across link types.

The alignment disambiguation can, in principle, be done by any search algorithm, and both Viterbi search (Vogel et al. 2000) and best-first search (Tiedemann 2003) have been used for that purpose. However, the previous alignment approaches made some simplifying assumptions:

1. parallel ordering: the order of translation units is assumed to roughly correspond to that of the source language units. This is especially true for sentence alignment approaches, but even within statistical word alignment procedures, parallel ordering is often at least partially assumed to reduce the computational load of the program. Often, parallel ordering is implicitly introduced into an alignment procedure by the use of a Viterbi search computing an alignment path along a *diagonal* from the first source and target language items to the last.

2. single-level alignment: a corpus is either aligned at the sentence level or at the word level but not at both levels simultaneously.

As ATLAS computes the corpus alignment with different levels of granularity that may not all show parallel ordering, i.e. as the system does not just compute sentence or paragraph but also word and phrase alignment, a search algorithm is required that does not however implicitly favour parallel ordering of source and target language.

It is without question that parallel ordering is a good alignment cue that should be used by an alignment module to generate hypotheses for phrase and sentence links. However, if a search procedure favours parallel ordering, it will probably yield good results for sentence and paragraph alignment, but unsatisfactory ones if the language pair in question has considerable word order differences. In order to compute alignment at all levels, hence, it is required that the search procedure is not biased towards a specific alignment cue.

Simultaneously, the algorithm will have to be adapted to allow for simultaneous multiple level alignment: the final alignment should be cohesive in that e.g. word links do not cross boundaries given by sentence links.

The first requirement can be met using a best-first search as this strategy is not based on any however implicit assumption of parallel ordering. The success of a best-first search exclusively depends on whether the system assigns high confidence values to correct hypotheses, and low confidence values to incorrect ones. Thus, using a best-first search for the alignment disambiguation shifts the focus of the system development to the choice of suitable statistical or heuristic alignment models rather than to the disambiguation problem[10].

Of course, a best-first search is sensitive towards the choice of the starting hypothesis: if it is incorrect, then it is highly probable that the final alignment contains many errors. Again it is vital that only alignment hypotheses that have high probabilities of being correct are computed prior to the alignment disambiguation, and additionally that confidences of unreliable alignment modules and their hypotheses are low.

[10] In order to optimize the alignment disambiguation, it should nevertheless be possible to use a search beam or any other possibility that ensures robustness but restricts computational cost.

The second requirement is more difficult to meet. In essence, a corpus alignment that gives information on the alignment of paragraphs as well as sentences, words, and phrases must be coherent and *must not* contain contradicting information. By coherence, it is meant that if two corpus units A and B are aligned, this is the same as saying that all items $a_1..a_n$ in unit A can only be aligned to those items $b_1..b_m$ of unit B, and vice versa[11]. The failure of one or more items $a_1..a_n$ to be aligned to some item $b_1..b_m$ is acceptable only if the respective item is aligned to null. In any other case, the alignment contains a contradiction and is erroneous.

Here, any hypothesis on units A and B is coherent with its *child hypotheses* covering $a_1..a_n$ and $b_1..b_m$ if the corpus items $a_1..a_n$ are either aligned to null or to $b_1..b_m$, and vice versa. Additionally, any hypothesis on units A and B is coherent with its *parent hypothesis* if either A and B are both contained within the larger corpus units that are aligned by the parent hypothesis, or if A or B is null. The failure to comply with both conditions is taken to indicate an erroneous hypothesis.

In order to achieve full coherence in a multi-level alignment, accordingly, it is necessary to constrain the alignment disambiguation such that for each hypothesis, i.e. for each link, it is tested whether the corpus alignment would still be coherent if the link was accepted. The addition of this constraint to the best-first search proved to be quite simple: starting with the hypothesis that has the highest confidence value, each hypothesis is tested whether

1. it contains *new* information, i.e. alignment information that covers corpus items not covered by previously accepted hypotheses,

2. it is *coherent*, i.e. if its both coherent with the child and parent hypotheses listed on the agenda.

If a hypothesis complies to these conditions, i.e. if it contains new information and is coherent with the information recorded on the agenda, it is added to the agenda. In all other cases, it is discarded. After all hypotheses have been tested and either added to the agenda or discarded, the agenda contains an unambiguous, coherent corpus alignment. If the alignment is not perfect, then this error is due to one or more hypotheses with inappropriately high confidence values, and it is possible to track down the error to the responsible alignment model. At this point, the confidence value assignment of the model can be adjusted. Or, if the mistake is due to the wrong choice of alignment clue, a new alignment model may be developed that performs better.

3.6 The Output: An Annotated, Aligned Parallel Corpus

After the alignment disambiguation, all alignment information computed by ATLAS is made available in the system-internal XML format. It lists the hypotheses of the alignment agenda, and each hypothesis is given along with information on its type and its confidence value (here called *certainty*, as in the XCES format.). The system-internal corpus indices are used as pointers to the corpus items (figure 3.6).

ATLAS also generates a bilingual dictionary on the basis of the alignment information. The dictionary is likewise encoded in XML and lists the lemma, language, and category information for each headword, along with a list of translations. The translation information does not only consist of the translation's lemma, but also includes information on the translation's syntactic category and its confidence in the translation pair (figure 3.7).

[11] Here, two corpus units A and B are taken to be sets of 0 or more paragraphs, sentences, words, or phrases, hence it is irrelevant whether an actual corpus unit instance is a 1:1 or any other n:m link.

```
<corpus>
  <l1 lang="German" files="de03106.crp de03121.crp ..." dir="/corpora/euNews/de">
  <l2 lang="English" files="en03106.crp en03121.crp ..." dir="/corpora/euNews/en">
</corpus>
<alignments>
  <aligned type="paragraph" l1="f:1-p:1" l2="f:2-p:1" certainty="0.91656375" />
  <aligned type="paragraph" l1="f:1-p:2" l2="f:2-p:2" certainty="0.9220715" />
  ...
</alignments>
```

Figure 3.6: System-internal format: example alignment of the EU NEWS corpus

```
<item>
   <lemma>point of order</lemma>
   <category>multiword</category>
   <language>English</language>
   <translations>
      <translation>
         <lemma>Geschäftsordnung</lemma>
         <category>noun</category>
         <language>German</language>
         <confidence>0.78571</confidence>
      </translation>
   </translations>
</item>
```

Figure 3.7: Example of an ATLAS lexicon entry

A high confidence value indicates that a specific translation is judged to be reliable by the alignment system, and hence the lexicon can be used like a probabilistic lexicon[12]. An example lexicon file can be found in appendix B.2.

3.7 An Example Alignment Process

As the architecture of ATLAS is quite complex and as the alignment disambiguation is a new technique, I provide an example. Imagine the input to the alignment program is a small sample corpus, consisting of 12 sentences per language, organized in two paragraphs per language. The corpus is annotated with lemma information and POS-tags.

Firstly, ATLAS reads in and indexes the corpus, generating a first alignment hypothesis stating that the German text file and the English one are translations of each other. This alignment hypothesis is used as input to all paragraph modules of the program, which in turn align the paragraphs within the corpus, i.e. they generate paragraph alignment hypotheses.

The newly generated paragraph hypotheses are returned to the task manager. It changes the confidence values of the hypotheses according to the reliability of the paragraph module. Afterwards, it updates the data bases with the new alignment information and starts a new cycle: the paragraph hypotheses serve as input to the sentence alignment strategies. Again, all of them generate alignment hypotheses and return them to the task manager.

[12] Note, however, that the confidence value *is not* a translation probability, and therefore the output lexicon cannot be considered probabilistic. The confidence value indicates only the *reliability* of a translation.

Again, the ATLAS task manager changes the confidence values of each hypothesis, depending on which module generated it and how reliable the module is. Additionally, it checks for hypothesis duplicates: As there are four sentence alignment strategies, three of which also generate word links (the cognate-based and dictionary-based sentence alignment strategies), the possibility exists that a single alignment hypothesis is generated several times, based on different clues, and by different strategies. For example, one sentence hypothesis is computed based on a cognate occurrence, and it has also been aligned based on the similarity of the respective sentence length.

Confidence	Type	L1 items	L2 items	Module
0.1045815	sentence	par1-sent2	par1-sent2	length-based
0.2065875	sentence	par1-sent2	par1-sent2	cognate

Hence duplicates exist that contain exactly the same alignment information, the only difference being the confidence values and the modules that generated the duplicates. Whenever the task manager discovers such duplicates, it merges them, and retains the highest confidence value that has been assigned to the duplicates. In the example given above, the confidence value of the second (cognate-based) hypothesis would be re-used, and the other would be ignored.

Again, the data bases are updated, and as some word hypotheses have been generated, they are not just listed among the available alignment hypotheses. Instead, they are also added to the system-internal dictionary.

The modules might also generate alignment hypotheses that have, in fact, be used as inputs in previous cycles. The task manager therefore checks for each hypothesis whether it has been generated before. If yes, then the hypotheses are merged as mentioned above, but they are *not* recycled as inputs to the alignment modules.

The whole cycle of submitting hypotheses to the relevant alignment modules, receiving newly-generated alignment hypotheses, updating the data bases, and removing of duplicates is repeated until no new alignment hypotheses are generated. Then, the task manager starts the alignment disambiguation: the alignment hypothesis with the highest overall confidence value, irrespective whether it is a word, sentence, or any other type of hypothesis, is retrieved and added to the list of *accepted* hypotheses.

In the next iteration, the hypothesis with the next-highest confidence value is retrieved. As the list of accepted hypotheses is no longer empty, ATLAS test if the list of accepted hypotheses would still be *coherent* if this hypothesis would also be accepted.

After several cycles, the list of accepted hypotheses might contain the following hypotheses:

Confidence	Type	L1 items	L2 items	Module
0.78987	sentence	par:1-sent:1	par:1-sent:1	cognates
0.7656375	paragraph	par:1	par:1	length-based
0.5209925	sentence	par:2-sent:4	par:2-sent:4	cognates
0.5209924	sentence	par:2-sent:5	par:2-sent:5	dictionary
0.5209915	sentence	par:2-sent:3	par:2-sent:3	cognates
0.5209915	sentence	par:2-sent:9	par:2-sent:9	cognates
0.5209915	sentence	par:2-sent:10	par:2-sent:10	cognates

and the hypothesis with the next-highest confidence value might be

```
0.324575 sentence        par1-sent2 par2-sent3 length-based
```

If this hypothesis is now compared to the ones that have already been accepted, then it is obvious that it would cause the list to become incoherent: if the first German and English paragraphs are already linked, according to the second-best hypothesis

```
        0.7656375  paragraph par:1 par:1 length-based
```

how can a sentence of the German first paragraph be aligned to a sentence that *does not* occur in the first English paragraph, but in the second? As such an alignment is both incoherent and implausible, ATLAS discards the sentence hypothesis and continues the disambiguation with the next-best hypothesis in the data base.

This alignment disambiguation cycle continues until all alignment hypotheses have been either accepted or discarded. The resulting *final* set of hypotheses is added to the corpus annotation as alignment information, and it is used to generate a bilingual dictionary.

3.8 Computational Issues

The text alignment system ATLAS has been implemented as a script collection, all of the scripts having been written in Perl version v5.8.8 on a Debian Linux operating system. Apart from the basic Perl distribution, the additional packages MLDBM and XML::TWIG were needed for data base handling and the parsing of the XML files.

The computation of alignment information for any two parallel texts has a complexity of $O(n^2)$, with n being the number of corpus items per language. In other words, if a corpus contains n segments per language, then there are n^2 ways to arrange them in 1:1 links. If other link types have to be assumed, the number of possible links within the corpus increases by the number of different link types, i.e. for hypothesis generation, the complexity is $O(t * n^2)$ in the worst case, t being the number of different link types. Thus, a sentence aligner assuming the six link types 1:1, 0:1, 1:0, 1:2, 2:1 and 2:2 can generate 6*n^2 hypotheses. Accordingly, the search through all possibly generated links, in order to compute an optimal alignment path also has a complexity of $O(t * n^2)$[13].

Hence, computing alignment information soon becomes infeasible for larger corpora, and most alignment systems use pruning strategies for two reasons: firstly, to decrease the number of alignment hypotheses: while it is theoretically possible to align each corpus item of the source language with every other corpus item of the target language, most alignment systems will only compute those alignment hypotheses that are probable. Secondly, the number of possible alignment paths also grows with the number of corpus items in both languages. Again, restricting the number of alignment hypotheses reduces computational complexity.

ATLAS is in no way better than other aligners. Moreover, as each alignment module runs independently of all others, the theoretically possible number of alignment hypotheses increases to $m * t * n^2$, m being the number of modules used. The alignment disambiguation, however, is slightly less expensive, as the information on generated hypotheses is compressed during the process such that the number of modules used becomes irrelevant.

Fortunately, this performance only constitutes the worst case. With respect to the generation of alignment hypotheses, each module will inspect, and may generate $O(n^2)$ hypotheses, discarding unreliable ones, but as will be shown for the cognate-based modules, they need not. It is up to the system developer to implement only modules that produce good hypotheses, i.e. hypotheses that are very likely to be correct, and thus restrict the computational load to a minimum. Furthermore, as the ATLAS task manager controls the alignment process, it will remove or merge hypotheses that have been generated by different modules, but that cover nevertheless the same corpus items. Thus, while hypothesis generation is be expensive, hypothesis storage is kept to a minimum. Finally, the alignment disambiguation will have a complexity of $O(n)$ in the best case: As soon as a hypothesis

[13] This constitutes the best case. In general, a hierarchical procedure like the one used in ATLAS will have a complexity of O(2**(2*n)) as in principle, any subset of words of L_1 may be aligned to any subset of words of L_2.

is added to the agenda, all competing hypotheses covering the same corpus items are removed from the search space.

Summed up, the alignment of corpora using ATLAS is computationally expensive, especially for larger corpora like the EUROPARL corpus. However, the alignment modules can be tuned to generate as few, but good, alignment hypotheses as possible, and the alignment disambiguation is achieved at less computational load than the generation of hypotheses.

In the future, I will optimize the alignment modules. Meanwhile, ATLAS can be cascaded with other tools, but also with itself (in a first run, it will compute the sentence alignment, in the second, phrase alignment etc.). Modules with a high computational load can be switched off and, moreover, a large corpus can be aligned in a stepwise fashion: Each text pair is aligned independently of the other text pairs in the same corpus, only connected by a shared dictionary.

Chapter 4

Alignment Phenomena

> Linguistic research is a bootstrapping process in which data leads to analysis and analysis leads to more and better-interpreted data.
>
> (Covington 1996, p. 485)

In the previous chapter, the central architecture of ATLAS was described. Now, I am describing experiments with alignment clues, and how they can be used the experiment results to implement and fine tune alignment modules in ATLAS. I have tried to test all strategies on at least two language pairs, or on texts of different genres. Furthermore, I have experimented with the most common types of linguistic corpus annotation, namely POS, lemmas, and syntactic information. POS-taggers and lemmatizers e.g. are available for many languages, and probably among the first NLP-tools that are developed for a new language.

Other types of annotation, e.g. morphological information may be hard to obtain for languages with fewer resources than English or German, or the morphological information may be too erroneous for a successful use in the alignment task. Still, whenever possible, I have also experimented with those clues: although computational linguistics is concerned with implementing efficient and highly accurate NLP-tools, it is also about finding out *which linguistic information helps* to succeed in a given task, and why. This is another research question that I followed during the development of the specific alignment strategies used in ATLAS.

As I mainly focused on the development of strategies that align *within* sentences, I have not tried to come up with new or creative alignment strategies that deal with paragraphs or sentences. However, I have implemented *one* strategy for aligning paragraphs or sentences in addition to those already suggested in the literature. Finally, the similarity measures used within the modules are first, sometimes naive ways to capture specific alignment cues. They demonstrate how a specific linguistic phenomenon can be used as an alignment cue, but they still leave substantial room for improvement.

4.1 Length-based Sentence Alignment

As a first module to both sentence and paragraph alignment, I have adapted the length-based alignment procedure by Gale and Church (1991b), but my adaptation only uses the similarity measure of the original approach, and does not compute a full-fledged sentence alignment. Additionally, it is implemented such that it generates sentence *or* paragraph hypotheses, depending on which kind of parent hypothesis it is started with. Hence it is possible to cascade the module, either with itself or with other modules, such that first, paragraph hypotheses are computed, and second, these paragraph hypotheses are used for computing sentence hypotheses.

Link Type	P(hypothesis)
1:1	0.89
1:0 or 0:1	0.0099
2:1 or 1:2	0.089
2:2	0.011

Table 4.1: Probabilities for n:m links, estimated by Gale and Church (1991)

In more detail, the module takes the paragraphs (or sentences) contained in the input hypothesis and aligns them in 1:0, 0:1, 1:1, 1:2, 2:1, or 2:2 links. In a second step, the newly generated hypotheses are given a probability using the similarity measure described by Gale and Church (1991b), i.e.

$$\delta = \frac{length_{L2} - length_{L1} * c}{\sqrt{length_{L1} * s^2}} \tag{4.1}$$

where $length_{L2}$ and $length_{L1}$ are the length of the L2 and L1 sentences, respectively. The language-pair specific parameters c (expected number of L2 characters per L1 character) and the variance s^2 are set to c=1 and $s^2 = 6.8$ as reported in Gale and Church (1991b).

Then, the probability for each hypothesis, given the similarity measure, is computed by

$$P(hypothesis|\delta) = \frac{2 * (1 - P(|\delta|) * P(hypothesis)}{P(|\delta|)} \tag{4.2}$$

with $P(|\delta|)$ being the similarity value δ under a normal distribution, and P(hypothesis) is taken from table 4.1 (reprinted from Gale and Church (1991b)), depending on the *link type*[1] of the hypothesis.

In order to derive the confidence value for each hypothesis, I have experimented with multiplying its probability with the confidence value of its parent hypothesis. However, this means that most confidences of the parent, paragraph hypotheses will make the new (sentence) confidence values drop to a value close to zero, and hence the usefulness of the newly generated hypotheses will be severely impaired. Accordingly, during the test runs, the module generated new alignment hypotheses without multiplying their confidence values with the one of the parent hypotheses.

Test Runs When testing the implementation on the German and English paragraphs of EUNEWS, it achieved a precision value of 36.7% and a recall of 37.0%. The main error source was that the search algorithm does not favour linear ordering, i.e. the alignment suggested by the length-based strategy includes many permutations. This issue is later addressed by a specific alignment module (section 4.4). Furthermore, the module was often wrong with respect to the link types, i.e. the probabilities that Gale and Church (1991b) use seem not to be optimal for the EUNEWS texts[2]. A minor error cause was that the link types are restricted to n:m-links with n, m $\leq$ 2.

In a second test run, I examined how well the strategy performed if it had to compute sentence hypotheses based on the gold paragraph links. This time, the precision was slightly higher with a value of 41.4%. Recall was higher too, being 48.1%. Again, most errors were due to permutations, and wrong link types occurred.

[1] i.e. whether it is an 1:1, 1:2 etc. alignment of sentences

[2] Ma (2006) also noticed that the originally suggested link type probabilities need not be suitable for each corpus, and accordingly re-estimated them.

	(Gale and Church 1991b)		EUNEWS paragraphs		EUNEWS sentences		merged information	
Link type	Frequency	P(link)	Frequency	P(link)	Frequency	P(link)	Frequency	P(link)
1:1	1167	0.89	236	0.9291	358	0.7033	1761	0.8487
1:0 or 0:1	13	0.0099	5	0.0197	25	0.0491	43	0.0207
2:1 or 1:2	117	0.089	11	0.0433	89	0.1749	217	0.1046
2:2	15	0.011	0	0	9	0.0177	24	0.0116
others	0	0	2	0.0079	28	0.0550	30	0.0145

Table 4.2: Probabilities for n:m links

Accordingly, I re-estimated the link type probabilities for 1:1, 0:1 and 1:0, 1:2 and 2:1 links, and for all other occuring link types. The link type probabilities for sentence links differ quite substantially from the ones given in Gale and Church (1991b), and hence, the bad performance of the module can be explained both by the choice of alignment search and the fact that the link type probabilities were not suitable for this particular corpus. In order to improve the alignment results, without overfitting the module to the development corpus, I merged the link type estimates for EUNEWS and the (Gale and Church 1991b) data and used the resulting probabilities (table 4.2). The re-estimated probabilities do not differ much from the original ones. However, the data on the EUNEWS suggests that link type probabilities vary considerably between corpora, and hence should be re-estimated for each corpus.

With these only slightly modified probabilities, the module achieves a precision of 35.7% and a recall of 36.2% on the German–English EUNEWS paragraph, i.e. no improvement was achieved. As the main error source, permutations due to the non-linearity of the alignment disambiguation, was not changed, nothing much could be expected.

With respect to sentence alignment quality, precision and recall decreased likewise, to 40.0% and 46.5%. This is due to the probabilities for the different sentence link types within EUNEWS being considerably different from the ones estimated on all available data.

When testing the alignment strategy on the German and French texts, the results were equally low for the paragraph alignment : it achieved a precision of 35.7% and a recall of 36.3%. In the sentence alignment task, the results were the best of all achieved with this module: precision was 44.4%, and recall was slightly higher with 49.6% (table 4.3).

language pair	precision (%)	recall (%)
German–English paragraphs	35.7	36.2
German–French paragraphs	35.7	36.3
German–English sentences	40.0	46.5
German–French sentences	44.4	49.6

Table 4.3: Test run results for length-based paragraph and sentence alignment

The difference between the German–English and German-French data sets reveals that the correlation between sentence length is much better for the language pair German–French than for German and English. The alignment performance can accordingly be enhanced by taking this correlation difference into account: the language-specific factors of the length-based approach, c and s^2 can be re-estimated. Alternatively, the *reliability* of the module can be made sensitive to the language pair in question. In this case, the length-based approach is assumed to be more reliable when aligning French and German than when it aligns English and German texts. Accordingly, the reliability of the module should be higher for the first language pair, and smaller for the second[3].

[3] As has been explained before, the reliability factors are determined manually during the tuning of the system, and

The results clearly show another characteristic of the length-based approach: its success is *not* due to the similarity measure itself. Rather, the viterbi search used in the original length-based approach introduces a strong bias towards *linear ordering*. This bias in the alignment strategy has been noticed before, especially in the context of deletions and insertions causing distortions in the alignment information: The similarity measure does not facilitate the detection of those translation gaps. Rather, it masks them.

In the case of insertions in the source language text, the probability is high that the "inserted" sentences have roughly the same lengths as their subsequent sentences. The translations of the inserted and subsequent sentences then would have the same lengths, too. Accordingly, in the absence of any other information, the gap in the translation *must* remain unnoticed by the algorithm, and result in distortions in the alignment starting from the point where an insertion (or deletion) occurred[4]

The only effect of the length-based similarity measure is to filter out larger mismatches in sentence or paragraph lengths, but the success of the strategy almost exclusively depends on i) the ability to generate n:m links, and ii) the linear ordering of the source and target language texts to be the same.

The open question then is, if it is more efficient to use linear ordering as *only* alignment clue, and augment it with the probabilities that Gale and Church (1991b) have used to generate n:m links. In order to answer that question, I have implemented and tested an alignment strategy that uses linearity as only clue to paragraph and sentence alignment, described in section 4.4.

4.2 Cognate-based Alignment

Two alignment modules use the notion of *cognates* to compute both word and sentence alignment hypotheses. The first, *resource-independent*, module defines cognates along the lines of Simard et al. (1992), i.e. it uses string similarity and word length as clues, while the second uses corpus annotation, i.e. POS-tags to compute cognates for verbs, adjectives, nouns and names, only.

Both compare the words within two aligned paragraphs, or sentences respectively. If two words are considered cognates, they are aligned. In a second step, the word hypotheses are used to derive sentence alignment information: The sentences that contain the aligned words *inherit* the alignment information. A side effect of this approach is that the cognate-based alignment modules generate 1:1 links *only*. However, it is easy to merge overlapping hypotheses, and ATLAS features such an operation (section 4.10).

4.2.1 Resource-independent Approach

The first, resource-independent approach is used whenever the corpus does not contain any annotation, i.e. it is segmented into paragraphs, sentences and words, but POS-tags or any other linguistic annotation is missing.

It computes a string comparison between all words of the aligned input paragraphs if

1. the 'words' are punctuation marks or numbers, or
2. the words are at least 3 characters long

in a two-pass process. In the first pass, all words are aligned linearly if they are orthographically identical and occur with the same frequency. Words that do not meet this requirement are aligned

an example for such a tuning is given in section 4.12.

[4] The algorithm *can*, by chance, accommodate deletions and insertions, e.g. if there is a larger length mismatch between the inserted sentences and their successors.

in the second pass if the source language word shares at least 90% of its characters with the target language word, and if the two words occur with exactly the same frequency. As in the first pass, these cognates are aligned in linear order.

The restrictions on the word alignment have not been chosen arbitrarily. The condition that two cognates must appear with the same frequency in order to be alignable, e.g. is derived from a certain *position uncertainty*: If two word types are aligned, it is uncertain which tokens at which corpus positions correspond most closely. A name e.g. could be translated either as a name, or, because the name has been mentioned before or for any other reason, by some pronoun or paraphrase. As a result, both type frequency and the linear ordering of the tokens may be misleading. However, I consider the frequency of a word to be more salient than the linear ordering of its tokens, based on a simple assumption: Especially rarely occurring items, like hapax legomena, will *have to* occur in some way in the translations, and replacing them by anaphora may not be possible. The exact position of an expression, however, depends on translation style etc., thus more variation is possible, and causing more alignment errors.

Additionally, following Simard et al. (1992), I have excluded short words from the cognateness computation as short words tend to be function words for which orthographic similarity may be spurious. I have conducted test runs on the EUNEWS corpus which threshold on the cognateness value yields best accuracy, calculated in terms of how many aligned word pairs are true cognates. I achieved the best results with a threshold of 90% of shared characters, and accordingly set the threshold.

Unlike probabilistic approaches, the hypotheses of this module do not receive a probability value. Rather, I use the degree of cognateness, computed using the longest-common substring measure. The similarity value is given by

$$\text{similarity value} = \frac{2 * \text{cognateness}}{\text{length of L1 word} + \text{length of L2 word}} \tag{4.3}$$

thus receiving a value between 0 and 1. The confidence value of each hypothesis is given by the multiplication of the similarity value with the confidence value of the parent hypothesis.

After the word hypotheses have been generated for likely cognates, a simple inheritance strategy is used to derive sentence hypotheses: two sentences are linked if they contain at least one linked pair of cognates.

Test Runs In preliminary tests on the German-English EUNEWS, I found that the strategy yields good alignment hypotheses, but at a high computational load as virtually every word of a paragraph in L1 is compared to virtually all words in L2. Additionally, the coverage of the module is relatively low: it produced only 359 hypotheses, while the gold standard contained 496 links. Still, the precision was 80.6%, and recall was 55.1% (table 4.4).

I also examined which kinds of cognates were computed by the strategy: most of them were names, acronyms or numbers, i.e. their orthographic forms were identical in both languages. But I also observed cognates such as *Markt* → *market* that are highly similar, but not spelled identically. Some lexicon entries even contained alternatives due to spelling or wrong POS-tagging, the German currency abbreviation *EUR*, tagged as a noun e.g. was correctly linked to its English translation *EUR*, irrespective of whether it was tagged as "adjective" or "name" *EUR*. Errors were basically due to spurious similarities between e.g. prepositions and nouns (example: German *mit*, preposition, and English *remit*, noun). In fact, 292 out of 308 cognates were correct.

For French and German, the results were almost the same, with precision having a value of 79.0% and a recall value of 48.8%. Only 22 out of 141 translation pairs were wrong in the automatically computed dictionary, and fortunately, the strategy is independent of POS-tagging: the acronym *FEMIP* e.g. was tagged as both noun and marked as abbreviation in the French text. Still, as the approach did not use information on word category membership, the acronym was correctly linked.

language pair	precision (%)	recall (%)
German–English sentences	80.6	55.1
German–French sentences	79.0	48.8

Table 4.4: Test run results for the resource-independent cognate alignment

4.2.2 Resource-dependent Approach

The second module that aligns words and sentences based on the notion of cognateness is resource-dependent in that it restricts the computation of orthographic similarity using the corpus annotation: Cognateness is only computed for the words belonging to lexical classes, i.e. nouns, names, adjectives and verbs[5].

The process itself, as with the resource-dependent module, proceeds in two passes: in the first, orthographically identical words are aligned if they are lexical items and belong to exactly the same word category. Additionally, they must occur with the same frequency. In the second pass, two words are aligned based on nearly the same restrictions: instead of requiring orthographic identity, the two words must share at least half of the characters of the source language word.

The requirement to align only words with the same frequency is based on the same considerations as for the resource-independent module. Unlike in the module above, however, I do not need to impose a length restriction on the words in question: functional words like determiners can be removed from the set of candidate cognates because they are *tagged* as functional words.

Using POS-information to remove functional words from the cognateness computation also had the effect that I could relax the cognateness threshold: As cognateness is not computed for function words, there is fewer noise and less spurious orthographic similarity that need to be filtered out using the threshold. In other words, the unusually low threshold is possible as false cognates are filtered out due to the POS-filter, and also due to the severe frequency restriction.

As before, the probability value of each hypothesis is replaced by a similarity value that is computed using the formula

$$\text{similarity value} = \frac{2 * \text{cognateness}}{\text{length of L1 word} + \text{length of L2 word}} \tag{4.4}$$

thus yielding a value between 0 and 1, words with identical spelling and category membership having a similarity of 1. Again, the confidence value is computed by multiplying the similarity value of each hypothesis with the confidence value of its parent hypothesis.

Test Runs As with the first cognate alignment module, I found that the strategy yields good alignment hypotheses on the German-English EUNEWS texts, but also at a high computational load; although the number of comparisons between L1 and L2 words is considerably lower than in the resource-independent approach, it is still high as the number of nouns, names, adjectives and verbs in a corpus is relatively high. Additionally, the coverage of the module is quite low: the automatic alignment contained only 359 links. However, precision was quite high with a value of 84.1% and recall was 43.7%.

[5] As ATLAS derives broad word classes like *adjective* or *name* from the more fine-grained POS-annotation of the corpora using its parameter files, it is possible to condition this approach to *lexical items only* or, if necessary, even stricter to *nominals only* etc.

language pair	precision (%)	recall (%)
German–English sentences	84.1	43.7
German–French sentences	89.4	36.9

Table 4.5: Test run results for the resource-dependent cognate alignment

Again, I also examined which kinds of cognates were computed by the strategy: most of them were names, and adjectives. Surprisingly, the dictionaries of the two cognate-modules differed: the resource-dependent approach aligned fewer acronyms and more adjectives than the resource-independent approach. The reason for the difference are probably the different threshold settings for the two approaches.

As the set of cognate candidates was rather restricted, it is not surprising that only 156 cognates were aligned. However, only six error occurred.

On the German–French data set, the resource-dependent module achieved a precision of 89.4% and a recall of 36.9%, i.e. the alignment quality was roughly the same, and the coverage of the module was still quite low. This is also reflected in the dictionary, which contained only 78 entries. However, all of these entries were correct, and some of them were even nice cognate pairs like

(1) Projekt ↔ projet,

both meaning *project* in English.

4.3 Dictionary-based alignment

This alignment strategy is inspired by the approach of Kay and Röscheisen (1993): it computes first word alignment hypotheses based on the information in the system-internal dictionary, and then uses these hypotheses to derive sentence alignment hypotheses[6].

The module computes alignment hypotheses for corpus items within aligned paragraphs or sentences: for each word in the respective source language unit, a lexicon lookup is performed: if the word is contained in the system-internal dictionary with its appropriate word category, then its translation or translations are retrieved and the module determines whether one or more of the word's translations occur in the corresponding target language paragraph or sentence. If so, then the source language word is aligned to all its possible translations, i.e. a word may be multiply aligned.

The probability of each word hypothesis is the same as the confidence value for the translation pair as found in the system-internal dictionary. Another probability function could have been chosen, as well. However, as the success of the alignment strategy depends more on the quality of the dictionary used, and less on the input text, the confidence values recorded in the dictionary provide better information with respect to the correctness of a hypothesis.

Each word hypothesis that is generated by the module also receives a confidence value, and this equals the probability multiplied by the confidence of the input paragraph hypothesis.

The dictionary information for the module is taken from two sources: firstly, as has been explained before, each word hypothesis that is generated by any module of the aligner is also added to the system-internal dictionary, hence the dictionary is built during the alignment process with information derived directly from the corpus. This way, cognates found in the parallel corpus in one paragraph can be reused in another section of the corpus. Secondly, an electronically available

[6] It may also be used on sentences or phrases to establish word links only

dictionary can be made available to ATLAS, in which case its dictionary information will be used for the sentence and word alignment task.

After the generation of word hypotheses, the module derives sentence hypotheses based on a simple inheritance strategy: if two words are aligned, then the sentences they are contained in are aligned, too. As many word hypotheses with possibly differing probabilities may lead to the same sentence hypothesis, the probability of the sentence hypothesis is computed as the product of all word hypotheses used to derive it. The confidence of a sentence hypothesis is again given by its probability multiplied by the confidence of the input paragraph hypothesis.

As a side effect, the strategy alone computes 1:1 links, only. But it has been combined with a merging operation described in section 4.10 in order to combine overlapping alignment hypotheses to n:m links.

4.4 Linear Ordering of Paragraphs and Sentences

During the implementation of the length-based approach by Gale and Church (1991b), I found out that much of its success is *not* due to its similarity measure, i.e. on the hypothesis that *longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences* (Gale and Church 1991b, p. 78). Instead, by computing a directed search path through the alignment matrix, the approach implicitly assumes that sentences and paragraphs are *linearly ordered.*

Hence, it should be possible to abandon the similarity measure suggested by Gale and Church (1991b), or better replace it with a heuristic that directly exploits the similarity between the corpus positions of a sentence and its translation.

Accordingly, I firstly hypothesise that

Linear ordering is preserved during translation,

and define the corresponding similarity measure. It relates the relative corpus position of a source language sentence s_{L1},

$$\text{relative position of } s_{L1} = \frac{\text{position of } s_{L1}}{\text{number of L1 sentences}} \tag{4.5}$$

to the relative corpus position of a sentences s_{L2} in the target language,

$$\text{relative position of } s_{L2} = \frac{\text{position of } s_{L2}}{\text{number of L2 sentences}} \tag{4.6}$$

such that the smaller the difference δ

$$\delta = \frac{\text{position of } s_{L1}}{\text{number of L1 sentences}} - \frac{\text{position of } s_{L2}}{\text{number of L2 sentences}} \tag{4.7}$$

between the two positions, the higher is the similarity between the two sentences,

$$similarity(s_{L1}, s_{L2}) = 1 - |\delta| \tag{4.8}$$

and accordingly the higher is the probability that the two sentences should be aligned[7].

[7] A very similar similarity measure has been suggested by Debili and Sammouda (1992). However, the authors do not test the similarity measure in isolation. So the effect of the linearity constraint on alignment quality is still unknown.

This similarity measure can be combined with the probabilities for n:m links as estimated by Gale and Church (1991b), I arrive at the probability of a given alignment $P(s_{l1}, s_{l2})$ as

$$P(s_{l1}, s_{l2}) = similarity(s_{l1}, s_{l2}) * P(hypothesis) \tag{4.9}$$

with P(hypothesis) being the probability of a link type where $0 <= n, m <= 2$, as described before (section 4.1). However, as the link probabilities have been problematic in previous test runs, they are not used here. Instead, two sentences are merged into one segment, and the segment's position then computed as the "average" of the sentences.

$$\text{relative position of } s_1, s_2 = \frac{\frac{\text{position of } s_1 + \text{position of } s_2}{2}}{\text{number of sentences in that language}} \tag{4.10}$$

This strategy has the advantage that it allows to compute any kind of n:m link, without being restricted to those link types for which probabilities have been estimated. Secondly, link type probabilities estimated on data from one corpus may produce errors when used on another corpus, hence not relying too much on these parameters is advantageous.

Either way, this similarity measure should allow to align sentences or paragraphs with roughly the same success as the original version of the approach by Gale and Church (1991b), especially if combined with the module using sentence lengths as only cue.

Test run After implementing the similarity measure, I tested it by aligning the corpus EUNEWS on the paragraph and sentence level. In a first test run, I used the module to align the paragraphs on the German–English EUNEWS texts, and encountered runtime problems: As the program generates many hypotheses, the alignment process is considerably slowed down for larger documents. This is either a result of an inefficient implementation, or a problem of the subsequent alignment disambiguation, having to disambiguate too many hypotheses. For the initial test, I then excluded the largest two texts from the test run.

On the remainder of the EUNEWS texts, most of the paragraph links were correct. After examining the results more closely, I decided to set a threshold and filter out all paragraph alignment hypotheses where the similarity between their corpus positions is rather low. I empirically set to 0.5, assuming that this threshold was sufficient to filter out most improbable hypotheses, but still able to keep a wide-enough range of competing hypotheses. This threshold enabled me to re-run the alignment on the complete corpus. Now, this alignment module achieved a precision of 60.6% and a recall of 67.2% (table 4.6).

language pair	precision (%)	recall (%)
German–English paragraphs	60.6	67.2
German–English sentences	26.7	33.1
German–French paragraphs	41.8	49.3
German–French sentences	25.2	31.6

Table 4.6: Test run results for the linear ordering sentence alignment

On the German–French texts, the module achieved a precision of 41.8% and a recall of 49.3% on the paragraph level. As this result is considerably lower than for the German–English texts, I examined the data and discovered that the French texts contained many more and smaller paragraphs than the English texts, which explains the low alignment quality to some extent.

Surprisingly, an error in the paragraph alignment does not lead to a misalignment of all "subsequent" links (as in the length-based original). Rather, correct links were sometimes randomly distributed over the corpus, with errors or even error clusters in between. Unfortunately, I could not conclusively find out why errors occurred. However, I have the intuition that the errors were partially caused by the probabilities for insertions, deletions, and generally non-1:1-links.

With respect to sentence alignment, I tested how well the module performs if given perfect paragraph alignments. On the German–English data, the precision was considerably lower than before, with a value of 26.7%. Recall was similarly low (33.1%). Much of this effect is due to the module generating too many 1:0 and 0:1 links, but also to its sensitivity to permutations. The results on the German–French data set are similar. Here, the module also performed badly, with a precision of 25.2% and a recall of 31.6%.

4.5 Aligning Nominals: Expression Length and Category Membership

A particular problem inherent in statistical word alignment models is their inability to recognize and links multiword sequences and rare words. Accordingly, a module was implemented that links nominals irrespective of their frequencies. Furthermore, the module automatically detects nominal compounds and nominal multiword sequences, based on a set of heuristics.

The module has been developed on the basis of a data analysis of hapax legomena and rare events, i.e. types occurring with a frequency of 10 or below. In this data analysis, I discovered that a large portion of rare types, especially of the hapaxes, are nominals. Moreover, most of them are, depending on the language, multiword sequences like

(2) marital status, submission to Congress

(3) état civil, soumise au Congrès (French)

as in English, compounds as in German

(4) Personenstand, Kongreßvorlage (German)

or either of the two possibilities,

(5) civilstånd, framläggande inför kongressen (Swedish)

as in Swedish. As a result, these nominals are notoriously difficult to align: they may be very infrequent, and in these cases there is insufficient statistical information to align them correctly. Additionally, they usually involve n:m links, i.e. one or more tokens of the source language have to be aligned with a sequence of tokens in the target language, or vice versa.

Both difficulties have been recognized and tackled before: With respect to rare events, Dejean et al. (2003) e.g. report that lemmatizing infrequent types improves the alignment results, although they do not give an explanation for the phenomenon. Multiwords, on the other hand, have been discussed to a greater extent: Brown et al. (1993) had to introduce a statistical parameter called *fertility* into their models to handle n:m links. Unfortunately, fertility has to be computed for each word type of the corpus, thus it increases the computational load of the translation models to a significant extent. Additionally, the fertility of a word, i.e. its ability to participate in n:m links, is affected by random co-occurrence: if two tokens, like e.g. *the* and *a*, co-occur within many sentence links, they may erroneously be taken to constitute a multiword unit. In other words: although fertility can lead to correct n:m links, it also contributes to the amount of alignment errors.

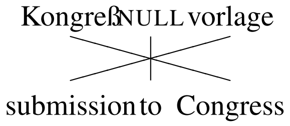


Figure 4.1: Example: Alignment of split-up compounds

In order to avoid fertility errors, many researchers have experimented with splitting compounds into their components, thus reducing the need to compute n:m links (Nießen and Ney 2001; Schrader 2002; Köhn and Knight 2003; Schrader 2004): if compounds like

(6) Kongreß|vorlage[8]
submission to Congress

are split up into their components (here, *Kongreßvorlage* is split up into *Kongreß* and *vorlage*), then these components can be aligned to their translations (here: *Congress* and *submission* with the preposition left unaligned) in a sequence of 1:1 links.

However, this approach to multiword alignment has several disadvantages: First, reliable morphological decomposers are hard to find, especially for languages with scarce resources, and moreover, they tend to compute several alternatives for most compounds. Thus, the splitting of compounds into their components adds the problem of disambiguating between several morphological structures. As examples like German

(7) Staubecken → Staub|ecken (dusty corners; literally: dust corners)
or
Staubecken → Stau|becken (reservoir; literally; dam basin)

or Swedish

(8) bildrulle → bil – drulle (bad driver; literally: car fool)
or
bildrulle → bild – rulle (roll of film; literally: picture roll)

(Swedish examples taken from Sjöbergh and Kann (2004))

show, the disambiguation of several morphological structures is not always an easy task. Thus it may introduce so many errors into the corpus that the strategy ceases to improve the alignment quality.

Moreover, splitting compounds into their components may lead to a *decrease* of alignment quality (Schrader 2002; Schrader 2004): the decrease of alignment quality is largely due to the splitting operation itself, even if the output of the morphology is perfect: The operation artificially increases the number of tokens and types in a corpus. As each token of a source language, whether it is part of a compound or not, can, *in principle*, be aligned with any token of the target language, the number of alternative links for compound components is higher than the number of alternative links for the original component. Accordingly, the amount of error increases.

The effect is obvious even in small examples. Consider the example sentence pair

[8] The | marks the morpheme boundaries.

(9) The submission to Congress constitutes a step in the right direction .
Die Kongreßvorlage stellt einen Schritt in die richtige Richtung dar .

It consists of 12 English and 11 German words, including the compound *Kongreßvorlage*, thus 132 1:1 links are possible. If the compound is split up into its components, i.e. if the compound is replaced by the sequence of its component words, the sentence pair will consist of

(10) The submission to Congress constitutes a step in the right direction .
Die Kongreß vorlage stellt einen Schritt in die richtige Richtung dar .

now 12 words in either language, and 144 alternative 1:1 links[9]. The splitting of the compound thus results in an increase in alignment alternatives, here of 9%[10]. Accordingly, the error rate may increase, too. A solution to this problem would be to somehow filter the results of the morphological decomposition, e.g. by allowing only those structures whose components can be translated, as done by Köhn and Knight (2003). However, this solution implies the use of a bilingual dictionary. Furthermore, the strategy suggested by (Köhn and Knight 2003) specifically deals with the alignment of "content" words and ignores intervening functional words occuring within multiword sequences.

There is also a theoretical disadvantage: splitting compounds into their components and aligning them implicitly assumes that a compound's meaning is made up compositionally, and that (the same degree of) compositionality also holds for its translation. However, this is not necessarily true, as examples like

(11) Personen|stand (marital status)
person status

readily show. Hence it would seem more advisable to *not* split compounds but finding means to align them correctly to the entire equivalent expression in the other language.

Tschorn and Lüdeling (2003), on the other hand, argue that many unknown words are due to productive word formation processes, i.e. that compositionality holds, and that translations for compound components can be found in existing dictionaries. In their approach, unknown words are morphologically analyzed, their components are looked up in a bilingual dictionary, and, if possible, aligned with their translations. This yields, in the best case, complete matches between a German compound and its English translation. Even partial matches between the compound and a part of its translation helps to improve sentence alignment quality, which is the task the authors have in mind. Unfortunately, partial matches are insufficient for word alignment, and the strategy of Tschorn and Lüdeling (2003) necessarily fails if compositionality does not hold.

A quite different approach has been suggested by Kupiec (1993): He aligned noun chunks in a French-English parallel corpus using POS-patterns for recognizing nominals including postmodifying prepositional phrases. Second, the EM-algorithm is used to statistically learn the correct alignment of the chunks. As Kupiec (1993)'s strategy relies on statistics, however, it has difficulties dealing with rare chunks.

Summed up, the problem persists despite numerous efforts: statistics do not suffice given the large amounts of rare nominals (and other words) and using morphological decomposers or POS patterns in combination with traditional statistical alignment computation does not improve the resulting alignment quality.

[9] Of course, deletions, insertions, and n:m links are also possible, thus the "real" number of possible links is still higher than 132 and 144, respectively.

[10] In sentences with more compounds, the compound splitting will of course increase the number of possible links even further.

However, neither multiword expressions nor compounds are random linguistic structures, and hence it should be possible to find clues that influence an algorithm towards making the right alignment decisions: if e.g. the structure of nominal multiword expressions is known, then it should be possible to exploit this structure for the alignment task. In this case, the focus will have to shift from word frequency statistics to statistics on the structures of compounds and multiword expressions. Simultaneously, correlations between the structure of a compound and the structure of its translation should be found, and vice versa the correlations between the structure of a multiword expression and the structure of its translation.

Data Analysis Accordingly, I have conducted a thorough data analysis of a total of 1113 German nouns and their English translations. As a first step, I extracted 512 German hapaxes from a 100,000 token subset of the EUROPARL corpus and analyzed them with respect to their word category membership, morphological complexity and word length. I also aligned them manually to their English correspondences in the corpus, and analyzed which categories they belonged to, what their morphological structure was, and how long they were. I also hypothesized which kinds of alignment problems are to be expected if the corpus is automatically aligned.

The analysis yielded that 353 of the 512 German hapax legomena, or 68.95%, are noun compounds, and that the translations of the noun compounds are chunk-like multiword expressions in 68% of all cases (table 4.7). These expressions are most often a sequence of nouns (52.41%), either followed by a PP (16,71%), or preceded by one or more adjectives (9.92%). Paraphrases (8.22%), or other nominal structures are rare (12.75%).

Structure of the Multiword	percentage
noun(s)	52.41%
noun(s) followed by PP	16.71%
adjective(s) noun(s)	9.93%
paraphrases	8.22%
other nominal structures	12.75%

Table 4.7: Patterns of English multiword nominals

Additionally, I found a strong correlation between nouns and their translations in terms of their morphological complexity : If a German noun contains *n* elements, then its translation most often also contains *n* elements (table 4.8). Moreover, most nominals in the data set consist of two components, irrespective of the language, and usually have to be aligned to an expression that, likewise, consists of two components. In the case that a multiword contains one more element than its compound translation, this is mainly due to the multiword containing an additional preposition or other functional element, as in the now familiar example

# elements	1	2	3	4	> 4
1	59	2	1	0	3
2	30	119	56	15	10
3	2	20	15	8	10
4	0	1	0	0	2

Table 4.8: Expression complexity. Rows show the number of components in English multiword units, columns give the equivalent numbers for German compounds.

(12) Kongreß|vorlage ↔ submission to Congress

Surprisingly, the German nominals in the data set and their translations show another correlation: the difference between the word lengths, counted in characters[11] is rather small.

word length	minimum	1^{st} quartile	median	average	3^{rd} quartile	maximum
German	3	12	15	15.43	19	32
English	3	11	16	16.03	20	58
German/English	0.2586	0.7857	1	1.1390	1.214	9

Table 4.9: Expression complexity: word lengths of German and English nominals correlate

Furthermore, the median of the length ratios is 1, and the average of the ratio

$$\text{length ratio} = \frac{\text{German compound length}}{\text{English multiword length}}$$

equals 1.139.

Parts of these findings could have been expected - given the large amounts of nouns in a corpus, their percentage among the hapax legomena has to be high, likewise. Additionally, it is plausible to assume that a nominal, at least in languages as closely related as English and German, can most often be translated by another, somehow "nounish" expression. These assumptions are, moreover, supported empirically in the case of multiword translations of German nouns: These multiword expressions roughly correspond to English nominal phrases, i.e. a German compound is translated by an English nominal phrase minus its determiner[12].

On the other hand, the strong correlations in terms of numbers of expression components and in terms of word lengths are surprising. The structural similarities indicate that what is a complex expression in one language, whether it is a compound or a multiword unit, almost necessarily is also a complex expression in the other language.

Several reasons are possible for these similarities: Firstly, the two languages are closely related and hence structural similarities with respect to word formation can be expected. Secondly, compounds containing more than two components may be dispreferred for reasons of optimal encoding of information. Or, the reason for the observed similarities stems from the two languages being spoken in a highly similar cultural context, in that concepts in both linguistic communities tend to be coded and combined along similar lines. However, these hypotheses are hard to prove, and very speculative. Before any considerations are undertaken, further data has to be examined: maybe the correlations between English and German nominals are spurious even for this language pair. More case studies for more language-pairs are needed.

Alignment Strategies In any case, the results of the data analysis can directly be exploited for the alignment task in a two-step fashion. Firstly, German compounds and English nominal expressions have to be identified in the corpus, and then the correlations between their lengths can be exploited to compute which German compound has to be aligned to which English expression.

The English nominals can easily be identified using the POS-patterns discovered in the data analysis: For each English noun, identified by its POS-tag, the alignment strategy seeks to construct several candidates: i) the noun or noun sequence itself (the nominal), ii) the nominal preceded by one or more adjective, iii) the nominal followed by a prepositional phrase, and finally iv) the nominal preceded by an adjective *and* followed by a PP.

[11] including the blanks within multiword nominals

[12] But including modifiers!

The identification of the German compound, however, depends on the corpus annotation: if the corpus is morphologically annotated, then each German noun that consists of more than one component is eligible to be aligned to an English multiword nominal. In the absence of morphological information, the word length of each German noun indicated whether it probably is a compound.

Likewise, the definition of the similarity measure depends on the corpus annotation. If morphological information on the German compound is available, then the similarity measure can relate the numbers of German compound components to the numbers of English multiword components. Otherwise, the German compounds and English multiwords have to be compared to each other using their "word" lengths[13].

4.5.1 Alignment of Compounds based on Word Length

In a first step, I have implemented an alignment module that exploits POS-patterns and word lengths to align German compounds to their English translations. The strategy is restricted to aligning only *complex* expressions, i.e. German simplex nouns are not aligned, at all. The reason for this restriction is that although there is a correlation between a German compound and its English translation in terms of expression length, I do not know whether the correlation exists for German non-compounds and their translations, as well[14]

The English multiword nominals are identified as described above. A German token is considered a compound if it is tagged as a noun and if it is at least 12 characters long. This threshold corresponds to the first quartile of the hapax noun lengths in the data set.

In a next step, the length ratios between each German compound and each English candidate translation of a sentence link are computed. If the similarity

$$\text{sim (compound,multiword)} = 1 - |\text{word length difference}|$$

is greater than zero, then an alignment hypothesis for this particular translation pair is generated. No further filtering, e.g. with respect to frequency of a compound, is employed, i.e. both hapax and non-hapax nominals are aligned.

Test Results After the implementation of the compound alignment strategy, I have conducted a test on the 100,000 token EUROPARL subset on which I carried out the data analysis: only this subset of the corpus has been submitted to ATLAS, including POS-annotation and sentence alignment information. Within this subset, the strategy aligned German noun compounds and their translations, and all results were used to construct a bilingual German-English dictionary. No other alignment module was used, nor did ATLAS compute a full text alignment or disambiguate between different link possibilities.

Afterwards, I semi-automatically evaluated whether the dictionary contained lexical entries for the 353 hapax nouns in the analysis, whether these lexical entries contained the correct translations, or partial translations. Furthermore, I analyzed why correct translations were missing in the lexicon entries. I also tested why lexicon entries were missing in the automatically generated dictionary. I did not evaluate the translation direction English→German, assuming that as the strategy is not directional, it is not necessary to examine both translation directions.

Results are that the dictionary contains lexical entries for 236 of the 353 compounds in the data set (66.86%), and more than 1600 additional entries with German headwords. This is not surprising given that the module does not use any frequency threshold, i.e. it aligns nouns irrespective

[13] This question is concerned entirely with the question whether a parallel corpus has been annotated with this particular type of corpus annotation; ATLAS can exploit this type of morphological information, if it is part of the corpus annotation.

[14] It would be interesting to examine on an aligned corpus whether the correlation can be found for German non-compounds and their translations, and to which degree it holds for words of other categories, as well.

of how often they occur in the input data. With respect to the 115 missing entries, I found out that in most cases, the German compounds did not pass the length threshold (66.96%), and hence no hypotheses were computed for them. Other error sources like tokenization problems (1.74%), compound recognition errors due to paraphrases or hyphenation (10.43%) occurred, but rarely. Fortunately, there were only 11 cases (9.57%) where I could not attribute the errors to tokenization or POS-patterns, and hence had to attribute them to the length-based similarity measure.

With respect to the hapaxes that received an entry in the dictionary, I found out that the entries contained the correct translation in 47% of all cases. Additionally, I found 306 translation suggestions where the correct translation was either partially present, or a substring of a suggestion. 121 entries, however, did not contain any correct translation. The error analysis showed that these fully incorrect lexicon entries were partially due to the similarity measure itself (46.28%), to POS-patterns that were not used in the module (33.06%) and partially due to the nominal recognition not working optimally. In detail, the nominal recognition failed because hyphenated words in both languages had been split into their components during tokenization, because the POS-tagger treated words that occurred sentence-internally *and* in upper case as names, but did not tag them as common nouns, because the module did not account for all POS-patterns, and because of paraphrasing, deletion or category changes during the translation process.

I further examined 423 lexicon entries randomly chosen from the 1600 additional lexicon entries that the module computed. Overall, the analysis yielded that in principle, the method works well, but that the recognition of the nominals can be improved. In more detail, 58.39% of the 423 lexicon entries contained correct translations, with many more partial translations, or cases where the correct translation was a substring of a suggestion. In the roughly 40% of examined cases where the correct translation was missing in the lexicon, the error was mostly due to either the similarity measure (21.51%), the fact that the nominal was not literally translated, but paraphrased (17.20%), or that the nominal recognition did not work well enough (26.88%). Again, errors occurred because hyphenated words had been split into their components by an over-eager tokenizer during corpus preprocessing[15].

Accordingly, I revised those parts of the implementation that dealt with the recognition of German and English nominals. In detail, I extended the POS-patterns to cover nominals tagged as names, and repaired the over-eager tokenization with respect to German and English hyphenated nominals. Now, a hyphen adjacent to a noun causes the module to reanalyse and complete the token sequence as a hyphenated noun, thus treating token sequences like

(13) Geldwäsche - Bekämpfungsrichtlinie
(English: anti-money laundering directive)

or

(14) anti - riot act (German: Antiterrorgesetz)

as compounds, or multiwords, respectively. In a subsequent test run, I discovered that the module correctly identified 986 hyphenated words, but only 167 errors occurred.

After the improvements in the alignment module were complete, I re-ran the testing. Overall, the performance increased: Now, 248 of 353 compounds (70.25%) were headwords in the automatically generated dictionary, with 175 entries containing the correct translations (table 4.10). With respect to the missing entries, error numbers decreased with respect to unaccounted-for POS-patterns for the English expressions, and with respect to the similarity measure, although I did not

[15] This was true for hyphenated words in both English and German. The EUROPARL version I used for the experiments reported here was the version originally prepared by Köhn (2003). The version now distributed via the OPUS homepage (Tiedemann and Nygaard 2004) has obviously been preprocessed using a different tokenizer that does not split hyphenated words.

language pair	precision
German–English 353 compounds	49.58%
German–English 760 nominals	73.81%
German–Swedish 224 compounds	45.98%

Table 4.10: Test results for heuristic Nominal Compound Alignment

change it at all. This could be an effect of the similarity measure actually discarding errors made during the nominal recognition in the first test run.

A final analysis was carried out on further 760 compounds that did not appear in the original data set, but had been aligned nevertheless: I found 561 correct translations, including multiple translations for some head words, as in

(15) Berufsausbildung ↔ vocational training, professional training

These 760 compounds come from all frequency ranges within the 100,000 token subset that I used for this evaluation, i.e. the set contains hapaxes as well as other rare events, but also frequent compounds like *Geschäftsordnung* (Rules of Procedure), which occurred 32 times in this subset. The lexicon entries of these frequent nouns contained many more translation candidates than those of rare words. Additionally, frequent compounds tended to have multiple translations in the corpus, whereas non-frequent words did not. In those cases where multiple translations were used, they were also listed in the dictionary.

Two final analysis were carried out to find out whether partially correct translations could be exploited to improve the results, and whether the similarity measure works reasonably well. According to my data, partially correct translations cannot be considered useful, as most compounds, and simultaneously most of their translations, are made up of two components, so contracting a set of partial translations in order to gain correct and complete ones is not possible. Room for improvement, however, is given for the similarity measure, as correct translation pairs generally do not receive a higher confidence than their alternatives. In other words, although the similarity measure is useful to align compounds with their translations, it does not work sufficiently well to ensure that correct word hypotheses are computed and used in the final alignment disambiguation step. However, it is unclear whether it is worthwhile to improve the similarity measure: Hypotheses aligning frequent compounds with recurring translations may still receive confidence values that are higher than erroneous alternatives, simply due to the correct translation pair occurring more often than each single error.

Extension to Swedish In order to test whether the strategy ports to another closely related language, I have repeated the data analysis for the language pair German-Swedish. I re-used the German hapax nouns described above, and manually aligned them with the help of Judith Degen, student research assistant at the institute. Additionally, we collected information on the morphological or, more generally, structural properties of the Swedish translations, and determined which kinds of alignment problems could be expected.

All in all, the data showed that Swedish made less use of compounding than German, but also that it used more multiword expressions than German. As an effect, complex German nominals were usually translated by complex Swedish expressions, which need not be compounds themselves, but could be multiword nominals, as well.

In more detail, 285 German nouns were morphologically complex, while the same could be said of only 254 Swedish expressions. Furthermore, 23 German nouns were not translated into

Swedish. Accordingly, I have excluded these 23 nouns from the analysis. In most of the cases, German complex expressions where translated by Swedish complex expressions, and the use of compounds prevailed: in only 31.33 % of all cases, a German nominal was translated as a multiword expression. In other words: whereas German makes heavy use of compounding, both alternatives, compounding and multiword expressions, can be used in Swedish. But the use of compounds is preferred in two-thirds of all cases in the data set.

These multiword units were basically nominals, and I found again a correlation between the number of components of an expression, and the number of components of its translation: Generally, a German noun containing *n* elements was translated by a Swedish expression that usually also contained *n* elements. (table 4.11).

# elements	1	2	3	4	>4
1	47	5	4	3	0
2	24	122	31	9	6
3	3	21	28	11	7
4	1	3	2	1	1
5	0	0	0	0	1

Table 4.11: Expression complexity. Rows show the number of components in Swedish multiword units, columns give the equivalent numbers for German compounds.

Given these similarities, it is of course interesting to see whether the same correlation can be found with respect to word length, counted in characters. None too surprisingly, German compounds and their Swedish equivalents were of roughly the same lengths, the Swedish nominals being slightly shorter (table 4.12).

word length	minimum	1^{st} quartile	median	average	3^{rd} quartile	maximum
German	3.00	12.00	15.00	15.62	19.00	33.00
Swedish	3.0	10.0	16.0	15.8	21.0	52.0
German/Swedish	0.1875	0.7816	1.0000	1.2530	1.2860	6.7500

Table 4.12: Expression complexity: word lengths of German and Swedish nominals correlate

As a result, it should be easy to extend ATLAS to also align German and Swedish compounds based on their word lengths, disregarding the cases where Swedish used multiword expressions: German compounds are detected as described above. Furthermore, the same routine is used for detecting Swedish compounds, the only difference being the length threshold: while any German noun counts as a compound if it is longer than 12 characters, Swedish nouns need only be longer than 10 characters.

As for German–English, I tested the length-based compound alignment on the 100,000 token EUROPARL subset on which I carried out the data analysis, and semi-automatically examined the automatically created dictionary for all translation pairs in the data set. The automatically created dictionary contained 20,579 entries, among them I found 103 correct translation pairs. 106 German nouns could not be aligned correctly due to their translations being multiword expressions. As I had used the automatic alignment information of the EUROPARL corpus, 16 failures could be attributed to alignment errors, and further 10 error were due to errors of the POS-annotation. The remaining 95 translation pairs were not included in the dictionary due to the length thresholds: in these cases, the German expression was not recognized by the module because it was too short, or its translation did not pass the length threshold. In other words, the module computed correct translation pairs in only 31.21% of all cases (table 4.10). However, as the module had deliberately been set up to ignore Swedish multiword expressions, the 106 translation pairs where the Swedish translation was a multiword have to be excluded from the analysis, thus reducing the data set to 224 translation pairs. From this perspective, the module correctly found translation equivalents in 45.98% of all cases. The length threshold is responsible for another 42.86% of all cases, and the remaining 13.73% are due to errors in the corpus annotation.

4.5.2 Alignment of Nominals based on Morphological Information

I also implemented and tested a modified version of the above-described alignment strategy: It uses the number of (free) morphemes in a nominal expression, instead of counting characters.

The strategy works in essentially the same way as before, with two changes: Instead of using word lengths, the morphological annotation is used to recognize German compounds, and to distinguish them from simple forms: any German token tagged as a noun and consisting of more than two morphemes is considered a compound.

Secondly, I adapted the similarity measure to the task. Instead of computing the ratio between the source and target language nominals, it uses the ratio between the *numbers of free morphemes* in two words w_{L1} and w_{L2}.

$$sim(w_{L1}, w_{L2}) = 1 - \frac{|\#\text{ morphemes in } w_{L1} - \#\text{ morphemes in } w_{L2}}{\#\text{ morphemes in } w_{L1} + \#\text{ morphemes in } w_{L2}} \tag{4.11}$$

For the German compounds, the number of morphemes is, obviously, given by the morphological annotation. The number of morphemes in the English nominals corresponds to the number of *words* in the expression[16]

Test results A first test run on the morphologically annotated CroCo corpus showed that the alignment strategy yields good results. In fact, the results were better than if no morphological information was used. As a result, I extended the strategy to align all kinds of nouns, i.e. it is no longer restricted to noun compounds. Again, the test results were very encouraging.

For a more thorough test run, and in order to compare the results to the nominal alignment strategy described above, I let the module align all nominals on the 100,000 token subset of the EUROPARL corpus that I had used for the initial data analysis: All nouns in the subset were morphologically analyzed using GERTWOL, a commercial morphological analyser (Haapalainen and Majorin 1994), and disambiguated with the method described in Volk (1999)[17]. Afterwards, the morphological information was added to the corpus annotation, and ATLAS was rerun to align all nouns in the subset.

After the alignment, the automatically generated dictionary contained 140,554 different lexical entries in English and German, including improbable or incorrect information. Among them, 193 of the 353 gold standard nouns (54.67 %) received a correct translation (table 4.13). This is considerably less than when aligning based on word length alone. The difference is only due to the different thresholds: in the first strategy, any token was taken to be a compound if it exceeded a certain length, hence many non-compounds were aligned. When morphological information is used to detect compounds, these long simplex forms are excluded from the alignment. Furthermore, some words were not decomposed at all, despite their being compounds. In sum, 57.72% of the nouns missing in the automatically generated lexicon were either morphologically simplex,

[16] In morphology, words are considered free morphemes, in contrast to *bound morphemes*, like derivational or inflectional affixes, clitics, etc.

[17] Thanks to Martin Volk for providing the morphological information.

language pair	precision
German–English 353 compounds	54.67%
German–English 760 nominals	63.33%
German–Swedish 224 compounds	36.16%

Table 4.13: Test results for Nominal Compound Alignment using Morphology

or compounds that were not decomposed by GERTWOL. Apart from this difference, the error sources were the same as in the length-based compound alignment: some compounds were not aligned because they were paraphrased or deleted in the translation, or because of tagging errors, etc. (24.83%). Others were translated, but using multiword patterns not covered by the strategy (17.45%).

With respect to the roughly 760 additional nouns, 494 received at least one correct translation (63.33%). The remaining words were either not compounds (22.95%) or taken to be simplex forms by GERTWOL (3.33%). Further failures at translation the compounds occurred because the translations followed different patterns than those used by the alignment module (6.92%), or the errors were due to tagging mistakes (2.05%) or various other errors (1.41%).

Extension to Swedish Fortunately, and also thanks to Martin Volk, I could obtain morphological analyses for the Swedish nouns in the EUROPARL subset. As a result, I could also test how the module performed for the language-pair German–Swedish. I repeated the compound alignment on the German–Swedish subset of EUROPARL on which I had carried out the data analysis and the previous compound alignment based on word length. This time, however, German and Swedish nominals were aligned if both of them consisted of at least two free morphemes. The similarity measure is the same as used previously to align German and English nominals based on their numbers of morphemes.

In this experiment, the automatically generated dictionary contained only 11,994 entries, among them 81 correct translation pairs of the data set. Again, 106 German nominals could not be aligned because their translations were multiword expressions. Further 16 nominals could not be aligned correctly due to errors in the sentence alignment, and 10 POS-tagging errors, of course, also decreased the alignment success. In the remainder of 116 cases, the nominals were either simplex forms or incorrectly annotated as simplex forms. If I again exclude the 106 nominals that are translated by multiword expressions, the module correctly found translation equivalents in 36.16% of all cases.

So it seems that the positive results from the German–English experiments do not easily carry over to another language pair. However, it should be kept in mind that the adaption to the new language pair was not optimal: although Swedish was observed to use more multiword units than German, and to have specific POS-patterns, I did not adapt the module to take these POS-patterns into account. A more thorough adaptation to the language pair, accordingly, will probably lead to better results.

4.6 Word Category Membership

When having a parallel corpus annotated with POS-information, one very obvious idea is to develop an alignment strategy that aligns words if they belong to the same word category, or at least to similar ones. The assumption behind this strategy is that

word category membership is not changed during translation.

Unfortunately, there are some caveats: Firstly, changes in word category *do occur* for various reasons: There may be more general morphological differences between the languages, the source language e.g. very explicitly distinguishing between nouns and adjectives, but not the target languages. In these cases, there will be very systematic word category changes. Even if the source and target languages are highly similar, like e.g. English and German, there may be lexical gaps or divergences, such that a word is translated using a periphrastic construction, or a support verb construction or idiom is used. There may be also structural or stylistic reasons to translate a nominalization e.g. with a verb. Or, anaphoric expression might replace noun phrases etc. in the translation.

In practice, the POS-tagsets used might differ with respect to the amounts of information they encode, and they may belong to different tagset. The two Italian tagsets e.g. for which the tree-tagger has been trained, e.g. differ with respect to verbal distinctions. Whereas one tagset has one tag for verbs, and several suffixes for more fine-grained morphological distinctions, the other additionally distinguishes between several verb classes such as auxiliaries versus modal and causal verbs versus “normal” verbs (table 4.14).

This second caveat can be met fairly easily by mapping specific word classes like *VVFIN* (finite full verb, STTS) to more general classes like *verb*. This is already parametrized in ATLAS, as the system automatically generalises from specific tagsets to word classes like *noun, verb, preposition*. Furthermore, the ATLAS parameters also specify which POS-tags can be attributed to functional and lexical category classes (appendix E).

Concerning the first, it is possible to refine the hypothesis in two directions. One is to make a difference between lexical and functional category classes, the former being the categories *noun, adjective,* and *verb*, and the latter including all other categories. The alignment clue could then be

> Words from lexical category classes are translated by words from lexical category classes, and words from functional category classes are translated by words from functional category classes.

Category changes between lexical category classes would be well accounted for by this hypothesis, as well as word category changes between functional ones. Still, category changes are possible in case where an anaphoric expression, being functional, has been translated by a word or expression from a lexical class, like a noun.

The second solution is to weaken the hypothesis to

> generally, word category membership is not changed during translation,

thus requiring a probabilistic model that predicts when word category changes *are likely to occur*. One way to estimate the probabilities of word order changes is to align a small subset of words from a variety of word categories, in a variety of contexts, with their translations.

In fact, I have tested all three hypotheses using one randomly chosen file of the English–German EUNEWS: I aligned the 357 German and 334 English tokens of the corpus, ignoring 1:0 and 0:1 links. For all other link types, I recorded which word category they and their translations belonged to, and whether these word categories are lexical or functional. If one token was part of a multiword sequence, or if it had to be aligned to a multiword sequence, then I recorded the relevant sequence of word categories.

If e.g. the German expression *in erheblichem Umfang* was translated by the English adverb *extensively*, I recorded that the German expression has the word category sequence *preposition adjective noun*.

In these cases, the distinction between lexical and functional word categories is hard to make, so I avoided this decision by assigning them membership in a *miscellaneous* category. In a second step, I distinguished between multiword sequences and single words, and decided to analyze the alignment patterns of the single words.

Italian 1	meaning	Italian 2	meaning
VER:cimp	verb conjunctive imperfect		
VER:cond	verb conditional		
VER:cpre	verb conjunctive present		
VER:futu	verb future tense		
VER:geru	verb gerund	VER2:geru VER2:geru:cli VER:geru VER:geru:cli AUX:geru AUX:geru:cli	gerundive form of modal/causal verb gerundive form of modal/causal verb with clitic gerundive form of verb gerundive form of verb with clitic gerundive form of auxiliary gerundive form of auxiliary with clitic
VER:impe	verb imperative		
VER:impf	verb imperfect		
VER:infi	verb infinitive	VER2:infi VER2:infi:cli VER:infi VER:infi:cli AUX:infi AUX:infi:cli	infinitival form of modal/causal verb infinitival form of modal/causal verb with clitic infinitival form of verb infinitival form of verb with clitic infinitival form of auxiliary infinitival form of auxiliary with clitic
VER:pper	verb participle perfect	VER2:ppast VER:ppast VER:ppast:cli AUX:ppast	past participle of modal/causal verb past participle of verb past participle of verb with clitic past participle of auxiliary
VER:ppre	verb participle present	VER2:ppre VER:ppre AUX:ppre	present participle of modal/causal verb present participle of verb present participle of auxiliary
VER:pres	verb present	VER2:fin VER2:fin:cli VER:fin VER:fin:cli AUX:fin AUX:fin:cli	finite form of modal/causal verb finite form of modal/causal verb with clitic finite form of verb finite form of verb with clitic finite form of auxiliary finite form of auxiliary with clitic
VER:refl:infi	verb reflexive infinitive		
VER:remo	verb simple past		

Table 4.14: Mapping between the two Italian tagsets for the tree-tagger

word category	category class	German	English
name	lexical	19	23
noun	lexical	63	54
nominal compound	lexical	0	18
adjective	lexical	24	16
number	lexical	7	9
verb	lexical	18	23
truncated word	lexical	3	0
adverb	functional	6	6
conjunction	functional	14	14
determiner	functional	14	7
preposition	functional	30	28
pronoun	functional	2	4
punctuation	functional	26	26
miscellaneous	functional	33	30
Σ		259	258

Table 4.15: Word category distribution in EUNEWS subset

	English category classes		
German category classes	functional	lexical	miscellaneous
functional	79	0	13
lexical	0	126	8
miscellaneous	8	17	8

Table 4.16: Alignment patterns in EUNEWS subset: category classes

According to this analysis, the German EUNEWS text contained 226 single tokens and 33 multiword expressions, while the English translation consisted of 210 single tokens, 18 nominal compounds according to the definition above (section 4.5), and 30 other types of multiword expressions (table 4.15).

92 German tokens belong to functional categories, while 134 German tokens are lexical. The numbers of lexical versus functional words are similar for the English texts: 87 tokens are functional, and 143 are lexical. Moreover, when analysing the alignment patterns, it is obvious that functional words almost exclusively align to functional words, the only exception being 1:m links where a functional word is aligned to a multiword sequence. The picture for lexical words is similar: they align either to lexical words, or to multiword sequences (table 4.16).

So obviously it is safe to assume that a word from a lexical class is virtually always translated as either another word from a lexical class, i.e. nouns, adjectives and verbs are translated using another noun, adjective or verb. Cases where lexical words are parts of multiword sequence occur, but rarely. Simultaneously, a word from a functional class is always translated as either another functional word, or as part of a multiword sequence.

These findings are intuitively plausible as there is no obvious reason why a conjunction e.g. should be translated as an adjective. Additionally, the data does not even include one case where an anaphora, being a functional word, is linked to a (lexical) noun. Of course, the data set is quite small with only 260 links, and may not give an accurate picture. However, the data is sufficient to notice that noun–pronoun translations are very rare, if they occur at all.

	English lexical categories						
German lex. cat.	adjective	compound	misc	name	noun	number	verb
adjective	**16**	1	2	0	3	2	0
misc	0	5	**0**	4	4	0	4
name	0	0	0	**19**	0	0	0
noun	0	12	5	0	**43**	0	3
number	0	0	0	0	0	**7**	0
verb	0	0	1	0	1	0	**16**
trunc	0	0	0	0	3	0	0

Table 4.17: Alignment patterns in EUNews subset: within lexical categories

A repetition of the data analysis on additional parallel data could verify whether the data set was simply too small, or too clean. In any case, the fact that the proportions of word categories in German and English are roughly equal already indicates that the translations are very clean.

A closer look at the data reveals that category changes occur rarely even within lexical categories (table 4.17). German adjectives, e.g. are usually linked to adjectives, the only exceptions being four cases where an adjective was linked to a noun or noun compound, and two cases where they were aligned to numbers. As numbers are often used as adjectives, however, one might argue that it is not necessary to distinguish between adjectives and numbers. Similarly, the distinction between nouns and compounds is rather artificial, and those truncated words that occurred in the data were components of German noun compounds in elliptic constructions, as in

(16) Kohäsions- und Strukturfonds
truncated conjunction noun
Cohesion and Structural Funds

Only three true cases of category changes occurred in the data, namely three cases where a German noun was aligned to an English verb, as e.g.

(17) Finanzierung ↔ by financing.

So with respect to lexical words, it is relatively safe to assume that word category changes take place in very rare cases. Hence, with reserving no or only marginal probabilities for very specific category changes, POS-information can be used to filter out inappropriate word links based on the assumption that

category changes do not occur for words from lexical categories.

Functional words change their category memberships slightly more often during translation, namely in roughly 30% of all cases. Still, the changes are not random: German determiners may be linked to determiners (42.86%) or prepositions (21.43%), not counting links to multiword sequences, and German prepositions may be linked to adverbs (3.33%), determiners (3.33%) or prepositions (76.67%). Furthermore, 14 functional words are linked to multiword sequences (16.28%). English adverbs may be linked to either adverbs (33.33%) or prepositions (16.67%), and determiners are translated as either determiners (85.67%) or prepositions (14.29%). English prepositions , finally, are aligned either to determiners (10.71%) or prepositions (82.14%). Finally, seven functional words are translated into German using multiword sequences (8.86%).

Although these absolute frequencies are too small to be reliable, it is possible to assume that except for prepositions and determiners, word category changes do *not* occur. Even though, the

	English functional categories								
German func. cat.	adverb	comma	conjunction	determiner	ignore	misc	preposition	pronoun	punctuation
adverb	**2**	0	0	0	0	4	0	0	0
comma	0	**10**	0	0	0	0	0	0	0
conjunction	0	0	**14**	0	0	0	0	0	0
determiner	0	0	0	**6**	0	5	3	0	0
ignore	0	0	0	0	**7**	0	0	0	0
misc	3	0	0	0	0	**0**	2	2	0
preposition	1	0	0	1	0	5	**23**	0	0
pronoun	0	0	0	0	0	0	0	**2**	0
punctuation	0	1	0	0	0	0	0	0	**8**

Table 4.18: Alignment patterns in EUNews subset: within functional categories

estimates are more reliable if I distinguish only between the probability of a category change, or a determiner or preposition having the same word category as its translation: the probability that an English preposition is translated by a German preposition is then 82.14%, and similarly, the probability of an English determiner having been translated as a German determiner is 85.71%. Estimates for the reverse translation direction are lower: the probability that *no* category change occurs is 76.76% for prepositions, and 42.86% for determiners. In other words,

except for prepositions and determiners, category changes do not occur within functional word classes.

A Filter These two heuristics on word category changes can both be used to filter out inappropriate word pairs, but it should be kept in mind that they are based on a relatively small and well-translated data set: the heuristics may be difficult to use on larger data sets, and the usefulness of the estimates is doubtful. Moreover, as links between anaphora and nouns did not occur in the sample, this alignment pattern, and possibly more, are not accounted for at all. Furthermore, the sample data was concerned only with 1:1 links, i.e. deletions and 1:m and n:m links must be accounted for using different means.

Still, when e.g. inducing a lexicon from parallel data, the findings that generally, word category changes do not occur, can be used to filter out inappropriate or at least improbable word pairings.

A Strategy Secondly, these findings can directly be used for a word alignment module: in this module, all words of a sentence link are aligned based on their word categories. Of course, the resulting word hypotheses of the module will include a lot of noise. However, the architecture of ATLAS will provide two additional filters: it is quite improbable that every sentence of the source language will be linked to every sentence of the target language, hence some word hypotheses, although theoretically possible, will never be generated. If two sentences are not linked, then those words that they contain are, likewise, never linked. Secondly, as each word hypothesis is fed into the system-internal dictionary, the recurrence of a word pair will lead to an increase in its confidence, and hence in the confidences of its accompanying word hypotheses. As a result, word pairs occurring over and over again will have a higher confidence than rarely paired words. This effect is quite similar to the one achieved by using co-occurrence scores for word alignment. However, it comes with the additional advantage that inappropriate word linkings either do not occur, or are highly improbable, depending on the degree of filtering employed.

As a test, then, I have written a simple alignment module aligning each and every word of a sentence link on the basis of its word category membership, where words are assigned a similarity score $sim_{restrictive}$

$$sim_{restrictive} = \frac{1}{\#\text{occurrences of pos in L1} + \#\text{occurrences of pos in L2}} \qquad (4.12)$$

language pair	precision
German–English	18.95%
German–French	26.92%

Table 4.19: Test results for POS Alignment

if they have the same word category, based on the frequency of that specific word category in the current sentence link. Using only this similarity measure, words are not linked at all if their word categories do not match, i.e. the resulting set of word hypotheses may contain gaps.

In order to fill these gaps, a second similarity score $sim_{non-restrictive}$

$$sim_{non-restrictive} = \frac{1}{\#words_{L1} + \#words_{L2}} \tag{4.13}$$

is used as a fall-back to generate all other possible word hypotheses. This time, the numbers of words with the sentence link are used to compute the similarity score, i.e. the resulting similarity scores will be considerably smaller than when the word categories of two words match.

Test Runs This module was tested on the English-German part of the EUNEWS corpus on which I have carried out the data analysis. Additionally, I have tested the module on the same data, for French and German. When using only those word hypotheses where the words had identical word categories, 306 entries were included in the automatically generated dictionary, 58 of them being correct (18.954%). Further 16 entries were at least partially correct, i.e. multiword units were involved, or the translation pairs were at least probable. When using both similarity measures, i.e. when all words within a sentence link were aligned with higher similarity scores for words with the same word categories, the automatically generated dictionary was considerably larger with 476 entries. However, only few of these entries, namely 66 translation pairs, were correct (13.866%). Partial or at least probable translations occurred 21 times[18].

When aligning the parallel German and French text, only 52 lexicon entries were included in the automatically generated dictionary, and only 14 of them were correct (26.923 %), if only words with the same word categories were aligned (table 4.19). When using both similarity measures, i.e. when aligning all words, favouring those links where words had the same word category, resulted in an automatically generated lexicon with 70 entries, among them 17 correct ones (24.286 %).

So it seems from these rather tentative results that word category membership is not a useful clue for the word alignment task. Basically, the bad results are due to two difficulties: first of all, the module so far generates only 1:1 links, thereby being very limited in scope. Thus it is necessary to somehow detect multiword sequences, either by using separate alignment modules like the one for aligning compounds (section4.5) or by using a segmentation and tagging scheme that can handle multiword sequences. Secondly, a close examination of the automatically generated dictionaries revealed that the disambiguation between various, equally plausible word hypotheses, is a problem: whenever a sentence contains two or more words of the same word category, then choosing the correct hypothesis depends crucially on whether support of it comes from elsewhere, e.g. from other occurrences of the particular hypothesis at some other corpus position. Other supporting evidence, such as additional dictionary information or cognateness may also be useful to disambiguate. Hence this module must be used preferably in combination with others. Still, it should be kept in mind that the module was tested on a very small text sample, and hence its success or non-success is no reliable indication for its general applicability. The strategy should,

[18] As in previous experiments, no alignment disambiguation was done, hence the bilingual dictionaries contain many alternative translation candidates per head word.

therefore, be thoroughly tested on a larger data set, and it should be used in combination with other strategies using different clues.

4.7 Lexicon Induction using Word Rank Correlation

As has been discussed before (section 2.2), word alignment is usually done using word cooccurrence statistics or statistical translation models, which both may be unreliable for rare events. Moreover, most approaches rely on prior sentence alignment information. Hence these approaches are not easy to incorporate into ATLAS: the sentence alignment information may not be available. Or, as it is generated simultaneously with the word alignment, it may not be complete. Either way, a different strategy must be used if ATLAS is to compute alignment hypotheses bases on the statistical properties of words.

However, it is possible to induce a lexicon even without any prior sentence alignment, as several approaches have shown (Fung and Church 1994; Fung and McKeown 1994; Debili and Sammouda 1992)). Unfortunately, Fung and McKeown (1994) and Fung and Church (1994) do not give an evaluation how their automatically induced lexicon influenced the sentence alignment, neither did (Debili and Sammouda 1992).

Accordingly, a number of experiments have been conducted in order to implement a statistical word alignment module that can induce lexical information from a parallel corpus, without requiring sentence alignment information, and without using co-occurrence statistics. Concerning rare events, a statistical cue was looked for that works both for low-frequency words and frequently occuring ones.

Rare Events and Statistics

As has already been mentioned above, co-occurrence statistics need not give reliable results for rare events (Dunning 1993; Evert and Krenn 2001). They may give reliable results for many words of a corpus, provided their frequencies are high enough. However, the word types in text corpora typically show a Zipfian frequency distribution, where rare types make up at least half of the vocabulary (figure 4.2).

wordforms (lang.)	types	hapax leg. (%)	rare events (%)	“normal´´ words (%)	frequent words (%)
EUNews (de)	2,590	1,546 (59.69)	909 (35.10)	118 (4.56)	18 (0.70)
EUNews (en)	2,223	1,199 (53.94)	878 (39.50)	133 (5.98)	14 (0.63)
EUNews (fr)	2,451	1,343 (54.79)	940 (38.35)	147 (5.99)	22 (0.90)
EUROPARL (de)	373,993	187,977 (50.26)	128,505 (34.36)	41,295 (11.04)	16,217 (4.34)
EUROPARL (en)	130,731	51,631 (39.49)	46,560 (35.62)	20,295 (15.52)	12,246 (9.37)
EUROPARL (fr)	153,727	57,626 (37.49)	56,125 (36.51)	25,616 (16.66)	14,361 (9.34)

Table 4.20: EUNEWS and EUROPARL: vocabulary size (word forms)

As can be seen in table 4.20 and in figure 4.2, hapax legomena[19] make up 37-59% of the word types, i.e. the vocabulary, of the development corpora EUNews and EUROPARL corpora, depending on the language. Furthermore, the amount of other rare events is also quite large, with 34-39% of all word types. On the other hand, “normal” words, i.e. words that are neither rare nor particularly frequent, make up only 4-17% of a corpus' vocabulary. Finally, frequent words constitute a relatively small portion of the corpus vocabulary (roughly 1-10%).

[19] Hapax legomena are words that occur exactly once in a given text. Here, words occuring between 2 and 10 times or less are called rare events, while frequent words are assumed to occur at least 80 times. These frequent words are often assumed to be stopwords, and they are often excluded in applications like information retrieval.

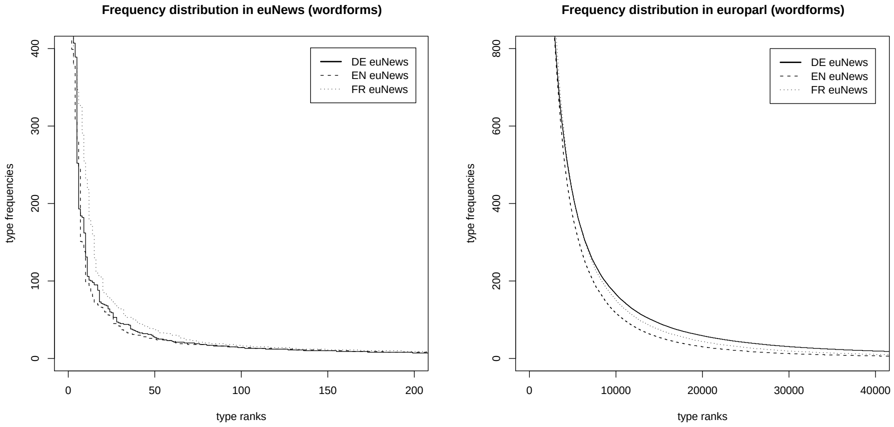


Figure 4.2: EUNEWS and EUROPARL: frequency distributions (word forms)

As can also be seen, the English and French texts have considerably smaller vocabularies than the German ones. This is probably an effect of the frequent use of compounding in German, as opposed to the intensive use of multiword sequences in the other two languages. As has been shown above (section 4.5), German compounds tend to be translated by English multiword units that often contain functional words. So while a German compound increases the number of types and tokens only once, an English multiword expression increases type and token numbers once *for each of its components*. An expression like *submission to Congress*, then, is counted as three types *submission, Congress,* and *to*, the latter being a very frequent type in the English texts[20]. So the difference between the percentages of hapax legomena in the German and English texts, as well as the differences between the frequent words in the same texts can be taken as indication that English uses more periphrastic constructions than German, the latter language using compounds instead[21] .

Still, all three languages show a Zipfian distribution, and the size of the vocabularies does not seem to matter. These frequency distributions are likely to cause misalignments: highly frequent types of the source language will occur in nearly every sentence of the corpus and will therefore co-occur with a large set of target language words, and vice versa. Rare events, on the other hand, will rarely co-occur, and as their frequencies are low, co-occurrence statistics will additionally yield unreliable results.

Accordingly, it would be advantageous to find a means that allows to align rare events correctly: aligning them correctly means aligning a large part of a corpus' vocabulary correctly, thus increasing its usefulness for lexicographic and MT purposes. Furthermore, as rare events are rare, it is highly probably that they are translated, and not paraphrased or referred to using an anaphoric expression. Moreover, there will be less translation variation, i.e. a rare event has a higher probability of having only one correct translation in the corpus than frequent words.

[20] In a 100.000 token subset of EUROPARL, e.g. *to* occurs 3,737 times, the complete multiword expression, on the other hand, is a hapax legomenon, as well as its translation *Kongreßvorlage*.

[21] This assumption needs to be tested more thoroughly: the frequency distributions of nouns e.g. could be compared to those of English noun-noun sequences. Furthermore, the English and French EUROPARL texts could be resegmented, such that multiword expressions are marked as single tokens rather than token sequences. A subsequent analysis of the frequency distributions in the resegmented texts will probably yield that the percentages of hapaxes and frequent words between the three languages will be more similar than they are now.

corpus (language)	types	hapax leg. (%)	rare events (%)	“normal´´ words (%)	frequent words (%)
EUNEWS (de)	2,110	1,192 (56.49)	777 (36.83)	124 (5.88)	18 (0.85)
EUNEWS (en)	1,933	982 (50.80)	804 (41.60)	134 (6.93)	14 (0.72)
EUNEWS (fr)	1,887	936 (49.60)	775 (41.07)	155 (8.21)	22 (1.17)
EUROPARL (de)	284,782	155,096 (54.46)	94,580 (33.21)	25,105 (8.82)	10001 (3.51)
EUROPARL (en)	60,505	23,167 (38.29)	19,622 (32.43)	10,195 (16.85)	7,521 (12.43)
EUROPARL (fr)	60,006	24,169 (40.28)	19,041 (31.73)	9,424 (15.71)	7,372 (12.29)

Table 4.21: EUNEWS and EUROPARL: vocabulary size (lemmas)

Lemmatization

An obvious strategy is to change the type frequencies of a corpus is to lemmatize it, and align lemmas instead of word forms. The idea behind this strategy is two-fold: on the one hand, lemmatization allows to find common translations for word forms that are clearly related. Aligning lemmas could e.g. yield that word forms sharing the same lemma, like *swim, swimming, swims, swam...*, all are aligned to word forms that, equally, share the same lemma, like *schwimmst, schwimmend, schwimmt, schwamm ....* This strategy comes with the additional benefit that information on morphological alternations can be learnt to a certain extent. Secondly, lemmatizing a corpus has the effect that the frequencies of at least some rare words are increased: if e.g. a word form occurring once belongs to a lemma that occurs more often, lemmatizing has the effect that this word form is attributed to the lemma, the word form frequency is subsumed in that of the lemma.

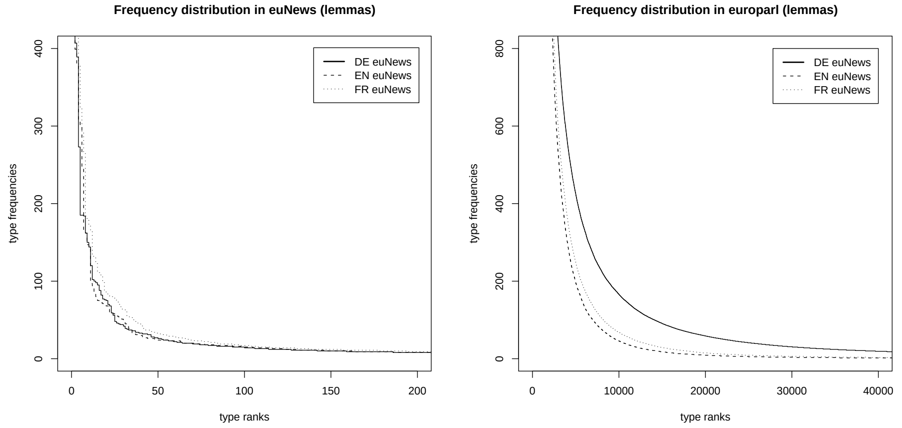


Figure 4.3: EUNEWS and EUROPARL: frequency distributions (lemma)

Unfortunately, lemmatized corpora also show a Zipfian distribution (table 4.21 and figure 4.3): Hapax legomena[22] still make up 39-56% of a corpus’ vocabulary, which is almost the same amount of hapaxes as in the unlemmatized case. Furthermore, the amounts of rare events (39-56%), normally frequent lemmas (6-15%), and frequent lemmas (1-10%) are highly similar to those in the unlemmatized cases, a well. In other words, the quality of the frequency distribution remains the same, and statistical co-occurrence measures will still give unreliable results for large parts of a corpus’ vocabulary.

The situation is slightly different if only parts of the corpus are lemmatized. If e.g. only those types occurring 5 times or less are lemmatized, then there is a chance that at least their lemmas

[22] The definitions used here are nearly the same as the ones above: Hapax legomena are lemmas with a frequency of 1, rare events occur 2-10 times, and frequent lemmas have a frequency of 80 or more.

corpus	rare types (frequency<5)	frequent lemmas (frequency>5)
EUNEWS (de)	2,305	764
EUNEWS (en)	1,912	658
EUNEWS (fr)	2,116	877
EUROPARL (de)	65,660	54,980
EUROPARL (en)	61,510	46,512
EUROPARL (fr)	71,266	58,572

Table 4.22: EUNEWS and EUROPARL: rare types and their (frequent) lemmas

occur sufficiently often for statistical analysis. This is what has been done e.g. by Dejean et al. (2003), although they could not attribute a reason for why partial lemmatization helps to overcome data sparseness. Moreover, they could not explain how to best set the frequency threshold.

The effect that partial lemmatization overcomes data sparseness better than complete lemmatization can at least partially be explained by corpus statistics: Generally, the number of lemma types in a corpus is lower than the number of word form types, which is an effect of the concept of lemmatization: if lemmas are used to subsume those word forms that have the same meaning, but differ due to inflectional morphology, then there *must* be fewer lemma types than word form types.

These considerations can be supported by empirical findings: using the corpora EUNEWS and EUROPARL, I have counted how many types occur with a frequency of 5 or below, the threshold used by Dejean et al. (2003). In addition, I counted how many of these tokens are attributed to a lemma that occurs more than five times. As can be seen (table 4.22), some rare word types can be attributed to lemmas that have higher frequencies, but not all. In fact, only 75.62% of all rare word types in the English EUROPARL belong to lemmas that are more frequent, i.e. that occur more than five times. In other words, lemma information need not necessarily yield better statistical information than word type information. Rather, lemmatization helps to overcome data sparseness for *most of the words*, especially in a large corpus such as EUROPARL.

The remainder of the words are likely ambiguous and can hence be attributed to several lemmas. In these cases, the frequency of a word type may be higher than those of its lemmas. In fact, 11 of the non-rare types (with a frequency higher than 5) in the German texts of EUROPARL, are more frequent than its lemma(s), and all of them are ambiguous. The German word *angebracht* e.g. occurred 8 times throughout the text. Half of the time, it was tagged and lemmatized as an adjective, meaning *appropriate*. The remainder of the occurrences were taken to be the present participle of the verb *anbringen*, which could be translated as either of the following, non-exhaustive list, depending on the context: *add, apply, attach, fix, install, mount*. Thus the choice of lemmas depends on the context, and need not help to overcome data sparseness.

Word Rank Correlation as Alignment Cue

So if using word co-occurrences to compute alignment is problematic, a new statistical approach to word alignment needs to be suggested. A statistical word alignment procedure that does not use word co-occurrence counts might, for example, be to align words having similar frequencies, based on the assumption that

> highly frequent words are translated by highly frequent words, and rare words by rare words.

This heuristic captures the intuition that a link between a highly frequent word and a hapax legomenon is not to be expected. Additionally, it correlates with the observation that highly frequent words tend to be function words, irrespective of the language in question. If, accordingly, the heuristic is used for alignment, it will link highly frequent function words of the source language to equally highly frequent function words of the target language, thereby avoiding the error to align function words to content words with fewer occurrences.

The heuristic can be translated into a similarity measure by using frequency rankings of the types in the source and target language texts of a parallel corpus: given the frequency rankings of the two languages, tow types are aligned if the difference between their respective ranks $rank_{w_{L1}}$ and $rank_{w_{L2}}$ is small.

Formally, the heuristic can then be defined as

$$sim(w_{L1}, w_{L2}) = 1 - \left| \frac{\text{rank}_{w_{L1}} - \text{rank}_{w_{L2}}}{\text{rank}_{w_{L1}} + \text{rank}_{w_{L2}}} \right| \tag{4.14}$$

With this heuristic, a bilingual lexicon can be induced from a parallel corpus prior to the alignment process[23]. Furthermore, no sentence alignment information is needed to induce the lexicon.

A side effect of the similarity measure is that each rare event of the source language text will appear highly similar to every rare event of the target language text, as these types do not show much difference in their frequencies. For word alignment on a sentence-aligned corpus, however, this effect is negligible: few rare words will be co-occurring within one and the same sentence pair, which means that many wrong word pairs will never be used for computing an *actual* word link.

Lexicon Induction Parameters

A problematic issue when inducing a lexicon is that the cross-product of the two vocabularies in a parallel corpus is high. A short text from EUNEWS, e.g. containing no more than 191 German and 185 English types already allows up to 35,335 different translation pairs, most of them being incorrect. Hence it is necessary to filter an automatically induced lexicon, discarding those translations pairs that have low to zero probability of being correct. One filter is, of course, to compare the automatically induced lexicon to an aligned corpus, and only keep those translation pairs that can be found in the corpus.

As an example, the text mentioned above was used to induce a lexicon which accordingly contained 191 entries, each having 185 translations, with an average confidence of 0.908. Despite the high confidence values, only 63 of these translation pairs were found to be correct. However, when using the lexicon for sentence alignment, the aligner achieves a precision of 87.5% and a recall of 48.2%. Given the huge amount of errors in the lexicon, these alignment results are surprisingly good. The reason for these good alignment result could be that the highly-frequent word pairs, as they occur more often in the corpus, are good sentence alignment cues[24].

However, when words with frequencies of 5 or less were excluded, the induced lexicon contained only 8 entries, with 8 translations each, and a much lower average confidence of 0.687. Furthermore, only 7 translation pairs were correct. Using this lexicon for sentence alignment led, accordingly, to a much lower recall (26.8%). Even precision decreased slightly to 85.0%. So while word frequencies are an important parameter in this approach, it is not necessarily true that frequently occurring word pairs constitute good word links.

If high rank similarities are correlated with the correctness of a translation pair, on the other hand, using only those translation pairs that have the highest similarity should lead to good alignment results. Based on the observation that words may be polysemous, the numbers of translation candidates per word should not be too restricted. One possibility to filter out many translation

[23] Frequency counts from paragraphs and sentences do not provide enough statistical information on types, hence the lexicon must be induced using the complete corpus.

[24] This assumption implies that if two words frequently co-occur, then they are likely to be translationally equivalent.

pairs, while allowing for translation alternatives, is to use only those translation pairs with e.g. the two highest similarity values. Or, a rank threshold may be used that filters out translation pairs if their rank difference is too big. This threshold may be empirically set or it can be based on the assumption that

> frequent words have many translation alternatives, and rare words have few.

When testing the compound alignment strategy mentioned above (section 4.5), e.g. I collected information on roughly 1100 German nouns, some of which where hapax legomena, while others occurred quite frequently. 963 of these nouns were hapax legomena, 207 nouns where dis legomena or occurred with a frequency between three and ten, and 22 were more frequent than that.

For 95 of the roughly 1100 German nouns, I found more than one possible translation in the text, i.e. they were not translated consistently with the same expression. The German word *Währungsunion* e.g. was equivalent both to *monetary union* and *currency union*. On average, each of these 95 nouns received 2.268 translations. When counting over the whole set of nouns, each noun is translated, on average, by 1.049 translations. Of course, this data sample is very biased towards hapax legomena and other rare events. Still, it is not reasonable to suppose that a word will have hundreds of different translations. Hence, the question is how to filter out incorrect translation pairs from the induced lexicon.

One possibility is by using the word ranks to dynamically set thresholds to the similarity between two words. The threshold should be defined such that it filters out few translation candidates for frequent words, and many translation candidates for rare words. This is e.g. achieved by dividing the rank number of a word $rank_w$ by the total number of ranks in a corpus.

$$\text{threshold} = 1 - \frac{\text{rank}_w}{\text{\#ranks in vocabulary}} \tag{4.15}$$

Word pairs with a similarity value below this threshold should be discarded.

Finally, word category information can also be used to distinguish between good translation pairs and erroneous ones, based on the empirical findings described above (section 4.6). Category changes are e.g. only allowed within the same category class, i.e. in order to be a translation pair, both words need to be lexical or functional. Translation pairs where one word is lexical, the other functional, are discarded. Or, no category changes were allowed at all, except in those cases when a translation pair consisted of a preposition and a determiner.

All of these parameters have been used in initial experiments on one EUNEWS text pair. As has been described above, a raw lexicon without any filters can be used successfully for sentence alignment (with a precision of 87.5% and a recall of 48.2%), although the lexicon itself contains many errors (freq-raw). Excluding rare word pairs from the lexicon yields to a decrease in sentence alignment quality (freq-5plus) with a precision of 85.0% and recall of 26.8%). A simple n-best list where only the translation pairs with the highest confidences are used already achieves a precision of 91.7% and a recall of 30.4% (1-best). Using the translation pairs with the two highest confidences results in perfect precision and a further decrease of recall down to a value of 28.0% (2-best). The dynamically set threshold causes the sentence alignment precision to be 87.5%, and the recall is 47.0% (dynamic-thresh). When filtering out translation pairs where one word belongs to a lexical category, whereas its translation does not results in a precision of 90.6% and a recall of 44.6% (lex-func). Allowing category changes only for translation pairs where one word is a prepositions and the other a determiner, finally, results in a sentence alignment precision of 89.3% and a recall of 37.5% (only prep-det). As can be seen from these results, summarised in table 4.23, restricting the lexicon size almost invariably leads to an increase in precision. However, recall will also decrease. Furthermore, there seems to be a correlation between the average confidence of the translation pairs within a lexicon and its precision.

experiment	precision (%)	recall (%)	lexicon size (translation pairs)	average confidence
freq-raw	87.5	**48.2**	35,335	0.908
freq-5plus	85.0	26.8	**64**	0.687
1-best	91.7	30.4	191	0.967
2-best	**100.0**	28.0	382	**0.971**
dynamic-thresh	87.5	47.0	34,571	0.920
lex-func	90.6	44.6	24,890	0.937
only prep-det	89.3	37.5	6,440	0.936

Table 4.23: Results of initial lexicon induction experiments

Furthermore, although these initial tests have been conducted on a very small text, word rank correlation seems to be a useful cue to induce bilingual lexica and use them for sentence alignment. Accordingly, the method is used by ATLAS to populate its system-internal dictionary with lexicon information.

Test Runs

In a second series of tests, three bilingual lexica were induced in order to sentence align the whole German–English EUNEWS corpus, and the different parameters were partially combined. In the first of these German–English lexica, words with a frequency of 5 or less were excluded. In addition, the second lexicon only contained those translation pairs with the two highest confidences. Thirdly, only the rarest words, with a frequency of 3 or less, were discarded from the lexicon, and no other filter was used.

Then, four lexica were induced for the German–French EUNEWS: the first lexicon again contained only those words with a frequency higher than 5, while the second additionally disallowed word category changes. For the last two German–French lexica, words with a frequency of 3 or less were excluded. This frequency restriction was combined first with the requirement that only the two best confidences were to be used, and secondly word category changes were forbidden.

As can be seen in table 4.24, the overall results were not as good as in the initial experiments: the sentence alignment precision was relatively low with values between 59.3% and 67.4%, and recall ranged between 29.6% and 32.5%. Including rare words into the lexicon almost inevitably resulted in an increase in precision, again. Furthermore, a combination of parameters does not necessarily result in an increase in alignment quality. Additionally, the inclusion of rare events in the dictionary does not seem to give better results for every language pair. Quite contrarily, the

language pair	condition	precision (%)	recall (%)
German–English	frequency >5	60.4	29.8
German–English	frequency >5, 2-best	59.3	29.6
German–English	frequency >3	**65.5**	**32.5**
German–French	frequency >5	**67.4**	**31.9**
German–French	frequency >5, POS	66.9	31.2
German–French	frequency >3, 2-best	67.0	31.5
German–French	frequency >3, POS	66.2	**31.9**

Table 4.24: Test results for lexical alignment

```
<item>
  <lemma>Beschluß</lemma>
  <category>noun</category>
  <language>German</language>
     <translations>
        <translation>
        <lemma>decision</lemma>
        <category>noun</category>
        <language>English</language>
        <confidence>0.97495</confidence>
     </translation>
  </translations>
</item>
```

Figure 4.4: Example of a correct lexicon entry

experiments carried out for the language pair German–French indicates that excluding rare words yields an increase in alignment quality.

One *filtered* lexicon per language pair, including only those word pairs that had been used to compute the sentence alignment, was also examined: The lexicon that lead to the best alignment quality on the English–German texts contained 482 different entries[25]. Most of these entries (328 entries, i.e. 68.05%), contained at least one correct translation. The numbers of translation candidates per head word varied considerably: Prepositions e.g. tended to have many translation candidates, including a large variety of prepositions from the other language. Nouns, on the other hand, had fewer translation candidates, and some of the lexicon entries were perfect, like German *Beschluß* being linked to its English equivalent *decision*.

Other lexical entries suggested that using word category information for filtering is important if the lexicon is to be used for word alignment[26]. Given these insights, then, the dictionary quality is not just high enough to be used for the sentence alignment task, but may also be used to align the words within sentence pairs.

The smallest filtered lexicon for German–French[27] contained only 277 lexicon entries, each with approximately 6 translations. 181 of these translation pairs were actually correct, hence there is reason to suppose that the dictionary is also useful for aligning at the word level.

4.8 Phrase Alignment

Another useful source of information for aligning corpora is, obviously, syntactic annotation. Firstly, because phrase alignments are useful for finding translation equivalents for multiword units, whether they are idioms, support verb constructions, or any other kind of collocations.

Secondly, unlike most previous attempts at phrase alignments, where recurring word sequences are taken to be units that should be aligned as such, syntactic annotations offer a very systematic approach to those multiword units that occur within syntactic constituents. Furthermore, syntactic

[25] Words occuring more than three times in the corpus had been included during the lexicon induction. Additionally, only those translation pairs with the two highest similarity values were included

[26] Due to the relatively high number of translation pairs in the lexicon, I have decided to give only insights into the quality of the induced, and filtered lexicon.

[27] It had been induced using all words that occurred more than three times, and using those translation alternatives with the two highest confidences per entry

information offers a way out of the problem of data sparseness: some multiword units may escape statistical phrase alignment approaches by virtue of the multiwords being rare, but the same is not necessarily true when they are syntactically annotated beforehand. Rather, a noun phrase e.g. is a noun phrase, and can be parsed as such, irrespective of the exact words used within it[28]. In other words, data sparseness is avoided because the phrase alignment is done without using the lexical material that the phrases contain.

Some researchers (cf. Cherry and Lin 2003; Gildea 2003) hence have chosen to use syntactic information, mostly generated by dependency parsers, to guide the alignments, and have achieved some success. However, these approaches have focused on finding *identical* structures of some kind, e.g. hypothesising that a functional relation like predicate-argument structure is carried over during translation from one language to the other. This may be true for most cases, especially for closely related languages, but even then, changes in word categories, lexical gaps, or simply different predicate-argument structures associated with the same concept will lead such a sophisticated alignment procedure off the right track.

The question then is, if it is not safe to rely too much in structural isomorphism, then how else can syntactic information aid the alignment? As an answer, I argue that there is more to syntactic information than its function. First of all, syntactic boundaries restrict the search space to some extent, in that translation equivalents should not wildly cross constituent boundaries. Rather, it is relatively safe to assume, and this assumption is used within *all* previous approaches to phrase alignment, that

if two phrases are linked, then the words that they contain must be linked within these two phrases.

Discontinuous multiword units like the so often cited German verbs and their particles may seem to contradict this assumption, but then, the question is simply which phrases to use for restricting the search space[29].

Secondly, phrases do not only have a *function* within a sentence, but they also have a *type* (or syntactic category), such as them being noun phrases, verb phrases, or prepositional phrases. Hence phrase alignment might exploit this type information, assuming that

a source language phrase of a specific type *t*, e.g. of type *noun phrase*, will have a translation equivalent in the target language that has the same type *t*.

This assumption is, of course, closely related to the assumptions behind phrase alignment approaches that use dependency annotations[30]. In these approaches,

a source language phrase with a specific function *f*, is taken to have a translation equivalent in the target language that has the same type *f*.

Lastly, any syntactic constituent can be seen as conveying bits of information that are closely tied together, using a very specific combination of concepts. The translation of the syntactic constituent then somehow has to convey the same combination of concepts, i.e. in terms of conceptual *complexity*, a constituent and its translation will be highly similar. Of course, conceptual complexity is hard to formalize. Moreover, this assumption need not hold for unrelated language pairs.

[28] Of course, parsers and chunkers may not cover every possible construction, nor are their vocabularies large enough to contain information on each and every possible word of a language. Still, they are better suited to detect constituents than mere word sequence statistics.

[29] In the worst case, the appropriate phrase boundaries will be those of the sentence, a sentence being just another type of linguistic phrase.

[30] As ATLAS currently does not support functional information, this alignment cue cannot currently be tested.

However as has been observed before, the complexity of an expression *is* a clue for the alignment task: as I have shown above (section 4.5), there is a high correlation between the *morphological* complexity of German nominals and their English, or even Swedish translations. Hence there is reason to suppose that a very similar correlation can be found for syntactic constituents. Accordingly, my hypothesis is that

A source language phrase of containing *n* elements will be translated by a target language phrase also containing *n* elements.

Depending on the languages and their grammars, this hypothesis is rather extreme, and variations in phrase length need to be accounted for. Additionally, the *depth* of a constituent may play a role, i.e. the number of embedded nonterminal nodes within a phrase may also serve as alignment clue. For initial experiments, however, I have tested only the most basic assumptions, namely that

1. phrase boundaries restrict the search space for word alignments,
2. source language phrases have the same type as their translation equivalents,
3. phrases that are translationally equivalent consist of roughly the same number of words.

The first of these assumptions need not be formalised within ATLAS at all: given syntactic corpus annotation and hypotheses on phrase alignments, the alignment disambiguation will rule out any links that violate this assumption.

Secondly, using phrase types as alignment clues is captured relatively straightforwardly with an alignment strategy that aligns phrases only if they have the same type. Phrases are aligned if they have the same type, e.g. if they are both noun phrases, and their similarity

$$sim_{restrictive} = \frac{1}{\text{\# occurrences of phrase type in L1 } + \text{\# occurrences of phrase type in L2}} \tag{4.16}$$

is defined in terms of the frequencies of that specific type within the respective sentence link. In order to allow structural divergences, it is possible to align phrases irrespective of their types, but giving lower similarities to phrase hypotheses where the phrase types do not match,

$$sim_{non-restrictive} = \frac{1}{\#phrases_{L1} + \#phrases_{L2}} \tag{4.17}$$

taking the numbers of *all* phrases of a sentence link into account.

Secondly, I define phrase similarities in terms of their complexity, counted as numbers of words contained within them. This similarity score,

$$sim = 1 - \left| \frac{\text{\# words in } phrase_{L1}}{\text{\# words in } sentence_{L1}} - \frac{\text{\# words in } phrase_{L2}}{\text{\# words in } sentence_{L2}} \right| \tag{4.18}$$

attributes a greater similarity to those phrase links where both are equivalent in length.

Finally, it is always possible to combine both definitions of phrase similarities, i.e. those phrases receive a high similarity score that have equivalent length *and* are of the same type.

These considerations are all based on empirical findings on words and their translation equivalents, rather than on empirical findings on the alignment of phrases. However, I rely on syntactic constituents sharing some properties with their head, i.e. if nominals have the same morphological complexity as their translations, then it is highly probable that the same is true for noun phrases. Furthermore, if word category changes are rare, then it is equally improbable that phrases change their types during translation. However, empirical tests are due in order to see whether phrase alignment as defined here works well.

Test Runs Accordingly, I have conducted three tests: Firstly, an alignment module is used that aligns phrases based on their complexity, or, in other words, based on their lengths counted in words. In a second experiment, phrases are aligned based on their types. Finally, both approaches to phrase alignment are combined.

This test has been carried out exclusively on the German–English LITERATURE corpus. More specifically, it has been tested on one short story of the corpus. Roughly 1200 phrases were aligned in each of these test runs. The best performing similarity measure used both the types of phrases and their lengths as alignment clues. In this test run, 375 of the phrases (30.74%) were aligned correctly. A major source of error (55.08%) was the remaining ambiguity: in most cases, especially noun and prepositional phrases occur in abundance within every sentence link so that neither phrase complexity nor phrase type provide sufficient clues to compute correct phrase hypotheses. The remaining errors occurred due to preprocessing errors, or because the modules aligned phrases with differing types.

The worst results were achieved using only phrase complexity as a clue (table 4.25). Here, only 15.58% of all phrases were aligned correctly, the major error source being mismatched phrase categories (51.28%). When phrases were preferred if their categories match, 22.91% of all phrases were aligned correctly, and the major error source was the ambiguity between phrases of the same type, but with differing lengths (62.73%). Surprisingly, errors of the POS-tagger and chunker only accounted for a small portion of the errors (on average around 3%).

German–English	precision
phrase complexity	15.59
phrase type	22.91
phrase type & complexity	30.74

Table 4.25: Test results for the phrase alignment

So, the available information on phrases is not sufficient to achieve a good alignment quality, basically because of the large amounts of ambiguity involved. This situation, however, can easily be remedied either by incorporation dependency information or by running the phrase alignment module in parallel with at least some word alignment module.

4.9 Inheritance

As ATLAS can simultaneously compute alignments for sentences and paragraphs as well as for words and phrases, it is possible to use *inheritance* as an alignment clue: If two linguistic items A and B are aligned, then it is possible to derive further alignment information for

- those smaller linguistic items that *are contained in* a and b, and
- those larger linguistic items that *contain* a and b.

If two sentences A and B e.g. have already been aligned by the system, it is reasonable to assume that all words contained within sentence A have to be aligned with some words in sentence B – or not at all –, thereby restricting the search space during the word alignment task. This is the underlying assumption of all those word alignment strategies that rely on a prior sentence alignment, although the inheritance assumption is usually not stated explicitly. As ATLAS uses sentence and paragraph hypotheses to compute word alignments, this type of inheritance is used throughout the alignment process.

In cognate- or dictionary-based sentence alignment algorithms, inheritance is used in the opposite direction: if two words A and B have been aligned, then the information can be used to align the sentences that A and B occur in, thus deriving new sentence alignment information. This strategy is well-known from the sentence alignment task, where cognateness or dictionary lookup are used to compute a rough word alignment, and derive sentence alignment information from it. I have implemented this type of inheritance as a separate alignment strategy that use hypotheses on words, phrases, or sentences to derive hypotheses about a next-higher types of linguistic items, i.e. a word or phrase hypothesis is used to derive a sentence hypothesis, and a sentence hypothesis is used to derive one covering paragraphs.

The confidence value of the "parent" is likewise inherited, without any further change. I have considered changing the confidence value, e.g. decreasing it to 90% of its original value, based on the assumption that even inheritance can go wrong. However, I have not found any sound reason to assume that if one hypothesis is correct, its "heir" might not. Accordingly, confidence values are inherited without any further modification.

As the success of the inheritance strategy crucially depends on the reliability of its input, it is hard to test whether the strategy yields good results. Hence I have *not* tested it separately from any other module. Instead, I use it as background information throughout the alignment process, and I will watch for inheritance effects during the full-fledged evaluation described in chapter 5.

4.10 Merge of Overlapping Hypotheses

Generally, an alignment strategy will find clues for 1:1 links, i.e. it is able to align one corpus item like a sentence with another. However, it will hardly find sufficient clues to propose other types of links, like deletions, insertions, or links where one word or sentences corresponds to two or more words (or phrases). These link types are usually predicted using probabilistic parameters: in sentence alignment tasks, the probabilities estimated on the UBS corpus data, by Gale and Church 1991b, have been re-used for the purpose of inserting n:m links (cf. Simard et al. 1992), or the *fertility* of a word, i.e. its ability to participate in n:m links, is estimated during the training of statistical translation models (Brown et al. 1993).

Here, I take another, more rule-based approach that has some advantages over probabilistic approaches: it is *not restricted to any specific type of n:m links*, i.e. unlike the approaches that re-use the probabilities of Gale and Church (1991b), it is not restricted to links involving at most two corpus items in either language. Secondly, it generates n:m links if a corpus item of the source language shows *positive evidence* that it should be aligned to more than two items in the target language, only. Furthermore, the resulting, merged hypothesis may be *discontinuous*.

Here, two alignment hypotheses are *merged* if they overlap, i.e. if at least one corpus item is covered by two (or more), competing alignment hypotheses: in set-theoretic terms, each alignment hypothesis is a set of corpus positions that can be intersected with another. If the intersection is non-empty, i.e. if at least one corpus position is an element of both sets, then there is evidence that the union between the two (or more) sets should be computed.

As a result, this merging operation generates alignment hypotheses of *any* type, depending on the link type and number of overlapping hypotheses. In the simplest case, it will combine two 1:1 alignment hypotheses to generate a new 1:2 or 2:1 hypothesis. If more alignment hypotheses are intersected, 2:2, 3:1, 1:3, or indeed any other link type can be generated. The same holds for the case when a 2:1 hypothesis is merged with a 1:1 alignment hypothesis, etc.

However, there are also disadvantages: as the strategy depends on positive information, it cannot generate deletion or insertion, as the only indication of a deletion or insertion is negative, i.e. that there is *no evidence* that a particular corpus position should be aligned. Furthermore, if too many overlapping alignment hypotheses exist, the strategy will generate alignment hypotheses that are too large. In the worst case, it will even generate an alignment hypothesis that aligns all words of a corpus with each other[31].

Fortunately, the disadvantages can be held at bay by a careful use of the strategy. If it is used locally, i.e. to merge only those hypotheses generated by a specific alignment module, it will merge only a few hypotheses, and there will be few clues, if any, to generate larger links than 2:2. Currently, the module is only used to merge the hypotheses generated by one of the two cognate-based sentence alignment modules.

4.11 Process of Elimination

As a final alignment module, I have implemented the *process of elimination* strategy: it aligns a source language item A and a target language item B if there are no other alternatives left, i.e. it generates the link ($a \leftrightarrow b$).

Unlike any other module, its success hence depends on the amount of previously generated alignment hypotheses: the more alignment hypotheses have already been generated, the fewer corpus items are still unaligned, and the fewer alignment alternatives exist per corpus item. In other words, if only one source and one target language sentence (or word) are left unaligned, then they are aligned in order to close the gap in the alignment[32].

In any other case, there is not enough indication to align based on the process of elimination, and hence all corpus items that are still unaligned are aligned to null.

As there is no possibility to estimate how reliable the module aligns, I have set the confidence value of each *process of elimination* hypothesis to 0.1. This value is low enough to indicate the low reliability of the module, while still being above zero. Based on further experiments with the full-fledged alignment system, I will be able to estimate how often the *process of elimination* module generates reliable hypotheses, and use this knowledge to revise the way in which this module attributes confidence values to its hypotheses.

Finally, to maximise the usefulness of the module, it is activated after any other alignment module has generated alignment hypotheses *and* after the alignment disambiguation. The reason for calling the module as very last alignment strategy is that during the alignment disambiguation, implausible hypotheses are *discarded* which means that the alignment may contain gaps that should be closed.

Additionally, using the *process of elimination* as a clue after the alignment disambiguation has the advantage that disambiguated alignment information is available to guide the process by imposing another restriction: no alignment hypothesis of this strategy may violate the coherence of the overall alignment.

Summed up, and slightly paraphrased, the *process of elimination* strategy is used as a fall-back option that ensures that each and every sentence, word, etc. of the source language is aligned to *some* expression in the target language, provided the target language expression is in an equivalent structural position. In all other cases, null-links are added to ensure that a minimum extent of explicit alignment information exists for every sentence, word, etc.

4.12 Tuning of the Alignment Strategies

So far, I have tested all alignment strategies in single test runs, i.e. there was no interaction between the different modules. Now, in order to correctly set the reliability factors, all modules that produce

[31] Which amounts to saying that all source language words in the parallel corpus are translated by all target language words in the corpus, or, much shorter, that the source language text is translated by its target language text. This assumption is the starting point of the alignment process, i.e. such an alignment hypothesis is less than useful.

[32] Informally, the alignment cue can be put as *nobody likes you two, so maybe you like each other.*

a specific kind of alignment hypotheses are run in parallel, and the results are evaluated.

Paragraph and Sentence Alignment I have rerun the paragraph alignment, using *all* implemented strategies, in order to fine-tune ATLAS and set the reliability factors. Due to the differences between the two language pairs, it was necessary to make the reliability factors for paragraph alignment dependent on the language-pair: for English–German, no preference between the two paragraph alignment modules is necessary. Here, the interaction between both modules causes an increase in precision and recall to 58.9% and 67.5%, respectively. The case is different for French–German. The best results are achieved if the length-based hypotheses are preferred over those based on linear ordering. In this case, precision can increase up to 53.6%, and recall is now 54.0%.

How much the reliability factors, and the interaction between different modules, affect the alignment quality can be seen nicely with the sentence alignment modules: assuming they are equally reliable yields a precision of 50.3% and a recall of 55.2% on the German–English EUNEWS. However, when ranking the most reliable module, the resource-dependent cognate module, best, precision increases to 56.3%, and recall increases, too, to a value of 60.8%. Ranking the resource-dependent cognate module slightly higher than the modules using sentence length and linear ordering also yields an increase in precision (57.9%) and recall (62.3%). Unfortunately, using reliability factors alone cannot increase the system's performance up to 100%. Finally, precision reaches its maximum at 58.4%, with a recall of 62.8%. The same reliability factors are optimal for the language pair French–German. However, the best possible precision seems to be 54.4% for this language pair, and the best recall is 58.3%.

One observation from these test runs is that modules that make *no* use of linguistic information perform considerably worse than those that do: both cognate-based strategies are more reliable than both the module using sentence length and that using linear ordering, irrespective of the language pair in question. Thus I assume that sentence alignment using dictionary information, although it depends on the size and quality of the dictionary in question, is as reliable as the cognate-based modules, hence it receives the same reliability factor as the best of the cognate modules (table 4.26)[33].

Unfortunately, setting the reliability factors manually is tiresome, and nothing guarantees that the final setting is the best possible. For this reason, it would be good to experiment with machine learning methods, and see in how far these methods can achieve an optimal parameter setting.

Phrase and Word alignment Concerning phrase and word alignment, a test run on the German–English EUNEWS, using all modules except the lexicon induction, already shows promising results: most lexicon entries contain at least one correct translation, and 30% of the resulting dictionary are correct. This percentage is bound to increase if an additional lexicon, whether pre-existing or automatically induced, is used.

The dictionary generated as an result of the alignment process reveals the strengths and weaknesses of the word alignment modules more detailed: especially the nominal alignment works well, including the correct alignment of multiword units like *research infrastructure*.

Even category changes, even if caused by wrong POS-tagging, do not hinder the use of correct word hypotheses. Especially with respect to German and English participles, or when accounting for German truncated nouns within elliptic constructions.

However, the dictionary also reveals that there are considerable problems with function words and multiword units that are not nominals. In these cases, the dictionary contains many wrong or incomplete translation pairs. This can be addressed nicely by exploring the errors more thoroughly

[33] If two reliability factors are given, then the first is valid for the English–German texts. The other is used when aligning French–German.

module	generates	precision/recall	languages	reliability
length-based	paragraphs	36-44%/36-50%	independent	1
linearity	paragraphs	35-60%/31-67%	independent	1/0.01
all paragraph modules	**paragraphs**	**53.6-58.9%/54-67.5%**	**independent**	
length-based	sentences	36-44%/36-50%	independent	0.01
linearity	sentences	35-60%/31-67%	independent	1/0.001
cognates 1	words & sentences	79-81%/49-55%	independent	0.1
cognates 2 (with resources)	words & sentences	84-89%/37-44%	POS information	1
dictionary	words & sentences	–	bilingual dictionary	1
all sentence modules	**words, sentences**	**54.4-58.4%/58.3-62.8%**	**various**	
nominals 1 (length)	words	70.25 %(31.21%)[34]	POS information	1
nominals 2 (morph.)	words	54.67% (36.16%)	morphology	0.1
word category	words	18.95% (26.92 %)	POS information	0.001
lexicon induction	words, sentences	59.3-67% / 29.6-31%	variable	0.1
phrase	phrases	30.74%	chunks/parses	0.001
inheritance	sentences, paragraphs	–	independent	1
merge	sentences, paragraphs	–	independent	1
process of elimination	all levels	–	independent	1
all word modules	**words, sentences**	**30%**	**various**	

Table 4.26: Interactive use of the implemented alignment strategies

```
<item>
   <lemma>Forschungseinrichtung</lemma>
   <category>noun</category>
   <language>German</language>
   <translations>
      <translation>
         <lemma>research infrastructure</lemma>
         <category>multiword</category>
         <language>English</language>
         <confidence>0.00646</confidence>
      </translation>
   </translations>
</item>
<item>
   <lemma>Forschungsgebäuden</lemma>
   <category>noun</category>
   <language>German</language>
   <translations>
      <translation>
         <lemma>laboratory</lemma>
         <category>noun</category>
         <language>English</language>
         <confidence>0.03834</confidence>
      </translation>
   </translations>
</item>
```

Table 4.27: EUNEWS lexicon after alignment: examples for noun entries

```
<item>
   <lemma>mitfinanziert</lemma>
   <category>adjective</category>
   <language>German</language>
   <translations>
      <translation>
         <lemma>part-financed</lemma>
         <category>verb</category>
         <language>English</language>
         <confidence>0.00328</confidence>
      </translation>
   </translations>
</item>
[...]
<item>
   <lemma>Therapie-</lemma>
   <category>component</category>
   <language>German</language>
   <translations>
      <translation>
         <lemma>therapeutic</lemma>
         <category>adjective</category>
         <language>English</language>
         <confidence>0.00101</confidence>
      </translation>
   </translations>
</item>
```

Table 4.28: EUNEWS lexicon after alignment: examples for category changes

and developing alignment modules that deal with specific kinds of alignment problems, e.g. with the alignment of verbal constructions or functional words like prepositions and determiners. However, the data is hard to examine without the additional support of a word-aligned gold standard, hence the reliability factors are tentative, and will have to be tuned more thoroughly.

4.13 Summary

In sum, thirteen different alignment modules are implemented, using a variety of alignment clues. Some of them, as the length-based paragraph and sentence alignment modules, and as the lexicon induction procedure, are language and annotation independent. Others use different kinds of corpus annotation, among them POS-tags, information on morphological composition, lemmas and syntactic constituency.

All of the strategies have various draw-backs: some of them achieve high-precision alignments at the cost of coverage. Others provide a large amount of hypotheses, but with a high degree of errors. However, by simultaneously using different alignment modules, and ranking them according to their reliabilities, performance improvements can be achieved.

Furthermore, the majority of alignment modules are language-pair independent, i.e. they either assume no corpus annotation, or only specific kinds of them, but they are not tailored to only a very specific language pair. The only exception to this is the alignment of nominal compounds which has so far been only tested and developed for the language pairs German–Swedish and German–English (table 4.26).

Chapter 5

Evaluation of Text Alignment Quality

> It is well known that manually performing a word alignment is a complicated and ambiguous task.
>
> (Och and Ney 2003, p. 33)

> These rates indicate that the gold standard is reasonably reliable and that the task is reasonably easy to replicate.
>
> (Melamed 1998b, p. 12)

Following practice in other areas of NLP, the performance of alignment systems is evaluated against a reference corpus that is, ideally, perfectly aligned. All evaluations are accompanied by at least a sketchy annotation scheme that describes how the annotation of the reference data came about. One such scheme exists for the evaluation of sentence alignment quality, and there are at least three annotation guidelines for word alignment, namely the guideline developed in the ARCADE project, the guideline of the BLINKER project (Melamed 1998a) and the PLUG annotation guidelines (Merkel 1999). The most common metrics to evaluate alignment quality are precision and recall for sentence alignment, and *alignment error rate* that is supposed to measure word alignment quality.

All approaches to measuring alignment quality encounter certain problems: text units, whether sentences or words, need to be aligned based on their translational equivalence, i.e. on their *meaning the same* irrespective of the languages used. However, this *translational equivalence* is usually not formally defined. A formal definition is not needed for the sentence alignment task as humans can determine whether two sentences are translations of each other without formal instructions than *align two sentences if they mean the same*. However, precise instructions are needed to manually perform a word alignment. Human annotators need to be told exactly when translational equivalence holds: does it hold only when two expressions mean the same in *all possible contexts?* Or is it context-dependent, i.e. two expressions are equivalent if, *in the given context,* they mean the same?

Another problem when measuring alignment quality is that translation pairs may consist of more than one text unit per language. A sentence in a source language, e.g. may be translated as two sentences in the target language. Other n:m links, where n, m ≥ 1 one are also possible. In these cases, an evaluation should not just take into account whether an automatically computed alignment is exactly as given in the reference. Instead, a match might be partial, and hence its partial correctness need to be measured somehow. Different evaluation metrics have been suggested to address this issue and measure alignment quality adequately. Additionally, different annotation schemes and reference data sets have been designed.

In the following, I will review the standard approach to evaluate NLP-applications in general (section 5.1) and alignment quality in particular (section 5.2). Then, I redefine precision and recall so that alignment quality can be assessed appropriately, and that furthermore, the evaluation facilitates error analyses (section 5.3). Finally, I evaluate ATLAS (section 5.4).

5.1 Formal Evaluation Requirements

In a quantitative evaluation, the output of an NLP-application is typically evaluated against some reference data, the so-called gold standard. It has been annotated with reference annotations that show what kind of output the application should produce. The reference annotation has been made with respect to an annotation scheme or guideline tailored to the specific application at hand. Finally, the evaluation results are measured using one or more well-defined metrics. Ideally, an evaluation is also accompanied by a detailed analysis of the errors of the evaluated system.

The need for a gold standard is obvious – it is not possible to measure a system's performance without comparing it to an ideal standard. For the same reason, an annotation scheme is required that defines what exactly is considered ideal. Last, but not least, an evaluation metric has to measure to what degree an NLP-system falls short of the ideal.

5.1.1 The Gold Standard

As the gold standard should show what, ideally, the application should produce, its annotations should be as unambiguous and consistent as possible. The data should also exceed a certain size, i.e. a gold standard of a few sentences, or a few sentence pairs, is not sufficient to determine how well a system performs. However, it is unclear how big the gold standard should be. It may be a collection of test cases of varying difficulty, each carefully chosen to test the system's performance with respect to a certain phenomenon. These test cases should also allow assessing the overall *real-life* performance of a system, i.e. the relation between *easy* and *tricky* test cases should be roughly the same as if the system was encountering *normal*, i.e. non-gold standard data. Or, the gold standard is a large random sample of *normal* data, without including specifically chosen test items. This approach implies that not every type of problem occurs in the gold standard, but the larger the gold standard is, the more frequent problems will be included, and hence missing problems are not an obstacle. Usually, gold standards consist of such random samples, as it is comparatively hard to collect error types and estimate the relation between difficult and easy data.

In order to avoid inconsistencies, gold standards should always be annotated at least twice, with an annotation-final resolution step where differences between the two annotations are resolved. This procedure enforces a relatively consistent annotation, and simultaneously enforces the removal of all those annotation errors, where one of the annotators happened to make a mistake. After the difference resolution step, consistency checks might be in order to detect where both annotators happened to have made the same annotation mistakes.

5.1.2 The Annotation Guideline

While the gold standard is important in itself, so that no evaluation can be conducted without at least *some* kind of reference data, an annotation guideline is a tool rather than a requirement. It defines how the reference data should be *annotated*, i.e. it defines which structures, whether alignment links, syntactic trees or POS-tag sequences, are considered to be correct. As such, it is tightly connected to the *purpose* of the evaluation.

If the alignment information is needed e.g. for lexicographic purposes, a good alignment of nouns, verbs, and adjectives, in short, of words from lexical classes, may be sufficient. While the set of function words in a language is typically finite, there is an infinite amount of open class words i.e. lexicon acquisition for open class words is imperative. But if the aligned data is used for training a MT system, then the mere alignment of lexical words is not sufficient. Finally, research in translation studies may require other or more fine-grained alignment information.

To a certain degree, the *inter-annotator agreement* describes the quality of the annotation guideline, but also of the task. High inter-annotator agreement indicates that either the annotation guidelines provided the annotators with enough instructions and examples to annotate with high reliability and correctness, or that the task was sufficiently easy to do. Usually, both is true. A good annotation guideline simplifies the task of the annotators so that the gold standard is consistent and relatively error-free. Low inter-annotator agreements indicate, on the other hand, that the task was not sufficiently clear to annotate consistently or correctly.

The inter-annotator agreement is usually calculated using the *kappa statistic*: it is used to correct for chance agreement, i.e. it indicates the reliability of the annotation. The kappa statistic is given by

$$\kappa = \frac{p_0 - p_c}{1 - p_c} \tag{5.1}$$

where p_0 is the number of times the two annotators agreed, and p_c the number of times that the agreement between the two annotators is expected by chance (Carletta 1996).

Unfortunately, it is hard to estimate p_c for the alignment task. When using the kappa statistic, inter-annotator agreement is essentially defined as a classification task that classifies a given set of *objects* into several categories. If the number of objects is unknown, as in the alignment task, then kappa cannot be used. Accordingly, the Dice-coefficient has been used (cf. Melamed (1998b, Véronis and Langlais (2000)) for alignment inter-annotator agreement, or inter-annotator agreement is measured by the amounts of overlap or mismatches between annotations. Bojar and Prokopová (2006) compute how many links are made by both annotators, and how many and which kinds of mismatches occurred. Unfortunately, this inter-annotator agreement does not give information whether the "matches" between the two annotators occurred by chance or not. (Kruijff-Korbayová, Chvátalová, and Postolache 2006) compute the *intra-annotator agreement* as the intersection between two annotations A_1 and A_2, divided by the union of the two[1].

5.1.3 The Evaluation Metrics

Finally, evaluation metrics show the degree to which a system makes mistakes, and they typically relate these error rates to *coverage*, i.e. whether a system computes structures for a large amount of the gold standard data or not. Additionally, the metrics allow *comparing* across systems, i.e. to determine which system performs best in a given task.

In order to define such metrics, evaluation judgements are usually based on the following four categories:

- true positives (tp): the number of data instances considered correct by the system, and also correct according to the gold standard

- true negatives (tn): the number of data instances considered incorrect by the system, and incorrect according to the gold standard

[1] The authors also compute the kappa statistic (Kruijff-Korbayová, Chvátalová, and Postolache 2006), but they do not give information as to how they estimated p_c.

- false positives (fp): the number of data instances considered correct by the system, but wrong according to the gold standard
- false negatives (fn): the number of data instances considered wrong by the system, but correct according to the gold standard

The best-known evaluation metrics, precision and recall, relate these categories in terms of errors (false positives and false negatives) and correct data instances (true positives and true negatives), to overall quality and quantity of a system's output. Precision indicates to which degree the automatic results are correct, and recall shows how much of the gold standard data is covered in the automatic results.

Precision is defined as the number of true positives divided by the sum of *all* data instances produced by a system

$$\text{precision} = \frac{\text{tp}}{\text{tp} + \text{fp}} \tag{5.2}$$

or, alternatively, as the percentage of correct instances in the system's output.

Recall, on the other hand, describes the coverage of a system, i.e.

$$\text{recall} = \frac{\text{tp}}{\text{tp} + \text{fn}} \tag{5.3}$$

the number of correct data instances divided by all data instances of the gold standard. The *F-measure*,

$$\text{F-measure} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{recall} + \text{precision}} \tag{5.4}$$

finally, is a harmonic combination of precision of recall[2].

In alignment tasks, the gold standard, i.e. a manually created or revised reference alignment consisting of a sequence of alignment links, is compared to an automatically generated alignment. Thus precision would be defined as

$$\text{precision} = \frac{\text{number of alignment links correctly found}}{\text{all alignment links found}} \tag{5.6}$$

and recall

$$\text{recall} = \frac{\text{number of alignment links correctly found}}{\text{all gold standard alignment links}} \tag{5.7}$$

with the F-measure definition remaining unchanged. However, these definitions have been shown to be problematic when used for evaluating alignment systems with respect to n:m links: in these cases, the decision on whether a link is correct or incorrect ceases to be binary. Instead, an alignment link may be partially aligned, i.e. some of its parts are correctly aligned, while others are either aligned incorrectly, or are even left unaligned. Precision and recall as given above, however, do not take partial correctness into account and thus give skewed evaluation results.

[2] The F-measure is in fact defined as

$$F = (1 + \alpha^2) \cdot (\text{precision} \cdot \text{recall}) / (\alpha^2 \cdot \text{precision} + \text{recall}) \tag{5.5}$$

but with $\alpha = 0.5$ the formula can be simplified to the version given above.

5.2 Former Approaches to Alignment Evaluation

5.2.1 ARCADE I: Evaluating Sentence Alignment

The issue of measuring alignment quality was first tackled in the 4-year ARCADE project (Langlais et al. 1998; Véronis and Langlais 2000) of AUPELF-UREF, a network of mostly French-speaking universities. The project consisted of two different phases, the first being devoted to setting up a gold standard and the evaluation of sentence alignment quality, the second taking first steps towards measuring word alignment quality.

The Gold Standard that was created in the ARCADE project is a parallel, English-French corpus collection that contains the *BAF* and a subset of the *JOC* corpus. The French-English subset of *JOC* that was used in the ARCADE project consists of 287,000 English and 245467 French words of political texts. The *BAF*, on the other hand, is an corpus collection of *institutional*, *scientific*, *technical*, and *literary* texts. Altogether, the ARCADE corpus consists of roughly 827 000 French and 716 000 English tokens. Most of the parallel data consists of 1:1 links (88%), but omissions (4%) and n:m links (8%) occur, too.

Annotation Guideline The reference alignment was created semi-automatically: an alignment system computed an initial sentence alignment that was subsequently checked and corrected by two human annotators. The annotation differences between the alignments of the two humans were resolved. In cases where the order of the sentences of source and target language differed, the annotation was done so that no *crossing links* occurred. Unfortunately, no information on the inter-annotator agreement is given.

Evaluation Metrics As evaluation metrics, precision and recall were adapted to the task and defined as

$$\text{precision} = \frac{\text{number of alignment links correctly found}}{\text{all alignment links found}} \tag{5.8}$$

and

$$\text{recall} = \frac{\text{number of alignment links correctly found}}{\text{number of reference alignment links}} \tag{5.9}$$

Additionally, the F-measure was computed.

Precision and recall, as defined above, do not take partially correct alignment links into account: if a system fails to align an n:m link in exactly the same way as given in the gold standard, this is counted as an error. Accordingly, such an error lowers precision and recall disproportionately. Given the high numbers of 1:1 links that are typical for sentence alignment, the problem seems negligible.

Nevertheless, as a workaround, the metrics were used to evaluate at finer levels of granularity: in addition to counting how many links were correct in a system's output, precision and recall were computed to show how many sentences, words, or even characters were correctly contained in a link. As a result, precision and recall were computed four times each for each system alignment, and the evaluation results were compared. Generally, evaluating a system by the degree to which characters are parts of correct links yielded the highest precision and recall values. The researchers argue hence that at this level, segmentation errors are irrelevant and accordingly do not have negative impact on the evaluation results (Langlais et al. 1998; Véronis and Langlais 2000).

Unfortunately, while computing precision and recall seems reasonable to assess the overall alignment quality, it does not facilitate the gaining of insights into the strengths and weaknesses of the system. If the evaluation directly assessed how many sentence links were computed correctly, it would also yield the information which links were erroneous and thus could help to qualitatively analyze why errors occurred. Alignment errors due to sentence segmentation failures, e.g. could lead to the development or use of better segmentation tools.

Moreover, the researchers report a strong correlation between the evaluation results computed at different segment levels (alignment link, sentence, word, character). Hence tokenization errors are relatively irrelevant for the evaluation results, thus there is no reason to compute precision and recall at the character level.

5.2.2 ARCADE I and II: Translation Spotting

The second phase of the ARCADE I alignment evaluation project was dedicated to the evaluation of word alignment systems on a French-English gold standard. However, the researchers noted that evaluating word alignment quality presented much more difficulties. The reason for the difficulties encountered in ARCADE is the underlying assumption that the translation relations found in the parallel corpora should be context-independent, i.e. two words are translations of each other irrespective of where and how they are used. However, context independence is not given for each translation pair within a sentence. Instead, most translation pairs convey the same meaning only with respect to a specific context, i.e. they are context-dependent.

As a result, the evaluation of word alignment quality was restricted to *translation spotting*, i.e. to evaluate in how far a word alignment system is able to correctly determine *context-independent* translation pairs. Context-dependent translation pairs are not part of the gold standard annotation and hence no full text alignment was evaluated (Véronis and Langlais 2000).

ARCADE II is the follow-up project of ARCADE I, this time aiming at evaluating word and sentence alignment quality in a multilingual setting. Thus the gold standards consist of multiple language pairs, including both western European and other languages. Additionally, the word alignment track has been restricted to the alignment of named entities, i.e. names of persons, organizations, locations etc. To a certain degree, this task is a subtask of *translation spotting*.

Gold Standard In the second part of the ARCADE I project, the gold standard did not comprise of texts of various genres. Instead, the *JOC* corpus was chosen as test corpus.

For the evaluation, the researchers decided on using those 60 French words that were also being used in the ROMANSEVAL word sense disambiguation task (Véronis 998a): they had been chosen for the word sense disambiguation task because they occurred in at least 60 different contexts each in the JOC corpus, and were judged polysemous. Twenty of each of these words were nouns, verbs and adjectives.

The advantage of reusing these 60 frequent, polysemous words obviously is that the statistical alignment systems had enough data to compute correct links for them. A failure to align these words would have indicated serious weak points of an alignment system. Additionally, as the target words were polysemous, a certain degree of translation variation had to be expected, and hence aligning the target words *correctly* to their translations would be a considerable success.

An annotation guideline was developed, and all occurrences of the 60 words chosen for the gold standard were aligned manually by two annotators. Inter-annotator agreement, computed as the Dice-coefficient,

$$\text{inter-annotator agreement} = 2 \cdot \frac{\text{Number of common words}}{\text{Total number of words for both annotators}} \tag{5.10}$$

ranged between 0.84 and 0.98, with the lower agreement rates achieved for verbs. Véronis and Langlais (2000) additionally report that many omissions occurred, and that many words were parts of multiword units.

For ARCADE II, the gold standard was extended to cover parallel texts French, English, German, Italian and Spanish, each language corpus consisting of roughly 1 million tokens. As a second gold standard, the news corpus MD, containing texts from *Le Monde diplomatique*, was aligned. It consists of 150 parallel texts in French and Arabic (roughly 500 000 tokens per language), with additional texts in Russian, Chinese, Japanese, Greek and Persian (appr. 170-240 000 tokens per language). A subset of the corpus was annotated with *named entity alignment information* (3,639 French and 2,924 Arabic named entities).

Annotation Guideline The ARCADE I annotators were instructed to align given target words so that the links annotated were as small as possible, but also as large as necessary to achieve translational equivalence (Véronis 1998). The annotators were allowed to use n:m links, but alignment at the sub-word level was forbidden. Additionally, information on the type of the alignment had to be included in the annotation along with a reliability judgement and optionally, a comment on the type of annotation problem encountered.

If a word could not be aligned to a proper translation, the annotation has to include a reason for the divergence. This could be

- an omission
- an anaphor or other referring expression
- a spelling error in the translation
- ellipsis (carte de credit ou de paiement ↔ credit-card or pay-card)
- a paraphrase

These instructions highlight that the focus was on *context-independent translational equivalence*, as an alignment that included a referring expression or other context-dependent translations had to be marked as divergent.

The annotation guideline gives further, more detailed instructions regarding the treatment of determiners, genitive constructions, prepositions, separable verb particles, relative pronouns, passives and auxiliary verbs, if relevant for the alignment of nouns, adjectives, and verbs. These guidelines are again tailored to the translation spotting task, i.e. on aligning items with their context-independent translations.

No formal annotation guideline is mentioned in ARCADE II that describes how the actual alignment should be done. Rather, the authors (Chiao et al. 2006) argue that what constitutes a good alignment is not sufficiently well-defined. Therefore, they restricted themselves to annotating named entities using monolingual guidelines. Unfortunately, it is not clear whether the named entities were only monolingually annotated or also *aligned*.

Evaluation Metrics In ARCADE I, precision and recall were adapted to the word alignment task: precision was defined as

$$\text{precision} = \frac{|\text{words correctly aligned}|}{|\text{all words aligned}|} \tag{5.11}$$

and recall analogously as

$$\text{recall} = \frac{|\text{words correctly aligned}|}{|\text{all gold standard words}|} \tag{5.12}$$

They were computed for each occurrence of a gold standard word, but not with respect to a single, unambiguous gold standard. Rather, they were computed with respect to those gold standard annotations that had the most overlap with a system's output. As a result, a system's performance is judged in a manner most favourable to the system. However, as the different systems are evaluated against different, however slightly, versions of the gold standard, comparability between the systems is lost. The overall recall and precision values were computed as the average of the single-occurrence evaluation values.

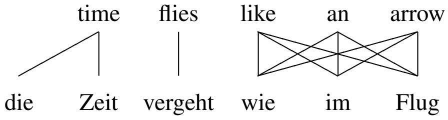


Figure 5.1: Toy gold standard

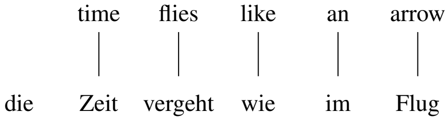


Figure 5.2: Toy automatic alignment

These metric definitions seem to severely punish n:m links. If a gold standard sentence consisted for instance of mostly n:m links as in the toy gold standard in figure 5.1, then an automatic alignment that fails to reproduce these n:m links (figure 5.2) exactly would receive disproportionately low precision and recall values. In this case, the system would have linked three words correctly (flies, Zeit, vergeht) out of 10 linked words, as opposed to 11 words that are linked according to the gold standard. Thus, it would score a precision of 33%

$$\text{precision} = \frac{|\text{3 correctly aligned words}|}{|\text{10 aligned words}|} = 0.33 \tag{5.13}$$

and recall would be 27%

$$\text{recall} = \frac{|\text{3 correctly aligned words}|}{|\text{11 aligned gold standard words}|} = 0.27 \tag{5.14}$$

resulting in an F-measure of 29%. It is intuitively clear that the automatic alignment quality is low. Still, these precision and recall values are too low and especially precision favours gaps in the alignment (words that are not aligned do not count as errors).

5.2.3 BLINKER: Creation of a Gold Standard

In the BLINKER project, an English-French aligned corpus was created to facilitate the development and testing of translation lexicons, statistical translation models, and word sense disambiguation methods, but also to allow contrastive linguistic research (Melamed 1998b). However, as it was not part of an evaluation exercise, a discussion of evaluation metrics or results is not given.

Gold Standard The corpus is a subset of two English and French Bible versions that were available online. This subset was chosen to contain all those verses, and their translation, in which

100 target words occur. These 100 words where chosen so that 25% each are hapax legomena, dis legomena, words occurring three times, and words with a frequency of four. In total, the gold standard consists of 250 verse pairs with 7510 English and 8191 French word tokens.

All types within the target set, however, are single word expressions, with a large amount of proper names and common nouns. Some target words are verbs, adjectives, or even adverbs, but clearly no function words like determiners or prepositions were chosen.

The annotation of all words in these 250 verses was done by seven annotators that aligned independently of each other, and each verse was annotated five times. After the annotation, the alignments were compared to rate the *inter-annotator agreement*: it is computed using a set-theoretic equivalent D(X,Y) of the Dice-coefficient. The inter-annotator agreement ranges between roughly between 80 and 85%, which is considerably lower than the agreement rates achieved in the ARCADE translation spotting task (section 5.2.2). However, if computed for content words (nouns, adjectives and verbs) only, the inter-annotator agreement is comparable to those achieved in the ARCADE word alignment track, ranging between roughly 88 and 92%.
Melamed (1998b) identifies three reasons for the low overall agreement rates, namely that

1. although parallel, the two Bible versions are not translations of each other,
2. the annotation guideline was based only on a small text sample
3. the annotation tool was equally improvable.

Annotation Guideline The BLINKER annotation guideline (Melamed 1998a) was developed in three steps: a preliminary version was created and used to align ten verse pairs. Subsequently, the variations in the test annotations were used to revise the guidelines. Afterwards, the revised guidelines were used for annotating the BLINKER gold standard. In case that annotators reported further alignment problems, these were resolved by discussions among the annotators, and the annotation guidelines remained unchanged.

Melamed (1998a)'s annotation guideline clearly deviates from the *translation spotting* task: if two expressions in source and target language mean the same in a specific context, then they have to be aligned (Melamed 1998a). Special attention is drawn to the fact that the two texts may not be translations of each other, but translations from a third text, and hence that the annotators should not focus on the question *is X a translation of Y* but *are X and Y both translations of an unknown Z*, i.e. do X and Y convey the same meaning.

Additionally, and again unlike the guidelines given in ARCADE, links should be as detailed as possible, even if this includes aligning below the word level. Idiomatic expressions, however, must be aligned as n:m links, and a comparable strategy is used for aligning pronouns to their translations: a pronoun should be aligned to the expression that refers to the same entity, whether that expression is a pronoun itself, or not. Resumptive pronouns should be aligned to the translation of their antecedent, as well. The same strategy is followed with *conjunctive non-parallelism*, as (Melamed 1998a) calls it, if a coordination is more explicit in the one than in the other language (figure 5.3). Other non-parallelisms, involving punctuation, are treated similarly.

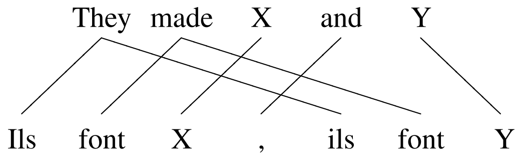


Figure 5.3: Example taken from the BLINKER annotation guideline, p. 12

Rules are also given for handling so-called extra determiners, i.e. determiners that need not be translated due to the grammatical structure of the target language, and possessives.

Both determiners and possessives are to be aligned to those parts of the translations that function in the same way. For determiners, this means that they may be aligned to non-function words as in figure 5.4 where the determiner is needed in English to denote the singular. In the French equivalent sentence, however, the singular is implicitly encoded in the noun *fermier*. In these cases, the English determiner has to be aligned to *fermier*. Interestingly, the annotation guideline includes a rule for handling situations when the translation of an active sentence is in passive voice. In these cases, the predicates are not to be aligned as wholes, but have to be torn apart as much as possible (figure 5.5). As a result, the annotation is unsatisfactory. The auxiliaries *ai* and *was* have significantly different functions as they convey different voices. Additionally, they are not literal translations of each other. Accordingly, it would have been better to treat both verb groups as multiword units (figure 5.6). Another interesting aspect of the annotation guideline is the treatment of bare NP versus PP. If the translation of a prepositional phrase is a bare nominal phrase, i.e. the preposition is missing in the translation, then the two phrases should be aligned as whole, i.e. both the preposition and the noun of the source language are aligned to the noun in the target language (figure 5.7). From a functional point of view, this annotation decision is justified in that both phrases, whether NP(*Moses*) or PP (*à Moise*) are equivalent and hence should be aligned as such. However, there are alternatives that can take the structural differences between the two languages into account. The structural difference could lead to the hypothesis that the two phrases are objects, i.e. that the PP is subcategorised by the verb *prescrire*. Incorporating this information into the alignment would yield a structure as in figure 5.8. Thus, subcategorization information is visible and the noun *Moise* is linked to its translation *Moses*. However, this annotation decision is rather risky from a linguistic point of view. No clear guidelines or even intuitions may exist to distinguish objects from adjuncts reliably. Accordingly, subcategorization information should only

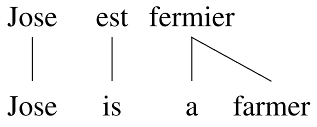


Figure 5.4: Example taken from the BLINKER annotation guideline, p. 18

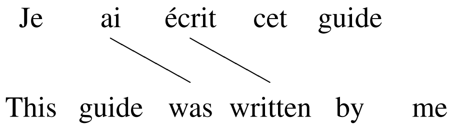


Figure 5.5: Example taken from the BLINKER annotation guide, p. 15

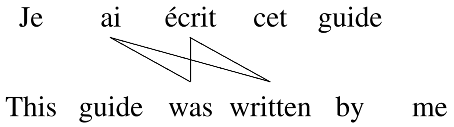


Figure 5.6: Example taken from the BLINKER annotation guideline, p. 15

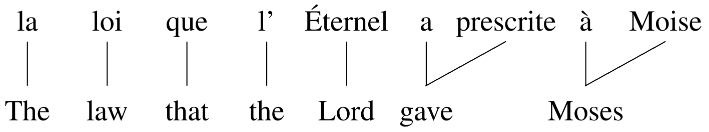

la loi que l' Éternel a prescrite à Moise
The law that the Lord gave Moses

Figure 5.7: Example taken from the BLINKER annotation guideline, p. 16

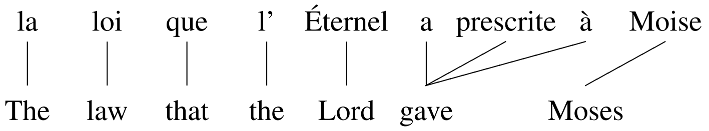

la loi que l' Éternel a prescrite à Moise
The law that the Lord gave Moses

Figure 5.8: Example: incorporating subcategorization information into the annotation

be added to the alignment annotation, if distinctions between adjuncts and objects are clear-cut and can be made explicit, and even testable, in the annotation guidelines.

The second alternative makes a distinction between context-dependent and context-independent translational equivalence. Context-dependent translational equivalence holds between the french PP *à Moise* and the English NP *Moses*, while context-independent translational equivalence holds only for the French NP *Moise* that is argument of the PP, and the English NP *Moses*. If preference is given to context-independent translational equivalence, the annotation guideline has to be adjusted so that whenever there is a choice between aligning context-independent and -dependent equivalent translations, then the context-independent alternative should be preferred. In the case of *Moses*, the annotation would have to treat the preposition as a deletion, and to align the context-independently equivalent nouns *Moses* and *Moise* (figure 5.9). However a preference of strong

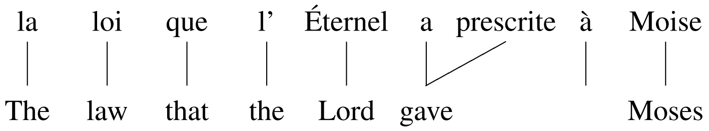

la loi que l' Éternel a prescrite à Moise
The law that the Lord gave Moses

Figure 5.9: Example: Marking context-independent translational equivalence in a specific case

translational equivalence has severe implications for the whole annotation: it is difficult to decide when to prefer which type of equivalence and may soon lead to a restriction of the alignment to *translation spotting*. A compromise would be to align at several levels of linguistic abstraction: on the word level, context-independent translational equivalence would be enforced, while allowing context-dependent correspondences at the phrase levels. An example solution to the *Moses* problem would align the words *Moses* and *Moise* as they are context-independently equivalent, as well as the NPs *Moses* and *Moise* (figure 5.10). Context-dependent translational equivalence, then, would add the alignment of the PP *à Moise* and the NP *Moses*. However, it may be hard to keep track which links have which type. As long as phrases are not aligned, however, the decision has to be made whether to follow Melamed (1998a)'s example and align based on functional equivalence, or to prefer context-independently equivalent links in specific contexts.

Although the annotation guideline is quite specific and well-designed, it does not give good advice for dealing with multiword expressions in general. Frequently, multiword expressions which have an obvious equivalent in the other language are only partially aligned, as e.g. in the example in figure 5.11 where the focused nominal *c'est lui* and *qui* should both be aligned to their equivalent *He*, as in figure 5.12.

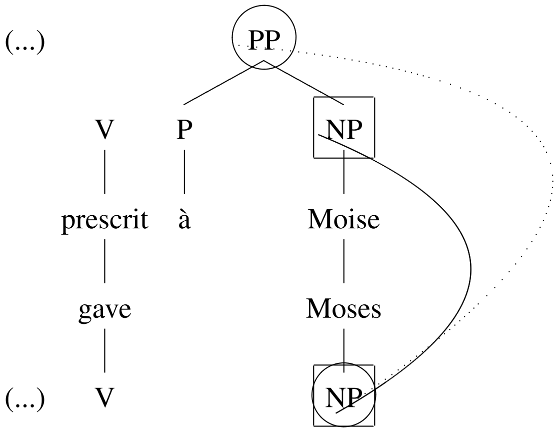


Figure 5.10: Example alignment with various levels and types of translational equivalence. Square boxes and lines indicate context-independent translational equivalence. Circles and dotted lines are used for context-dependence translational equivalence.

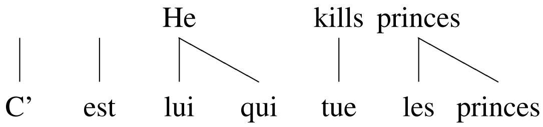


Figure 5.11: Example taken from the BLINKER annotation guideline, p. 11

Other examples include the above-mentioned treatment of prepositional objects and the case of "divergent" prepositions, i.e. cases where quasi idiomatically the choice of prepositions differs largely between two languages.

In general, the guidelines are well-designed and easily adaptable to language pairs other than English-French. Unlike the *translation spotting* task, it focuses on complete text alignment and context-dependent translational equivalence, i.e. expressions have to be aligned if they mean the same in a specific context. Additionally, the alignment is supposed to be function-sensitive in that functional elements have to be aligned to those entities that observe the same function, whether the translations themselves are functional elements or not. The best examples for this approach to word alignment are the rules for aligning pronouns, and for aligning prepositional objects. However, the treatment of multiword expressions is inconsistent as some examples given above highlight. The reasons for this inconsistency are probably the limited data set that was basis for the annotation guideline, and the lack of a theory on multiword expressions.

5.2.4 The PLUG Approach: Taking Partial Links into Account

In the word alignment project PLUG – *Parallel corpora in Linköping, Uppsala and Göteborg* of the three Swedish universities Linköping, Uppsala, and Göteborg, the researchers developed the PLUG *Link Annotator* together with the PLUG Link Scorer, in order to facilitate both manual annotation of the gold standard and comparing the gold standard to an automatically derived word alignment. Although both tools have been developed for evaluations done within the PLUG project, they can be reused for word alignment evaluation in other contexts.

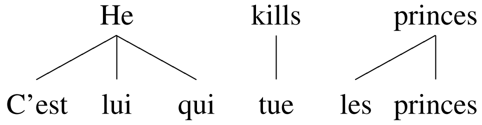


Figure 5.12: Example: alignment of focused French NP to English pronoun

Gold Standard Within PLUG, the distinction is made between *textual links* and *lexical links*, the latter being what one would expect to find in a dictionary, while the former is the type of links found in actual texts, and which may, due to none too literal translations or paraphrases, differ considerably from dictionary information. As PLUG deals with the alignment of parallel corpora, the focus is on textual links. However, the distinction between *regular* and *fuzzy* links serves to indicate where it is easy to derive lexical links – this would be the regular case – and where the correspondence deviates from regularity for some reason.

As in ARCADE, word alignment evaluation is done as *translation spotting*, i.e. not all words in the reference set are aligned, but only a previously defined subset. According to Ahrenberg et al. (1999), 500 tokens were randomly sampled from each subcorpus used in PLUG[3] to create the gold standards for the subcorpora. However, no information is give whether these tokens are instances of frequent or infrequent types, or how many of these tokens are content or function words.

The gold standards have been annotated by four different annotators, where two of them worked in a team and annotated jointly, while the other two annotated independently of each other. The inter-annotator agreement that Ahrenberg et al. (1999) report ranges between 89.8% and 95.4% which is similar to the agreement rates reported in ARCADE (Véronis and Langlais 2000). However, the authors do not state how they computed the inter-annotator agreement. Additionally, the differences between the three different annotation versions were not resolved to create a single, consistent gold standard. Rather, the authors chose to use one annotation by "the most experienced annotator" (Ahrenberg et al. 1999, p. 4), and to ignore the rest. No information is given how the most experienced annotator was chosen, nor why his annotation was preferred.

Intuitively, it seems plausible to use the annotation created by the most experienced person. However, the person may have developed annotation practices different from what the annotation guideline demanded, hence the gold standard might contain deviations from the guidelines that could have been avoided.

Annotation Guideline The annotation guidelines closely follow the example set by the ARCADE guidelines in many ways: the links have to be as large as necessary and as small as possible to ensure symmetric translational equivalence and they are not concerned with full text alignment, but with partial alignment done for a small set of types. Unlike in ARCADE, the PLUG guidelines include many rules and test for annotating omissions, multiword expressions, verbal constructions etc. English verbal constructions with "to", e.g. are considered a multiword expression and have to be aligned as such, as in the Swedish-English example in figure 5.13. All links have to be are typed. They may be "regular", i.e. an expression has exactly the same meaning as its translation. Alternatively, a link can be "fuzzy", because there is a shift in meaning, a paraphrase, or some other deviation from the norm, Or, a link may be "null", i.e. a token is not translated at all.

Many rules are not concerned with how to align specific tokens or idioms, but whether these links should be considered "regular" or "fuzzy". Inflectional differences are for instance considered minor and do not cause an alignment to be "fuzzy" whereas category changes are always causing a "fuzzy" alignment.

[3] A sample of only 100 tokens was taken from the smallest subcorpora (Ahrenberg et al. 1999).

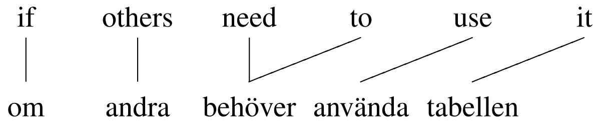


Figure 5.13: Example taken from the PLUG annotation guideline, p. 14, and adapted to this presentation style.

In sum, although the PLUG annotation guidelines are very detailed and provide the annotators with many tests and examples for deciding on the correct alignment, they cannot easily be reused for full text alignment. The reason for this restriction is that most details are much more concerned with the typing of links than with the way the links *themselves* should be created.

Evaluation Metrics As evaluation metrics, precision, recall and the F-measure are used. However, Merkel et al. (2002) define their evaluation metrics to account for *partial matches*, i.e. links that overlap with links contained in the gold standard, but fail to be identical. In order to weigh partial matches differently from fully correct ones, (Merkel et al. 2002) make use of two different precision measures: Precision I,

$$\text{Precision I} = \frac{\#\text{ correct links found} + \#\text{ partial links found}}{\#\text{ all links found}} \tag{5.15}$$

is calculated as the sum of all correct and partial links, divided by the number of all links found. This means that partial links count as correct, although they do not fully overlap with the links given in the reference set. Defined this way, precision generously ignores gaps in the alignment. Within the second precision measure,

$$\text{Precision II} = \frac{\#\text{ correct links found} + (0.5 \cdot \#\text{ partial links found})}{\#\text{ all links found}} \tag{5.16}$$

the word aligner is punished for having generated partial links – the number of partial links is divided by two before added to the number of complete links. Although the strategy of giving less weight to partial links than to complete ones seems plausible, it is unclear why they divided by two and not some other factor.

The definition of recall,

$$\text{Recall} = \frac{\#\text{ all links found}}{\#\text{ all reference links}} \tag{5.17}$$

as reported in Merkel et al. (2002), however, deviates significantly from what has been used in other approaches[4]. This deviation results in the metric no longer showing to which extent the system is able to reproduce the reference alignment. Instead, even *alignment errors* are used to increase this value.

Another formula for computing recall is given in (Ahrenberg et al. 1999) where recall

$$Recall = \frac{\#\text{ incorrect links} + \#\text{ correct links} + \#\text{ partial links}}{\#\text{ incorrect links} + \#\text{ correct links} + \#\text{ partial links} + \#\text{ links missed by the system}}$$

[4] The original formula in (Merkel et al. 2002) is given as

$$\text{Recall} = \frac{\#\text{ all tried links found}}{\#\text{ all reference links}} \tag{5.18}$$

where the *tried links* are the sum of partial, correct, and all other links computed by the system. As this is the sum of all links produced by the system, I abbreviate the definition here for clarity reasons.

is computed as the number of all links found, divided by the sum of system links that were incorrect, partial, correct, and missing in the system's output. However, this recall definition looks even more dubious than the first. According to this formula, recall seems to be meant as the sum of all links that the system produced, divided by the sum of all links that the system produced *plus* the number of links that the system missed, rather than the number of partial or fully correct links produced by the system, divided by the number of all links in the gold standard. Either the formula given by Ahrenberg et al. (1999), and subsequently by Ahrenberg et al. (2000) is misleading, and the authors have the same definition in mind that was used in (Merkel et al. 2002). Or, the authors really compute the union of the links in the gold standard and the links produced by the alignment system[5]. Displayed schematically, the set of links that is used to compute recall is the union of gold standard and system links as in figure 5.14.

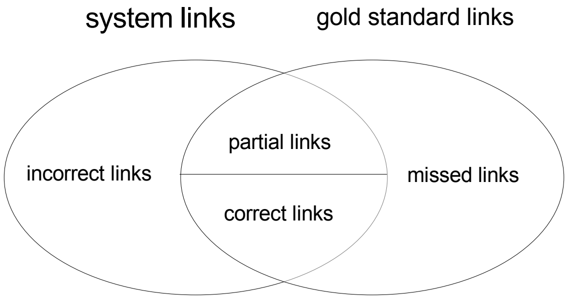


Figure 5.14: Factor of the PLUG recall definition

As the numbers of correct, incorrect, partial and missed links reported in Ahrenberg et al. (1999) add up exactly to the number of tokens in the gold standard, the quality of the gold standard can be doubted. However, given the careful design of the gold standards, I suppose that the description of recall as given in (Ahrenberg et al. 1999) and (Ahrenberg et al. 2000) is misleading.

A recall definition that corresponds to the two precision formulae given above would be

$$\text{Recall I} = \frac{\text{\# correct links found} + \text{\# partial links found}}{\text{\# all reference links}} \tag{5.19}$$

or, if partial matches are again weighted, the following recall definition would have to be used:

$$\text{Recall II} = \frac{\text{\# correct links found} + (0.5 \cdot \text{\# partial links found})}{\text{\# all reference links}} \tag{5.20}$$

These recall definitions take partial matches into account. However, *alignment errors* are not considered as true positives. Using precision and recall I on the toy gold standard and automatic alignment given above (figures 5.1 and 5.2) clearly shows that they yield good results despite insufficient alignment quality.

[5] Making sure that there are no doubles, i.e. that no correct gold standard link is contained in this union that is partially represented by one of the system's partial links.

Precision I would be 100%

$$\text{Precision I} = \frac{\text{1 correct link found} + \text{2 partial links found}}{\text{all 3 links found}} = 1 \tag{5.21}$$

as would recall and F-measure.

$$\text{Recall I} = \frac{\text{1 correct links found} + \text{1 partial links found}}{\text{all 3 reference links}} = 1 \tag{5.22}$$

Note that as the metrics are designed to work for n:m links, too, the toy gold standard would consist of three links, namely a 1:2 link (time ↔ die Zeit), a 1:1 link (flies ↔ vergeht) and a 3:3 link (like an arrow ↔ wie im Flug). Thus, the automatic alignment would consist of a correct link (flies ↔ vergeht) and two partial links (time ↔ die Zeit and like an arrow ↔ wie im Flug).

Precision and recall II would reduce the positive impact of the partial links somewhat, and both would be 67%, as would the F-measure.

$$\text{Precision II} = \frac{\text{1 correct links found} + (0.5 \cdot \text{2 partial links found})}{\text{all 3 links found}} = 0.67 \tag{5.23}$$

$$\text{Recall II} = \frac{\text{1 correct links found} + (0.5 \cdot \text{2 partial links found})}{\text{all 3 reference links}} = 0.67 \tag{5.24}$$

However, while decreasing the impact of the partial links is certainly called for, it seems somehow arbitrary to set the weight to 0.5.

5.2.5 The Standard Exercise: Precision, Recall and Alignment Error Rate

Franz Josef Och and Hermann Ney (2003, 2000a and 2000b) suggest the evaluation metric *alignment error rate*, that has subsequently become accepted and reused as a standard for assessing word alignment quality.

Gold Standard Their German-English evaluation corpus consists of 1.4% or 354 sentences of their training corpus, the VERBMOBIL corpus, which contains roughly 34 500 sentences in German and English. The 354 reference sentences, the authors report, correspond to 3109 German and 3233 English tokens. No information is given on whether the number of 354 test sentences is the overall number of sentences in the test set, irrespective of the language, or whether the test set actually consists of 354 sentence links, or 354 English sentences aligned to their German translations, or vice versa[6].

Additionally, not all 354 sentences are used for evaluation purposes, as the authors use the first 100 sentences of the reference alignment for parameter optimization (Och and Ney 2003, Section 6, p35). As this fragmentation of the reference corpus leaves only 254 sentences for evaluation purposes, it would be interesting to know the corresponding token and type numbers of the English and German sentences. However, this corpus information is not given.

With respect to the English-French evaluation corpus HANSARDS, the same considerations hold: it is unclear whether the evaluation corpus consists of 500 sentences per language, or if it consists of 250 sentence per language.

[6] The table format, however, indicates that the evaluation corpus consists of overall 354 sentences, i.e. 177 sentences per language.

The reference set was manually aligned by two human annotators independently of each other, and after the annotation, the alignments were merged to create the reference set. As annotation help, the annotators were instructed to indicate how reliable they judged their alignments are: The annotation included whether they judged a word link to be *sure* (S) or merely *possible* (P), the latter remark being used to indicate alignment uncertainties within idiomatic expressions or translational paraphrases[7]. The annotators can use n:m links, but apart from the instruction to type the alignment with respect to their reliability, no further help for annotators such as an annotation guideline is mentioned. Hence it is reasonable to suppose that the annotators did not receive any more specific instruction. This absence of annotation guidelines indicates that little importance is attached to the consistency and – linguistic or translational – adequacy of the annotation.

After the annotation, the annotators are presented with their so-called *mutual errors*, i.e. each annotator is shown the parts of his annotation that differ from what his colleague has annotated, and he is asked *to improve the alignment if possible* (Och and Ney 2003, section 5, p34). The authors do obviously not insist on fully resolving the annotation differences. Additionally, it is unclear whether the two annotators are revising their annotations together and hence can discuss the differences and resolve them so that annotation errors are corrected and the resulting reference annotation is unambiguous. Or, whether the annotators have been asked to revise their annotations independently of each other, so that it is possible that even after error correction, *mutual errors* remain in the gold standard[8]. After this error correction, the final set of *sure* links is computed as the intersection of what the two annotators considered sure, with the set of *possible* links being the union of what the annotators considered possible. Additionally, the *sure* links constitute a subset of the *possible* links ($S \subseteq$ P).

The slackness with which the authors seem to handle the difference resolution has severe consequences for the quality of the resulting gold standard: even after the annotators have revised their alignments, differences can remain. Additionally, the merging of the two manual annotations introduces further ambiguity and, possibly, even gaps into the gold standard: If the set of *sure* links is the intersection of what the annotators considered *sure*, do the links where the annotators disagreed with respect to this reliability typing remain in the corpus, but typed as *possible*, or are they simply left out of the gold standard? If the set of *possible* links is the union of what the two annotators think *possible*, then it is possible that the gold standard contains *possible* links where the annotators *disagreed.* As a result of this annotation procedure, the gold standard may contain links that contradict each other. As no information on the inter-annotator agreement is given, it is not even possible to have an intuitive expectation how much the quality of the reference alignment is affected by the annotation union.

Other researchers (Lambert et al. 2005) argue that including *possible* links and ambiguities in the gold standard are necessary as they consider the task was highly ambiguous and hence the resolution of annotation differences difficult. Furthermore, the authors argue that ambiguities between annotations should remain in the gold standard as they *could make sense*. Finally, high-recall applications might profit from ambiguities in the gold standard. However, even Lambert et al. (2005) note that ambiguities in the gold standard should be avoided as the comparability between evaluation results suffers.

Annotation Guideline Apart from the instruction to judge links either sure or possible, no annotation guideline is mentioned.

[7] An obvious question to a linguist or lexicographer is at this point, whether it makes actually sense to align words *within* an idiomatic expression, or whether it would actually be more important to identify such an idiomatic expression, aligning it, without further attempts at refining the alignment.

[8] This would be the case if annotator A edits his annotation to match what annotator B did, while annotator B revises his annotation in favour of the original annotation of his colleague A.

Evaluation Metrics For the evaluation, Och and Ney (2000a) use

$$\text{precision} = \frac{|A \cap P|}{|A|} \tag{5.25}$$

here given in set-theoretic terms[9], and a redefinition of

$$\text{recall} = \frac{|A \cap S|}{|S|} \tag{5.26}$$

along with a new evaluation metric, the *alignment error rate*. Precision is defined as the percentage of word links that occur both in the reference alignment and in the automatically generated alignment A, irrespective of whether it is to be considered *sure* or merely *possible*. Recall, however describes only the percentage of *sure* word links contained both in the gold standard and in the automatically computed alignment, i.e. recall is only computed for those links that have been easy to align manually. The difficult links, rated *possible* by the annotators, are left out of this part of the evaluation. Additionally, no particular attention is paid to partial matches between an automatic alignment and the gold standard. As the annotation may include n:m links, and as hence partial matches are possible, it is likely that the evaluation metrics are too coarse in that they will punish partial matches as errors, despite their being at least partially correct. Or, precision and recall are computed using 1:1 links only, and that each n:m word in the automatic alignment and in the gold standard has been transformed into a number of 1:1 alignment with overlapping source or target words.

However, this evaluation approach gives skewed results: Precision is calculated using the full reference corpus, including *possible* errors, and recall is computed for the *non-difficult*, i.e. *sure* subset of the alignment, only. Without knowing the amounts of *sure* and *possible* links in the annotation, it is not possible to tell whether the recall value reflects coverage of a large part of the corpus, or whether recall is computed using e.g. only 20% of the corpus annotation. Corpus characteristics given by Lambert et al. (2005) indicate that the French–English evaluation corpus consists of 23% sure and 77% possible links (roughly 4000 sure vs 20,000 possible links). Thus 100% recall would indicate that overall, only a fifth of the corpus has been aligned as sure, and thus recall values are only relevant for this corpus subset.

This approach may seem wise given that no annotation guideline is used that gives clear instructions on how to deal with problem cases, and as it is furthermore difficult to find *automatic* means for aligning these problem cases. However, if the percentage of *sure* links in a gold standard is too small, then recall ceases to indicate alignment quality.

Additionally, the reliability typing of the gold standard annotation in combination with the evaluation metrics may have serious side effects. Uninstructed annotators might type links too liberally as *possible* even if they were quite sure of their annotations. As an effect, the gold standard annotation may not meet the expectation to be consistent and reliable.

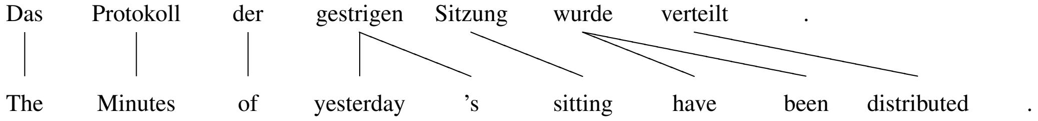


Figure 5.15: Example gold standard with sure links

In the example gold standard in figure 5.15, all links might be considered sure. A word alignment system that performs well on the 1:1 links, but that cannot align n:m links, might then align

[9] Note that $S \subseteq P$. $|A \cap P|$ is the number of *true positives*, and $|A|$ is the number of all data instances found.

the sentence pair as in figure 5.16 and would hence achieve 55% precision and 71% recall. These results would have to be considered encouraging rather than good. However, if the gold standard distinguished between sure and possible links, then the two annotators could have arrived at typing the genitive construction *possible*. Either the two annotators could have been uncertain how to align the genitive construction. Or, while one had decided to align (gestrigen ↔ yesterday, 's) as a 1:2 link, the other might have decided to align only (gestrigen ↔ yesterday), and leave the "'s" unlinked. By computing the union between the two annotation, both alternatives would be included in the gold standard, i.e. the reference data would be ambiguous.

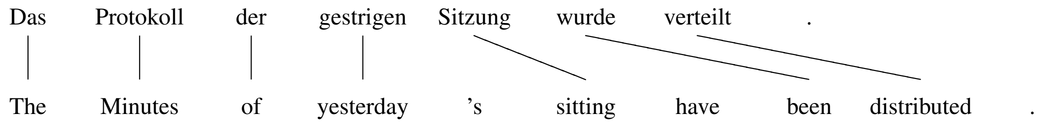


Figure 5.16: Example automatic standard with only sure links

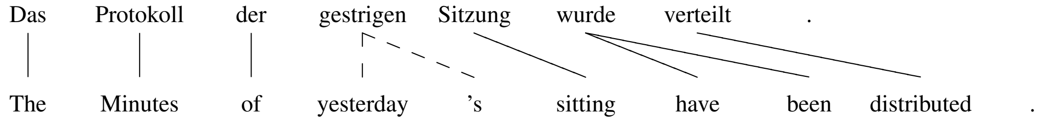


Figure 5.17: Example gold standard with sure and possible links. Possible links are denoted with dashed lines.

Evaluated against this gold standard, the system would show considerable improvement and score 77% precision and 100% recall. Note that nothing was done to improve the performance of the system, the effect is due only to the difference in gold standard annotation. The effect is more pronounce the more links are judged *possible* rather than sure.

This example is not quite an exaggeration. For comparison's sake, it is likely that the idiomatic expression of the previously mentioned toy gold standard (figure 5.1, repeated here as 5.18) would contain several possible links: while the links (time ↔ die Zeit) and (flies ↔ vergeht) are comparatively easy to align, the annotator(s) might have had a hard time figuring out how to link the remainder of the sentence, and might have settled on linking (like ↔ wie) as *sure*, while the other eight 1:1 links of the multiword expression are typed as *possible*.

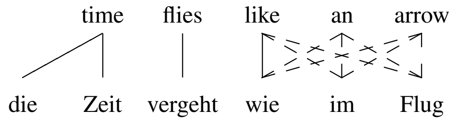


Figure 5.18: Toy gold standard

In this case, an automatic alignment as seen previously (figure 5.2, repeated here as 5.19) would look almost perfect. All five links of the automatic alignment would be true positives, and as there are no further links in the automatic alignment, precision would be 100%.

$$\text{precision} = \frac{5}{5} = 1 \tag{5.27}$$

$$\text{recall} = \frac{2}{3} = 0.67 \tag{5.28}$$

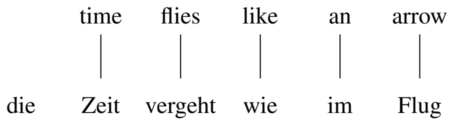


Figure 5.19: Toy automatic alignment

For computing recall, the links between (like an arrow, wie im Flug) would be ignored, thus there would be two true positives and three sure links to consider, thus recall would be 67%. The F-measure would be 80%.

Obviously, these evaluation metrics do *not* allow insights into the quality of an automatically generated word alignment as long as the percentages of *sure* versus *possible* links are not known. Even then, a reference annotation without guidelines but with *possible* links might not be as golden as it looks. Or, as other researchers have phrased it, *too many possible alignments in the golden data weaken the metrics* (Bojar and Prokopová 2006).

The *alignment error rate* (AER), finally, is defined as

$$AER(S,P;A) = 1 - \frac{|A \cap S| + |A \cap P|}{|A| + |S|} \tag{5.29}$$

finally, is said to be derived from the F-measure, and as it uses the modified recall-definition given above, will also give skewed evaluation results. In our toy example, the AER would be 12.5%.

$$AER(S,P;A) = 1 - \frac{2+5}{5+3} = 0.125 \tag{5.30}$$

Thus, precision, recall, and alignment error rate do not just give skewed results, they allow ignoring possibly big portions of the reference data. In fact, the more reference data is linked as *possible*, the higher are recall and AER and hence the less meaningful is the evaluation.

Note that just the typing of links as *sure* and *possible* is not harmful in itself. In fact, this distinction allows assessing whether a system performs well with respect to the set of *easy* or *sure* links, and how it performs with respect to the *possible*, but problematic, links. Only the combination of typing reference data as *sure* and *possible* in combination with AER and these problematic precision and recall values should be avoided.

Summed up, neither is the annotation of the reference set sufficient to be used for a thorough evaluation of a word alignment tool, nor is the suggested evaluation measure impartial and well-defined enough to give a clear picture of word alignment quality. The annotation scheme only satisfies minimal requirements. It is easy to follow, but does not ensure annotation quality. The suggested evaluation metric is bound to give as good results as possible despite the numbers of errors both in the alignment produced by a word aligner, and in the reference corpus. More disturbing is that the impact of problematic alignment decisions on overall alignment quality is decreased using these evaluation metrics, and if the gold standard is annotated so that data that is handled poorly by an alignment system is marked to be left out of the evaluation (i.e. marked as *possible*), evaluation results will be good despite a bad system performance. Thus, if a system is evaluated using AER and a sloppily set up gold standard, the evaluation results have to be treated with caution.

5.2.6 The PWA approach: Refining Precision and Recall

Drawing from experience both with the ARCADE and the PLUG project, Ahrenberg et al. (2000) thoroughly discuss how to evaluate word alignment quality from a variety of perspectives.

Gold Standard and Annotation Guideline The reference alignment that (Ahrenberg et al. 2000) use is the same that has been previously used within PLUG, i.e. each gold standard consists of 500 tokens chosen randomly from a PLUG subcorpus (see (Ahrenberg et al. 1999) for a description). Hence they also use the same annotation scheme as in PLUG.

Evaluation Metrics Ahrenberg et al. (2000) and Ahrenberg et al. (1999) adapt the evaluation metrics used in PLUG further in defining a factor Q,

$$Q = \frac{C_{src} + C_{trg}}{max(S_{src}, G_{src}) + max(S_{trg}, G_{trg})} \tag{5.31}$$

which corresponds to the number of words that have correctly been assembled correctly in the same link, divided by the maximum length of the link. As an example, I could compare the partial link (enquiries had been ↔ einigem Drängen übergeben) contained in the sentence pair given in figure 5.20. to the gold standard link

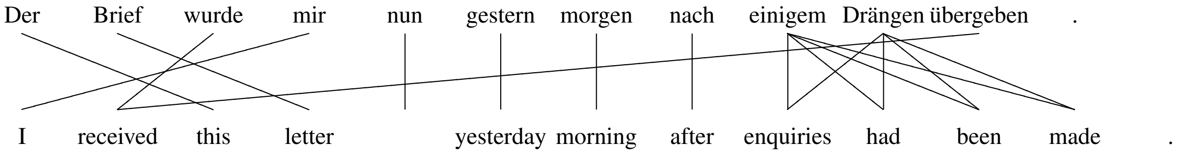


Figure 5.20: Example: PWA evaluation metrics

(enquiries had been made ↔ einigem Drängen)

In this case, the numerator of Q (5) would be the number of correctly aligned words in English (3), plus the number of correctly aligned German words (3). The factor would be the maximum length of the English expression participating in the link (4) plus the maximum length of the German expression participating in the link (2), i.e. it would be 6, and

$$Q = \frac{3+2}{4+2} = \frac{5}{6} = 0.83 \tag{5.32}$$

Intuitively, Q may be interpreted as the percentage of correctness of this particular link. Summed over all links in a system's output, Q will give the percentage of correctness of the whole word alignment. However, the factor is set up arbitrarily: why should the maximum number of the system's (S_src) and gold standard's (G_src) tokens be added to the maximum number of the system's (S_trg) and gold standard's (G_trg) tokens? It would make more sense to compute Q by dividing the amount of words correctly assembled in a link by the number of words that *should* be assembled in the link, i.e. by the number of words that are part of the corresponding gold standard link.

The sum over all Q's in a system's output is used to compute

$$\text{precision} = \frac{\Sigma Q}{I + P + C} \tag{5.33}$$

i.e. the degree of correctness is averaged by the number of links that the system produced[10]. Recall,

[10] The capital letters stand for the numbers of incorrect (I), partial (P), missing (M) and correct (C) links.

$$\text{recall} = \frac{\sum Q}{I + P + C + M} \tag{5.34}$$

however, is again computed averaged by the union of gold standard and system output instead of being divided by the number of links in the gold standard. However, the definition of recall is misleading, as done before for the PLUG definition of this metric. Despite the formula, the authors probably intended to compute recall as

$$\text{recall} = \frac{\sum Q}{\text{\# number of gold standard links}} \tag{5.35}$$

Clearly, Q is intended to capture the degree to which a partial link is correct. However, if it is used to compute precision and recall on the toy example (figures 5.1 and 5.2, repeated in figures 5.21 and 5.22), it gives again precision and recall values that may be intuitively implausible.

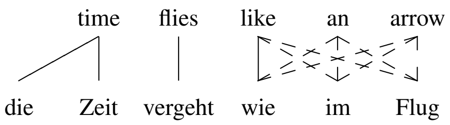


Figure 5.21: Toy gold standard

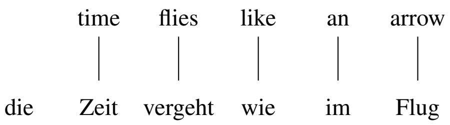


Figure 5.22: Toy automatic alignment

The factor Q would be 2.67 for precision and recall,

$$\sum Q = \frac{2}{3} + 1 + \frac{6}{6} = 2.67 \tag{5.36}$$

thus precision would be 53% and recall would be 89%. The F-measure, finally, would be 66%. These values seem counter-intuitive. All links in the automatic alignment are correct, if incomplete, so precision seems quite low. On the other hand, rather a lot of the gold links is missing in the automatic alignment, so it would seem recall is too high.

	precision (%)	recall (%)	F-measure (%)	AER	features
ARCADE	33	27	29	–	sentence + word alignment
PLUG I	100	100	100	–	translation spotting
PLUG II	67	67	67	–	translation spotting
AER	100	67	80	12.5	precision/recall on data subsets
PWA	53	89	66	–	translation spotting

Table 5.1: Comparison of the proposed evaluation metrics on a toy gold standard

5.3 Definition of a New Alignment Evaluation Approach

Some success has been achieved at designing annotation guidelines for both word and sentence alignment, and several gold standards have been set up. However, annotating a gold standard still seems to be a difficult task, as most approaches to word alignment evaluation restrict themselves to translation spotting. The only approach to measuring full word alignment does not define clear annotation principles, nor is the accompanying gold standard well documented. Additionally, with the exception of the BLINKER project, it seems unclear what constitutes a *good* alignment. Furthermore, a reliably annotated English–German gold standard does not seem to exist.

A variety of evaluation metrics have been suggested, none of which seems adequate to measure alignment quality. The problem is inherent in the task. Word alignment is difficult as it involves annotation possibly many n:m links, and it may not always be clear which words to link how. Hence it is vital that an evaluation metric does not categorise alignment decisions binarily into correct versus incorrect. Rather, an evaluation metric should take partial matches between the gold and an automatic alignment into account. On these grounds, all evaluation metrics suggested so far are insufficient. The metrics first proposed in ARCADE punishes partial matches severely. One set of the PLUG metrics assess alignment quality too generously, while the second set is arbitrarily set up. AER gives skewed results as precision and recall are computed over different subsets of the gold and automatic alignments. The PWA metrics, finally, seem counter-intuitive (table 5.1).

All of these issues need to be addressed, and the flaws in the current approaches need to be remedied. Accordingly, I have set up a new German–English gold standard, accompanied with annotation guidelines. The alignment has been done based on the notion of *context-dependent translational equivalence* in order to arrive at a complete, unambiguous word and sentence alignment that can be inspected for lexicographic and corpus linguistic purposes, and that serves as training material for statistical MT systems.

5.3.1 Gold Standard

As a first step towards a sound alignment evaluation, I have developed a new gold standard: In order to ensure comparability to existing evaluations, I have chosen to use data from the EUROPARL corpus. This way, the data is comparable to the *Canadian Hansards* in terms of genre. Simultaneously, using the EUROPARL corpus, it is possible to construct a gold standard not just for one language pair, but also to add gold annotations for more language pairs.

As a first start, I have randomly chosen a protocol file, available in German and English, that consists of all debates that took place in the European Parliament on May 5th, 2000. I had two annotators, native speakers of German with excellent knowledge of English, correct the automatic sentence alignment[11]. Subsequently, the annotators added word alignment information in 242 randomly chosen sentence links taken from the same protocol file.

[11] In fact, one of the annotators was me.

Sentence Alignment As the corpus has already been aligned at the sentence level, adding this information was not necessary. However, I had the two annotators check and correct, if necessary, the automatic alignment. The segmentation of the sentences was not changed. The annotators needed 6 to 7 hours to correct the alignment of the 4252 German and 4200 English sentences in the protocol file, in 4,261 sentence links. On average, then, they checked approximately 11 links every minute. Afterwards, the annotators compared their annotations: only 54 links differed between the annotators, i.e. they agreed on 98.61% of the links.

Language	Tokens	Types	Sentences	Paragraphs	Sentence Links	Word Links
English	110,046	8,818	4,199	1,350	4,261	4,807
German	104,267	13,714	4,251	1,350	4,261	4,807

Table 5.2: Characteristics of the gold standard

Word Alignment After the sentence alignment had been corrected, I randomly chose 242 sentence links containing 4788 German and 5336 English tokens for further annotation. These 242 sentence links are taken from three different text passages in the protocol, and one additional singleton link. The reason for choosing text passages instead of randomly choosing singleton links from all over the protocol was made to simplify the task for the annotators: As they have to word-align whole passages, they know, and have easy access to, the context of each sentence link and hence can resolve ambiguities more easily.

The manual word alignment is based on a guideline described in the following section (section 5.3.2), and was done independently of each other by the two annotators. On average, they needed 1.6 minutes per sentence link. After the annotation, the annotators resolved annotation differences by discussion, thereby creating an unambiguous and consistently annotated gold standard. These discussions were done in additional 17 hours, i.e. on average, the two annotators needed 4 1/2 minutes per link. This high number indicates that the guidelines were not well-designed enough to help the annotators in all cases. In fact, the annotation guideline was revised during the process of resolving annotation differences. Another reason for long discussion times might have been that the annotators had little or none practice before the gold standard annotation began. As a side effect, their annotations might show more variability than if they had received special annotation exercises beforehand.

In order to compute the inter-annotator agreement, I used Melamed (1998b)'s definition of the Dice-coefficient[12],

$$D(X,Y) = \frac{2}{\frac{1}{\text{precision}(X|Y)} + \frac{1}{\text{recall}(X|Y)}} = 0.644 \tag{5.37}$$

that compares the annotations of the two annotators X and Y, assuming Y has produced a gold standard. This Dice-coefficient corresponds to a precision of 0.57 and a recall of 0.72. The numbers again indicate that the guidelines were not sufficient help for the annotation task. However, the difference between the inter-annotator agreements reported for BLINKER and the inter-annotator agreement achieved here may in part be due to the different language pairs. English and French may show more structural similarities than do English and German. Thus, a lower inter-annotator agreement for the English–German gold standard might have been expected. More research should go into the question how the degree of difficulty relates to the respective language pair.

[12] For calculating precision and recall, I used the metric definitions given in 5.3.4 rather than the scheme used by Melamed (1998b).

Moreover, the fact that recall is higher than precision can be taken to show that most of the problems of the annotators were due to the alignment of multiword expressions: In these cases, many and larger annotation differences occurred, so that the precision with which each multiword expression was aligned suffered. Recall, however, was not affected as each multiword expression *was* aligned by both annotators.

In order to have a first impression what the upper limits of any automatic word alignment could be, I also compared each annotator to the final, unambiguous gold standard.

annotators	precision	recall
annotator 1	0.596	0.778
annotator 2	0.612	0.771

Table 5.3: Gold standard: upper limits for automatic alignment

According to these results, the human aligners achieve precision values up to roughly 60%, and recall is at 77%. Furthermore, the values correspond roughly to the inter-annotator agreement, i.e. it seems that both annotators aligned the gold standard roughly in the same way. Thus, the gold standard annotation should be reasonably consistently annotated.

However, as the annotation guidelines were not optimal, these upper limits can only be taken as a rough indication of optimal performance. Given other guidelines, and possibly more annotation experience, I assume that higher precision and recall values are possible, as well as a higher consistency in the gold standard annotation.

With only 242 sentence links, the word-aligned gold standard is relatively small. However, it is large enough to compute precision and recall, and gain first insights into the strengths and weaknesses of the system. Additionally, it is possible to align further parts of the protocol, thereby enlarging the gold standard, which is planned for the future. I also plan to add more languages, primarily French, to the evaluation data, in order to allow evaluation for German–English as well as German–French or other language pairs.

5.3.2 Annotation Guideline

Before the gold standard was annotated with word alignment information, I designed an annotation guideline based that is based on the principles and annotation decisions of both the ARCADE and the BLINKER project. Additionally, I aligned the first 100 sentence links of a randomly chosen protocol of the EUROPARL corpus manually, and used this experience to devise a detailed set of annotation rules and examples. During the resolution of alignment differences, I further revised the annotation guideline based on the discussions of the two annotators. Some of the initial alignment rules were changed, e.g. on how to align German and English genitive constructions. Other examples were added to show how to align problematic cases of translation paraphrases, larger structural divergences, and idiomatic expressions.

Basically, links should contain as few words as possible, and as many words as necessary in order to ensure *context-dependent translational equivalence*: words or expressions should be aligned if, *in the given sentence pair*, they are used to convey the same meaning. Translational equivalence should be symmetric, i.e. if an expression X is translated by X_t, then the translation of X_t should be X, as well.

Additionally, I distinguish between

- *context-independent translational equivalence* as a translation relation that would be expected to hold in many different contexts. Context-independent equivalent translation pairs could typically be included in a bilingual dictionary, as e.g. the translation pair (cat ↔ Katze);
- and *context-dependent translational equivalence*; Context-dependent translational equivalence holds in a specific context but usually not in other contexts; i.e. two expressions mean something different literally, but may be used to convey the same message.

These definitions are still incomplete as they do not define when *independence of context* is given, nor do they clarify what *meaning* is. However, they are useful to direct the manual alignment process towards a certain degree of consistency: if only those items that are equivalent context-independently were aligned, then the annotator would have to check for every translation pair whether it would be acceptable in other randomly chosen contexts. If the annotation is not restricted to context-independent translational equivalence, however, all those boundary cases can be aligned that are acceptable as translations, but would not be included in a dictionary.

Context-dependent translational equivalents thus gives the possibility to link expressions that are opposites of each other, like (niemanden (no one) ↔ everyone) and (misfallen (displease) ↔ please) in figure 5.23.

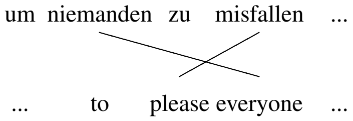


Figure 5.23: Example: context-dependent translational equivalence

The distinction I draw here is not absolutely new. In the ARCADE word alignment track (section 5.2.2), the adopted annotation scheme is restricted to *translation spotting*, i.e. only those items are aligned that are equivalent context-independently. In the PLUG project (section 5.2.4), the distinction is drawn on one axis between *textual* and *lexical* links, i.e. between translation pairs found in the text, and translation pairs to be included in a dictionary. Textual links are given if translational equivalence is context-dependent, with lexical links being context-independent. However, there is an important difference between the categorisation done in PLUG and the one here – the PLUG terminology indicates that the translational equivalence holds between words, and that, in the case of lexical links, there is reason to suppose the translation relation is or should be *lexicalized*. However, context-independent translational equivalence may as well hold between phrases or even sentences, and I do not hypothesise about the *lexicality* of a translation pair.

The second axis used in the framework of PLUG is given by the definition of *fuzzy* versus *regular* links, where *fuzziness* indicates category changes, or semantic or grammatical shifts between an expression and its translation. If the translation relation is not *fuzzy*, it is considered *regular*. This distinction is rather awkward for several reasons – the *regularity* of a link can only be defined if seen in opposition to its *fuzziness*. Or, in other words, if a link is not fuzzy, then it must be regular. Secondly, fuzziness judgements are both gradual[13] and highly subjective in that one annotator might perceive less fuzziness in a link than another, possibly depending on his or her language skills and even annotation training. A detailed annotation guideline can overcome this subjectivity in ensuring a certain consistency between annotators. However, the decisions taken in

[13] Which is true for context-dependent versus independent translational equivalence, as well.

the annotation guideline will still be arbitrary and subjective. Finally, there is no clear benefit from typing links in this way as degrees of or sources for fuzziness are not annotated, and hence cannot be exploited for system evaluations.

5.3.3 Guiding Principles

As guiding principles, I have adopted the rules used both in ARCADE and BLINKER, namely that links should be as small as possible while containing as many words as necessary. Additionally, non-translated items or punctuation should not be aligned at all. Translation errors, however, have to be aligned, based on context-dependent translational equivalence.

As few words as possible The word alignment[14] has to be as fine-grained as possible, i.e. words should be aligned in 1:1 links (figure 5.24). Additionally, only **entire words** are to be aligned, i.e. there is no alignment at the sub-word level.

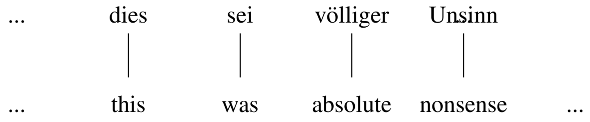


Figure 5.24: Example: as few words as possible

As many words as necessary In order to maintain translational equivalence, as many words as necessary have to be included in a link. This condition typically causes n:m links, as in figure 5.25. **In all cases, the resulting alignment should be symmetric**. A symmetric alignment is given if all words of the source expression are linked to all words of the equivalent target expression, as in figure 5.25. Asymmetric alignment is forbidden (figure 5.26).

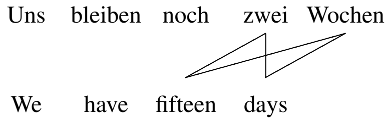


Figure 5.25: Example: as many words as necessary

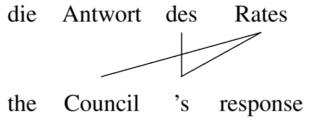


Figure 5.26: Example: asymmetric alignment

[14] For the sake of space and clarity, I often highlight only how a specific construction should be aligned, instead of linking all elements of an example. Additionally, I shortened the sentences. All German–English examples are taken from the EUROPARL corpus (Koehn 2005), and the English–French examples have been taken from the BLINKER guidelines (Melamed 1998b)

Deletions and Insertions If an expression is not translated for whatever reason, it should be left unaligned[15] (figure 5.27).

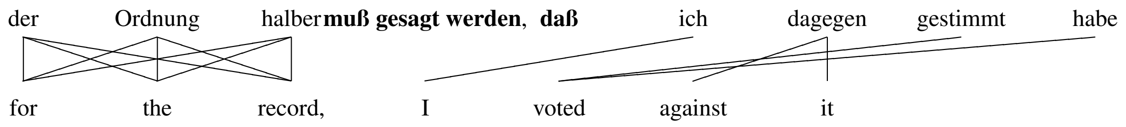


Figure 5.27: Example: deletions and insertions

Punctuation Punctuation should not be aligned. An exception is the hyphen (figure 5.28). In these cases, the hyphen has to be treated like a word and has to be aligned to its "translation"[16]

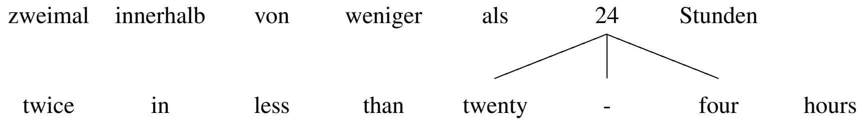


Figure 5.28: Example: aligned hyphen I

In coordinations, the hyphen indicates that one element has been elided (figure 5.29). In these cases, the hyphen has to be aligned to the translation of the elided element. The resulting alignment appears to be asymmetric. However, it is not. Instead, it consists of overlapping links in the same way as for other repeated elements and cases of ellipsis (see below).

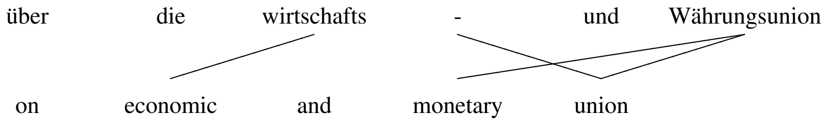


Figure 5.29: Example: aligned hyphen II

Translation errors Sometimes, translation errors occurred. In these cases, the errors should be treated like an omission in the translation (figure 5.30). If this strategy would leave too many parts of the sentence pair unaligned, the error has to be treated as if it was correct, i.e. as if, in the specific context, it has its intended meaning (figure 5.31). In this case, both annotators agreed that *angesichts des Umstandes* (English: under these circumstances) never means *to recognize* (German: erkennen). However, they agreed on aligning the two expressions as if they were translation equivalents.

[15] In other alignment annotation approaches, untranslated items have to be linked to a special *null word*. However, as it can be inserted automatically, if necessary, I feel free to omit it.

[16] In hyphenated words, I treat the hyphen as a morpheme that was irregularly torn apart by an over-eager tokenizer. In elliptic constructions, it can be considered an abbreviation and hence somewhat deserves word-status.

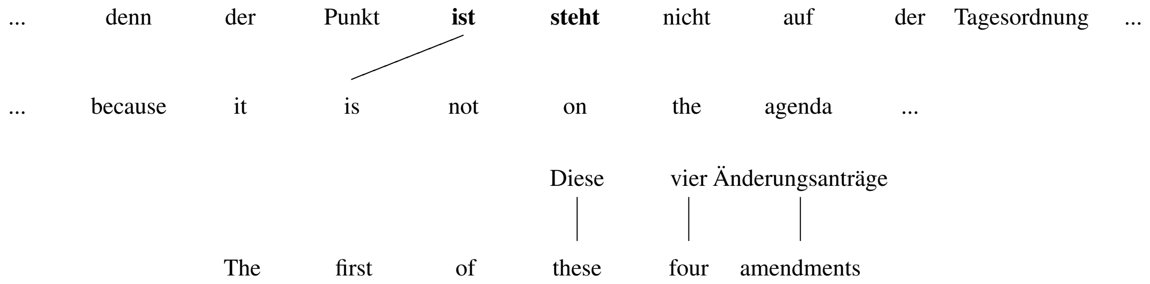

... denn der Punkt **ist** **steht** nicht auf der Tagesordnung ...
... because it is not on the the agenda ...
Diese vier Änderungsanträge
The first of these four amendments

Figure 5.30: Example: translation errors I

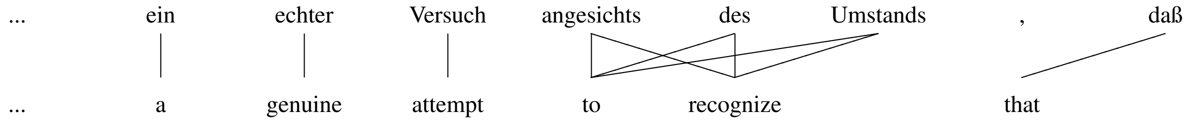

... ein echter Versuch angesichts des Umstands , daß
... a genuine attempt to recognize that

Figure 5.31: Example: translation errors II

Alignment of Nominals

Often, nominals can be aligned straightforwardly as a 1:1 links. Others are more complex and, moreover, the translation of a nominal may not be derivable from the translations of its elements. In other words, the nominal is translated non-compositionally. In these cases, the complete nominals have to be aligned as n:m links (figure 5.32)..

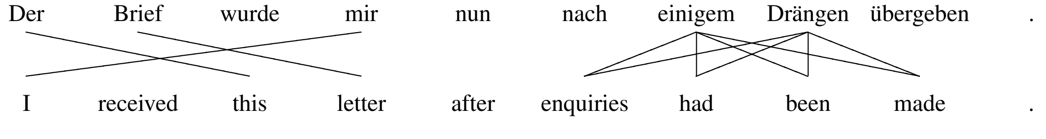

Der Brief wurde mir nun nach einigem Drängen übergeben .
I received this letter after enquiries had been made .

Figure 5.32: Example: alignment of nominals I

Sometimes, a quantifier or (indefinite) determiner may not be translated overtly, as the information is encoded morpho-syntactically on the noun. In these cases, the overt quantifier or determiner should be aligned to the noun (figure 5.33).

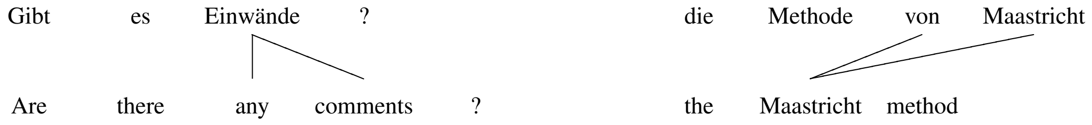


Figure 5.33: Example: alignment of nominals II

Pronouns should be aligned to those expressions that refer to the same entity, irrespective of whether the translation is in itself a pronoun or a nominal expression (figure 5.34).

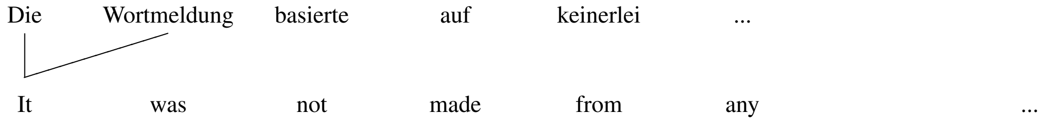


Figure 5.34: Example: alignment of pronouns

Expletive pronouns like *es*, *it*, *there* etc should be aligned with the corresponding expletive in the translation (figure 5.35).

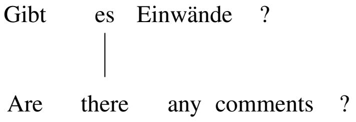


Figure 5.35: Example: alignment of expletive pronouns

Postnominal Genitives The postnominal genitive is encoded differently in the German and English grammar. Whereas it is encoded using a determiner in genitive case in German, the English construction involves a prepositional phrase. In these cases, the determiner of the German genitive should be aligned to the English preposition, and, if present, also to the subsequent determiner of the English translation (figure 5.36).

Prenominal Genitives The English prenominal genitive, however, is more difficult to align, as the clitic *'s* is treated as an independent word by the tokenizer. In these cases, it should be aligned as if it was still attached to the preceding word, i.e. it is aligned to the preceding word's translation (figure 5.37).

Pre- versus Postnominal Genitives There may also be instances where a prenominal genitive is translated by a postnominal one, or vice versa (figure 5.38). In these cases, the whole postnominal prepositional phrase has to be aligned to the whole prenominal genitive. This may cause a 2:2 link (figure 5.39).

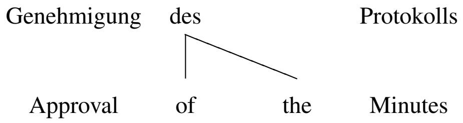


Figure 5.36: Example: postnominal genitive

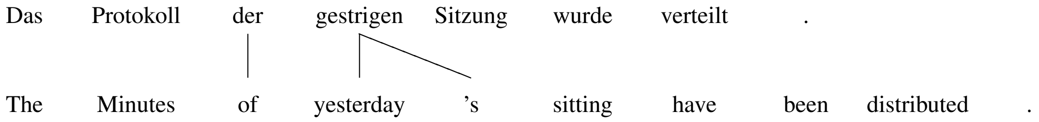

Das Protokoll der gestrigen Sitzung wurde verteilt .
The Minutes of yesterday 's sitting have been distributed .

Figure 5.37: Example: prenominal genitive

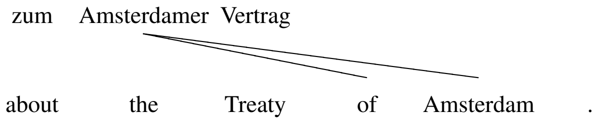


Figure 5.38: Example: pre- and postnominal genitive I

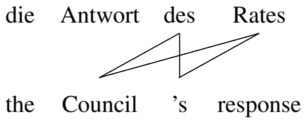


Figure 5.39: Example: pre- and postnominal genitive II

Alignment of prepositions and prepositional phrases

Prepositional phrases are most often translated as prepositional phrases, hence the prepositions can be aligned with each other (figure 5.40). The guidelines for aligning the succeeding NPs are shown above.

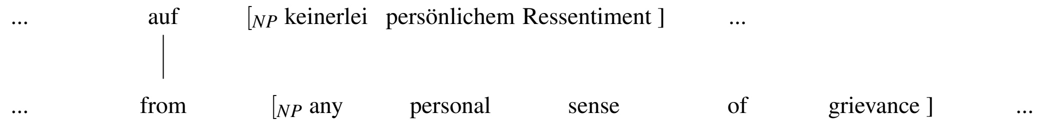


Figure 5.40: Example: prepositions I

Frequently, a German contracted form like *zum*, consisting of a preposition and an article, has to be aligned to a single English preposition (figure 5.41).

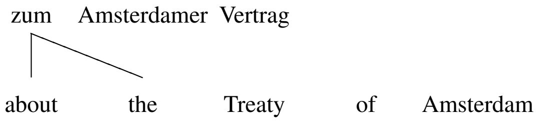


Figure 5.41: Example: prepositions II

One observation that the annotators made during the annotation was that they often copied this alignment pattern to occasions when no contracted preposition was present, although it violated both the annotation guideline and the rule to annotate context-dependent translational equivalence. A reason for this misalignment was that the pattern is easy to learn and obey, and I suppose that this patterns is equally fast learnable for a word alignment system.

However, it is linguistically inadequate, and the correct pattern can be learnt as easily, even with automatic means[17].

Alignment of pronominal adverbs

Pronominal adverbs like *darauf* may refer to a complete clause as in figure 5.42. In these cases, the pronominal adverb should be left unaligned. However, pronominal adverbs may also have non-clause, prepositional translations. Then, they should be aligned in 1:m links (figure **??**).

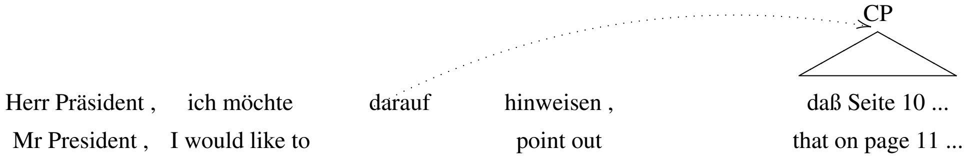


Figure 5.42: Example: pronominal adverbs I

[17] The German contracted forms like *zum*, *zur* can be tagged reliably by a POS-tagger.

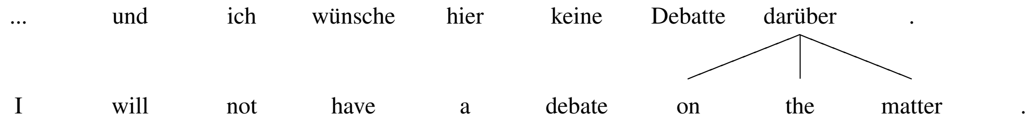

... und ich wünsche hier keine Debatte darüber .
I will not have a debate on the matter .

Figure 5.43: Example: pronominal adverbs II

Alignment of verbs and verbal groups

Verbs may be part of a verbal group, i.e. of a – possibly discontiguous – sequence of modal, auxiliary, main verbs, and particles. In these cases, the alignment should be as fine-grained as possible, i.e. each verb should be aligned to the element that is its closest translation in terms of *function* and *meaning* (figure 5.44).

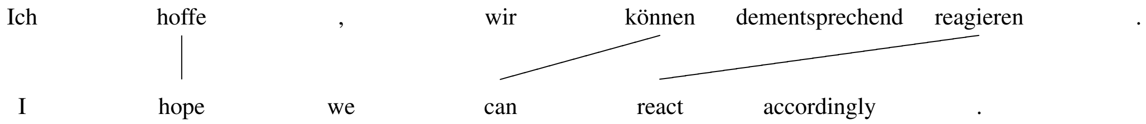

Ich hoffe , wir können dementsprechend reagieren .
I hope we can react accordingly .

Figure 5.44: Example: verbs

As a rule, auxiliaries used to code tense and aspect should be aligned to those verbs that equally encode tense and aspect, while the main predicate is aligned based on its meaning (figure 5.45).

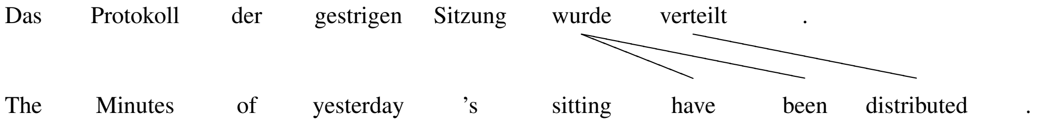

Das Protokoll der gestrigen Sitzung wurde verteilt .
The Minutes of yesterday 's sitting have been distributed .

Figure 5.45: Example: auxiliary verbs

If the main verb of one language encodes functions and meaning that are distributed over the whole verb cluster in the target language, then this source language verb has to be aligned to the whole target language verbal cluster (figure 5.46). In some cases, an active sentence will be translated by a passive construction (or vice versa). Then, the verbal group should be treated as one single multiword unit (figure 5.47).

The subject of the active clause may also be omitted in the passive translation[18] (figure 5.47). Particle verbs may be discontiguous. Still, the particle[19] has to be aligned to the translation of its verb. Subject-NPs occurring with imperatives should be aligned to the imperative verb form (figure 5.48).

[18] The English-French example was taken from the BLINKER annotation guideline, but adapted to conform to my guidelines.

[19] Often also called separable verb prefix

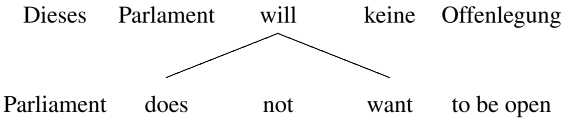


Figure 5.46: Example: verb cluster

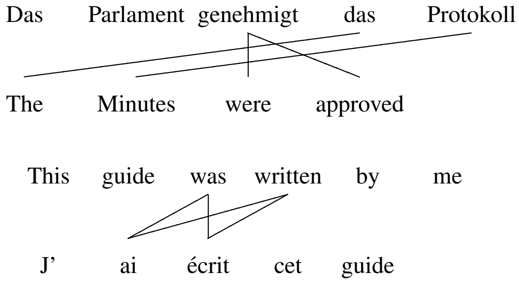


Figure 5.47: Example: active versus passive

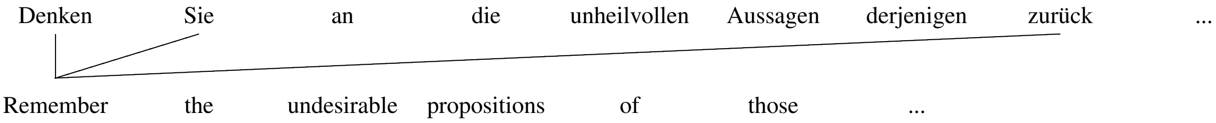


Figure 5.48: Example: imperatives

Negation

Negations may involve more than one element per language, and consequently, all parts of a negation have to be aligned in an n:m fashion (figure 5.49).

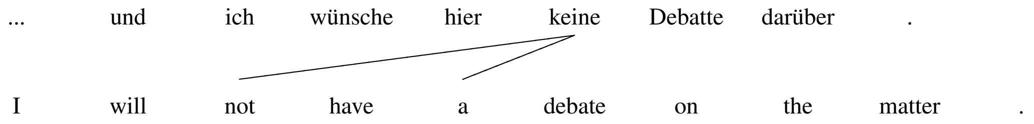

... und ich wünsche hier keine Debatte darüber .
I will not have a debate on the matter .

Figure 5.49: Example: negation

Coordination

Most coordinations can be aligned using 1:1 links (figure 5.50).

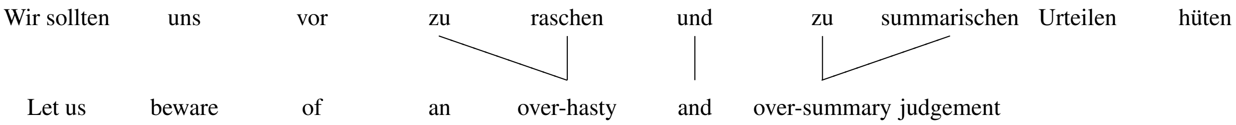

Wir sollten uns vor zu raschen und zu summarischen Urteilen hüten
Let us beware of an over-hasty and over-summary judgement

Figure 5.50: Example: coordination

Resumptive pronouns, correlatives and ellipsis Resumptive pronouns and other repeated elements should be aligned to the translation of their antecedent, *unless the antecedent is a phrase*. The antecedent itself has to be aligned to its translation, too[20]. Elliptic constructions are to be aligned along similar lines (figure 5.51).

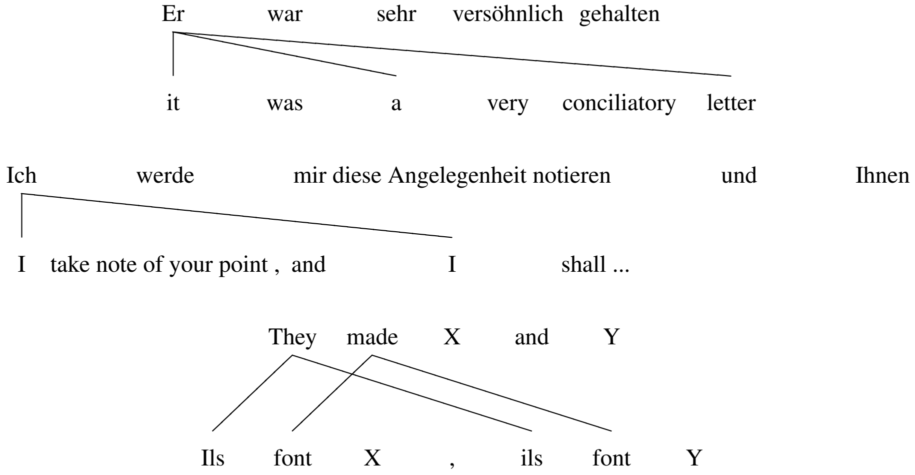

Er war sehr versöhnlich gehalten
it was a very conciliatory letter
Ich werde mir diese Angelegenheit notieren und Ihnen
I take note of your point , and I shall ...
They made X and Y
Ils font X , ils font Y

Figure 5.51: Example: resumptive pronouns

[20] Note that the antecedent of *er* is the nominal *a letter*, not just the noun *letter*

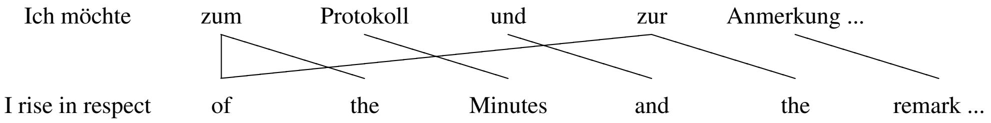


Figure 5.52: Example: repeated elements

In figure 5.52, two contracted prepositions (zum, zur) are aligned to their corresponding determiners, and to the corresponding preposition. This alignment looks asymmetric, but is not. Instead, the alignment information has to be interpreted as *two links*, (zum ↔ of the) and (zur ↔ of the), respectively. Generally, elided or repeated elements complicate the manual alignment, and the alignment strategy in those cases should be similar to the one adopted for resumptive pronouns etc[21].

Structural divergences and Multiword Expressions

Often, minor structural divergences occur: a prepositional phrase with a nominal as argument e.g. is translated as a prepositional phrase plus verbal construction. In these cases, the verb has to be aligned to the nominal (figure 5.53). In predicative constructions, adjectives should be aligned to the corresponding prepositional phrases 8figure 5.54).

Such structural divergences typically effect very few words and both source and target language constructions have the same function as in the above example. However, larger structural divergences occur as well. Often, idioms or subordinate clauses are involved (figure 5.55).

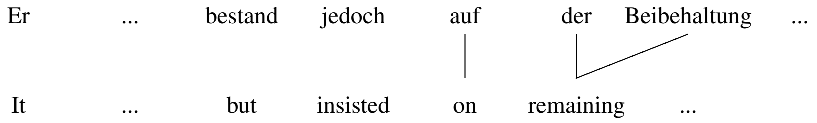


Figure 5.53: Example: structural divergences I

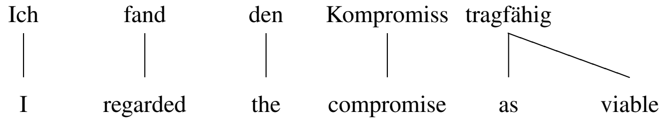


Figure 5.54: Example: structural divergences II

Subordinate clauses present considerable alignment problems when the structures of a subordinate clause and its translation differ, i.e. when one subordinate clause is infinite while its translation is not. In these cases, it is important to align based on *context-dependent translational equivalence* and refrain from elements that are obviously, due to the different syntactic structures, not overtly translated.

[21] It also unclear how an automatic aligner should arrive at such an annotation except when treating repeated elements as multiword units, which is clearly not an optimal solution. One possibility is to type the links as alternatives that both have to be present (Kuhn, p.c.). However, this approach can complicate the alignment, and later the evaluation of alignment systems, to a degree where it is no longer feasible.

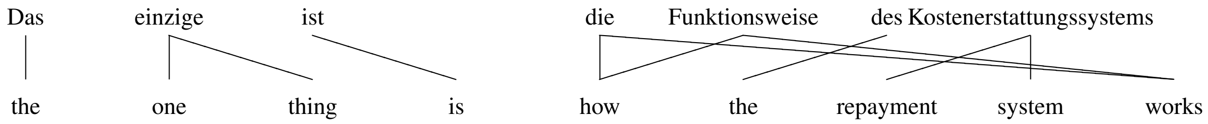

Das einzige ist die Funktionsweise des Kostenerstattungssystems
the one thing is how the repayment system works

Figure 5.55: Example: structural divergences III

If it is not possible to determine which source language words of an expression or construction should be aligned to which target language words, then it is necessary to align the whole construction in a multiword, n:m fashion (figures 5.56 and 5.57). It is also possible that all words of a source language sentence have to be aligned to all words of the target language sentence.

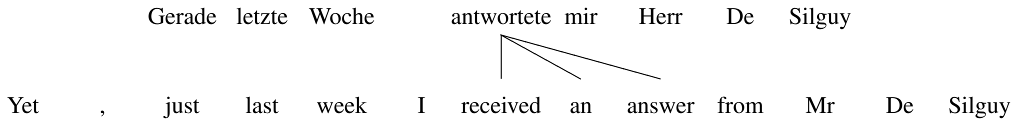

Gerade letzte Woche antwortete mir Herr De Silguy
Yet , just last week I received an answer from Mr De Silguy

Figure 5.56: Example: structural divergences IV

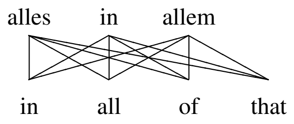


Figure 5.57: Example: structural divergences V

5.3.4 Evaluation Metrics

As has been discussed, the metrics that have been used before to measure alignment quality are insufficient on various grounds. Hence, they should be replaced with metrics that can be shown to give plausible and intuitively correct results in a variety of different alignment comparisons. These new metrics should be applicable to any alignment task, whether at the sentence or word level. Maybe they could even be defined such that they also facilitate a semi-automatic, quantitative error analysis.

A first step towards defining such metrics is to analyze which types of (correct) matching patterns between a gold standard and an automatic alignment can occur. In principle, these patterns fall into four different categories.

1. The gold standard and the automatic alignment are identical, i.e. the automatic alignment is *perfect* (case 1).

2. The gold standard and the automatic alignment do not coincide at all, i.e. the automatic alignment is *completely wrong* (case 2).

3. The automatic alignment does not include alignment information on all expressions that are aligned in the gold standard, i.e. the automatic alignment is *incomplete* (case 3).

4. The automatic alignment includes alignment information on all expressions that are aligned in the gold standard, but some part of the correct alignment information is missing, i.e. the automatic alignment is *partially correct* (case 4).

It is straightforward to define how the evaluation metrics precision and recall should assess alignment quality in these four cases.

- If the automatic alignment is perfect, precision and recall (or indeed any other metric) should both be 100% (case 1).
- If the automatic alignment is completely mismatched, precision and recall (or any other metric) should be zero (case 2).
- If the automatic alignment is incomplete or partially correct, then precision should be 100%, while recall should be lower and reflect the degree to which the corpus has been aligned (cases 3 and 4).

The cases 3 and 4 should have a similar effect on precision and recall, and hence it may be tempting to conflate them into one. However, the distinction between the two cases is important. In case 3, the overall alignment contains gaps that count as neither correct nor incorrect. Moreover, it does not matter whether a 1:1 link is missing or an n:m link (n, m, or both $\geq$1). In case 4, an n:m link is incomplete, i.e. it is neither fully correct nor fully erroneous or missing. Thus, precision and recall need to reflect the *degree* to which a link has been aligned.

There is a final case to consider. However, there may be no clear intuitions how exactly precision and recall should reflect the alignment quality.

5 The automatic alignment includes alignment information on all expressions that are aligned in the gold standard, but while some parts of the alignment information are correct, other parts are wrong, i.e. the automatic alignment is *partially incorrect* (case 5).

Intuitively, both precision and recall should have values that are neither 100% nor zero ($0 <$ precision, recall $< 100\%$). However, depending on the situation, it may be hard to define which precision and recall values to expect for this case. Hence, case 5 will not be discussed here[22].

With the first four matching patterns in mind, it is straightforward to set up an example gold standard and define "automatic" alignments that match perfectly, not at all, incompletely, and partially correctly.

Let a gold standard consist of four tokens *a, b, c, d* in the source language, and the tokens *k, l, m, n* in the target language. These eight tokens are aligned in three links. One of these links is a 2:2 link (a,b ↔ k,l), and the other words are aligned in two 1:1 links, namely (c ↔ m) and (d ↔ n) (figure 5.58).

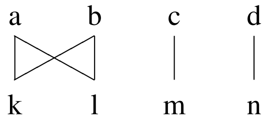


Figure 5.58: An example gold standard

Additionally, there may be four automatic alignments, corresponding to cases 1-4 (figures 5.59 - 5.62). Table 5.4 shows how precision and recall should react to these four automatic alignments.

[22] Intuitions about partially incorrect matches depend largely on the specific example used. Thus, it is hard to define how an evaluation metric should behave when confronted with partially incorrect matches. One should hope, however, that metrics that give plausible results for cases 1-4 will give at least acceptable ones for case 5.

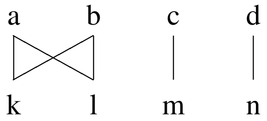


Figure 5.59: Example automatic alignment (case 1)

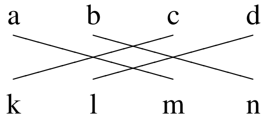


Figure 5.60: Example automatic alignment (case 2)

All evaluation metrics discussed before (section 5.2) can be shown to give correct results for cases 1 - 3. However, they differ with respect to partially correct matches (case 4).

The metrics used in ARCADE I punish partially correct links as errors. The partial linking of the n:m link (a,b ↔ k,l) would count as an error while the two 1:1 links would be true positives. However, the n:m link is neither missing nor erroneous, it is simply not complete. Thus, a precision below 100% seems implausible.

$$\text{precision} = \frac{\text{number of alignment links correctly found}}{\text{all alignment links found}} = \frac{2}{3} = 0.67$$

$$\text{recall} = \frac{\text{number of alignment links correctly found}}{\text{number of reference alignment links}} = \frac{2}{3} = 0.67$$

The refinement of ARCADE II to count how many words were correctly aligned is either impractical to handle for n:m links, or gives wrong results, too[23].

$$\text{precision} = \frac{|\text{words correctly aligned}|}{|\text{all words aligned}|} = 1$$

$$\text{recall} = \frac{|\text{words correctly aligned}|}{|\text{all gold standard words}|} = 1$$

Partial links are also a problem for the metrics used in PLUG, given here as PLUG I and PLUG II. Although these metrics explicitly take partial matches into account, PLUG I is too generous in that partial matches do not count as errors. Furthermore, due to the definition, partial matches are much more likely to decrease precision than have an effect on recall. In case 4, precision would be 75% due to the simple fact that the n:m link has been divided up into two 1:1 links. Thus the automatic alignment contains more links than necessary. However, as all of the available alignment information on the n:m link is correct, it is implausible why precision should suffer. Nor is it plausible that recall should not.

[23] In a partially correct match, all source language tokens may be correctly aligned to *some* part of the equivalent target expression. The deciding question is whether a word counts as correctly aligned if it is aligned to *all* tokens of the translational equivalent, or if it is at least correctly aligned to some of them.

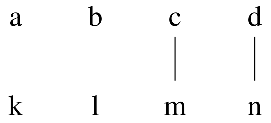


Figure 5.61: Example automatic alignment (case 3)

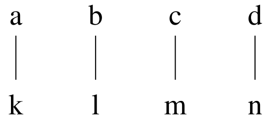


Figure 5.62: Example automatic alignment (case 4)

$$\text{Precision I} = \frac{\text{\# correct links found} + \text{\# partial links found}}{\text{\# all links found}} = \frac{2+1}{4} = 0.75$$

$$\text{Recall I} = \frac{\text{\# correct links found} + \text{\# partial links found}}{\text{\# all reference links}} = \frac{2+1}{3} = 1$$

PLUG II, as a remedy to the problem, decreases the impact of partial matches onto precision and recall by an arbitrary factor. However, the underlying problem remains in that precision (63%) is still lower than recall (83%).

$$\text{Precision II} = \frac{\text{\# correct links found} + (0.5 \cdot \text{\# partial links found})}{\text{\# all links found}} = \frac{2+0.5}{4} = 0.63$$

$$\text{Recall II} = \frac{\text{\# correct links found} + (0.5 \cdot \text{\# partial links found})}{\text{\# all reference links}} = \frac{2+0.5}{3} = 0.83$$

Unfortunately, precision and recall as defined in the *alignment error rate*-approach also gives implausible results. As this evaluation approach differs from the above-mentioned, some additional parameters need to be defined. First, in order to have comparable precision and recall values, all links of the gold standard are assumed to be *sure*. Second, as precision and recall are not defined to work for n:m links, the 2:2 link (a,b ↔ k,l) need to be decomposed into four 1:1 *sublinks*, i.e. the 2:2 link is replaced by four 1:1 links (a ↔ k), (a ↔ l), (b ↔ k), and (b ↔ l). However, loosing the information on the n:m link does not result in an adequate assessment of alignment quality. The missing information on the n:m link decreases precision and recall to the same extent, resulting in the precision and recall values observed for the ARCADE I metrics.

$$\text{precision} = \frac{|A \cap P|}{|A|} = \frac{4}{6} = 0.67$$

$$\text{recall} = \frac{|A \cap S|}{|S|} = \frac{4}{6} = 0.67$$

Accordingly, although it is not explicit in the definitions of precision and recall, they seem to treat partial links as errors.

metric	perfect	mismatch	incomplete	partially correct
precision	100%	0	100%	$0 \leq$ precision $\leq 100\%$
recall	100%	0	$0 \leq$ recall $\leq 100\%$	$0 \leq$ recall $\leq 100\%$

Table 5.4: Expected behaviour of precision and recall

Finally, the Q factor developed within PWA in order to account for partial matches, causes precision and recall to give the same results as the much simpler PLUG I definitions. Moreover, it is difficult to compute. $\sum Q$ over all matches would be 3

$$\sum Q = \frac{2}{2} + \frac{2}{2} + \frac{2}{4} + \frac{2}{4} = 1 + 1 + 0.5 + 0.5 = 3$$

and thus precision would be 75% and recall 100%.

Summed up, neither of the previously suggested evaluation metrics assess alignment quality adequately, the reason being their poor accounting for partial matches (all example precision and recall values can be found in table **??**). Another possibility to account for partial matches and to intuitively plausibly assess alignment quality is to *micro-average*. Precision and recall need to be computed *for every link* of the alignment. Thus, the degree to which an n:m link is aligned correctly, erroneously, or not at all, is directly measured.

In order to compute a *micro* or *local* precision, it is necessary to first decompose every n:m link in the corresponding number of *sublinks* that link each and every token of the source language expression to each and every token of the target language expression. This decomposition has already been done for the n:m link in the example alignment: the 2:2 link (a,b ↔ k,l) is decomposed into the four 1:1 *sublinks*, i.e. it is replaced by four 1:1 links (a ↔ k), (a ↔ l), (b ↔ k), and (b ↔ l). Second, the sublinks are used to compute precision and recall for the translation pair[24] given the following definitions:

$$\text{precision}_{local} = \frac{\text{number of sublinks correctly found}}{\text{all sublinks found}} \tag{5.38}$$

and

$$\text{recall}_{local} = \frac{\text{number of sublinks correctly found}}{\text{all gold subalignment links}} \tag{5.39}$$

The sums of all local precision and recall values, divided by the numbers of links in the gold and automatic alignments in a corpus (links, *not* sublinks!) will give an overall assessment of the alignment quality.

$$\text{precision}_{global} = \frac{\sum \text{precision(local)}}{\text{\# automatic links found}} \tag{5.40}$$

and

$$\text{recall}_{global} = \frac{\sum \text{recall(local)}}{\text{\# gold standard links}} \tag{5.41}$$

These precision and recall definitions finally give results that correspond to our intuitive assessments on the example alignments in all cases 1-4 (table 5.5).

[24] The sublinks are used to compute the precision and recall for the n:m link.

metrics	perfect	mismatch	incomplete	partially correct	precision	recall
ARCADE I	√	√	√	!	67%	67%
ARCADE II	√	√	√	!	100%	100%
PLUG I	√	√	√	!	75%	100%
PLUG II	√	√	√	!	63%	83%
AER	√	√	√	!	67%	67%
PWA	√	√	√	!	75%	100%
ATLAS	√	√	√	√	100%	75%

Table 5.5: Actual behaviour of precision and recall

An additional advantage of the approach is that the local precision and recall values may be used to quantitatively measure which types of errors occur: does the automatic alignment contain many gaps, does it compute 1:1 links correctly most of the time, or does it fail to produce error-free n:m links? Moreover, the local precision and recall values make it possible to quantify whether bigger n:m links like 4:5 or 6:5 links make up a bigger part of the alignment errors than comparatively small but probably more common 2:2 links, and whether a high number of alignment errors are "near misses" in that only one or two sublinks may be missing or erroneous. Thus, the metrics do not just assess alignment quality, they indicate how alignment approaches should be improved.

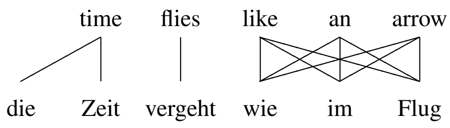


Figure 5.63: Toy gold standard

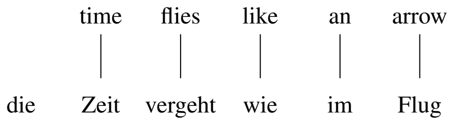


Figure 5.64: Toy automatic alignment

With respect to the toy gold standard mentioned in section 5.1 (figures 5.63 and 5.64), the local precision and recall values would indicate that each link only contains correct information, i.e. all local precisions and the average precision value would be 100%. Additionally, recall would indicate that on average, a value of 60% was achieved. A quantitative analysis would further reveal that while the 1:1 link was aligned correctly, the automatic produced incomplete alignments on all n:m links with n, m, or both >1. Thus, the evaluation would show that in order to improve alignment quality, the treatment of n:m links should be refined.

5.4 Evaluating ATLAS

With the previously described gold standard and the redefined evaluation metrics precision and recall, the performance of the ATLAS alignment system can be evaluated. The evaluation is accompanied by an error analysis to show which errors are produced by ATLAS, and how can they be avoided.

The evaluation is carried out in three different tracks: in the first, ATLAS is tested on a relatively small data set with a previously not used language-pair. This evaluation shows that a new language pair can be aligned with ATLAS even without prior tuning (section 5.4.1). The other tracks are carried out on the previously described gold standard on EUROPARL data. The first of these shows how well ATLAS aligns on the sentence and paragraph level, and how it scales to a larger data set. The second evaluates how it aligns at the word level (sections 5.4.2 and 5.4.3).

5.4.1 Sentence and Paragraph Alignment Quality on the EuNews Corpus

In the first track, ATLAS is evaluated on the EUNEWS corpus described previously (section 3.3.3). This corpus has been used to develop and test in particular the sentence and paragraph modules of ATLAS. However, the development has been done exclusively using the two language pairs English–German and French–German, having been aligned manually at the sentence and paragraph level by a single annotator. Thus, the third language pair, English–French, consisting of 259 paragraph and 517 sentence links, is available for evaluating how ATLAS performs

- on a small data set,
- for sentence and paragraph alignment only,
- using a familiar genre, i.e. one that ATLAS, in principle, has been fine-tuned to during development,
- with a previously unseen language pair, and accordingly without any initial fine-tuning.

Thus the evaluation focuses on the capabilities of ATLAS as a sentence and paragraph aligner for a new language pair. Simultaneously, because the data and the task are closely related to tests during development, the evaluation allows insights into

- performance characteristics,
- required fine-tuning, and
- robustness with respect to new language pairs.

In particular, only the sentence- and paragraph modules of ATLAS are used, but no word alignment modules. Nor does ATLAS use a dictionary or induce a lexicon.

The results reveal that there is room for improving ATLAS, as the paragraph alignment quality alone is quite low with a precision and recall value of 37.5%. Within the correctly aligned paragraphs[25], sentence alignment achieves a precision of 54.26% and a similar recall with a value of 54.99%. Furthermore, the alignment takes several hours to complete, due to the amount of hypothesis overgeneration. In the initial parameter setting, each alignment module may generate all possible alignments.

As can be seen in table 5.6, the most important source for sentence alignment errors is that small permutations occurred, i.e. cases of partial links where the erroneously linked sentences are adjacent or very close to those words that should have been linked (10 cases). The other error sources consist in sentence links being either too small, i.e. cases where the gold standard link is an 1:3 or other type of n:m link, but where the automatic alignment suggests a 1:1 link (7 cases), or the opposite case (2 cases).

[25] As the alignment is cohesive, correct sentence links within incorrectly linked paragraphs are not to be expected. Hence, incorrectly linked paragraphs are excluded before carrying out the sentence alignment evaluation.

Error Source	Frequency	Percentage
unaligned sentences	0	0
permutations	10	50
links too small	7	35
links too large	3	15

Table 5.6: Sentence alignment evaluation: error analysis

These results clearly show the limits of the system. Due to the interaction between the alignment modules, there is a substantial amount of error propagation. Furthermore, hypothesis overgeneration is slowing down the alignment process considerably. In a second run, hypothesis overgeneration is restricted as each module may only submit the two best hypotheses per corpus item to the task manager. As an effect, the performance of ATLAS increases considerably. The alignment process is complete after roughly 4 minutes. The resulting alignment quality is also higher with 73.77% precision and 32.14% recall for paragraph alignment quality, and 36.95% precision and 20.47% recall sentence alignment quality (computed over *all* sentence links, irrespective of whether they are contained in an incorrect paragraph link or not). The error analysis reveals that permutations still make up a big portion of the sentence alignment errors (75 cases), while unaligned sentences (55 cases) and sentence links that were either too small or too big (45 cases) are smaller error sources.

These results indicate that limiting hypothesis overgeneration has a very positive effect on alignment speed, while it decreases recall considerably. Simultaneously, the precision values show that the sentence and paragraph modules seem to not have enough cues to produce high precision link hypotheses. If more information on word alignment was available to the system, more correct link hypotheses might have been generated. Another possibility is that the interaction between the modules is not yet optimal. Correct hypotheses generated by one module may have been ruled out during the alignment disambiguation because their confidences were lower than the (bad) hypotheses of another module. Finally, the evaluation also shows that the choice of language-pair seems to be irrelevant. The addressed error sources, slow alignment, hypothesis overgeneration and error propagation, are entirely caused by the system architecture and have been addressed and partially solved by reducing the amount of hypotheses. The system has not in any way been fine-tuned to the language pair. The performance can equally not be correlated to language-pair specific information like a dictionary, or to the quality of a particular dictionary, because this information has not been available to the system. Hence, it is reasonable to suppose that once dictionary information is available, the alignment quality will increase further.

5.4.2 Sentence and Paragraph Alignment Quality on Europarl

The second valuation investigates how well ATLAS scales to bigger data sets. Accordingly, the system is used to align the gold standard data on the sentence and paragraph level. As before, no additional lexicon information is used, and the hypothesis generation is restricted again to using only the two best hypotheses per corpus item.
Thus, this evaluation allows insights into

1. how well ATLAS scales to bigger data sets,
2. using a language pair that was worked on during development, i.e. for which the system was fine-tuned, and
3. using a genre that has only been used during some word alignment experiments.

Error Source	Frequency	Percentage
unaligned sentences	689	25.76
permutations	603	22.54
links too small	6	0.22
links too large	1377	51.48

Table 5.7: Europarl Sentence Alignment Evaluation

Accordingly, precision and recall values are expected to be similar to those reported in section 5.4.1. This alignment process, due to the large amount of data, is completed after roughly 3 hours. On this data set of 4.261 sentence links, ATLAS achieves a precision of 42.82% and a recall of 26.46%. Unlike before, however, the error analysis shows that most of the errors are that the system generated too many n:m links (51.48% of all cases, see table 5.7).

Both the relatively high number of unaligned sentences and the low recall further suggest that there is a substantial amount of error propagation. It is likely that the paragraph alignment contains errors, and that hence it is difficult to generate any sentence hypothesis within these erroneous paragraph links. Thus, the permutation errors may be attributed to error propagation. An informal inspection of a very small sample of these errors supports that hypothesis. Other sentence link errors, however, cannot be related to an erroneous paragraph link.

Overall, the evaluation results are similar to those of section 5.4.1 which suggests that ATLAS aligns consistently. Furthermore, the errors can be attributed to the same sources, namely the massive hypothesis-overgeneration inherent in the system. Accordingly, to increase the alignment quality and reduce the computational load, it is necessary to restrict the alignment computation even further to only high-precision hypotheses. This restriction will inevitably result in low recall values. However, gaps in the alignment may help to identify further alignment clues that may be exploited for hypothesis generation. Finally, as word alignment information has not been used so far, there is reason to suspect that even very limited but high-precision word alignment information may increase the alignment quality considerably.

5.4.3 Word Alignment Quality on Europarl

Unfortunately, word alignment quality may also suffer from error propagation. Hence, as quite a big number of erroneous sentence and paragraph links persist in the automatic alignment, it is reasonable to suspect that word alignment, too, will contain a large number of errors, and that the overall alignment quality of ATLAS is quite low.

Without assessing word alignment quality in isolation, and without reducing error-propagation even further, it thus does not seem reasonable to assess the overall alignment quality of ATLAS. Hence, the word alignment capabilities of ATLAS are evaluated in isolation.
Thus, the evaluation allows to

- assess word alignment quality,
- interpolate what its influence was if the system aligned at all levels simultaneously, and
- to investigate into weaknesses of the word alignment modules.

Accordingly, ATLAS word alignment modules are assessed on the basis of the existing gold sentence alignment annotation. Furthermore, only the word-aligned section of the gold standard is

Error Source	Frequency	Percentage
unaligned sentences	0	0
permutations	3761	60.87
links too small	1541	24.94
links too large	872	14.11
wrong types	5	0.08

Table 5.8: Word alignment evaluation: error analysis

used. This restriction to a small data set of only 242 sentence links is made to reduce the computational load, i.e. to increase the alignment speed[26].

In a first step, a lexicon is induced from the 242 word-aligned sentence links. This lexicon induction is restricted to word types that occur more than 5 times, a parameter settings that has been most promising during the module development (section 4.7). Furthermore, only the translation candidates with the highest confidence values per lexicon entry are retained. This automatically induced lexicon consists of 5554 entries, each having on average 56 translations. On this data set, ATLAS performs poorly. The system achieves a precision of only 3.72%, and recall is even lower with 0.72%. Furthermore, only 1035 word links are generated, i.e. 78.32% of the data has not been aligned. Fortunately, the alignment computation takes only approximately 10 minutes.

Most often, permutations occurred, thus indicating that the heuristics of the word alignment modules do not work well. Secondly, ATLAS suggested links that are either too big (like 2:3 links where 1:1 links would have been correct), or too small. Finally, a few errors occurred where the link type was completely wrong, but cannot be categorised as too big or too small. In these cases, the correct gold links were 2:3 links, but the word hypotheses suggested are 1:4 links (table 5.8).

These values seem to indicate that the word alignment information generated by ATLAS is completely useless. However, a close inspection of the automatically generated lexicon yields better results. This lexicon consists of 563 lexicon entries, i.e. most of the lexicon entries that have been induced previously have been discarded during the alignment. Furthermore, each lexicon entry contains on average 2.8 translations , with an average confidence of 0.7234. Among these 563 lexicon entries, 93 lexicon entries are correct, i.e. each translation of the lexicon head lemma is correct (16.52%). Additionally, the lexicon contains 80 partially correct lexicon entries, i.e. entries for which at least one of the listed translations was correct (14.21%).

[26] As it is already obvious that alignment errors are enormously propagated, using a bigger data set would increase the error rate considerably, but other than that, it would not bring new insights into the performance of the aligner.

Again, these numbers are encouraging rather than good. However, most of the (partially) correct lexicon information is interesting. Among the correct entries are many noun translation pairs like

(1) Gegenteil ↔ contrary (confidence: 1)

(2) Glaubwürdigkeit ↔ credibleness (confidence: 0.75)

(3) Vergütung ↔ allowance (confidence: 0.09)

Furthermore, nominals, i.e. noun multiword units, have also been correctly aligned:

(4) Mitgliedsstaat ↔ Member States (confidence: 1)

(5) Personalpolitik ↔ personnel policy (confidence: 0.93)

(6) Sicherheitspolitik ↔ security policy (confidence 0.8)

Of course, the lexicon also contains cognates like

(7) Name ↔ name (confidence: 1)

but also correct translation pairs that are neither cognates nor nouns.

(8) stimulieren ↔ encourage (confidence: 1)

(9) jetzt ↔ now (confidence: 1)

Finally, some partially correct translation pairs indicate the potential usefulness of ATLAS:

(10) Haushaltspolitik ↔ strict budgetary policy (confidence: 0.69)

(11) nur ↔ remedy (confidence: 0.85), how (confidence: 1), merely (confidence: 1)

Some lexicon entries are, unfortunately, errors:

(12) Finanzverwaltung ↔ problem of fraud (confidence: 1)[27].

Another striking peculiarity of the lexicon is that many words are linked despite their belonging to rather different word categories, i.e. determiners may be aligned to nouns, verbs, etc. An obvious improvement of the alignment might accordingly be to incorporate information during the lexicon induction, and thus to discard these word pairs from the initial lexicon. However, an induced lexicon that uses such a POS filter has not resulted in an improvement on word alignment or lexicon quality. Precision is still low with a value of 3.19%, whereas Recall increases slightly to 0.74%. The automatically generated lexicon equally consisted of 563 entries, 93 of them correct (16.52%) and 86 at least partially correct (15.26%).

[27] "Finanzverwaltung" should have been aligned to "financial management", and "problem of fraud" is equivalent to German "Problemen im Zusammenhang mit Betrug".

5.5 Summary

Obviously, ATLAS is not yet ready for full-fledged corpus alignment, i.e. alignment at all levels simultaneously. Nor does it yet compute alignment information with a decent quality. Paragraph and sentence alignment quality is low, and a substantial amount of the errors must be attributed to error propagation: an incorrect alignment hypothesis erroneously receives a high confidence and thus is preferred over other, correct hypotheses during the alignment disambiguation. As a result, the error causes other correct paragraph and sentence alignment hypotheses to be discarded.

Furthermore, the interaction between alignment modules of the complete system leads to worse alignment quality than if a single module is used. One reason for this is, again, error propagation. Second, a global measure of confidence is missing. Rather, alignment modules are competing in a “the higher confidence value wins”-fashion. Thus, an incorrect alignment hypothesis with a high confidence value, generated by one module, may prevail against many competing, correct hypotheses of other modules simply because none of them have a high enough confidence value, and because *majority votes* between the modules are not taken into account. Designing a global confidence assessment mechanism that does not just compare confidence values of hypotheses, but also takes majority votes into account should have a positive effect on the alignment quality. Of course, exchanging the naive similarity measures of the modules by more sophisticated and well-defined ones would also have a positive effect, as would the focus on generating high-precision alignment hypotheses.

Another likely reason for the bad alignment quality of ATLAS is the absence of strong constraints on linear ordering as known from length-based approaches to sentence alignment, or locality constraints known from some of the statistical word alignment approaches. Rather, these constraints are either absent within ATLAS, or, in the case of the linearity principle, weakened to one cue amongst many. If such constraints could be built in and strengthened in ATLAS, or if the system would be combined with one of these approaches, the alignment quality might improve. However, integrating strong linearity or locality constraints into the system would have lead well beyond the scope of this thesis.

Some strengths of the alignment system have also emerged. Its modular architecture and flexibility allows easy integrating of previously unused language pairs (like English–French). Furthermore, general error reduction techniques like the restriction to n-best hypotheses can be added without requiring major modifications. Furthermore, using the evaluation data, it is not just possible to assess alignment quality. Rather, the error analyses can directly lead to improvements that may affect either the core functionalities, and leave the alignment modules unmodified. Or new modules can be added that address specific weaknesses of the system.

The high processing time has been reduced using very simple means, even with a positive impact on alignment quality. Finally, and despite the bad word alignment quality, the analysis of the automatically generated dictionary has been shown to contain translation pairs that may be useful or interesting, at least from a lexicographic point of view.

Much needs to be done to increase word alignment quality: the alignment disambiguation must be made more robust against erroneous hypotheses with high confidences, or, alternatively, incorrect hypotheses should not receive high confidence values. Secondly, ATLAS has to be adapted to link multiword sequences appropriately, and link permutations need to be avoided, possibly by including positional similarity among the word alignment cues.

Chapter 6

Conclusion

> Anyone attempting to build an optimal translation model should infuse it with all available knowledge sources, including syntactic, dictionary, and cognate information.
>
> (Melamed 2000, p. 222)

6.1 Summary

The aim of this thesis has been to develop a text alignment system that is highly flexible and may align sentences as well as words within a bilingual, parallel corpus. Further requirements are that the system

- is language-pair independent, i.e. not restricted to align texts of only a single language pair,
- exploits linguistic corpus annotation for computing alignments,
- uses a variety of techniques, such as linguistic rules, statistics, or heuristics,
- is modular to allow for the easy incorporation of new language pairs and alignment cues.

Most importantly, the alignment system has to align hierarchically, i.e. it must be able to compute alignment information simultaneously for paragraphs, sentences, words and phrases. As a result of the alignment task, the system has to provide both corpus alignment information and a bilingual dictionary, at preferably high precision. Processing speed and recall, however, are not considered to be as important.

These design decisions have been made based on observations of the different approaches to text alignment, as discussed previously. Hybrid sentence alignment approaches that use a variety of cues may be considered superior to those using only a single alignment cue. However, the restriction that these systems only compute sentence links is unfortunate. Approaches to aligning below the sentence, i.e. at the word level suffer from the limited variation of techniques used. Especially standard word alignment systems rely too strongly on statistics. Thus the alignment of multiword units, rare events, and when language pairs show considerable word order differences, is problematic.

The alternative alignment system ATLAS has been designed to address these issues. It is highly modular, with core functionalities like hypothesis management and alignment disambiguation separated from the generation of alignment hypotheses. The alignment hypothesis generation is done within encapsulated modules, each using at least one alignment cue. Due to the encapsulation, the architecture facilitates adding new alignment modules as they can be plugged in without having to implement corpus management methods or search algorithms.

The core functionalities include a mechanism that allows interactions between modules, i.e. hypotheses from one alignment module are typically re-used by another. Comparisons between different hypotheses from different alignment modules are facilitated by the use of confidence values that are derived from the heuristics or statistics of the modules. These confidence values can be modified to take the reliability of a module into account, i.e. the confidence values also reflect the extent to which the modules are known to generate “good”, i.e. probably correct alignment hypotheses. Furthermore, a constrained best-first search, the alignment disambiguation, has been designed so that ATLAS can combine alignment hypotheses of different granularities (paragraphs, sentences, words, and chunks) into a single, cohesive alignment path.

All alignment modules currently implemented in ATLAS have been developed based on experiments. These experiments show how different kinds of corpus annotation may be used for the alignment task, and how. Some experiments use alignment cues that have been introduced elsewhere, like using sentence lengths to indicate probably good sentence links. Other alignment strategies are extensions of well-known approaches, like using POS-filtered cognates for sentence alignment, or inducing a bilingual lexicon from a parallel corpus without using sentence alignment information. Yet again others are original, like using more fine-grained morphological information as an alignment cue, especially for aligning nouns and nominal multi-words.

The linguistic information currently supported covers the most basic and surface-oriented corpus annotations, namely information on parts-of-speech, lemmas, syntactic constituency and morphological structure. Deep linguistic analysis like dependency or semantic information is not used. The similarity measures used by the modules are first naive implementations and thus can be improved considerably.

As a first step towards an evaluation of ATLAS, current alignment evaluation practices have been examined. Unfortunately, they have been found lacking. Accordingly, the evaluation metrics precision and recall have been redefined in order to capture alignment quality adequately. Furthermore, a gold standard has been created for the word alignment subtask, created manually by two annotators, and with the help of an annotation guideline.

In the evaluation, the sentence alignment quality achieved on the English–French EUNEWS data is disappointing. The processing time is initially very high, and the alignment quality is extremely low, both due to the huge number of alignment hypotheses produced by the system. Moreover, most of these hypotheses are erroneous. Restricting the generation from all possible alignment hypotheses to only the n-best ones increases the performance considerably. Processing speed on the English–French EUNEWS data is reduced from several hours to four minutes. Precision and recall are still quite low, with values of roughly 74% and 32% on the EUNEWS, and approximately 43% precision and 26% recall on EUROPARL. While these results are bad, they are also relatively similar despite having been achieved on data from different language pairs and on very different amounts of data. Thus, it may be hedged that the alignment capabilities of ATLAS are basically independent of the language-pair in question, and that the aligner should scale well to larger corpora than those used during the evaluations.

Concerning word alignment, the results are also disappointing. Precision and recall are both approaching zero, with the main error source being permutations, i.e. aligment hypotheses that combine sentences or words that are close to what would have been correct alignment links. However, when examining the bilingual dictionaries that are generated along with the corpus alignment information, nearly a third of their entries was found to be at least partially correct. Thus, while the alignment information is useless for corpus annotation and alignment, the automatically generated lexicon may still provide interesting translation pairs for lexicographic purposes.

Apart from the above-mentioned permutations, a major error source of the alignment computation is error propagation. A single, erroneous alignment hypothesis with a high confidence value may set the alignment disambiguation on a wrong track, so that more and more erroneous hypotheses are included in the process-final alignment. As a solution to this problem, it is necessary to focus on high-precision alignment modules and clues, i.e. to restrict the hypothesis generation to only those hypotheses that are most likely to be correct, and to discard the others as soon as possible.

6.2 Future Improvement Directions

From the evaluation results that have been achieved, it is obvious that there is much room for improvement. First of all, some aspects of the system architecture have to be re-examined and enhanced. Secondly, the experiment results open possibilities for further research and modifications. Finally, ATLAS should be extended to support more languages and language combinations in order to allow for more cross-linguistic experiments.

An obvious target for modifications of the system architecture is the alignment disambiguation. It should be made more robust with respect to erroneous alignment hypotheses, e.g. by

- the timely discarding of erroneous alignment hypotheses, e.g. by using n-best restrictions like those used during the evaluation of ATLAS;
- the use of a search beam to ignore those hypotheses that are probably erroneous, based on heuristics on linear ordering;
- the extension of the alignment disambiguation to pursue alternative alignment paths.

Especially the last modification would give robustness to the overall system. If the alignment disambiguation was able to compute two or more distinct alignment paths, then these different paths could be compared in order to rule out improbable hypothesis sequences. A hypothesis sequence e.g. that consists of a few high-confidence hypotheses, and a lot of low-confidence hypotheses or many 0:1 links may be erroneous, based on the assumption that on average, the confidences in an alignment path should be high. Additionally, deletions or insertions should be rare in a correct alignment.

As a comparatively minor point, the corpus management should be extended in order to exploit additional linguistic information, e.g. information on dependency relations between syntactic constituents. Another obvious modification is to enable the system to cope with different styles for encoding syntactic information in order to use outputs of different parsers or chunkers. Along with extending the support of different kinds of corpus annotations, further experiments should be conducted in order to find additional clues that use these new corpus annotations.

The experiments conducted during the development of the system also provide opportunities for improvements and further research. As has been said before, the heuristics used are sometimes naive first implementations. So one focus should lie on improving these heuristics to yield high-precision alignment hypotheses. Furthermore, the confidence values should be re-inspected with respect to the comparability between alignment modules. So far, there is reason to suspect that the alignment modules do not use the range of confidence values in exactly the same way. If e.g. a probable hypothesis by a module A can only achieve a maximum value of 0.7, while another module assigns confidence values between 0.7 and 0.9, it is impossible to compare the hypotheses and use them for the alignment disambiguation. Additionally, the reliability factors that take the credibility of the different alignment modules into account need to be tuned more precisely, e.g. with the help of machine learning algorithms.

With the focus on high-precision alignment clues and modules, recall will certainly suffer. As a result, the corpus alignment will contain gaps. It is thus vital to inspect these gaps and use them to arrive at new alignment clues and modules which can then be plugged into ATLAS. This research direction can also be expected to yield interesting insights for linguistics and translation studies.

One conclusion to draw from the evaluation is that especially in a hybrid, modular system like ATLAS, preference must be given to alignment strategies that produce high precision hypotheses. If the high-precision alignment contains gaps, these can then be filled by adding more modules to the system, using additional, or maybe only highly-specialized alignment clues. This strategy will, as a side effect, limit error propagation. Furthermore, as the restriction to only high-precision hypotheses also reduces hypothesis generation to a minimum, processing speed will also increase.

A final point concerns the support for additional languages. So far, ATLAS has been developed and tested for only four languages (German, English, French, Swedish) in various combinations. For the development, this approach has had the advantage that differences in character encodings and tokenization could be ignored. But as these languages belong to the same language family, one should expect more similarities between their grammars and vocabularies than between languages from unrelated family groups. Experiments should be conducted that use a wider range of languages, and typologically more different languages. It should be interesting to work on aligning agglutinative languages like Finnish or Inuktitut with languages like English or German. It might also be interesting to use languages that use different scripts and hence present more challenges to corpus annotations, like Chinese or Thai.

The choice of EUROPARL as a development and evaluation corpus has partially been made with these considerations in mind. As this corpus already contains parallel texts from eleven European languages, including Finnish, it facilitates working on a wider range of different language pairs without changing genre. Furthermore, by including annotations for more language pairs in the evaluation set, the results achieved for the different language pairs are directly comparable. To this end, the gold standard data will be made available to the research community. Finally, after refactoring the code and possibly including a first set of improvements in the alignment modules, ATLAS will be made available to the research community, thus hopefully supporting especially cross-linguistic research.

Appendix A

Technical Details

A.1 Adding New Languages

Adding a new language that uses the Latin script can typically be done without making any changes to the system. Specific types of corpus annotations, namely information on lemmas, or equivalent information, is automatically supported. However, other kinds of corpus annotation, like POS-tags, need to be parametrized (section A.3).

A.2 Input and Format Specifications

An options file is to inform the system on the kinds of annotation present, and on the format of the corpus (An example of an options file can be found in appendix B.1). The corpus information has to be divided into at least two files, one per language, and it has to be available in a system-internal XML format (figure A.1). The system-internal format can also be used to add alignment information to the input (section A.5).

```
<s id=3>
<nonterminal id=1 type=NC>
<terminal id=1 lemma="d" category="ART">Die</w>
<terminal id=2 lemma="Wissenschaft" category="NN">Wissenschaft</w>
</nonterminal>
<nonterminal id=2 type=NC>
<terminal id=3 lemma="d" category="ART">der</w>
<terminal id=4 lemma="Deduktion" category="NN">Deduktion</w>
</nonterminal>
</s>
```

Figure A.1: System-internal format: sentence taken from the LITERATURE corpus

The input format of the IMS tree-tagger (figure A.2) is also supported, and further filters for tiger-xml (Volk et al. 2006), and the XCES instantiation of the *Corpus Encoding Standard* (Ide et al. 2000) will be added in the future.

Parameter files are used to translate from the detailed POS-information included in the supported tagsets STTS (German, Thielen et al. 1999), PENN TREEBANK (English, Santorini 1990), IMS FRENCH (French), SUC (Swedish, Ejerhed et al. 1992), IMS ITALIAN (Italian), and IMS SPANISH (Spanish) to more general word classes such as *noun, verb, particle, punctuation, etc*[1].

[1] See section 3.4 for details on the tagsets, and appendix D for the generalizations.

```
<s>
<NC>
Die     ART     d
Wissenschaft    NN      Wissenschaft
</NC>
<NC>
der     ART     d
Deduktion       NN      Deduktion
</NC>
</s>
```

Figure A.2: Tree-tagger format: sentence taken from the GUTENBERG corpus

The syntactic constituents may also be labelled with respect to their type, i.e. whether they are nominal or prepositional phrases, or whether they are chunks, verb clusters, etc. However, it is not yet possible to generalize from parser- or chunker-specific syntactic tags to broader syntactic constituent classes.

The translations from POS-tags to larger category classes are necessary for several reasons: firstly, alignment on the basis of syntactic category is only possible if the system can translate between the different tagsets, i.e. if it "knows" that both a JJ-tag (PENN TREEBANK) and an *ADJA*-tag (STTS) are used for adjectives. Secondly, not every distinction made in a tagset is relevant to the alignment process, and hence the translation of tags into larger syntactic classes allows to ignore unnecessary information.

Furthermore, a machine-readable dictionary may be added to the input, following the XML-style dictionary format used by ATLAS (An example lexicon file can be found in appendix B.2).

A.3 Extending the Set of Supported Corpus Annotations

A side effect of the conversion of language-specific tagsets to more general word categories is that it simplifies the porting of ATLAS to new languages or tagsets: It is not necessary to define which specific tag of one tagset corresponds to which tag of another tagset[2]. Rather, it is only required to define which tag corresponds to which larger syntactic classes like adjective, name, noun, verb, conjunction, determiner, particle, preposition, pronoun, citation, foreign, punctuation.

If the corpus annotation makes use of a *new* tagset for an already supported language, as is e.g. the case if a corpus is not using any of the tagsets described above, or if the tagset is relevant for a *new*, i.e. previously unsupported language, a new parameter file for this tagset has to be written. This takes approximately 20-30 minutes depending on the extent to which the programmer is familiar with the language and tagset in question.
A parameter file should include the following information:

1. Each tag of the tagset,

2. A grouping of the tags into the three category classes *lexical* (for nouns, names, adjectives, adverbs[3], and verbs, including modals and auxiliaries[4]), *functional* (for prepositions, conjunctions, determiners, particles, and pronouns), and *various* (for punctuation, symbols, foreign words, and other language or corpus-specific tags),

[2] This approach would become tedious if more than three or four different tagsets are supported.

[3] Adverbs are included in this class although they are functional words rather than lexical ones. The reason is that adjectives are often used *adverbially*, hence adverbs and adjectives should belong to the same word category class.

[4] No distinction is drawn between different verb types.

3. A description which is the more general syntactic category of words with such a tag. Words that are tagged *VAFIN* in the German tagset STTS (Thielen et al. 1999), e.g. are verbs (among other things such as finite and an auxiliary.).

The category descriptions (adjective, verb, noun etc.) of a new parameter file *have to be* already in use in the other, existing and tested parameter files, or, if a new general category is necessary in one parameter file, then the other parameter files should be adjusted to include the new general category as well, for the appropriate language-specific tags. Otherwise, the system is not be able to find equivalents for these categories among the parameters of the other languages.

In some cases, i.e. when the person writing the parameter file is not sufficiently familiar with the tagset and its language, an expert has to be consulted. When expanding the parameter files to cover Spanish, e.g. tags for nouns, names, etc. did not present any difficulties, but an expert was needed to reliably categorize the tag SE: It is used to tag the Spanish particle *se* that is either as reflexive or as a means to construct a passive, among other functions[5]. Finally, that tag was parametrized to be a particle.

So far, ATLAS supports the following types of corpus information: word form, lemma, POS-tag, syntactic constituent, and morphological information for compounds. In order to make a new type of annotation available to the alignment system, the following modifications have to be made:

1. the indexing script *Indexing.perl* and the subroutine *readInternal* has to be changed in order to include the new information in the index,

2. the internal format has to be adapted, along with its DTD,

3. the corpus data base has to be extended to include a slot for the new information.

So far, the corpus data base is not fully flexible, hence the required changes on this data base are difficult to make. A future version of ATLAS will feature another data base organization, and then, it will be easier to add any new kind of corpus annotation, including dependency information or semantic annotations.

A.4 New Alignment Strategy

New alignment modules can be added comparatively easily as long as they do not require any previously unsupported corpus annotation[6]. The module call should be included in the appropriate section of the task manager: it sorts module calls according to which type of input hypothesis (corpus, paragraph, sentence, phrase or word) is required. The new module takes as input the name of a parent hypothesis and it returns a list of new, child hypotheses. Interface subroutines are provided that query the system's data bases for relevant information such as information on the category, morphological structure or lemma of a word. So implementing a new module is facilitated by the modular architecture of ATLAS.

However, if an alignment module has to use corpus annotation that has previously *not* been used by ATLAS, then additional interface subroutines are required. Adding these involves deep knowledge of the alignment program, a situation which will be improved with the next version.

[5] Thanks to Gemma Boleda for her help with the Spanish parameter file.

[6] In this case, the interface to access the new information will simply be missing.

A.5 Output and Output Formats

All alignment information computed by ATLAS is made available in the system-internal XML format. It lists the hypotheses of the alignment agenda, and each hypothesis is given along with information on its type and its confidence value (here called *certainty*, as in the XCES format.). The system-internal corpus indices are used as pointers to the corpus items.

```
<corpus>
  <l1 lang="German" files="de03106.crp de03121.crp ..." dir="/corpora/euNews/de">
  <l2 lang="English" files="en03106.crp en03121.crp ..." dir="/corpora/euNews/en">
</corpus>
<alignments>
  <aligned type="paragraph" l1="f:1-p:1" l2="f:2-p:1" certainty="2.1656375" />
  <aligned type="paragraph" l1="f:1-p:2" l2="f:2-p:2" certainty="2.20715" />
  ...
</alignments>
```

Figure A.3: System-internal format: example alignment of the EU NEWS corpus

In order to facilitate the manual correction and checking of the alignment information, ATLAS can augment the pure list of alignment links with the respective paragraph or sentence pairs in plain text. This format is only available as output format for sentence and paragraph alignment information. In the future, further output formats (XCES and tiger-xml) will be added.

```
  ###########################################
    <aligned type="sentence" l1="f:1-p:7-s:18" l2="f:2-p:7-s:12" certainty="0.63542" />
  Vorrangige Ziele der FEMIP sind :

  Ses priorités d’ action sont :
  ###########################################
```

Figure A.4: System-internal format: example alignment of the EU NEWS corpus

The dictionary that is also generated by ATLAS may be *unfiltered*, i.e. it contains all word alignment hypotheses that have been generated during the alignment process, or it may be filtered according to two different criteria: the dictionary may contain only those word alignment hypotheses that have been used to produce an unambiguous text alignment, i.e. those word alignment hypotheses that remain valid after the *alignment disambiguation* step. Or, the dictionary also includes the information given by the pre-existing dictionary that has been used during the alignment process. This option is useful to gradually acquire lexicon information during a series of independent alignment processes.

The lexicon is encoded in XML and lists the lemma, language, and category information for each headword, along with a list of translations. The translation information does not only consist of the translation’s lemma, but also includes information on the translation’s syntactic category and confidence (See example A.5).

```
<item>
   <lemma>point of order</lemma>
   <category>multiword</category>
   <language>English</language>
   <translations>
      <translation>
         <lemma>Geschäftsordnung</lemma>
         <category>noun</category>
         <language>German</language>
         <confidence>0.78571</confidence>
      </translation>
   </translations>
</item>
```

Figure A.5: Example of an ATLAS lexicon entry

Appendix B

Example XML Files

B.1 Options File

Below is an example of an ATLAS options file. The header information states which corpus (TITAN) has to be aligned, where the input dictionary (MNEMOSYNE-DE-EN.XML[1]) is found and where to write the output dictionary (TITAN-DE-EN.XML). The option

```
<filter>alignedInput</filter>
```

indicates that all information of the input dictionary *and* all word pairs used in the final text alignment should be included in the dictionary. Another filter option,

```
alignedOnly
```

causes the output dictionary to contain *only* those word pairs to be included that were used in the final alignment. The option

```
unfiltered
```

however, causes *all possible and generated* word hypotheses to be included in the dictionary.

It is also possible to specify which output format is to be used by the system: either a text version suitable for manually checking the alignment (txt) and a system-internal format (internal). In case no supported option is chosen, the system automatically switches to the internal format.

The index file TITAN.INDEX stems from a previous run of the program on the corpus TITAN, and is a compact corpus representation that allows faster reading in of the corpus than do the various input formats. Finally, a sentence alignment file is given indicating that the corpus TITAN should be enriched with additional, i.e. word alignment information.

The body information on the corpus TITAN states where to find the corpus files, separated by language. Note that a corpus may contain multiple files per language, but that the order of the files has to be the same for both languages.

In principle, the corpus files can be available in different formats. Here, the German corpus files use the system-internal format, and the English corpus files are available in the tree-tagger format. For each language, a single format must be chosen.

Further information includes the listing of the relevant parameter files, here the files for the English Penn Treebank tagset and the German STTS-tagset, and which types of structural annotation is present (paragraph and sentence boundaries).

[1] Mnemosyne (sometimes confused with Mneme) was the personification of memory in Greek mythology. This titaness was the daughter of Gaia and Uranus and the mother of the Muses by Zeus. (http://en.wikipedia.org/wiki/Mnemosyne)

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
<!DOCTYPE options SYSTEM "/xml/options.dtd">
<options>

<header>
<name>titan</name>
<directory>/corpora/titan/</directory>
<InputDictionary>
   <file>mnemosyne-de-en.xml</file>
   <directory>/dictionary/</directory>
   <format>xml</format>
</InputDictionary>
<OutputDictionary>
   <file>titan-de-en.xml</file>
   <directory>/dictionary/</directory>
   <format>xml</format>
   <filter>alignedInput</filter>
</OutputDictionary>
<OutputFormat>internal</OutputFormat>
<index>/corpora/titan/titan.index</index>
<sentenceAlignment>
   <files>/corpora/titan/titan.aligned</files>
   <format>ces</format>
</sentenceAlignment>
</header>

<!-- Information on l1 -->
<languageOne>
<name>German</name>
<directory>/corpora/titan/de/</directory>
<files>de_titan1.crp de_titan2.crp</files>
<format>internal</format>
<parameters>sttsGerman.xml</parameters>
<structure>
<annotation type="sentence"> s </annotation>
<annotation type="paragraph"> p </annotation>
</structure>
</languageOne>

<!-- Information on l2 -->
<languageTwo>
<name>English</name>
<directory>/corpora/titan/en/</directory>
<files>en_titan1.pos en_titan2.pos</files>
<format>tree-tagger</format>
<parameters>pennEnglish.xml</parameters>
<structure>
<annotation type="sentence"> s </annotation>
<annotation type="paragraph"> p </annotation>
</structure>
</languageTwo>
</corpus>
```

B.2 Lexicon File

This is a small example of a lexicon file. Each lexicon entry includes information on the headword, namely its lemma, word category, language, and one or more translations. Each translation includes information on the lemma of the translation along with the translation's category, language, and confidence value. The latter indicates how reliable the translation is. High confidence values indicate high reliability.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
<!DOCTYPE dictionary SYSTEM "/xml/dictionary.dtd">
<dictionary>
<item>
   <lemma>Geschäftsordnung</lemma>
   <category>noun</category>
   <language>German</language>
   <translations>
      <translation>
         <lemma>condition</lemma>
         <category>noun</category>
         <language>English</language>
         <confidence>0.11111</confidence>
      </translation>
      <translation>
         <lemma>Rules of Procedure</lemma>
         <category>multiword</category>
         <language>English</language>
         <confidence>0.94444</confidence>
      </translation>
   </translations>
</item>
<item>
   <lemma>Rules of Procedure</lemma>
   <category>multiword</category>
   <language>English</language>
   <translations>
      <translation>
         <lemma>Tätigkeitsbereich</lemma>
         <category>noun</category>
         <language>German</language>
         <confidence>1.00000</confidence>
      </translation>
      <translation>
         <lemma>Geschäftsordnung</lemma>
         <category>noun</category>
         <language>German</language>
         <confidence>0.94444</confidence>
      </translation>
   </translations>
</item>
</dictionary>
```

Figure B.1: Example of a Lexicon File

B.3 Corpus Files

B.3.1 Monolingual Corpus File

The following is a short example file of the EUNEWS corpus, containing English text. As can be seen, words are tagged as <TERMINAL>, the information including a word id, a possibly <UNKNOWN> lemma, word category information and the word form itself. Additionally, a MORPH feature can be used to show the morphological structure of a word.

If a corpus is syntactically encoded, the constituents will be tagged as <NONTERMINAL>.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
<!DOCTYPE corpus SYSTEM "/xml/corpus.dtd">
<corpus>
<p id=1>
<s id=1>
<terminal id=1 lemma="<unknown>" category="NP">BEI/03/128</terminal>
</s>
</p>
<p id=2>
<s id=2>
<terminal id=1 lemma="Luxembourg" category="NP">Luxembourg</terminal>
<terminal id=2 lemma="," category=",">,</terminal>
<terminal id=3 lemma="4" category="CD">4</terminal>
<terminal id=4 lemma="December" category="NP">December</terminal>
<terminal id=5 lemma="@card@" category="CD">2003</terminal>
</s>
</p>
<p id=3>
<s id=3>
<terminal id=1 lemma="Hungarian" category="NP">Hungarian</terminal>
<terminal id=2 lemma="Prime" category="NP">Prime</terminal>
<terminal id=3 lemma="Minister" category="NP">Minister</terminal>
<terminal id=4 lemma="<unknown>" category="NP">P&eacute;ter</terminal>
<terminal id=5 lemma="Medgyessy" category="NP">Medgyessy</terminal>
<terminal id=6 lemma="pay" category="VVD">paid</terminal>
<terminal id=7 lemma="a" category="DT">a</terminal>
<terminal id=8 lemma="visit" category="NN">visit</terminal>
<terminal id=9 lemma="to" category="TO">to</terminal>
<terminal id=10 lemma="the" category="DT">the</terminal>
<terminal id=11 lemma="EIB" category="NP">EIB</terminal>
<terminal id=12 lemma="on" category="IN">on</terminal>
<terminal id=13 lemma="<unknown>" category="JJ">3&nbsp;December</terminal>
</s>
</p>
[...]
</corpus>
```

Figure B.2: Example of a Corpus File

B.3.2 Alignment File

All alignment information is encoded in relation to the monolingual corpus files. Each link gives information on its type, e.g. *paragraph* and on the reliability or *certainty* of the link. Most importantly, pointers are used to link the source (*l1*) and target language items (*l2*). One file may contain alignment information linking several file pairs.

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="no" ?>
 <!DOCTYPE alignment SYSTEM "/xml/alignment.dtd">
 <alignment>
 <header>
   <l1 lang="German" files="de03128.crp" dir="/home/bschrade/diss/korpus/euNews/de/"/>
   <l2 lang="English" files="en03128.crp" dir="/home/bschrade/diss/korpus/euNews/en/"/>
 </header>
 <alignments>
   <aligned type="paragraph" l1="f:1-p:1" l2="f:2-p:1" certainty="1.78" />
   <aligned type="paragraph" l1="f:1-p:2" l2="f:2-p:2" certainty="1.39362" />
   <aligned type="paragraph" l1="f:1-p:3" l2="f:2-p:3" certainty="1.32116" />
   <aligned type="paragraph" l1="f:1-p:4" l2="f:2-p:4" certainty="0.86915" />
   <aligned type="paragraph" l1="f:1-p:5" l2="f:2-p:5" certainty="1.71379" />
   <aligned type="paragraph" l1="f:1-p:6" l2="f:2-p:6" certainty="0.0396" />
   <aligned type="paragraph" l1="f:1-p:7" l2="f:2-p:7" certainty="1.01696" />
   <aligned type="paragraph" l1="f:1-p:8" l2="0" certainty="0.92197" />
   <aligned type="paragraph" l1="f:1-p:9" l2="f:2-p:8" certainty="1.19587" />
   <aligned type="paragraph" l1="f:1-p:10" l2="f:2-p:9" certainty="0.0396" />
   <aligned type="sentence" l1="f:1-p:1-s:1" l2="f:2-p:1-s:1" certainty="1.5842" />
   <aligned type="sentence" l1="f:1-p:2-s:2" l2="f:2-p:2-s:2" certainty="0.97109" />
   <aligned type="sentence" l1="f:1-p:3-s:3" l2="f:2-p:3-s:3" certainty="0.87273" />
   <aligned type="sentence" l1="f:1-p:4-s:4" l2="f:2-p:4-s:4" certainty="0.37771" />
   <aligned type="sentence" l1="f:1-p:5-s:5 f:1-p:5-s:6 f:1-p:5-s:7" l2="f:2-p:5-s:5 f:2-p:5-s:6"
    certainty="0.65372" />
   <aligned type="sentence" l1="f:1-p:6-s:8" l2="f:2-p:6-s:7" certainty="0.00671" />
   <aligned type="sentence" l1="f:1-p:6-s:9 f:1-p:6-s:10" l2="f:2-p:6-s:8" certainty="0.03123" />
   <aligned type="sentence" l1="f:1-p:6-s:11 f:1-p:6-s:12" l2="f:2-p:6-s:9" certainty="0.00078" />
   <aligned type="sentence" l1="f:1-p:7-s:13" l2="f:2-p:7-s:10" certainty="0.65525" />
   <aligned type="sentence" l1="f:1-p:7-s:14 f:1-p:7-s:15" l2="f:2-p:7-s:11" certainty="0.18297" />
   <aligned type="sentence" l1="f:1-p:7-s:16" l2="f:2-p:7-s:12 f:2-p:7-s:13" certainty="0.41161" />
   <aligned type="sentence" l1="f:1-p:8-s:17" l2="0" certainty="0.41161" />
   <aligned type="sentence" l1="f:1-p:9-s:18 f:1-p:9-s:19 :1-p:9-s:20 f:1-p:9-s:21" l2="f:2-p:8-s:14"
    certainty="0.02701" />
   <aligned type="sentence" l1="f:1-p:10-s:22" l2="f:2-p:9-s:15" certainty="0.00165" />
 </alignments>
 </alignment>
```

Figure B.3: Example of an Alignment File

Appendix C

DTDs

```
<!ELEMENT options (header,languageOne,languageTwo)>
<!ELEMENT header (name, directory, InputDictionary?, OutputDictionary?, OutputFormat, index?,
 sentenceAlignment?)>
<!ELEMENT OutputDictionary (file, directory, format, filter)>
<!ELEMENT InputDictionary (file, directory, format)>
<!ELEMENT OutputFormat (#PCDATA)>
<!ELEMENT filter (#PCDATA)>
<!ELEMENT directory (#PCDATA)>
<!ELEMENT index (#PCDATA)>
<!ELEMENT sentenceAlignment (files, format)>
<!ELEMENT files (#PCDATA)>
<!ELEMENT format (#PCDATA)>
<!ELEMENT languageOne (name, directory, files, format, parameters, structure?)>
<!ELEMENT languageTwo (name, directory, files, format, parameters, structure?)>
<!ELEMENT structure (annotation+)>
<!ELEMENT file (#PCDATA)>
<!ELEMENT parameters (#PCDATA)>
<!ELEMENT annotation (#PCDATA)>
<!ATTLIST files #PCDATA #REQUIRED >
<!ATTLIST languageOne #PCDATA #REQUIRED >
<!ATTLIST languageTwo #PCDATA #REQUIRED >
<!ATTLIST annotation type CDATA #REQUIRED >
<!ELEMENT name (#PCDATA)>
```

Figure C.1: Options DTD

```
<!ELEMENT dictionary (item*)>
<!ELEMENT item (lemma, category, language, translations)>
<!ELEMENT translations (translation+)>
<!ELEMENT translation (lemma, category, language, confidence)>
<!ELEMENT lemma (#PCDATA)>
<!ELEMENT category (#PCDATA)>
<!ELEMENT language (#PCDATA)>
<!ELEMENT confidence (#PCDATA)>
```

Figure C.2: Lexicon DTD

```
 <!ELEMENT corpus (p+ | s+)>
 <!ELEMENT p (s+)>
 <!ELEMENT s (terminal+|nonterminal*)>
 <!ELEMENT terminal (#PCDATA)>
 <!ELEMENT nonterminal (terminal|nonterminal)*>
 <!ATTLIST p id CDATA #REQUIRED>
 <!ATTLIST s id CDATA #REQUIRED>
 <!ATTLIST terminal id CDATA #REQUIRED
  lemma CDATA #REQUIRED
  category CDATA #REQUIRED
  morph CDATA #IMPLIED
 >
 <!ATTLIST nonterminal id CDATA #REQUIRED
  type CDATA #REQUIRED
 >
```

Figure C.3: Corpus DTD

```
<!ELEMENT alignment (header,alignments)>
<!ELEMENT header (l1,l2)>
<!ELEMENT l1 EMPTY>
<!ELEMENT l2 EMPTY>
<!ELEMENT alignments (aligned)+>
<!ELEMENT aligned EMPTY>
<!ATTLIST l1 lang CDATA #REQUIRED
  files CDATA #REQUIRED
  dir CDATA #REQUIRED
>
<!ATTLIST l2 lang CDATA #REQUIRED
  files CDATA #REQUIRED
  dir CDATA #REQUIRED
>
<!ATTLIST aligned type CDATA #REQUIRED
  l1 CDATA #REQUIRED
  l2 CDATA #REQUIRED
  certainty CDATA #REQUIRED
>
```

Figure C.4: Alignment DTD

```
<!ELEMENT parameters (lexClasses, funcClasses, varClasses)>
<!ELEMENT lexClasses (tag+)>
<!ELEMENT funcClasses (tag+)>
<!ELEMENT varClasses (tag+)>
<!ELEMENT tag (name,function)>
<!ELEMENT name (#PCDATA)>
<!ELEMENT function (#PCDATA)>
```

Figure C.5: Parameter DTD

Appendix D

Tagsets

D.1 English: Penn Treebank Tagset

category	generalized tag	original tag	description
lexical	adjective	CD	Cardinal number
	adjective	JJ	Adjective
	adjective	JJR	Adjective, comparative
	adjective	JJS	Adjective, superlative
	adjective	RB	Adverb
	adjective	RBR	Adverb, comparative
	adjective	RBS	Adverb, superlative
	adjective	WRB	WH-adverb
	name	NP	proper noun, singular
	name	NPS	proper noun, plural
	noun	NN	Noun, singular or mass
	noun	NNS	Noun, plural
	verb	MD	Modal
	verb	VB	Verb, base form
	verb	VBD	Verb, past tense
	verb	VBG	Verb, gerund or present participle
	verb	VBN	Verb, past participle
	verb	VBP	Verb, non-3rd person singular present
	verb	VBZ	Verb, 3rd person singular present
functional	conjunction	CC	Coordinating conjunction
	determiner	DT	Determiner
	determiner	PDT	predeterminer
	determiner	WDT	WH-determiner
	preposition	IN	Preposition or subordinating conjuction
	particle	RP	Particle
	particle	TO	*to*
	pronoun	EX	Existential *there*
	pronoun	POS	possessive ending
	pronoun	PP	personal pronoun
	pronoun	PP$	possessive pronoun
	pronoun	WP	WH-pronoun
	pronoun	WP$	possessive wh-pronoun
	foreign	FW	Foreign word
	ignore	LS	List item marker
	social noise	UH	Interjection
	symbol	SYM	Symbol

D.2 French: Tree-Tagger Tagset

category	generalized tag	original tag	description
lexical	adjective	ADJ	adjective
	adjective	ADV	adverb
	adjective	NUM	numeral
	name	NAM	proper name
	noun	NOM	noun
	verb	VER:cond	verb conditional
	verb	VER:futu	verb futur
	verb	VER:impe	verb imperative
	verb	VER:impf	verb imperfect
	verb	VER:infi	verb infinitive
	verb	VER:pper	verb past participle
	verb	VER:ppre	verb present participle
	verb	VER:pres	verb present
	verb	VER:simp	verb simple past
	verb	VER:subi	verb subjunctive imperfect
	verb	VER:subp	verb subjunctive present
	verb	VER:aux:subp	auxiliary verb subjunctive present
	verb	VER:aux:subi	auxiliary verb subjunctive imperfect
	verb	VER:aux:pres	auxiliary verb present
	verb	VER:aux:futu	auxiliary verb futur
	verb	VER:aux:impf	auxiliary verb imperfect
	verb	VER:aux:infi	auxiliary verb infinitive
	verb	VER:aux:simp	auxiliary verb simple past
	verb	VER:aux:cond	auxiliary verb conditional
	verb	VER:aux:impe	auxiliary verb imperative
	verb	VER:aux:ppre	auxiliary verb present participle
	verb	VER:aux:pper	auxiliary verb past participle
functional	conjunction	KON	conjunction
	determiner	DET:ART	determiner
	determiner	DET:POS	determiner
	preposition	PRP	preposition
	preposition	PRP:det	preposition plus article (au,du,aux,des)
	preposition	PRE:1st	preposition
	pronoun	PRO	pronoun
	pronoun	PRO:DEM	demonstrative pronoun
	pronoun	PRO:IND	indefinite pronoun
	pronoun	PRO:PER	personal pronoun
	pronoun	PRO:POS	possessive pronoun (mien, tien, ...)
	pronoun	PRO:REL	relative pronoun
	pronoun	DET:POS	possessive pronoun (ma, ta, ...)
various	citation	PUN:cit	punctuation citation
	ignore	ABR	abreviation
	sentence punctuation	PUN	punctuation
	sentence punctuation	SENT	sentence tag
	social noise	INT	interjection
	symbol	SYM	symbol

D.3 German: Stuttgart-Tübingen-Tagset

category	generalized tag	original tag	description
lexical	adjective	ADJA	attributive adjective
	adjective	ADJD	predicative adjective
	adjective	ADV	adverb
	adjective	CARD	number
	name	NE	proper noun
	noun	NN	noun
	verb	VAFIN	Verb, auxiliary finite
	verb	VAINF	Verb, auxiliary infinitive
	verb	VAIMP	Verb, auxiliary imperative
	verb	VAPP	Verb, auxiliary past participle
	verb	VVIZU	Verb, infinitive with *zu*
	verb	VMFIN	Verb, modal finite
	verb	VMINF	Verb, modal infinitive
	verb	VMPP	Verb, modal past participle
	verb	VVFIN	Verb, main finite
	verb	VVINF	Verb, main infinitive
	verb	VVIMP	Verb, main imperative
	verb	VVPP	Verb, main past participle
functional	conjunction	KON	conjunction, e.g. *und (and) , oder (or)*
	conjunction	KOUI	conjunction, subordinating with infinitive
	conjunction	KOUS	conjunction, subordinating with subordinate sentence
	determiner	ART	determiner
	determiner	PDAT	demonstrative determiner
	determiner	PIAT	indefinite determiner
	determiner	PIDAT	indefinite pronoun
	determiner	PPOSAT	possessive determiner
	determiner	PRELAT	Relative pronoun, attributive
	determiner	PWAT	Interrogative Pronoun
	preposition	APPR	preposition
	preposition	APPRART	preposition with determiner
	preposition	APPO	postposition
	preposition	APZR	circumposition
	preposition	PAV	pronominal adverb
	particle	PTKA	Particle next to adjective or adverb
	particle	PTKNEG	negation particle
	particle	PTKVZ	separable verb prefix
	particle	PTKZU	Particle *zu* next to infinitive
	pronoun	PDS	demonstrative pronoun
	pronoun	PIS	indefinite pronoun
	pronoun	PPER	personal pronoun
	pronoun	PPOSS	possessive pronoun
	pronoun	PRELS	Relative pronoun
	pronoun	PRF	Reflexive pronoun
	pronoun	PWAV	Interrogative Pronoun, adverbial or relative
	pronoun	PWS	Interrogative Pronoun
	comparison	KOKOM	comparison particle
various	citation	$(	sentence-internal punctuation: (), {}, []
	foreign	FM	foreign word
	ignore	$,	comma
	sentence punctuation	$.	sentence-final punctuation: . ; : ! ?
	social noise	ITJ	interjection
	social noise	PTKANT	answer particle
	symbol	XY	Symbol
	compound	TRUNC	word component

D.4 Italian: Tree-Tagger Tagset

category	generalized tag	original tag	description
lexical	adjective	ADJ	adjective
	adjective	ADV	adverb
	adjective	NUM	numeral
	name	NPR	name
	noun	NOM	noun
	verb	VER:cimp	verb conjunctive imperfect
	verb	VER:cond	verb conditional
	verb	VER:cpre	verb conjunctive present
	verb	VER:futu	verb future tense
	verb	VER:geru	verb gerund
	verb	VER:impe	verb imperative
	verb	VER:impf	verb imperfect
	verb	VER:infi	verb infinitive
	verb	VER:pper	verb participle perfect
	verb	VER:ppre	verb participle present
	verb	VER:pres	verb present
	verb	VER:refl:infi	verb reflexive infinitive
	verb	VER:remo	verb simple past
functional	conjunction	CON	conjunction
	determiner	DET:def	definite article
	determiner	DET:indef	indefinite article
	preposition	PRE	preposition
	preposition	PRE:det	preposition+article
	pronoun	PRO	pronoun
	pronoun	PRO:demo	demonstrative pronoun
	pronoun	PRO:indef	indefinite pronoun
	pronoun	PRO:inter	interrogative pronoun
	pronoun	PRO:pers	personal pronoun
	pronoun	PRO:poss	possessive pronoun
	pronoun	PRO:refl	reflexive pronoun
	pronoun	PRO:rela	relative pronoun
various	ignore	ABR	abbreviation
	sentence punctuation	PON	punctuation
	sentence punctuation	SENT	sentence marker
	social noise	INT	interjection
	symbol	SYM	symbol

D.5 Spanish: Tree-Tagger Tagset

category	generalized tag	original tag	description
lexical	noun	PNC	noun
	noun	NC	Common nouns (mesas, mesa, libro, ordenador)
	name¡	NP	Proper nouns
	adjective	ORD	Ordinals (primer, primeras, primera)
	adjective	ADJ	Adjectives (mayores, mayor)
	adjective	ADV	Adverbs (muy, demasiado, cÂ´omo)
	adjective	CARD	Cardinals
	verb	VEfin	Verb estar. Finite
	verb	VEger	Verb estar. Gerund
	verb	VEinf	Verb estar. Infinitive
	verb	VE	Verb estar. Past participle
	verb	VHfin	Verb haber. Finite
	verb	VHger	Verb haber. Gerund
	verb	VHinf	Verb haber. Infinitive
	verb	VHadj	Verb haber. Past participle
	verb	VLfin	Lexical verb. Finite
	verb	VLger	Lexical verb. Gerund
	verb	VLinf	Lexical verb. Infinitive
	verb	VLadj	Lexical verb. Past participle
	verb	VMfin	Modal verb. Finite
	verb	VMger	Modal verb. Gerund
	verb	VMinf	Modal verb. Infinitive
	verb	VM	Modal verb. Past participle
	verb	VSfin	Verb ser. Finite
	verb	VSger	Verb ser. Gerund
	verb	VSinf	Verb ser. Infinitive
	verb	VS	Verb ser. Past participle
functional	determiner	ART	Articles (un, las, la, unas)
	determiner	QU	Quantifiers (sendas, cada)
	conjunction	CSUBF	Subordinating conjunction that introduces finite clauses (apenas)
	conjunction	CSUBI	Subordinating conjunction that introduces infinite clauses (al)
	conjunction	CSUBX	Subordinating conjunction underspecified for subord-type (aunque)
	conjunction	CC	Coordinating conjunction (y, o)
	conjunction	CCAD	Adversative coordinating conjunction (pero)
	preposition	CQUE	que (as conjunction)
	preposition	PREP	Negative preposition (sin)
	preposition	PREP	Preposition
	preposition	PAL	Portmanteau word formed by a and el (preposition + det?)
	preposition	PDEL	Portmanteau word formed by de and el (preposition + det?)
	preposition	ALFP	Plural letter of the alphabet (As/Aes, bes)
	preposition	ALFS	Singular letter of the alphabet (A, b)
	particle	CCNEG	Negative coordinating conjunction (ni)
	particle	NEG	Negation
	particle	SE	Se (as particle)
	pronoun	DM	Demonstrative pronouns (Â´esas, Â´ese, esta)
	pronoun	INT	Interrogative pronouns (quiÂ´enes, cuÂ´antas, cuÂ´anto)
	pronoun	PPX	Clitics and personal pronouns (nos, me, nosotras, te)
	pronoun	PPO	Possessive pronouns (tuyas, tuya)
	pronoun	REL	Relative pronouns (cuyas, cuyo)
various	punctuation	FS	Full stop punctuation marks
	symbol	SYM	Symbols
	ignore	CODE	Alphanumeric code
	ignore	FO	Formula
	interjection	ITJN	Interjection (oh, ja)
	foreign	PE	Foreign word

D.6 Swedish: SUC Tagset

category	generalized tag	original tag	description
lexical	name	PM	name
	noun	NN	noun
	adjective	JJ	adjective
	adjective	AB	adverb
	adjective	RG	cardinal number
	adjective	RO	ordinal number
	verb	VBINF	
	verb	VBFIN	
	verb	VBIMP	
	verb	VBSUP	supinum
	verb	PL	verb particle
	participle	PC	verb, participle
functional	determiner	DT	determiner
	preposition	PR	
	preposition	PP	preposition
	particle	IE	infinitival marker
	pronoun	PN	pronoun
	pronoun	HA	interrogative,relative adverb
	pronoun	HD	interrogative,relative determiner
	pronoun	HP	interrogative,relative pronoun
	pronoun	HS	interrogative,relative possessive
	pronoun	PS	possessive
	conjunction	KN	conjunction
	conjunction	SN	subjunction
	social noise	IN	interjection
	particle	IE	infinitive marker
various	sentence	DL	sentence delimiter
	symbol	XY	symbol
	foreign	UO	foreign word

Appendix E

Parameter Files

E.1 English Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for English, Bettina Schrader, 10.12.2004-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>
<lexClasses>
<tag><name>NN</name><function>noun</function></tag>
<tag><name>NNS</name><function>noun</function></tag>
<tag><name>NP</name><function>name</function></tag>
<tag><name>NPS</name><function>name</function></tag>
<tag><name>JJ</name><function>adjective</function></tag>
<tag><name>JJR</name><function>adjective</function></tag>
<tag><name>JJS</name><function>adjective</function></tag>
<tag><name>CD</name><function>adjective</function></tag>
<tag><name>RB</name><function>adjective</function></tag>
<tag><name>RBR</name><function>adjective</function></tag>
<tag><name>RBS</name><function>adjective</function></tag>
<tag><name>WRB</name><function>adjective</function></tag>
<tag><name>MD</name><function>verb</function></tag>
<tag><name>VB</name><function>verb</function></tag>
<tag><name>VBD</name><function>verb</function></tag>
<tag><name>VBG</name><function>verb</function></tag>
<tag><name>VBN</name><function>verb</function></tag>
<tag><name>VBP</name><function>verb</function></tag>
<tag><name>VBZ</name><function>verb</function></tag>
<tag><name>VVH</name><function>verb</function></tag>
<tag><name>VH</name><function>verb</function></tag>
<tag><name>VHD</name><function>verb</function></tag>
<tag><name>VHZ</name><function>verb</function></tag>
<tag><name>VVP</name><function>verb</function></tag>
<tag><name>VVD</name><function>verb</function></tag>
<tag><name>VVN</name><function>verb</function></tag>
<tag><name>VVG</name><function>verb</function></tag>
<tag><name>VHP</name><function>verb</function></tag>
<tag><name>VHN</name><function>verb</function></tag>
<tag><name>VHG</name><function>verb</function></tag>
<tag><name>VV</name><function>verb</function></tag>
<tag><name>VVZ</name><function>verb</function></tag>
</lexClasses>
<funcClasses>
<tag><name>IN</name><function>preposition</function></tag>
<tag><name>DT</name><function>determiner</function></tag>
<tag><name>PDT</name><function>determiner</function></tag>
<tag><name>WDT</name><function>determiner</function></tag>
<tag><name>EX</name><function>pronoun</function></tag>
<tag><name>PP</name><function>pronoun</function></tag>
<tag><name>PP$</name><function>pronoun</function></tag>
<tag><name>POS</name><function>pronoun</function></tag>
<tag><name>WP$</name><function>pronoun</function></tag>
```

```
<tag><name>WP</name><function>pronoun</function></tag>
<tag><name>CC</name><function>conjunction</function></tag>
<tag><name>RP</name><function>particles</function></tag>
<tag><name>TO</name><function>particles</function></tag>
</funcClasses>
<varClass>
<tag><name>LS</name><function>ignore</function></tag>
<tag><name>,</name><function>ignore</function></tag>
<tag><name>:</name><function>ignore</function></tag>
<tag><name>(</name><function>citation</function></tag>
<tag><name>)</name><function>citation</function></tag>
<tag><name>’’</name><function>citation</function></tag>
<tag><name>‘‘</name><function>citation</function></tag>
<tag><name>SENT</name><function>sentence punctuation</function></tag>
<tag><name>SYM</name><function>symbol</function></tag>
<tag><name>$</name><function>symbol</function></tag>
<tag><name>FW</name><function>foreign</function></tag>
<tag><name>UH</name><function>interjection</function></tag>
</varClass>
</parameters>
```

E.2 French Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for French, Bettina Schrader, 1.3.2005-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>
<lexClasses>
<tag><name>NOM</name><function>noun</function></tag>
<tag><name>NAM</name><function>name</function></tag>
<tag><name>ADJ</name><function>adjective</function></tag>
<tag><name>ADV</name><function>adjective</function></tag>
<tag><name>NUM</name><function>adjective</function></tag>
<tag><name>VER:cond</name><function>verb</function></tag>
<tag><name>VER:futu</name><function>verb</function></tag>
<tag><name>VER:impe</name><function>verb</function></tag>
<tag><name>VER:impf</name><function>verb</function></tag>
<tag><name>VER:infi</name><function>verb</function></tag>
<tag><name>VER:pper</name><function>verb</function></tag>
<tag><name>VER:ppre</name><function>verb</function></tag>
<tag><name>VER:pres</name><function>verb</function></tag>
<tag><name>VER:simp</name><function>verb</function></tag>
<tag><name>VER:subi</name><function>verb</function></tag>
<tag><name>VER:subp</name><function>verb</function></tag>
<tag><name>VER:aux:subp</name><function>verb</function></tag>
<tag><name>VER:aux:pres</name><function>verb</function></tag>
<tag><name>VER:aux:futu</name><function>verb</function></tag>
<tag><name>VER:aux:impf</name><function>verb</function></tag>
<tag><name>VER:aux:infi</name><function>verb</function></tag>
<tag><name>VER:aux:simp</name><function>verb</function></tag>
<tag><name>VER:aux:cond</name><function>verb</function></tag>
<tag><name>VER:aux:impe</name><function>verb</function></tag>
<tag><name>VER:aux:ppre</name><function>verb</function></tag>
<tag><name>VER:aux:pper</name><function>verb</function></tag>
<tag><name>VER:aux:subi</name><function>verb</function></tag>
</lexClasses>

<funcClasses>
<tag><name>PRP</name><function>preposition</function></tag>
<tag><name>PRP:det</name><function>preposition</function></tag>
<tag><name>PRE:1st</name><function>preposition</function></tag>
<tag><name>DET:ART</name><function>determiner</function></tag>
<tag><name>DET:POS</name><function>pronoun</function></tag>
<tag><name>PRO</name><function>pronoun</function></tag>
<tag><name>PRO:DEM</name><function>pronoun</function></tag>
<tag><name>PRO:IND</name><function>pronoun</function></tag>
<tag><name>PRO:PER</name><function>pronoun</function></tag>
<tag><name>PRO:POS</name><function>pronoun</function></tag>
```

```
<tag><name>PRO:REL</name><function>pronoun</function></tag>
<tag><name>KON</name><function>conjunction</function></tag>
</funcClasses>

<!-- collect all tags that are seldom or serve specialized purposes -->
<!-- all tags should be present in the second language parameter file, too. -->
<varClass>
<tag><name>ABR</name><function>ignore</function></tag>
<tag><name>INT</name><function>interjection</function></tag>
<tag><name>PUN</name><function>sentence punctuation</function></tag>
<tag><name>SENT</name><function>sentence punctuation</function></tag>
<tag><name>PUN:cit</name><function>citation</function></tag>
<tag><name>SYM</name><function>symbol</function></tag>
</varClass>
</parameters>
```

E.3 German Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for German, Bettina Schrader, 10.12.2004-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>
<lexClasses>
   <tag><name>NN</name><function>noun</function></tag>
   <tag><name>NE</name><function>name</function></tag>
   <tag><name>ADJA</name><function>adjective</function></tag>
   <tag><name>ADJD</name><function>adjective</function></tag>
   <tag><name>ADV</name><function>adjective</function></tag>
   <tag><name>CARD</name><function>adjective</function></tag>
   <tag><name>VAFIN</name><function>verb</function></tag>
   <tag><name>VAINF</name><function>verb</function></tag>
   <tag><name>VAIMP</name><function>verb</function></tag>
   <tag><name>VVFIN</name><function>verb</function></tag>
   <tag><name>VVINF</name><function>verb</function></tag>
   <tag><name>VVIMP</name><function>verb</function></tag>
   <tag><name>VMFIN</name><function>verb</function></tag>
   <tag><name>VMINF</name><function>verb</function></tag>
   <tag><name>VVPP</name><function>verb</function></tag>
   <tag><name>VAPP</name><function>verb</function></tag>
   <tag><name>VMPP</name><function>verb</function></tag>
   <tag><name>VVIZU</name><function>verb</function></tag>
   <tag><name>PTKVZ</name><function>verb</function></tag>
</lexClasses>
<!--   -->
<funcClasses>
   <tag><name>APPR</name><function>preposition</function></tag>
   <tag><name>APPRART</name><function>preposition</function></tag>
   <tag><name>APPO</name><function>preposition</function></tag>
   <tag><name>APZR</name><function>preposition</function></tag>
   <tag><name>PAV</name><function>preposition</function></tag>
   <tag><name>ART</name><function>determiner</function></tag>
   <tag><name>PPOSAT</name><function>determiner</function></tag>
   <tag><name>PDAT</name><function>determiner</function></tag>
   <tag><name>PIDAT</name><function>determiner</function></tag>
   <tag><name>PIAT</name><function>determiner</function></tag>
   <tag><name>PRELAT</name><function>determiner</function></tag>
   <tag><name>PWAT</name><function>determiner</function></tag>
   <tag><name>PPER</name><function>pronoun</function></tag>
   <tag><name>PRF</name><function>pronoun</function></tag>
   <tag><name>PPOSS</name><function>pronoun</function></tag>
   <tag><name>PDS</name><function>pronoun</function></tag>
   <tag><name>PIS</name><function>pronoun</function></tag>
   <tag><name>PRELS</name><function>pronoun</function></tag>
   <tag><name>PWS</name><function>pronoun</function></tag>
   <tag><name>PWAV</name><function>pronoun</function></tag>
   <tag><name>KOUI</name><function>conjunction</function></tag>
   <tag><name>KOUS</name><function>conjunction</function></tag>
   <tag><name>KON</name><function>conjunction</function></tag>
   <tag><name>PTKANT</name><function>social noise</function></tag>
   <tag><name>ITJ</name><function>social noise</function></tag>
   <tag><name>PTKVZ</name><function>particle</function></tag>
   <tag><name>PTKZV</name><function>particle</function></tag>
   <tag><name>PTKZU</name><function>particle</function></tag>
   <tag><name>PTKNEG</name><function>particle</function></tag>
   <tag><name>PTKA</name><function>particle</function></tag>
</funcClasses>
<!-- collect all names that are seldom or serve specialized purposes -->
<!-- all names should be present in the second language parameter file, too. -->
<varClass>
<tag><name>KOKOM</name><function>comparison</function></tag>
<tag><name>$,</name><function>ignore</function></tag>
<tag><name>$(</name><function>citation</function></tag>
<tag><name>$.</name><function>sentence</function></tag>
<tag><name>TRUNC</name><function>component</function></tag>
<tag><name>XY</name><function>symbol</function></tag>
<tag><name>FM</name><function>foreign</function></tag>
<tag><name>FW</name><function>foreign</function></tag>
</varClass>
</parameters>
```

E.4 Italian Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for German, Bettina Schrader, 2.3.2005-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>

<lexClasses>
<tag><name>NOM</name><function>noun</function></tag>
<tag><name>NPR</name><function>name</function></tag>
<tag><name>ADJ</name><function>adjective</function></tag>
<tag><name>ADV</name><function>adjective</function></tag>
<tag><name>NUM</name><function>adjective</function></tag>
<tag><name>VER:cimp</name><function>verb</function></tag>
<tag><name>VER:cond</name><function>verb</function></tag>
<tag><name>VER:cpre</name><function>verb</function></tag>
<tag><name>VER:futu</name><function>verb</function></tag>
<tag><name>VER:geru</name><function>verb</function></tag>
<tag><name>VER:impe</name><function>verb</function></tag>
<tag><name>VER:impf</name><function>verb</function></tag>
<tag><name>VER:infi</name><function>verb</function></tag>
<tag><name>VER:pper</name><function>verb</function></tag>
<tag><name>VER:ppre</name><function>verb</function></tag>
<tag><name>VER:pres</name><function>verb</function></tag>
<tag><name>VER:refl:infi</name><function>verb</function></tag>
<tag><name>VER:remo</name><function>verb</function></tag>
</lexClasses>

<funcClasses>
<tag><name>PRE</name><function>preposition</function></tag>
<tag><name>PRE:det</name><function>preposition</function></tag>
<tag><name>DET:def</name><function>determiner</function></tag>
<tag><name>DET:indef</name><function>determiner</function></tag>
<tag><name>PRO</name><function>pronoun</function></tag>
<tag><name>PRO:demo</name><function>pronoun</function></tag>
<tag><name>PRO:indef</name><function>pronoun</function></tag>
<tag><name>PRO:inter</name><function>pronoun</function></tag>
<tag><name>PRO:pers</name><function>pronoun</function></tag>
<tag><name>PRO:poss</name><function>pronoun</function></tag>
<tag><name>PRO:refl</name><function>pronoun</function></tag>
<tag><name>PRO:rela</name><function>pronoun</function></tag>
<tag><name>CON</name><function>conjunction</function></tag>
</funcClasses>

<!-- collect all tags that are seldom or serve specialized purposes -->
<!-- all tags should be present in the second language parameter file, too. -->
```

```
<varClass>
<tag><name>ABR</name><function>ignore</function></tag>
<tag><name>INT</name><function>interjection</function></tag>
<tag><name>PON</name><function>sentence punctuation</function></tag>
<tag><name>SENT</name><function>sentence punctuation</function></tag>
<tag><name>SYM</name><function>symbol</function></tag>
</varClass>
</parameters>
```

E.5 Spanish Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for Spanish, Bettina Schrader, 4.5.2006-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>
<lexClasses>
<tag><name>PNC</name><function>noun</function></tag>      <!-- Unclassified word (probably noun?)-->
<tag><name>ORD</name><function>adjective</function></tag> <!-- Ordinals (primer, primeras, primera)-->
<tag><name>ADJ</name><function>adjective</function></tag> <!-- Adjectives (mayores, mayor)-->
<tag><name>ADV</name><function>adjective</function></tag> <!-- Adverbs (muy, demasiado, cÂ´omo)-->
<tag><name>CARD</name><function>adjective</function></tag><!-- Cardinals-->
<tag><name>NC</name><function>noun</function></tag>       <!-- Common nouns (mesas, mesa, libro, ordenador)-->
<tag><name>NP</name><function>name</function></tag>       <!-- Proper nouns-->
<tag><name>VEfin</name><function>verb</function></tag>    <!-- Verb estar. Finite-->
<tag><name>VEger</name><function>verb</function></tag>    <!-- Verb estar. Gerund-->
<tag><name>VEinf</name><function>verb</function></tag>    <!-- Verb estar. Infinitive-->
<tag><name>VE</name><function>verb</function></tag>       <!-- Verb estar. Past participle-->
<tag><name>VHfin</name><function>verb</function></tag>    <!-- Verb haber. Finite-->
<tag><name>VHger</name><function>verb</function></tag>    <!-- Verb haber. Gerund-->
<tag><name>VHinf</name><function>verb</function></tag>    <!-- Verb haber. Infinitive-->
<tag><name>VHadj</name><function>verb</function></tag>    <!-- Verb haber. Past participle-->
<tag><name>VLfin</name><function>verb</function></tag>    <!-- Lexical verb. Finite-->
<tag><name>VLger</name><function>verb</function></tag>    <!-- Lexical verb. Gerund-->
<tag><name>VLinf</name><function>verb</function></tag>    <!-- Lexical verb. Infinitive-->
<tag><name>VLadj</name><function>verb</function></tag>  <!-- Lexical verb. Past participle-->
<tag><name>VMfin</name><function>verb</function></tag>  <!-- Modal verb. Finite-->
<tag><name>VMger</name><function>verb</function></tag>  <!-- Modal verb. Gerund-->
<tag><name>VMinf</name><function>verb</function></tag>  <!-- Modal verb. Infinitive-->
<tag><name>VM</name><function>verb</function></tag>       <!-- Modal verb. Past participle-->
<tag><name>VSfin</name><function>verb</function></tag>  <!-- Verb ser. Finite-->
<tag><name>VSger</name><function>verb</function></tag>  <!-- Verb ser. Gerund-->
<tag><name>VSinf</name><function>verb</function></tag>  <!-- Verb ser. Infinitive-->
<tag><name>VS</name><function>verb</function></tag>  <!-- Verb ser. Past participle-->
</lexClasses>

<funcClasses>
<tag><name>ART</name><function>determiner</function></tag> <!-- Articles (un, las, la, unas) -->
<tag><name>CSUBF</name><function>conjunction</function></tag> <!-- Subordinating conjunction, finite clauses (apenas) -->
<tag><name>CSUBI</name><function>conjunction</function></tag> <!-- Subordinating conjunction, infinite clauses (al) -->
<tag><name>CSUBX</name><function>conjunction</function></tag> <!-- Subordinating conjunction underspecified (aunque) -->
<tag><name>CC</name><function>conjunction</function></tag> <!-- Coordinating conjunction (y, o) -->
<tag><name>CCAD</name><function>conjunction</function></tag> <!-- Adversative coordinating conjunction (pero) -->
<tag><name>CCNEG</name><function>particle</function></tag> <!-- Negative coordinating conjunction (ni) -->
<tag><name>CQUE</name><function>preposition</function></tag> <!-- que (as conjunction) -->
<tag><name>DM</name><function>pronoun</function></tag> <!-- Demonstrative pronouns (Â´esas, Â´ese, esta) -->
<tag><name>INT</name><function>pronoun</function></tag> <!-- Interrogative pronouns (quiÂ´enes, cuÂ´antas, cuÂ´anto) -->
<tag><name>NEG</name><function>particle</function></tag> <!-- Negation -->
<tag><name>PPX</name><function>pronoun</function></tag> <!-- Clitics and personal pronouns (nos, me, nosotras, te) -->
<tag><name>PPO</name><function>pronoun</function></tag><!-- Possessive pronouns (tuyas, tuya) -->
<tag><name>PPO</name><function>pronoun</function></tag><!-- Possessive pronouns (tuyas, tuya) -->
<tag><name>PREP</name><function>preposition</function></tag><!-- Negative preposition (sin) -->
<tag><name>QU</name><function>determiner</function></tag><!-- Quantifiers (sendas, cada) -->
<tag><name>REL</name><function>pronoun</function></tag><!-- Relative pronouns (cuyas, cuyo) -->
<tag><name>PREP</name><function>preposition</function></tag><!-- Preposition -->
<tag><name>PAL</name><function>preposition</function></tag>
     <!-- Portmanteau word formed by a and el (preposition + det?) -->
<tag><name>PDEL</name><function>preposition</function></tag>
     <!-- Portmanteau word formed by de and el (preposition + det?)-->
```

```
<tag><name>ALFP</name><function>preposition</function></tag>
     <!-- Plural letter of the alphabet (As/Aes, bes) -->
<tag><name>ALFS</name><function>preposition</function></tag>
     <!-- Singular letter of the alphabet (A, b) -->
<tag><name>SE</name><function>particle</function></tag><!-- Se (as particle) -->
</funcClasses>

<!-- collect all tags that are seldom or serve specialized purposes -->
<!-- all tags should be present in the second language parameter file, too. -->
<varClass>
<tag><name>FS</name><function>sentence punctuation</function></tag>
     <!-- Full stop punctuation marks-->
<tag><name>SYM</name><function>symbol</function></tag><!-- Symbols-->
<tag><name>CODE</name><function>ignore</function></tag><!-- Alphanumeric code-->
<tag><name>FO</name><function>ignore</function></tag><!-- Formula-->
<tag><name>ITJN</name><function>interjection</function></tag><!-- Interjection (oh, ja)-->
<tag><name>PE</name><function>foreign</function></tag><!-- Foreign word-->
</varClass>
</parameters>
```

E.6 Swedish Parameter File

```
<?xml version="1.0" encoding="ISO-8859-1" standalone="yes" ?>

<!-- Parameters for Swedish, Bettina Schrader, 05.10.2006-->
<!-- Note that all Classes should be defined for all languages used -->
<parameters>
<lexClasses>
   <tag><name>PM</name><function>name</function></tag>
   <tag><name>NN</name><function>noun</function></tag>
   <tag><name>JJ</name><function>adjective</function></tag>
   <tag><name>AB</name><function>adjective</function></tag> <!-- adverb -->
   <tag><name>RG</name><function>adjective</function></tag> <!-- cardinal number -->
   <tag><name>RO</name><function>adjective</function></tag> <!-- ordinal number -->
   <tag><name>VBINF</name><function>verb</function></tag>
   <tag><name>VBFIN</name><function>verb</function></tag>
   <tag><name>VBIMP</name><function>verb</function></tag>
   <tag><name>VBSUP</name><function>verb</function></tag> <!-- supinum -->
   <tag><name>PL</name><function>verb</function></tag>
   <tag><name>PC></name><function>participle</function></tag><!-- possessive-->
</lexClasses>
<!--   -->
<funcClasses>
   <tag><name>PR</name><function>preposition</function></tag>
   <tag><name>PP</name><function>preposition</function></tag>
   <tag><name>DT</name><function>determiner</function></tag>
   <tag><name>PN</name><function>pronoun</function></tag>
   <tag><name>HA</name><function>pronoun</function></tag><!-- interrogative,relative adverb-->
   <tag><name>HD</name><function>pronoun</function></tag><!-- interrogative,relative determiner-->
   <tag><name>HP</name><function>pronoun</function></tag><!--  interrogative,relative pronoun-->
   <tag><name>HS</name><function>pronoun</function></tag><!--  interrogative,relative possessive -->
   <tag><name>PS</name><function>pronoun</function></tag><!-- possessive-->
   <tag><name>IE</name><function>particle</function></tag><!-- infinitival marker-->
   <tag><name>KN</name><function>conjunction</function></tag>
   <tag><name>SN</name><function>conjunction</function></tag> <!-- subjunction -->
   <tag><name>IN</name><function>social noise</function></tag><!-- interjection -->
   <tag><name>IE</name><function>particle</function></tag><!-- infinitive marker -->
</funcClasses>
<!-- collect all names that are seldom or serve specialized purposes -->
<!-- all names should be present in the second language parameter file, too. -->
<varClass>
   <tag><name>DL</name><function>sentence</function></tag>
   <tag><name>XY</name><function>symbol</function></tag>
   <tag><name>UO</name><function>foreign</function></tag>
</varClass>
</parameters>
```

References

Ahrenberg, L., M. Andersson, and M. Merkel (1998). A simple hybrid aligner for generating lexical correspondences in parallel texts. In *Proceedings of the 19th international conference on computational linguistics and 36th annual meeting of the Association for computational linguistics (Coling/ACL)*, Montreal, Canada, pp. 29–35.

Ahrenberg, L., M. Merkel, A. S. Hein, and J. Tiedemann (1999). Evaluation of LWA and UWA. Report from the PLUG project, available as Working Paper No. 15 in Working Papers in Computational Linguistics and Language Engineering. Department of Linguistics, Uppsala University, Sweden.

Ahrenberg, L., M. Merkel, A. S. Hein, and J. Tiedemann (2000). Evaluation of word alignment systems. In *Proceedings of the Second International Conference on Linguistic Resources and Evaluation (LREC)*, Athens, Greece, pp. 1255–1261.

Al-Onaizan, Y., J. Curin, M. Jahr, K. Knight, J. Lafferty, I. D. Melamed, F. J. Och, D. Prudy, N. A. Smith, and D. Yarowsky (1999). Statistical machine translation. Technical report, Johns Hopkins University 1999 Summer Workshop on Language Engineering, Center for Language and Speech Processing (CLSP), Baltimore, MD, USA.

Bannard, C. and C. Callison-Burch (2005). Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, USA, pp. 597–604.

Bentivogli, L. and E. Pianta (2004). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus. *Natural language engineering 11*(3), 247–261.

Böhmová, A., J. Hajič, E. Hajičová, and B. Hladká (2001). The Prague Dependency Treebank: Three-level annotation scenario. In A. Abeillé (Ed.), *Treebanks: Building and Using Syntactically Annotated Corpora*, Chapter 7, pp. 103–125. Kluwer Academic Publishers.

Bojar, O. and M. Prokopová (2006). Czech-English word alignment. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, pp. 1236–1239.

Borin, L. (2000). You'll take the high road and I'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany, pp. 97–103.

Brants, S., S. Dipper, S. Hansen, W. Lezius, and G. Smith (2002). The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, Bulgaria, pp. 24–41.

Brown, P. F., J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin (1990). A statistical approach to machine translation. *Computational Lingusitics 16*(2), 79–85.

Brown, P. F., S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer (1993). The mathematics of machine translation: Parameter estimation. *Computational Linguistics 19*(2), 263–311.

Brown, P. F., J. C. Lai, and R. L. Mercer (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp. 169–176.

Brown, R. D., J. D. Kim, P. J. Jansen, and J. G. Carbonell (2005). Symmetric probabilistic alignment. In *Proceedings of the ACL workshop on building and using parallel texts*, Ann Arbor, Michigan, USA, pp. 87–90.

Callison-Burch, C. and M. Osborne (2003). Bootstrapping parallel corpora. In *Proceedings of the NLT-NAACL workshop "Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, pp. 44–49.

Callison-Burch, C., D. Talbot, and M. Osborne (2004). Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Meeting of the Association for computational linguistics (ACL)*, Barcelona, Spain, pp. 175–182.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics 22*(2), 249–254.

Ceauşu, A., D. Ştefănescu, and D. Tufiş (2006). Acquis communautaire sentence alignment using support vector machines. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, pp. 2134–2137.

Chang, J. S. and M. H. Chen (1997). An alignment method for noisy parallel corpora based on image processing techniques. In *Proceedings of the 35th Annual Meeting of the Association for Computational LinguisticsACL proceedings*, Madrid, Spain, pp. 297–304.

Chen, S. F. (1993). Aligning sentences in bilingual corpora using lexical information. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 9–16.

Cherry, C. and D. Lin (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Sapporo, Japan, pp. 88–95.

Chiao, Y.-C., O. Kraif, D. Laurent, T. M. H. Nguyen, N. Semmar, F. Stuck, J. Véronis, and W. Zaghouani (2006). Evaluation of multilingual text alignment systems: the ARCADE II project. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, pp. 1975–1978.

Church, K. W. (1993). Char_align: A program for aligning parallel texts at the character level. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 1–8.

Conley, E. S. (2002). Seq_align: A parsing-independent bilingual sequence alignment algorithm. Master's thesis, Department of Computer Science, Bar-Ilan University, Ramat-Gan, Israel.

Covington, M. (1996). An algorithm to align words for historical comparison. *Computational Linguistics 22*(4), 481–496.

Dagan, I., K. Church, and W. Gale (1993). Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora (WVLC)*, Ohio, USA, pp. 1–8.

de Gispert, A. and J. B. Marino (2005). Linguistic knowledge in statistical phrase-based word alignment. *Natural language engineering 12*(1), 91–108.

Debili, F. and E. Sammouda (1992). Appariement des phrases de textes bilingues français-anglais et français-arabes. In *Proceedings of the 14th International Conference on Computational Linguistics (COLING)*, Nantes, France, pp. 517–524.

Dejean, H., E. Gaussier, C. Goutte, and K. Yamada (2003). Reducing parameter space for word alignment. In R. Mihalcea and T. Pedersen (Eds.), *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, pp. 23–26.

Ding, Y., D. Gildea, and M. Palmer (2003). An algorithm for word-level alignment of parallel dependency trees. In *Proceedings of the 9th Machine Translation Summit*, New Orleans, USA, pp. 95–101.

Drábek, E. F. and D. Yarowsky (2004). Improving bitext word alignments via syntax-based reordering of english. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, pp. 146–149.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics 19*(1), 61–73.

Dyvik, H. (2004). Translations as semantic mirrors: from parallel corpus to wordnet. *Language and Computers 49*(1), 311–326.

EAGLES (2000). Corpus encoding standard. http://www.cs.vassar.edu/CES/. Version 1.5.

Eisenberg, P., J. Peters, P. Gallmann, C. Fabricius-Hansen, D. Nübling, I. Bartz, T. A. Fintz, and R. Fiehler (2005). *Duden. Die Grammatik* (7 ed.), Volume 4 of *Der Duden in zwölf Bänden*. Mannheim, Leipzig, Wien, Zürich: Duden.

Ejerhed, E., G. Källgren, O. Wennstedt, and M. Åström (1992). The linguistic annotation system of the Stockholm-Umeå corpus project. Technical report, Department of General Linguistics, University of Umeå.

Erjavec, T., C. Ignat, B. Pouliquen, and R. Steinberger (2005). Massive multilingual corpus compilation: Acquis Communautaire and totale. *Journal Archives of Control Sciences 15*(4), 529–540.

Evert, S. and B. Krenn (2001). Methods for the qualitative evaluation of lexical association measures. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, Toulouse, France, pp. 188–195.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis*, pp. 1–32. Oxford: The Philological Society.

Foster, G., S. Gandrabur, P. Langlais, P. Plamondon, G. Russell, and M. Simard (2003). Statistical machine translation: Rapid development with limited resources. In *Proceedings of the 9th Machine Translation Summit*, New Orleans, USA, pp. 110–119.

Fox, H. J. (2005). Dependency-based statistical machine translation. In *Proceedings of the Student Research Workshop at the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, Ann Arbor, Michigan, USA, pp. 91–96.

Fraser, A. and D. Marcu (2006, July). Semi-supervised training for statistical word alignment. In *Proceedings of the 21st international conference on computational lingusitics and 44th annual meeting of the Association for computational linguistics (Coling/ACL)*, Sydney, Australia, pp. 769–776. Association for Computational Linguistics.

Fung, P. and K. W. Church (1994). K-vec: a new approach for aligning parallel texts. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING)*, Kyoto, Japan, pp. 1096–1102.

Fung, P. and K. McKeown (1994). Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. In *Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA)*, Columbia, Maryland, USA, pp. 81–88.

Gale, W. A. and K. W. Church (1991a). Identifying word correspondences in parallel texts. In *Fourth DARPA workshop on speech and natural language*, Asimolar, California, USA, pp. 152–157.

Gale, W. A. and K. W. Church (1991b). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, California, USA, pp. 177–184. Reprinted 1993 in Computational Linguistics.

Gale, W. A., K. W. Church, and D. Yarowsky (1992). A method for disambiguating word senses in a large corpus. *Computers and the Humanities 26*(5-6), 415–439.

Germann, U. (2001). Aligned hansards of the 36th parliament of Canada – release 2001-1a. `http://www.isi.edu/natural-language/download/hansard/`.

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

Goutte, C., K. Yamada, and E. Gaussier (2004). Aligning words using matrix factorisation. In *Proceedings of the 42nd Meeting of the Association for computational linguistics (ACL)*, Barcelona, Spain, pp. 502–509.

Grappin, P. (Ed.) (1990). *Standardwörterbuch Deutsch – Französisch. Dictionnaire Général Français – Allemand*. Paris: Larousse.

Haapalainen, M. and A. Majorin (1994). GERTWOL*: Ein System zur automatischen Wortformerkennung deutscher Wörter*. Lingsoft, Inc.

Hansen-Schirra, S., S. Neumann, and M. Vela (2006). Multi-dimensional annotation and alignment in an English-German translation corpus. In *Proceedings of the workshop on Multidimensional Markup in Natural Language Processing (NLPXML-2006) held at the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Trento, Italy, pp. 35–42.

Harris, B. (1988). Bi-text, a new concept in translation theory. *Language Monthly 54*, 8–10.

Haruno, M. and T. Yamazaki (1996). High-performance bilingual text alignment using statistical and dictionary information. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pp. 131–138.

Henderson, J. C. (2003). Word alignment baselines. In *Proceedings of the HLT-NAACL 2003 Workshop on building and using parallel texts: data driven machine translation and beyond*, Edmonton, Canada, pp. 27–30.

Hiemstra, D. (1996). Using statistical methods to create a bilingual dictionary. Master's thesis, Universiteit Twente.

Hiemstra, D. (1998). Multilingual domain modeling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. In P. Coppen, H. van Halteren, and L. Teunissen (Eds.), *Proceedings of the eighth CLIN Meeting*, pp. 41–58.

Hiemstra, D., F. de Jong, and W. Kaaij (1997). A domain-specific lexicon acquisition tool for cross-language information retrieval. In *Proceedings of the RIAO '97 Conference on computer-assisted information searching on Internet*, Montreal, Canada, pp. 255–270.

Huang, C.-R., I.-J. E. Tseng, and D. B. S. Tsai (2002). Translating lexical semantic relations: the first step towards multilingual wordnets. In *Proceedings of the COLING-Workshop on Building and Using Semantic Networks*, Taipei, Taiwan.

Hull, D. A. (2001). Software tools to support the construction of bilingual terminology lexicons. In D. Bourigault, C. Jacquemin, and M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*, Chapter 11, pp. 225–244. John Benjamins.

Hwa, R., P. Resnik, A. Weinberg, C. Cabezas, and O. Kolak (2004). Bootstrapping parsers via syntactic projection across parallel texts. *Natural language engineering 11*(3), 311–325.

Ide, N., P. Bonhomme, and L. Romary (2000). XCES: An XML-based standard for linguistic corpora. In *Proceedings of the Second Language Resources and Evaluation Conference (LREC)*, Athens, Greece, pp. 825–30.

Imamura, K. (2001). Hierarchical phrase alignment harmonized with parsing. In *Proceedings of the 6th NLP Pacific Rim Symposium*, Tokyo, Japan, pp. 377–384.

Investment Bank (BEI), E. (2004). `http://www.bei.europa.eu/publications/`. The website originally used for retrieving the documents was `http://europa.eu.int/rapidstart/welcome.html`.

Isabelle, P., M. Dymetman, G. Foster, J.-M. Jutras, E. Macklovitch, F. Perrault, X. Ren, and M. Simard (1993). Translation analysis and translation automation. In *Proceedings of the 5th International Converence on theoretical and methodological Issues in Machine Translation*, Kyoto, Japan, pp. 201–217.

Kay, M. and M. Röscheisen (1993). Text-translation alignment. *Computational Linguistics 19*(1), 121–142.

Klavans, J. and E. Tzoukermann (1996). Combining corpus and machine-readable dictioanry data for building bilingual lexicons. *Machine Translation 10*, 185–218.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the 10th Machine Translation Summit*, Phuket, Thailand, pp. 79–86.

Kondrak, G., D. Marcu, and K. Knight (2003). Cognates can improve statistical translation models. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada, pp. 46–48.

Kruijff-Korbayová, I., K. Chvátalová, and O. Postolache (2006). Annotation guidelines for Czech-English word alignment. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, pp. 1256–1261.

Kuhn, J. (2004a). Experiments in parallel-text based grammar induction. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, pp. 470–477.

Kuhn, J. (2004b). Exploiting parallel corpora for monolingual grammar induction – a pilot study. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 54–57. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Kuhn, J. (2005). Parsing word-aligned parallel corpora in a grammar induction context. In *Proceedings of ACL-Workshop on Building and Using Parallel Text*, Ann Arbor, Michigan, USA, pp. 17–24.

Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio, pp. 17–22.

Köhn, P. (2003). Europarl: A multilingual corpus for evaluation of machine translation. Draft.

Köhn, P. and K. Knight (2003). Empirical methods for compound splitting. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Budapest, Hungary, pp. 187–193.

Köhn, P., F. J. Och, and D. Marcu (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003)*, Edmonton, Canada, pp. 48–54.

Lambert, P. and N. Castell (2004). Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 26–29. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Lambert, P., A. de Gispert, R. Bancs, and J. B. M. no (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation 39*(4), 267–285.

Langlais, P., M. Simard, and J. Véronis (1998). Methods and practical issues in evaluating alignment techniques. In *Proceedings of 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistic*, Montréal, Canada, pp. 711–717.

Lin, D. and C. Cherry (2003). Linguistic heuristics in word alignment. In *Proceedings of the Conference of the Pacific Association for Computational Linguistics (PACLING*, Halifax, Canada.

Ma, X. (2006). Champollion: a robust parallel text sentence aligner. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, pp. 489–492.

Manning, C. D. and H. Schütze (1999). *Foundations of statistical natural language processing*. Cambridge, Massachusetts, London: MIT Press.

Marcu, D. and W. Wong (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, PA, USA, pp. 133–139.

Marcus, M. P., B. Santorini, and M. A. Marcinkiewicz (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics 19*(2), 313–330.

Martin, J., H. Johnson, B. Farley, and A. MacLachlan (2003). Aligning and using an English-Inuktitut parallel corpus. In *Proceedings of the HLT-NAACL workshop Building and Using parallel texts. Data driven Machien translation and beyond*, Edmonton, pp. 115–118.

Martin, J., R. Mihalcea, and T. Pedersen (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, Ann Arbor, Michigan, pp. 65–74.

McTait, K. (2001). Linguistic knowledge and complexity in an EBMT system based on translation patterns. In *Proceedings of the EBMT workshop at the 8th Machine Translation Summit*, Santiago de Compostela, Spain.

Melamed, I. D. (1995). Automatic evaluation and uniform filter cascades for inducing n-best translation lexicons. In *Third Workshop on Very Large Corpora (WVLC3)*, Boston, MA., pp. 184–198.

Melamed, I. D. (1996). A geometric approach to mapping bitext correspondence. Technical Report 96-22, Institute for Research in Cognitive Science (IRCS), University of Pennsylvania. A Revised Version of the Paper presented at the First Conference on Empirical Methods in Natural Language Processing (EMNLP'96), Philadelphia, PA, May.

Melamed, I. D. (1997a). Automatic discovery of non-compositional compounds in parallel data. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Providence, RI, USA, pp. 97–105.

Melamed, I. D. (1997b). A portable algorithm for mapping bitext correspondence. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain, pp. 305–312.

Melamed, I. D. (1998a). Annotation style guide for the BLINKER project. Technical Report 98-06, Institute for Research in Cognitive Science, University of Pennsylvania.

Melamed, I. D. (1998b). Manual annotation of translational equivalence: The BLINKER project. Technical Report 98-07, Institute for Research in Cognitive Science (IRCS), University of Pennsylvania.

Melamed, I. D. (2000). Models of translational equivalence among words. *Computational Linguistics 26*(2), 221–249.

Melamed, I. D. (2001). *Empirical Methods for exploiting parallel texts*. Cambridge, MA.: MIT Press.

Merkel, M. (1999). Annotation style guide for the PLUG link annotator. Technical report, Linköping university, Linköping.

Merkel, M., M. Andersson, and L. Ahrenberg (2002). The PLUG link annotator – interactive construction of data from parallel corpora. In L. Borin (Ed.), *Parallel corpora, parallel worlds. Selected papers from a symposium on parallel and comparable corpora at Uppsala University, Sweden, 22-23 April, 1999*, pp. 151–168. Amsterdam/New York, NY: Rodopi.

Mihalcea, R. and T. Pedersen (2003). An evaluation exercise for word alignment. In *NHLT-NAACL 2003 Workshop: Building and Using parallel Texts. Data Driven Machine Translation and Beyond*, Edmonton, Canada, pp. 1–10.

Mitkov, R. and C. Barbu (2004). Using bilingual corpora to improve pronoun resolution. *Languages in Contrast 4*(2), 201–211.

Moore, R. C. (2001). Towards a simple and accurate statistical approach to learning translation relationships among words. In *Proceedings of the ACL workshop on data-driven machine translation*, Toulouse, France, pp. 79–86.

Moore, R. C. (2002). Fast and accurate sentence alignment of bilingual corpora. In *Machine translation: from resarch to real users*, pp. 135–144. Springer. Proceedings of the 5th conference of the Association for Machine Translation in the Americas (AMTA).

Moore, R. C. (2003). Learning translations of named-entity phrases from parallel corpora. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 259–266.

Moore, R. C. (2004). Improving IBM word-alignment model 1. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Barcelona, Spain, pp. 518–525.

Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, Vancouver, Canada, pp. 81–88.

Neumann, S. and S. Hansen-Schirra (2005). The CroCo project. Cross-linguistic corpora for the investigation of explicitation in translations. In *Proceedings from the Corpus Linguistics Conference Series*.

Nießen, S. and H. Ney (2001). Morpho-syntactic analysis for reordering in statistical machine translation. In *Proceedings of the 8th Machine Translation Summit*, Santiago de Compostela, Spain, pp. 247–252.

Nießen, S. and H. Ney (2001b). Toward hierarchical models for statistical machine translation of inflected languages. In *Proceedings fo the ACL Workshop on Data-Driven Machine Translation*, Toulouse, France, pp. 47–54.

Och, F. J. (1999). An efficient method for determining bilingual word classes. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Bergen, Norway, pp. 71–76.

Och, F. J. (2000). Giza++: Training of statistical translation models. http://www-i6.informatik.rwth-aachen.de/ och/software/GIZA++.html.

Och, F. J. and H. Ney (2000a). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING)*, Saarbrücken, Germany, pp. 1086–1090.

Och, F. J. and H. Ney (2000b). Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, China, pp. 440–447.

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

Och, F. J. and H. Ney (2004). The alignment template approach to statistical machine translation. *Computational Linguistics 30*(4), 417–449.

Och, F. J., C. Tillmann, and H. Ney (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Joint Conference of Empirical Methods in Natural Language Processing and Very large Corpora*, College Park, MD, pp. 20–28.

Ozdowska, S. (2004). Identifying correspondences between words, an approach based on a bilingual syntactic analysis of French/English parallel corpora. In *Proceedings of the Coling workshop on Multilingual linguistic resources*, Geneva, Switzerland, pp. 55–62.

Ozdowska, S. (2005). Using bilingual dependencies to align words in English/French parallel corpora. In *Proceedings of the ACL student Research Workshop*, Ann Arbor, Michigan, USA, pp. 127–132.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, Philadelphia, pp. 311–318.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik (1985). *A Comprehensive Grammar of the English Language*. Longman.

Rahman, S. A. and N. A. Azih (2004). Improving word alignment in an English–Malay parallel corpus for machine translation. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 22–25. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Resnik, P. (1999). Mining the web for bilingual text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*, College Park, Maryland, USA, pp. 527–534.

Ribeiro, A., G. L. Gaël Dias, and J. Mexia (2001). Cognates alignment. In B. Maegaard (Ed.), *Proceedings of the 8th Machine Translation Summit*, Santiago de Compostela, Spain, pp. 287–292.

Ribeiro, A., G. Lopes, and J. Mexia (2000). Using confidence bands for parallel texts alignment. In *Proceedings of the 38th Conference of the Association for Computational Linguistics (ACL 2000)*, Hong Kong, pp. 432–439.

Sahlgren, M. and J. Karlgren (2004). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. *Natural language engineering 11*(3), 327–341.

Samiotou, A., L. Kranias, G. Papadopoulos, M. Asunmaa, and G. Magnusdottir (2004). Exploitation of parallel texts for populating MT & TM databases. In *Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 1–4. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Samuelsson, Y. and M. Volk (2004). Automatic node insertion for treebank deepening. In *Proceedings of the Third Workshop on Treebanks and Linguistic Theories (TLT)*, Tübingen, Germany, pp. 127–136.

Santorini, B. (1990). Part-of-speech tagging guidelines for the Penn Treebank Project. Technical Report MS-CIS-90-47, Department of Computer and Information Science, University of Pennsylvania.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, England, pp. 44–49.

Schmid, H. (1995). Improvements in part-of-speech tagging with an application to german. Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Schmitt, E.-E. (2003). *Monsieur Ibrahim et les fleurs du Coran*. Ditzingen: Reclam.

Schrader, B. (2002). Verbesserung von Wortalignment durch linguistisches Wissen. Master's thesis, Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart.

Schrader, B. (2004). Improving word alignment quality using linguistic knowledge. In *Workshop proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 46–49. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

Simard, M. (1998). The BAF: a corpus of English–French bitext. In *Proceedings of the First International Language Resources and Evaluation Conference (LREC)*, Granada, Spain, pp. 489–496.

Simard, M. (1999a). Text-translation alignment: Aligning three or more versions of a text. In J. Véronis (Ed.), *Parallel Text Processing*, Text, Speech and Language Technology Series, Chapter 3, pp. 49–67. Dordrecht: Kluwer Academic Publishers.

Simard, M. (1999b). Text-translation alignment: Three languages are better than two. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, College Park, MD, pp. 2–11.

Simard, M., G. F. Foster, and P. Isabelle (1992). Using cognates to align sentences in bilingual corpora. In *Proceedings of the Fourth International Conference on theoretical and methodological issues in Machine translation*, Montreal, Canada, pp. 67–81.

Simard, M. and P. Langlais (2003). Statistical translation alignment with compositionality constraints. In *Proceedings of the HLT-NAACL Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond*, Edmonton, Canada, pp. 19–22.

Simard, M. and P. Plamondon (1998). Bilingual sentence alignment: Balancing robustness and accuracy. *Machine Translation 13*(1), 59–80.

Sjöbergh, J. and V. Kann (2004). Finding the correct interpretation of Swedish compounds. a statistical approach. In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, pp. 899–902.

Smadja, F., K. R. McKeown, and V. Hatzivassiloglou (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics 22*(1), 1–38.

Spencer, A. (1991). *Morphological Theory*. Cambridge, Massachusetts, USA: Blackwell.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, pp. 2142–2147.

Talbot, D. (2004). Constrained em for parallel text alignment. *Natural language engineering 11*(3), 263–277.

Taskar, B., S. Lacoste-Julien, and D. Klein (2005). A discriminative matching approach to word alignment. In *Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, Vancouver, Canada, pp. 73–80.

Thielen, C., A. Schiller, S. Teufel, and C. Stöckert (1999). Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, Institut für maschinelle Sprachverarbeitung (IMS), Universitä Stuttgart und Seminar für Sprachwissenschaft (SfS), Universitä Tübingen.

Thurmair, G. (2006). Using corpus information to improve MT quality. In *Proceedings of the LREC workshop on Language Resources four Translators*, Genoa, Italy, pp. 45–48.

Tiedemann, J. (1999). Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics*, Trondheim, Norway, pp. 216–227.

Tiedemann, J. (2001). Predicting translations in context. In *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 240–244.

Tiedemann, J. (2003). Combining clues for word alignment. In *Proceedings of the 10th Conference of the European Chapter of the ACL (EACL)*, Budapest, Hungary, pp. 339 – 346.

Tiedemann, J. (2004). Optimisation of word alignment clues. *Natural language engineering 11*(3), 279–293.

Tiedemann, J. and L. Nygaard (2003). OPUS - an open source parallel corpus. In *Proceedings of the 13th Nordic Conference on Computational Linguistics (NODALIDA)*. University of Iceland, Reykjavik.

Tiedemann, J. and L. Nygaard (2004). The OPUS corpus – parallel and free. In *Proceeding of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 1183–1186. http://omilia.uio.no/opus/.

Toutanova, K., H. T. Ilhan, and C. D. Manning (2002). Extensions to HMM-based statistical word alignment models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, Philadelphia, USA, pp. 87–94.

Tschorn, P. (2004). Automatically aligning English-German parallel texts at sentence level using linguistic knowledge. Technical Report 2, Institute of Cognitive Science, University of Osnabrück. PICS - Publication Series of the Institute of Cognitive Science.

Tschorn, P. and A. Lüdeling (2003). Morphological knowledge and alignment of English-German parallel corpora. In *Proceedings of the 2003 Corpus Linguistics Conference*, Lancaster, UK, pp. 818–827.

Tufiş, D., R. Ion, A. Ceauşu, and D. Ştefănescu (2005). Combined word alignments. In *Proceedings of ACL-Workshop on Building and Using Parallel Text*, Ann Arbor, Michigan, USA, pp. 107–110.

Tufiş, D. (2002). A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING)*, Taipei, Taiwan, pp. 1030–1036.

Tufiş, D. and A.-M. Barbu (2002). Lexical token alignment: Experiments, results, and applications. In *Proceedings of the Third International Conference on Linguistic Resources and Evaluation(LREC)*, Las Palmas, Spain, pp. 458–465.

van der Eijk, P. (1993). Automating the acquisition of bilingual terminology. In *Proceedings of the 6th conference of the European Chapter of the Association for Computational linguistics (EACL)*, Utrecht, the Netherlands, pp. 113–119.

Čmejrek, M., J. Cuřín, J. Havelka, J. Hajič, and V. Kuboň (2004). Prague Czech-English dependency treebank: syntactically annotated resources for machine translation. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Portugal, pp. 1597–1600.

Véronis, J. and P. Langlais (2000). Evaluation of parallel text alignment system - the ARCADE project. In J. Véronis (Ed.), *Parallel Text Processing*, Chapter 19, pp. 369–388. Dordrecht: Kluwer.

Vogel, S., H. Ney, and C. Tillmann (1996). HMM-based word alignment in statistical translation. In *Proceedings of the International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 836–841.

Vogel, S., F. J. Och, C. Tillmann, S. Nießen, H. Sawaf, and H. Ney (2000). Statistical methods for machine translation. In W. Wahlster (Ed.), *Verbmobil: Foundations of Speech-to-Speech Translation*, pp. 377–393. Berlin: Springer Verlag.

Volk, M. (1999). Choosing the right lemma when analysing German nouns. In *Multilinguale Corpora: Codierung, Strukturierung, Analyse. 11. Jahrestagung der GLDV*, Frankfurt, pp. 304–310. Enigma Corporation.

Volk, M., S. Gustafson-Capková, J. Lundborg, T. Marek, Y. Samuelsson, and F. Tidström (2006, April). XML-based phrase alignment in parallel treebanks. In *Proceedings of EACL Workshop on Multi-dimensional Markup in Natural Language Processing*, Trento, pp. 93–96.

Volk, M., Y. Samuelsson, S. Gustafson-Capkov´a, and J. Lim (2006). Alignment guidelines for the stockholm parallel treebank 2006. Version: 3rd October 2006.

Véronis, J. (1998, April 26). ARCADE. tagging guidelines for word alignment. http://www.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/. Version 1.0.

Véronis, J. (1998a). A study of polysemy judgements and inter-annotator agreement. In *Programme and advanced papers of the Senseval workshop*, Herstmonceux Castle, England.

Wang, W. and M. Zhou (2002). Structure alignment using bilingual chunking. In *Proceedings of the 19th International Conference on computational linguistics (COLING)*, Taipei, Taiwan.

Wermke, M., K. Kunkel-Razum, and W. Scholze-Stubenrecht (Eds.) (2004). *Duden. Die deutsche Rechtschreibung* (23 ed.), Volume 1 of *Der Duden in zwölf Bänden*. Mannheim, Leipzig, Wien, Zürich: Duden.

Wu, D. (1994). Aligning a parallel english-chinese corpus statistically with lexical criteria. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Las Cruces, New Mexico, USA, pp. 80–87.

Wu, D. (1995). An algorithm for simultaneously bracketing parallel texts by aligning words. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, Cambridge, MA., USA, pp. 244–251.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics 23*(3), 378–403.

Wu, D. (1999). Bracketing and aligning words and constitutents in parallel text using stochastic inversion transduction grammars. In J. Véronis (Ed.), *Parallel Text Processing*, Text, Speech and Language Technology Series, Chapter 7, pp. 139–167. Dordrecht: Kluwer Academic Publishers.

Wu, D. (2000). Alignment. In R. Dale, H. Moisl, and H. Somers (Eds.), *Handbook of natural language processing*, Chapter 18, pp. 415–458. Marcel Dekker Ltd.

Wu, D. and X. Xia (1995). Large-scale automatic extraction of an english–chinese translation lexicon. *Machine Translation 9*(3-4), 285–313.

Yarowsky, D., G. Ngai, and R. Wicentowski (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first International Conference on Human Language Technology Research (HLT)*, San Diego, California, USA, pp. 200–207.

Zhang, H. and D. Gildea (2005). Stochastic lexicalized inversion transduction grammar for alignment. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan, USA, pp. 475–482.

Zifonun, G., L. Hoffmann, and B. Strecker (1997). *Grammatik der deutschen Sprache*. Number 7 in Schriften des Instituts für deutsche Sprache. Berlin, New York: de Gruyter.