SEVENTH FRAMEWORK PROGRAMME

PANACEA

SEVENTH FRAMEWORK PROGRAMME
THEME 3
Information and communication Technologies

PANACEA Project

Grant Agreement no.: 248064

Platform for **A**utomatic, **N**ormalized **A**nnotation and
Cost-**E**ffective **A**cquisition
of Language Resources for Human Language Technologies

D5.1

Parallel technology tools and resources

Dissemination Level:	Public
Delivery Date:	July 16th 2010
Status – Version:	Final v1.0
Author(s) and Affiliation:	Pavel Pecina (DCU), Antonio Toral (DCU), Gregor Thurmair (LinguaTec), Andy Way (DCU)

UNIVERSITAT POMPEU FABRA · IEA ILSP · UNIVERSITY OF CAMBRIDGE · linguatec language technologies · DCU · elda

Table of contents

1 Introduction

The main objectives and tasks of WP5 "Parallel corpus and derivatives" are:

1) Developing word-aligned and chunk-aligned data from the parallel corpora induced in WP4 for training MT models. This task will involve sentence alignment of parallel corpora, parallel sentence extraction from comparable corpora, and consequent sub-sentential alignment on world, chunk, and subtree level.
2) Using the produced sub-sentential aligned data for deriving bilingual dictionaries. This task will include filtering the bilingual dictionaries obtained from the alignments carried out in the previous task. This will involve exploiting confidence measures provided by the alignment algorithms and frequency characteristics of the aligned terms in the corpora.
3) Using the produced sub-sentential aligned data and dictionaries for extracting transfer grammars. This task will involve exploring several approaches to transfer selection: topic identification, definition of grammatical contexts (morphosyntactic and semantic tests), definition of conceptual contexts (conceptual clustering, co-occurrence interpretation)

This report defines these tasks in Section 3, describes the current state of the art in the relevant areas in Section 4, provides analysis of tools that are available and specifies the tools and resources to be developed in this WP in Section 5 and Section 6, respectively, and proposes the solution path to be followed during the rest of the lifetime of the WP in Section 7.

2 Terminology

2.1 Definitions

A **corpus** is a (large) set of texts. In PANACEA, we assume the texts are stored electronically, in a given file format and character encoding, without any formatting information, eventually provided with metadata and/or linguistic annotation. Often, the texts are referred to as documents, in which case the texts are assumed to be topic-coherent.

A **monolingual corpus** is a corpus of texts in one language.

A **bilingual corpus** is a corpus of texts in two languages.

A **parallel corpus** is a bilingual corpus consisting of texts organized in pairs which are translations of each other, i.e. they include the same information (parallel texts). Usually, the pairs are identified at least for documents (parallel documents) and the corpus described as document-aligned parallel corpus. If the translation pairs are identified also for sentences (parallel sentences) we talk about sentence-aligned parallel corpus. Usually, one half of the parallel corpus (the texts in one of the two languages) is called the source language side (or source side) and the other half (in the other language) is called the target language side (or target side). This refers only to the intended translation direction (from the source language to the target language) and does not affect the corpus itself.

A **comparable corpus** is a bilingual corpus consisting of texts organized in pairs (comparable documents) which are only approximate translations of each other, i.e. they include similar information.

A **domain-specific corpus** (or in-domain corpus) is a corpus of texts from a given domain.

A **general domain corpus** is a corpus of texts containing general language texts, i.e. texts from no specific domain.

A **bilingual dictionary** is a specific kind of dictionary which contains correspondences of terms (words, multiwords, or phrases) between two languages, and hence is used to translate these terms from one language to the other.

A **transfer grammar** is a set of rules which are applied to translate source language structures into target language structures. Such grammar rules can be divided into structural rules (with no reference to lexical material) and lexical transfer rules (whereby the selection of a certain lexical item depends on contextual configurations.

2.2 Acronyms

CRF – Conditional Random Field
EBMT – Example-Based Machine Translation
EM – Expectation Maximization
HMM – Hidden Markov Model
ITG – Inversion Transduction Grammar
LFG – Lexical Function Grammar
LF – Lexical Function
LMF – Lexical Mark-up Framework
MRD – Machine Readable Dictionary
MT – Machine Translation
MWE – Multi-Word Expression
NLP – Natural Language Processing
NP – Noun Phrase
PBMT – Phrase-Based Machine Translation
POS – Part of Speech
PP – Prepositional Phrase
PTD – Probabilistic Translation Dictionary
RBMT – Rule-Based Machine translation
SL – Source Language
SMT – Statistical Machine Translation
SVM – Support Vector Machine
TL – Target Language
WSD – Word Sense Disambiguation

2.3 Related documents

D3.1 – Architecture and design of the platform
D4.1 – Technologies and tools for corpus creation, normalization and annotation
D7.1 – Criteria for evaluation of resources, technology and integration
D8.1 – Analysis of industrial user requirements

3 Task description

3.1 Alignment

Alignment is the identification of the corresponding parts (mutual translations) in parallel texts. In PANACEA, the term alignment refers to *automatic alignment* performed by a computer algorithm (often based on statistical methods) unless it is explicitly specified otherwise (e.g. *manual alignment* refers to alignment performed by a human being). *Sentential alignment* (or *sentence alignment*) refers to alignment of sentences and *sub-sentential alignment* refers to alignment of sub-sentential elements, such as words, chunks, phrases, and even more complex structures such as syntactic trees.

3.1.1 Sentential alignment

Sentence alignment is the identification of parallel sentences in parallel texts. Prior to this step, sentence boundaries must be identified (*sentence segmentation*) in both sides of the parallel documents. A sentence-aligned parallel corpus is one of the two essential data resources required for training SMT systems (the other one is a TL monolingual corpus used for language modelling). In general, all possible alignment combinations are allowed: 1-1 when one sentence in one language fully corresponds to one sentence in the other language. 1-0 or 0-1 in case a sentence is not translated on the other side, or M-N when M>0 sentences on one side correspond to N>0 sentences on the other one.

Sentence alignment is usually applied on a parallel corpus where the parallel texts are assumed to be reliable translations of each other. In PANACEA, this assumption cannot generally be made because the bilingual resources acquired in WP4 are more likely to have a make-up more similar to comparable corpora, as the parallel texts may not be accurate translations of each other (for example Wikipedia articles in multiple languages; they can but may not be accurate translations of each other). Therefore, an additional consequent step involving *parallel sentence extraction from comparable corpora* will have to be carried out if it is required. In this task, translation quality of each aligned sentence pair is estimated and those pairs with low translation quality are discarded.

3.1.2 Sub-sentential alignment

The basic approach to sub-sentential alignment is *word alignment*. Word alignment is the identification of corresponding words in parallel sentences. It is a fundamental component of all modern SMT systems where it is used in order to extract a set of translation phrase pairs into a translation table. Word alignment is also employed in other NLP applications, such as

translation lexicon induction and cross-lingual projection of linguistic information. Prior to word alignment, word boundaries must be identified (*tokenization*) in both sides of the parallel sentence. In general, all possible alignment combinations are allowed: 1-1 if one word on one side exactly corresponds to one word on the other side, M-N where M>0 sentences on one side correspond to N>0 sentences on the other one. 1-0 and 0-1 alignments are used when for a given word there is no translation equivalent on the other side of the parallel sentence (a word is deleted or inserted on the target side, respectively).

Translation phrase pairs extracted from word-aligned sentences without other linguistic knowledge need not, in fact, form grammatical phrases. In order to overcome this limitation, other approaches to sub-sentential alignment operating on syntactically annotated data have been introduced:

Chunk alignment is the identification of corresponding chunks (syntactic constituents, such as noun phrases, verb phrases, etc.) in parallel sentences. Chunking must be applied prior to this step, e.g. by shallow parsing or according to the Marker hypothesis (Green, 1979). The assumption is that the number of chunks in both sides of the parallel sentence is more or less the same and they can be aligned in a (more or less) 1-1 manner, although in general, 1-n chunk alignments are allowed too. Chunk alignment employed in SMT better captures local reordering and reduces the size of a translation table (e.g. Sikuan and Yanquan, 2009) (translation pairs are more linguistically motivated than in case of phrases extracted from word-aligned sentence pairs).

Subtree alignment is the identification of correspondences in parallel sentences where full syntactic information (on either or both source and target sides) is taken into account. This is another step in introducing deeper linguistic knowledge into SMT. Translation phrase pairs extracted from subtree-aligned parallel sentences are grammatical phrases. Both dependency (e.g. Meyers et al., 1998; Menezes and Richardson, 2003) and constituency syntax (e.g. Wu, 2000; Zhechev, 2009) can be used in this context. Syntactic analysis (parsing) is usually performed prior to tree alignment, but some approaches do not require this and can produce parallel trees from unannotated data (see Section 4.1.2).

3.2 Bilingual dictionary induction

A *bilingual dictionary* for two languages *a* and *b* contains correspondences (translations) of terms of *a* and *b*. The kind of terms covered range from single words to multi-word expressions (these in turn can be limited to noun phrases or deal with the more general concept of collocations).

A *probabilistic translation dictionary* (PTD) is a specific type of bilingual dictionary where each term in the source language is associated with its frequency in the corpus and each of its translations has a probability.

By *induction* we refer to the automatic derivation of bilingual dictionaries from bilingual corpora (be it comparable or parallel) or, more generally, from existing language resources. The most common measures used to evaluate the induction of bilingual dictionaries are: precision

(Karlgren and Sahlgren, 2005) (both precision of the 1st and 10 first correspondences are used), Mean Reciprocal Rank (MRR) (Yu and Tsujii, 2009), and average accuracy (Gamallo, 2008; Yu and Tsujii, 2009).

3.3 Transfer grammar induction

3.3.1 Problem description

As the terms *transfer grammars*, *transfer rules* etc. describe quite different phenomena, reaching from simple lexical replacement being called *transfer rule* to structural changes, it may be advisable first to define more precisely what is meant in this PANACEA task.

Once MT systems become more mature, their dictionaries will grow. Instead of the problem of dictionary gaps, which is more relevant for early phases of MT development, another problem needs to be solved: there are many possible translations for a given source term, and the system has to choose which translation should be selected in a given constellation. The challenge is to find clues to determine when a given translation should be selected.

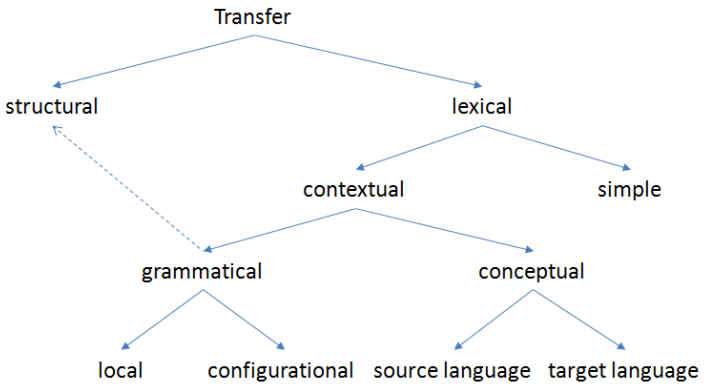

Figure 3.3.1: Classification of transfer domain

In a classification of transfer actions, the problem would be contextual lexical transfer, as transfer selection will depend on certain contextual constellations. Such constellations usually are described as annotations in a bilingual lexicon entry, e.g. de ‘bestehen’ (test: PObj-auf) -> en ‘insist’ (upon), de ‘bestehen’ (test: PObj-aus) -> en ‘consist’ (of).

3.3.2 Scope and limitations

PANACEA does not aim at learning structural transfer as the focus is on lexical selection; nor is simple lexical transfer of interest as lexical selection faces the situation where several translation options exist, and some disambiguation effort is required. As a result, the means of contextual transfer need to be explored. This section describes the means which are available to describe such contextual conditions, used by many current MT systems.

The transfer tests used by current systems fall into three types:

Global feature settings: Transfer is selected if a certain global feature is set, describing a domain, a locale, or a special target customer-specific translation (example: de 'Läufer' -> en 'runner' (SPORTS), de 'Läufer' -> en 'bishop' (CHESS)) .

Grammatical tests: Transfer is selected according to a specific structural constellation of the input sentence, such as: presence of a prepositional object with 'from'; usage as a compound specifier, etc. (example: de 'König' -> en 'king', de 'König' (test: compound-specifier) -> en 'royal' (Königshof -> royal court).

Conceptual tests: Transfer is selected according to a specific conceptual context (either on the source or the target side), defined as cooccurrence relations in various ways (example: de 'Gericht' (context: Anwalt Gesetz Recht) -> en 'court'; de 'Gericht' (context: Gemüse Zwiebeln Kochen) -> en 'dish').

Not all current MT systems support all options, but the three strategies define the scope of the current technology, and will form the basis of the PANACEA work. The tool to produce transfer annotations in PANACEA will focus on the automatic extraction of the following issues:

1) *global feature settings*:
   - automatic domain flag setting/topic tests
2) *grammatical tests* (including multiword tests):
   - local tests on the (values of the) following features: lemma, number, gender (for German), compound specifier (for German)
   - configurational tests on arguments, argument types, and their fillers (prepositions for verbs and nouns, reflexives, semantic types, lemmata)
3) *conceptual tests*
   - identification of conceptual contexts for certain transfers

Other options, and additional tests/actions (SL/TL constraints), will not be supported in this version. The extracted information will be presented in a generic bilingual dictionary representation, defined in Figure 6.3.2.

4 Current state of the art

This section presents a survey of the existing literature on relevant tasks, the different approaches, and results and consequences for the project.

4.1 Alignment

4.1.1 Sentential alignment

A. Sentence alignment in parallel corpora

Methods for sentence alignment in (document-aligned) parallel corpora can be classified into three groups based on the assumptions they make about parallel sentences:

Sentence length-based methods (e.g. Kay and Röscheisen, 1988; Brown et al. 1990; Church and Gale, 1993) assume that the length of a source sentence is highly correlated with the length of the target sentence. The sentence length (measured as the number of words or characters) information is then used to guide the alignment process which tries to find the best possible match over a text (document). Additional structural information can be used to delimit the search space (e.g. headline, paragraph delimiters etc.)

Word-correspondence-based methods (e.g. Melamed, 1997) assume that if two sentences are mutual translations their words must be translation of one another too. These word correspondences are used to guide the alignment process through the text (document). Word correspondences can be discovered based on their cooccurrence in the texts to be aligned, by their presence in a bilingual dictionary or by identification of cognates (graphically identical or similar tokens), such as dates, symbols, names, untranslated abbreviations etc.).

Hybrid methods (e.g. Moore, 2002) combine the two approaches mentioned above and use both sentence length and word correspondences to guide the process of sentence alignment.

B. Parallel sentence extraction from comparable corpora

Parallel sentence extraction from (document-aligned) comparable corpora is applied on sentence pairs identified as candidates of parallel sentences (assumed to express the same information). Usually, some parallel data or bilingual dictionary is used to determine word correspondences in the sentence pairs. This information is then used in a classifier which identifies candidate sentences as parallel or not parallel.

Sentence alignment can be employed to identify the set of candidate parallel sentences, but most sentence aligners expect monotonicity in sentence order and do not deal with any changes in sentence order.

Eventually, all possible sentence pairs can be considered as candidate parallel sentences, which makes this task an $O(n^2)$ problem. Therefore, several approaches have been proposed to reduce this complexity: e.g. Munteanu and Marcu (2005) filtered out pairs with high length difference or low word correspondences (based on a dictionary), and Smith et al. (2010) used two different approaches for extracting parallel sentences from aligned Wikipedia articles: first, they trained a ranking model which, for each source sentence, assigns at most one (or no) target sentence; second, they built a global sentence alignment model based on a first order Conditional Random Field with a hidden variable indicating the corresponding target sentence for each source sentence.

4.1.2 Sub-sentential alignment

A. Word alignment

Three different models have been proposed for word alignment so far: generative models, discriminative models, and association-based models.

Generative models describe the alignment (translation) as a process where a sentence in one language (source sentence f) generates a sentence in another language (target sentence e) and the actual alignment is only an artefact of this process. This approach is based on modelling the conditional probability p(e,a|f) in a HMM, where a is a hidden variable of the generation process. The optimal alignment a (which maximizes p(e,a|f)) is usually searched for by the Viterbi algorithm.

In general, the relation between source and target words can be arbitrary, but in practice most models constrain the alignment in such a way that each source word can only be aligned to exactly one target word. Therefore, only asymmetric alignments are produced (Och and Ney, 2003). The generating process may include word insertion or deletion, word reordering (or distortion, which indicates a change in relative position when generation a target word from a source word), and eventually also source word fertility (reflecting one-to-many generation) (Brown et al., 1993).

The most frequently used models which do not explicitly model fertility are IBM Model 1 and IBM Model 2 (Brown et al., 1993). They assume the generative process proceeds as follows: first, a source position is selected for each position in the target sentence and then a target word is produced as a translation of the selected source word. In IBM Model 1, the position is selected from uniform distribution and in IBM Model 2, the selection is conditioned by the target position.

Models employing fertility, such as IBM Model 3 and IBM Model 4 (Brown et al., 1993) assume a different generation process. First, it is decided how many target words each source word should generate (source word fertility). Then, target words are produced according to the distortion models. In IBM Model 3, each target position is chosen independently for the target words generated by each source word. In IBM Model 4, this decision is based on positioning the previous target words. IBM Model 5 (Brown et al., 1993) is a modification of Model 4 with a suitably refined distortion model to avoid the problem of deficiency. However, no efficient training and search algorithm exists for these models (they are implemented by using only approximate hill-climbing methods, not guaranteed to find the optimal solution), they can produce high-quality alignments applicable in various types of data-driven MT systems. A thorough evaluation of various generative word alignment models can be found in Och and Ney (2003).

Discriminative models, unlike the generative models, model the probability p(a|e,f) directly by decomposing it to a log-linear combination of a set of various different features. The optimal alignment a is searched for by maximising this log-linear combination. These models, however, require a certain amount of manually annotated word-aligned data for training. The model

parameters (feature weights) are trained in a supervised manner using various machine learning techniques including Averaged Perceptron (Moore, 2005), Maximum-Entropy (Liu et al., 2005), Support Vector Machines (Taskar et al., 2005), Conditional Random Fields (Blunsom and Cohn, 2006), etc.

Association-based models obtain word alignments by applying association measures on source and target words (Smadja et al., 1996; Melamed, 2000) and eventually combine them with other features (e.g. syntactic information, POS tags, chunk labels (Tiedemann, 2003) and dependency trees (Cherry and Lin, 2003) in a similar manner to the discriminative alignment models but only in a heuristic fashion. The search procedure is often implemented as a greedy algorithm (e.g. Cherry and Lin, 2003).

Further details of different word-alignment approaches can be found for example in the thesis of Ma (2009).

B. Chunk alignment

One method for extraction of chunks is *shallow parsing* which identifies syntactic constituents (noun phrases, verb phrases, etc.) but does not analyse their internal structure nor specifies their role in the sentence.

Chunking is an intermediate step towards full parsing. There are three main groups of shallow parsing methods: *Rule-based methods,* e.g. Abney (1996) who used hand-crafted cascaded Finite State Transducers. *Generative methods,* e.g. Ramshaw and Marcus (1995) who defined shallow parsing as a tagging problem, they labelled words as being inside an NP, outside of an NP, or between the end of one and the start of another NP. Skut and Brants (1998) extended this approach to other types of chunks. Molina and Pla (2002) proposed HMM based shallow parsing. *Discriminative methods,* e.g. Zhang et al. (2002) who applied a generalized version of the Winnow algorithm (similar to the Perceptron).

Another method for the extraction of chunks is based on the *Marker Hypothesis*, a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific set of lexemes or morphemes (Green, 1979). *Marker-based chunking* employs a set of closed-class ("marker") words, such as determiners, conjunctions, prepositions, possessive and personal pronouns, and split a sentence into chunks at each occurrence of a marker word. Each chunk, however, must contain at least one non-marker word (Gough and Way, 2004).

Chunk alignment itself is performed after the chunks are identified in both sides of parallel sentences. The chunk aligner relies on the identification of relationships between chunks which can be defined in different ways. Stroppa and Way (2006) use three features combined in a log-linear framework:

1) combination of word-to-word translation probabilities obtained from word alignment (word-based lexicon model),
2) ratio between the number of cognates identified between the source and the target words of a chunk and the total number of source words,

3) boolean feature indicating if the chunks have the same label. The most likely chunk alignment is computed by a simple dynamic programming algorithm based on the classical edit-distance algorithm (Levenstein, 1966) in which distances are replaced by opposite-log-conditional probabilities.

C. Subtree alignment

Existing approaches to automatic subtree alignment can be grouped according to whether they align dependency structures or phrase-structure trees. A detailed overview of them can be found in Zhechev (2009).

One of the first attempts on *alignment of dependency structures* is Matsumoto et al. (1993). In this work, the authors used Lexical-Functional Grammars (Kaplan and Bresnan, 1982) to parse English and Japanese sentences and then converted the LFG parses to the specific type of dependency structures (called decompositions) and employed a structural matching algorithm motivated by the branch-and-bound top-down backtracking algorithm. The algorithm relies on association of pairs of content words measured with the help of a Japanese–English dictionary and a thesaurus.

The work of Meyers et al. (1998) used a greedy search-based algorithm to align regularised parses, similar to the F-structures of LFG but using dependency relations. The algorithm employed features derived from the dependency structures in addition to word-level correspondences obtained from external bilingual dictionaries and allowed many-to-many alignments between trees in order to extract the aligned substructures used as them as transfer rules for MT.

Eisner (2003) proposed Synchronous Tree Substitution Grammars (based on dependency syntax) for training of SMT systems on pairs of trees. First, he considered all possible source and target trees in a pair and all possible alignments between the resulting trees. Then, he applied an inside-outside algorithm and the Expectation Maximisation algorithm to calculate occurrence statistics of elementary tree-pairs and finds the joint probability of the occurrence of a source—target pair of trees, summed over all possible alignments between the trees.

Ding et al. (2003) proposed an algorithm that uses dependency structures to constrain word alignments. It uses links between nodes in the dependency trees which are decomposed into sub-graphs (treelets) and the output of this system consists of word and phrase alignments derived from linked treelets, rather than aligned syntactic trees.

Menezes and Richardson (2003) described a rule-based system for alignment of sentence logical forms structured in a dependency fashion. The authors used a probabilistic bilingual dictionary to identify word correspondences (alignments) to form initial hypothetical tree alignments. Then, a list of several rules was applied to confirm or reject each hypothetical link and to add new links that were not suggested in the word-alignment step.

One of the first algorithms for alignment of phrase-structure trees was proposed by Kaji et al. (1992). First, both the source and target sentences were syntactically analysed using a chart

parser. Then, a bilingual dictionary was used to find potential correspondences between the source and target content words (function words were ignored) and finally, a heuristic procedure based on existing word-level correspondences was used to align the phrases (a target phrase was aligned to a source phrase if and only if it contained correspondences for all the content words in the source phrase and no correspondences to words outside of the source phrase).

Wu (2000) proposes Stochastic Inversion Transduction Grammars (ITG) to be used for phrasal alignment. Here, an ITG model is used to produce parse trees for both the source and target sentences. Nodes of these trees are marked to allow inversion of the surface order of their subtrees when transitioning from the source tree to the target tree. The leaf nodes then contain source/target word pairs (allowing also insertion and deletion). The phrasal alignments are obtained directly from the nodes which span both a source and a target phrase.

Imamura (2001) applied another approach to the alignment of phrase structure trees. He also employed word alignment to locate translationally equivalent phrases by identifying phrases which have the same or very similar syntactic categories and include the same semantic information.

Zhechev (2009) published a subtree alignment tool which can be used both in cases in which monolingual phrase-structure parsers exist for both languages and in cases in which such parsers are not available. First, a word alignment tool is used to obtain word-alignment probabilities for both language directions. If parsers are available for both languages, they are used to parse both sides of the parallel corpus. The resulting parse trees and the word-alignment probabilities are then used to obtain links between nodes in corresponding trees according to their translational equivalence scores (based on the word-alignment). If there is no parser available for one of the languages, the word-aligned parallel corpus is used directly by a modified version of the subtree aligner producing aligned trees from plain data.

4.2 Bilingual dictionary induction

Research in this area started in the nineties, and can be grouped according to the data used for term extraction:

Approaches based on parallel corpora use a combination of statistical and/or linguistic procedures to extract terms, both single and multiword terms; for multiword detection cooccurrence, POS patterns, or term similarity are used. An overview of different techniques is given in Cabré et al. (2001).

Approaches that focus on comparable corpora; they follow the assumption that there is a correlation between the patterns of word cooccurrences in texts (even if unrelated) of different languages (Rapp, 1995). They try to identify conceptual contexts of a translation candidate, translate those contexts using a dictionary of seed terms, and search in the target language for the most similar contextual clusters. Works belonging this paradigm study different variations of this basic idea. (Fung, 1995) used context heterogeneity. (Fung, 1998) applied an Information Retrieval approach . Gamallo (2008) defined syntactic rules to get lexico-syntactic contexts of

words and evaluated the efficiency of different coefficients. Finally, Yu and Tsujii (2009) extended Fung's idea to dependency heterogeneity.

Approaches that exploit structured resources. These can be classified in two groups, those that acquire dictionaries from MRDs (Neff and McCord, 1990; Helmreich et al., 1993; Copestake et al., 1994) and those that rely on Web 2.0 resources such as Wikipedia (Yu and Tsujii, 2009) or Wiktionary (Etzioni, 2009).

As the focus in PANACEA is on parallel corpora, we present a more detailed description of the state-of-the-art of bilingual dictionary induction from this kind of resource. Work in this area can be divided in two main approaches: hypothesis testing, and estimating.

The hypothesis testing approach (Gale and Church, 1991; Eijk, 1993; Smadja et al., 1996) has an important disadvantage; it needs a minimum number of observations to derive a valid hypothesis, so a limited amount of translation examples needs to be found with high accuracy. Conversely, the estimating approach makes it possible to find the most probable translations for each example. Work in this approach use directional translation models (Wu and Xia, 1995) or symmetric models (Hiemstra, 1998).

An important aspect to be considered regards the suitability of the corpora to induce bilingual dictionaries. In this regard, Santos and Simoes (2008) discuss the connection between bilingual dictionary quality, corpus genre and languages. They introduce two concepts that characterise a parallel corpus with respect to be potentially used for alignment purposes. The first is translation fertility, which characterises a parallel corpus by the average number of translation candidates in the induced dictionary. The second is alignment density, i.e. the ratio of aligned tokens in a parallel corpus.

There is also literature on extracting bilingual dictionaries of multi-word expressions (or collocations) from parallel corpora. There are approaches using re-estimation algorithms (Kupiec, 1993), frequencies (Van der Eijk, 1993), word alignment (Dagan and Church, 1994), similarity measures (Smadja et al., 1996) and translation patterns (Simoes and Almeida, 2008).

4.3 Transfer grammar induction

4.3.1 Learning structural transfer rules

This research identifies transfer rules without referring to lexical material, i.e. pure structural rules. Winiwarter (2004a, 2004b) describes a Japanese-to-German RBMT system using Prolog predicates for transfer. It is structural transfer, but no learning component is involved. In data-driven contexts, research even in structural transfer rules starts from the lexical level; the different approaches are described in Section 4.3.3. As explained the PANACEA transfer selection does not intend to extract structural transfer rules.

4.3.2 Learning simple lexical transfer: Extracting transfer entries from corpora

Term extraction is a special focus in PANACEA, and described in Section 4.2. The result of term extraction is a list of bilingual terms, with minimal linguistic annotation (usually POS

information). In contrast to this research, the current PANACEA tool will be concerned with multiple dictionary translations resulting from term extraction, and means to select proper transfers using disambiguation means.

4.3.3 Selecting the best transfer for a given context

This section describes the research in the approaches which PANACEA intends to follow: domain tag assignment, definition of grammatical contexts, and definition of conceptual contexts.

A. Research on domain tag assignment to terms

There is significant literature on automatic document classification. Classification uses feature vectors (usually words, sometimes lemmata) to describe the document classes, and computes the most similar document class for an incoming document at runtime. Classes and associated features can either be developed by hand, or by machine learning, using supervised or unsupervised methods. Overviews can be found in Goller et al. (2000) and Sebastiani (2002).

The focus of this research is to determine how well a given term describes the document class; the fewer occurrences outside the target class the better the term describes the document class. The point of view is from the class to the term, i.e. the contribution of a term to the definition of the class. This may even not be a property of the term: If 'Afghanistan' happens to occur only in the drugs class then it is considered to be a drug term, despite being just a country. Moreover, as document classification is a monolingual task, no translation issues are involved.

In the case of PANACEA, however, the point of view is from term to classes. The starting point is a term which is already known to be ambiguous (i.e. a 'bad' candidate for a classifier), and the goal is to find out if a given translation can be selected on the basis of the topic of the context, i.e. the contribution of the topic to the identification of the (translation of the) term. Automatic classification tries disambiguation of classes by terms; here disambiguation of (translations of) terms by classes is required, to be located in a multilingual setup.

A related topic is the discussion on the notion of *termhood* in the domain of term extraction. Termhood defines the significance of a concept for a given domain, and is is often decided by comparing the frequency rank of a term candidate in the special-domain corpus to its rank in a background baseline corpus (Kit, 2002; Drouin, 2006; Vu et al., 2007).

The question of termhood (i.e. whether a candidate really is a term, namely a meaningful concept in the domain at hand) is often decided by comparing the frequency rank of a term candidate in the special-domain corpus to its rank in a background baseline corpus (Kit, 2002; Drouin, 2006; Vu et al., 2007).

Although termhood implies assigning a topic to a term, this is not exactly the point of interest in PANACEA: the interest is not to find out whether a term is an important concept to a predefined domain, but rather to find out if one (or more) of the available domains can help in disambiguating its translation. Again the starting point is that a translation candidate already has a multi-domain assignment.

As a result, PANACEA will use techniques of document classification as a first step, namely to assign classes/domains to incoming documents; the intention is to investigate whether this assignment can help in the disambiguation of multiple translations.

B. Survey on the creation of grammatical transfer tests and actions

Recent research started looking into automatically building transfer rules; most of them attempt structural transfer, some however also have influence on lexical transfer.

The ReTraRos project (Caseli et al., 2008) does both bilingual dictionary extraction and structural transfer rule extraction based on a bilingual parallel corpus which is word-aligned, lemmatised and tagged. The bilingual dictionary exploits the word alignment, performed in both directions and merged, and morphosyntactic attributes are recognised and added as feature-value pairs (gender, number). The transfer rule extraction is based on 'alignment blocks', i.e. phrases which follow certain alignment types (omissions, reordering, same-order). For each type, rules are extracted. Rules follow POS patterns; they are enriched by constraints over the number and gender features (monolingual and bilingual), filtered for frequency, disambiguated in case of several rules for the same source-side POS sequence), and sorted according to frequency and weight (probability from the phrase table). In translation, the rules are tried according to their order.

A similar scenario is used in Sánchez-Martínez et al. (2007, 2009a). It aims at the extraction of (shallow) transfer rules from parallel corpora. Unlike Caseli et al. (2008), an existing bilingual dictionary to filter 'impossible' phrase pairs is used, and the strategy follows an 'alignment template' approach. It does word and phrase alignment, and extracts bilingual phrase pairs based on morphologically processed parallel corpora (SL and TL side). Learning is based on the use of POS tags instead of words; enriched by 'lexicalised' tags containing for example frequent prepositions and auxiliary verbs which often undergo lexical change in translation. In addition, restrictions can be specified on local features (e.g. gender, number). The resulting alignment templates are filtered by frequency. Transfer rules consist of a 4-tuple of <SL-pattern, TL-pattern, alignment information, TL restrictions>. They are applied at transfer time using the most frequent alignment template that satisfies the TL restrictions.

Both approaches have been embedded in the Apertium MT environment (Ginestí Rosell, 2010), and show better results than simple word-by-word translation. Neither really takes lexical selection into account (besides from preposition etc.), and offer transfer disambiguation (number, gender) only on the local node level, not on the configurational level. In case other translations have to be selected, frequency is the criterion used. So the disambiguation means are: (1) Local node filter on gender and number, followed by (2) template frequency checks.[1]

1 Current work in Apertium shows that a level of chunking is being introduced where chunks can be addressed as units. (Ginestí Rosell, 2010); this allows chunk translation in an example-based context, cf. Sánchez-Martínez et al. (2009b). Without this, only closely related languages can be translated in Apertium.

The learning focus, based on abstraction from word patterns to POS patterns made by both approaches shows that non-lexicalised transfer rules are in the focus.

The MT group in Microsoft has worked on another aspect of transfer rule production. Their focus in Menezes and Richardson (2001) is the mapping of transfers, mainly for the purpose of dictionary entry extraction. Unlike approaches based on (single or multi) word alignment of terms, they create analysis structures (Logical Forms) which they align in order to extract transfer mappings. Starting from known alignments (based on bilingual dictionary lookup) they collect mappings in a best-first strategy (supporting also n:m word mappings), and enrich them by contexts (e.g. head nouns for adjectives, main verbs for modals, etc.). The result is correspondences of LFs, enriched by contextual markers.

While the primary effect is to extract lexical items, it should be noted that using (structural) context information for transfers is intended to be a means of disambiguation of transfer items. The evaluation results show that using this context has significant influence on the overall translation quality.

Like the Microsoft team, the work of Jellinghaus (2007) uses semantic representations as an SL-TL interface structure to extract (structural) transfer rules. His work is based on minimal recursion semantics, and uses semantic predicates, produced from SL and TL aligned sentences, as the basis of alignment. Such predicates contain lexical elements, and can be compared to lexical mappings in the simplest cases. More complex cases could include operations on the arguments of the predicates; as the data basis is rather small it is difficult to say how more complex phenomena could be dealt with. Problems of lexical selection are not in the focus of this work.

Another approach towards transfer rule learning is researched in the context of resource-limited MT in the AVENUE project (Probst et al., 2002; Probst, 2005), where transfer rules are extracted from carefully designed user-aligned parallel elicitation corpora. Seed rules abstract from the words into basically the POS level, and try to induce some of the major (i.e. more resourced) language c-structure annotations for the minor language (i.e. lesser resourced). Later[2] steps generalise over the seed rules, by introducing nonterminal nodes (like NPs and PPs) based on the TL c-structure, and interpreting the attribute constraints. The third step is version space learning, which tries to merge two transfer rules into a more generic one, based on the analysis of their attributes, and deleting and/or merging operations of attributes. Details of the learning approach are given in Probst (2005).

2 This procedure presupposes a degree of isomorphism of the major and the minor language.

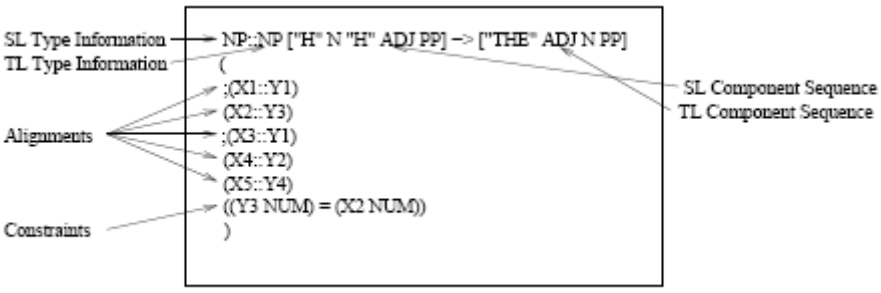

Figure 4.3.1: Example of AVENUE transfer rule (Probst, 2005, p.45).

```
;;SL: H N$IM IHIW AIN@LIGN@IWT
;;TL: THE WOMEN WILL BE INTELLIGENT
;;SL(alt1): H AI$H THIH AIN@LIGN@IT
;;TL(alt1): THE WOMAN WILL BE INTELLIGENT
;;SL(alt2): H AN$IM IHIW AIN@LIGN@IM
;;TL(alt2): THE MEN WILL BE INTELLIGENT
;;SL(alt3): H AI$ IHIH AIN@LIGN@I
;;TL(alt3): THE MAN WILL BE INTELLIGENT
S::S [NP "HIH" ADJP] -> [NP "WILL" "BE" ADJP]
(
(X1::Y1)
;(X2::Y2)
;(X2::Y3)
(X3::Y4)
(X0 = X2)
((Y1 GEN) = (X1 GEN))
((Y1 NUM) = (X1 NUM))
((Y1 PER) = (X1 PER))
(Y0 = Y2)
)
```

Figure 4.3.2: Example of a transfer rule in Hebrew-English MT (Probst, 2005: p.45), covering the example sentences above.

The transfer rules have several parts: 1. The rule-head (type, and components, in phrase-structure notation), 2. The alignments of the rule parts, 3. Constraints (consisting of SL side constraints, transfer constraints, and TL side constraints). An example is given in Figure 4.3.1.

In translation, the rules are applied top-down until a terminal node is reached, which then undergoes lexical replacement. Finally, a statistical decoder is used to find the best path through the translation alternatives.

The basic assumption in this approach is that translation goes from a ‘minor’ into a ‘major’ language, and that no syntactic analysis component is available for the minor language. The means for transfer disambiguation in this setup are target language grammatical annotations, projected from the (major) target into the minor (source) language. Unlike the approaches in the Apertium context, the learning here also relates to non-terminals and on transfer on phrase level (NP, AP, PP etc.).

The same transfer formalism as for AVENUE is also used in later developments in the Stat-XFER project, where the restriction to translate a minor into a major language does not hold any more, and parallel analysis on source and target side is possible.

This work has been extended to other languages and generalised (Lavie 2008; Lavie et al., 2008; Hannemann et al., 2008; Hannemann et al., 2009), and aims at extracting syntactically-labelled phrases for phrase-table translations. Starting from word-aligned sentences, matching parse tree nodes are identified, and consequently broken down into minimal phrase pairs, which are used in the decoding phase. The original number of rules is about 16 thousand for De-En in Hannemann et al. (2008), most of them singletons. Hannemann et al. (2009) show that syntactically annotated phrases can improve system performance.

Similar research to extract transfer rules has been carried out by Lavoie et al. (2002) on Korean-to-English. The approach intends to extract transfer rules, using deep-syntactic dependency structures on both sides (instead of Logical Forms only), and creating transfer rules from seed nodes on the word level, which are aligned using a bilingual dictionary. From the seed nodes they search alignment patterns by identifying alignment and attribute constraints to the subtrees of the (source and target) parse constituents. The resulting transfer rule candidates are sorted and filtered, with about 2000 rules remaining for a training set of about 1400 sentences.

Compared to the PANACEA task, this work does not have lexical disambiguation in its focus; this problem is handled only implicitly, by assigning probabilities to phrase rules which go into the decoding process as one of its parameters. Accordingly, it is left to the decoder to select the proper lexical translation, as is usual in statistical MT. Matching of subtrees to identify ‚meaningful' transfer constraints, however, is a common subtask with the PANACEA task, although the focus of Stat-XFER is on structural transfer, not lexical one.

There is research in the context of EBMT (Brown, 2001, 2003), where two approaches are combined: first the transfer rule induction, by trying to identify a kind of translation templates for sentences, containing variable elements to be filled by smaller phrases; and second a clustering of seed terms, taken from bilingual dictionaries, with extended contexts. In the present context, the clustering is more relevant than the rule induction: similar to the Microsoft research, the idea to enrich a given translation with contextual information improves the overall translation result significantly.[3] The difference, of course, is that Brown (2001) considers local context windows (of 3 words both directions) while Menezes and Richardson (2001) use contexts derived by linguistic analysis results (logical forms).

As a result of this part of the survey, the following conclusions can be drawn:

- There is no publication which explicitly focuses on learning lexical transfer selection, i.e. on the automatic extraction of complex lexical transfer conditions from bilingual corpora. This may be due to the fact that it takes a certain dictionary size for this problem to be visible.
- Most of the papers do not make a clear distinction between structural and lexical transfer, and in fact treat both types of rules: often transfer rules still have lexical elements in them,

3 It even seems to be more important than the rule induction (Brown, 2001).

be it on the POS level (Sánchez-Martínez, 2007, 2009a; Hannemann et al., 2009), be it on a semantic/logical level (Menezes and Richardson, 2001; Jellinghaus, 2007). As a result there could be context available which could be used to disambiguate different translations.

- The papers which use small test corpora (Probst, 2005; Jellinghaus, 2007) rely on careful data preparation, and tiny lexica; it is not obvious to see how transfer selection strategies could be introduced there. Large corpora are used by Caseli et al. (2008), Sánchez-Martínez (2007, 2009a) and Hannemann et al. (2009) to extract rules on the POS level, Lavie (2008) on the deep-syntactic dependency level, Menezes and Richardson (2001) and Jellinghaus (2007) to extract rules on the semantic/logical level, and Brown (2001) for clustering contexts around known transfers. For solving problems in lexical transfer selection, large corpora are inevitable.
- As far as the representation of transfer rules is concerned, it seems that the following information plays a role:
  - The rule itself (formulated as a phrase structure part for both SL/TL side)
  - The alignment information (which SL part aligns with which TL part) (note that this allows for SL deletions and TL insertions)
  - Conditions on the SL side
  - Conditions on the translation side (carrying information from SL to TL side)
  - Conditions on the TL side

An example is given above (cf. Figure 4.3.2). A proposal for a more general transfer entry description in the context of LMF can be found in (Francopoulo et al., 2009); it allows for source and target tests as additional elements of an entry, among others, like examples.[4]

- The means to disambiguate different transfers for a given SL candidates depend on the capabilities of the systems; this is
  - local morphology (number, gender) (Calessi et al., 2008; Sánchez-Martínez, 2007, 2009a),
  - all kinds of syntactic annotations (Lavie, 2008),
  - logical form contexts (Menezes and Richardson, 2001) (related to the lexical elements contained in them),
  - word contexts (Brown, 2001)

A frequently used means of MT systems, namely to use subcategorisations/ grammatical relations, has been proposed by (Lavie, 2008). There are examples of successful extraction of such structures from monolingual text (cf. also Harper et al., 2001; Korhonen, 2002; Preiss et al., 2007).

Using word contexts to disambiguate translation selection (both in Menezes and Richardson, 2001 and Brown, 2001, 2003) was maybe not the intention of their work but is an important

4 The SL-TL conditions would have to be split into SL and TL conditions, and the alignment information would have to be added.

side-effect. Putting this aspect into the focus leads to the third topic of MT means for transfer selection, which is word sense disambiguation.

C. Survey on the creation of conceptual contexts for translation

Work in this area has mainly been carried out in the context of word sense disambiguation (WSD).

In WSD, the standard technique is to create context vectors for a candidate term, usually consisting of local contexts (a window of 2-3 words or concepts[5]), and topical context, usually documents or paragraphs. Different clusters of context vectors indicate different senses. Training is usually done in a supervised way.[6] At runtime, SVMs or hierarchical decision lists (Yarowsky, 2000) are used to assign a sense to a given candidate word, according to its context.

The standard approach has been modified and extended in different directions. Martínez et al. (2002) showed that adding syntactic features can improve the sense recognition; similarly, Dang and Palmer (2002) and Chen and Palmer (2009) apply linguistically rich models for improved sense disambiguation. Klein et al. (2002) state that standard classification techniques produce similar results, and propose a hierarchical combination of several classifiers. An overview of these activities, in the context of the SensEval and SemEval campaigns, is given in Agirre et al. (2009).

In a position paper on WSD, Resnik and Yarowsky (1997), stated that problems with WSD evaluation result from a low inter-rated agreement, which in turn results from the non-existence of predefined sense-inventories. They propose to use multilingual material for WSD: "The essence of the proposal is to restrict a word to restrict a word sense inventory to those distinctions that are typically *lezicalized cross-linguistically*" (Resnik and Yarowsky, 1997: p. 84).

In consequence, a number of papers using bilingual or even multilingual material for sense disambiguation appeared. However, it should be noted that target language material is only used to detect *monolingual* senses.[7]

Ide at al. (2002) use a parallel corpus of Orwell's '1984' in six languages to investigate whether the number of senses for given words which can be detected in such contexts could come close to the number of senses detected by humans. Apidianaki (2008) forms sense clusters based on different transfers of a candidate word, and classifies the target language context vectors to

5 For problems with overfitting in local contexts, cf. Hoste et al. (2002).

6 For a semisupervised approach cf. Yarowsky (1995); seed clusters are built to which additional features are added in a bootstrap method.

7 Just like in Tsang et al. (2002) Chinese material (some special particles) is used to detect English semantic verb classification.

detect similarities in the source word senses, which in turn are mapped into the source language.[8]

As the approach assumes that word senses correlate to different translations, it has been shown (Specia et al., 2006) that this relation does not really hold: they give examples showing that the number of translations often does not match the number of senses. Both cases exist:[9]

A given sense cluster translates into just one target word as in:

de *Zelle* (biology)	→	en *cell*
de *Zelle* (cloister)	→	en *cell*
de *Zelle* (terrorist)	→	en *cell*
de *Zelle* (battery)	→	en *cell*

One sense translates into several target words as in:

de *ausschlafen*	→	en *sleep in*
de *ausschlafen*	→	en *sleep out*

Their conclusion is that word senses should be defined on a task basis, and that "applying monolingual methods for multilingual WSD can either imply unnecessary work, or result in disambiguation errors" (Specia et al., 2006: p. 39).[10]

In the context of WSD, there is also research which relates WSD closely to the translation task. Here the objective is not to detect (monolingual) senses using translations, but to detect translations using different senses.

These approaches differ in the way they link word senses to translation equivalents. Some of them use (SL) word senses as an intermediate level for translation equivalent selection, and map the source words to senses first, and then look for translations for these senses. Kikui (1999) uses a sliding window to detect features to create clusters describing (predefined) word senses; the features are translated using a bilingual dictionary, and the translation is selected according to its similarity to one of the clusters; the translation points from one source language sense to its target equivalence. Similarly, Lee and Kim (2002) also map (source) words to senses, and senses to (target) words; they use a bilingual Korean—English dictionary both as a sense inventory and sense description context. Also, Miháltz (2005) creates senses from sense-tagged corpora; he reduces the number of senses by looking at the Hungarian translations, and assigns equivalents to the remaining senses based on (SL) conceptual clusters.

There has been a debate whether or not sense disambiguation improves statistical machine translation. Carpuat and Wu (2005) did not find improvements at first, perhaps because they

8 This raises the question as to whether it is best to stay in the source language right away and try translation/sense disambiguation there.

9 They use examples from English and Portuguese, esp. the high frequency verbs (get, come, give etc.).

10 However, this is not quite true: In contexts like term extraction from bilingual corpora, there is the need to form contextual clusters which are coherent in themselves to allow for searching for translation equivalents.

used a predefined sense inventory. Vickrey et al. (2005) try do without such an inventory, and just create sentential contexts for the translation candidates (in both directions); they report on improvements, e.g. significant reduction of the search space. Chan et al. (2007) also do not use senses as intermediate step but use the translations directly as senses. They report quality improvements due to the fact that sense disambiguation provides additional (non-local) contexts to the decoder which is not used otherwise. Other, more recent approaches of integrating state-of-the-art WSD methods into SMT to improve the overall translation quality were also successful (Carpuat and Wu, 2007; Giménez and Màrquez, 2007, 2009).

Stroppa et al. (2007), for example, added source-side contextual features to a state-of-the-art log-linear PB-SMT system by incorporating contextdependent phrasal translation probabilities learned by using decision trees. Up to two words and/or POS tags on either side of the source focus word were considered as contextual features. Bangalore et al. (2008) employed an SMT architecture based on stochastic finitestate transducers that addresses global lexical selection, i.e. dedicated word selection. Specia et al. (2008) use dedicated predictions for the re-ranking of n-best translations, limited to a small set of words from different grammatical categories. Significant improvements were observed in both approaches. Hasan et al. (2008) present target context modeling into SMT using a triplet lexicon model that captures long-distance (global) dependencies. Their approach is evaluated in a reranking framework; slight improvements are observed over IBM Model 1 (Snover et al., 2006).

Recently, Haque et al. introduced dependency relations (2009a) and supertags (2009b) in his PBMT, to exploit source similarity in addition to target similarity, as modelled by the language model. However, it has been observed that the improvement gained through source context modelling tends to diminish with the increase in the training data size. But, for language pairs suffering from the scarcity of large amount of parallel corpus, source context modelling proves to be very useful.

In rule-based environments, approaches differ in the way contexts are used. Thurmair (2006) reports on a disambiguation procedure which uses source language corpus material. For each possible translation of a candidate term, context vectors are created in a supervised learning step; at runtime the best context for a translation is computed using a standard similarity measure. Results show over 90% correct disambiguation. Jassem et al. (2000) also use context vectors for translation disambiguation, however built on the target side (like in SMT). While source language disambiguation is easier to integrate into the workflow, target language disambiguation based on a language model may be stronger in using local contexts, e.g. for near-synonyms or collocations which have similar contexts in the source but specific translations in the target context. Systems like METIS (Carl et al., 2005; Carl, 2008) therefore delay the transfer selection decision, provide all alternatives to the generation, and use target language models for disambiguation.

As a result, identification of conceptual context for transfers is possible along three lines:

- Building source term clusters using a parallel corpus, by creating subsets based on the possible translations of the candidate source term; clustering of the source term contexts,

and using these clusters at runtime as source language indicators for a specific translation. Transfer selection can then be done during analysis (in fact, as a preprocessing step based on contexts larger than sentences (cf. Mihálta, 2005; Thurmair, 2006).

- Building the same subset of translations but building context vectors on the target side. In a given sentence, transfer selection would then have to be delayed until the target context is available.

- Not building clusters at all but using a target language model for disambiguation. All transfer decisions would be disambiguated by the decoder using the target LM, like in standard SMT. This approach requires a target LM as additional resource in a rule-based context.

The PANACEA development will compare these approaches and try to find the most adequate solution.

5 Existing tools

This section discusses tools and components for the relevant tasks available to the consortium (either as publicly available software or in-house products). Detailed information on the tools that will be integrated in the platform can be found in the Appendix A of this document.

We will integrate in the platform at least one tool for each of the tasks tackled in WP5. However, for several tasks covered in this WP there is more than one tool available. In any such case it would be premature to choose one of them to be integrated in the platform at this step. Instead, the decision will be taken by looking at their comparative performance when they are incorporated in the solution path (see section 7). This procedure will guarantee that the decisions taken are based on solid criteria.

5.1 Alignment

5.1.1 Sentential alignment

Several tools for sentence alignment are publicly available. The most widely used ones are:

Hunalign (Varga et al., 2005) can work in two modes. If a bilingual dictionary is available, this information is combined with sentence-length information (Gale and Church, 1991) and used to identify sentence alignment. In the absence of a bilingual dictionary, Hunalign first falls back to sentence-length approach and identify the aligment using this information only. Then, It builds an automatic dictionary based on this alignment. Finally, it realigns the text in a second pass, using the automatically induced dictionary.

GMA – Geometric Mapping and Alignment (Argyle et al., 2004) is an implementation of the Smooth Injective Map Recognizer (Melamed, 1997) algorithm for mapping bitext correspondence and the Geometric Segment Alignment (Melamed, 1996) post-processor for converting general bitext maps to monotonic segment alignments. The tool employs word correspondences, cognates, as well as information from bilingual dictionaries.

BSA – Bilingual Sentence Aligner (Moore, 2002) is a three-step hybrid approach. First, sentence-length based alignment is performed; second, statistical word alignment model is trained on the high probability aligned sentences and third, all sentences are realigned based on the word alignments. It generates only 1:1 alignments.

The quality of sentence alignment depends on the translation quality of parallel data. Most of the recent tools perform with comparable results. Varga et al. (2005) reported results of Hunalign and BSA on a good quality parallel corpus. Both achieved precision in the range of 97–99% and recall in the range of 97–98%.

5.1.2 Sub-sentential alignment

For sub-sentential alignment we will use the current state-of-the-art tools which are publicly available:

Giza++ (Och and Ney, 2003) is a statical machine translation toolkit that is used to train IBM Models 1-5 and an HMM word alignment model.

BerkeleyAligner is an alternative to Giza++. It is a word alignment toolkit combining unsupervised as well as supervised approaches to word alignment developed at University of Berkeley (Haghighi et al., 2009; DeNero and Klein, 2007; Liang et al., 2006). It features joint training of conditional alignment models (cross-EM), syntactic distortion model, as well as posterior decoding heuristics.

OpenMaTrEx (Dandapat et al., 2010) is a free/open-source example-based machine translation system based on the marker hypothesis. It comprises a marker-driven chunker, a collection of chunk aligners and two decoders. For WP5 purposes, only the chunker and chunk aligners will be used.

Subtree aligner (Zhechev, 2009) is an open-source system for fast and robust automatic generation of parallel treebanks. It implements the algorithm proposed by Zhechev and Way (2008) (see section 4.1.2).

5.2 Bilingual dictionary induction

Apart from word aligners like **GIZA++** and **BerkeleyAligner** which can also be used to induce bilingual lexica, we will consider these tools to be integrated in the platform:

NATools (Simoes, 2003) is a workbench for parallel corpora processing. It includes a PTD extractor based on (Hiemstra, 1998) .

K-vec++ (Varma, 2002) implements an extended version of the K-vec algorithm (Fung and Ward Church, 1994). This is based on the fact that if two words are translations of each other, then they occur almost an equal number of times and approximately in the same region in the parallel text.

Uplug (Tiedemann, 2003) is a collection of tools for linguistic corpus processing, word alignment and term extraction from parallel corpora.

Word packing (Ma et at., 2007) is a method implemented in an Yanjun Ma's aligner that can handle 1-to-n alignments, and therefore it can be used to build bilingual dictionaries of asymmetric (non-compositional) MWEs.[11]

5.3 Transfer grammar induction

There are no publicly available tools for transfer grammar selection, or automatic lexical transfer selection, to our best knowledge. Closest to our needs are some of the Apertium tools.

11 Documentation forms for this tool will be added to this document in a forthcoming version, as it is being packaged at the moment of writing this report.

5.3.1 Apertium

There are not really tools available for automatic transfer rule creation. Commercial RBMT system providers do not release them, and the only Open Source RBMT system that provides help is the Apertium system (http://www.apertium.org/) (cf. Ginestí Rosell, 2010). It has a toolbox for dictionary development, called dixtools. They provide the following functionality:

```
Tasks:
  cross:              cross 2 language pairs (using linguistic res. XML file - see Cross Model)
  cross-param:        cross 2 language pairs (using command line parameters) Crossdics
  merge-morph:        merges two morphological dictionaries (monodix) Merge dictionaries
  equiv-paradigms:    finds equivalent paradigms and updates references
  list:               lists entries in a dictionary - see Dictionary reader
  dix2trie:           create a Trie from an existing bilingual dictionary
  dix2tiny:           create data for mobile platforms (j2me, palm) from bidix
  reverse-bil:        reverses a bilingual dictionary
  sort:               sorts (and groups by category) a dictionary - see Sort a dictionary
  format:             Format dictionaries (according to Generic Options)
  fix:                fix a dictionary (remove duplicates, convert spaces)
```

Figure 5.3.1: Apertium dictionary tools.

Wit respect to bilingual dictionary creation, the authors state on the webpage:

“We also need a bilingual dictionary. If they aren't available, we have tools available to help construct them automatically: Crossdics as I mentioned in my article, and ReTraTos which can build Apertium-format dictionaries from the same alignments generated by GIZA++ - the output of this should be manually checked, however, as it can output many questionable entries, particularly with multiword expressions. Crossdics (part of apertium-dixtools) is a program that can be used to "cross" language pairs. That is, given language pairs aa-bb and bb-cc it will create a new language pair for aa-cc”.

As the Apertium dictionaries have a somewhat idiosyncratic format, and cannot describe non-local transfer tests and actions it remains to be evaluated whether the tools can be used in PANACEA.

5.3.2 Required tools

For the three types of tasks in WP 5.3, the following tools are required:

1) For topic tests, a topic identification component is needed. We will use an adapted Linguatec tool for this purpose (LT-TopicIdentifier), with the taxonomy provided by this tool (we will not re-train the classifier)
2) For grammatical tests, a parser for German and English is required. We will use a modified version of the Linguatec parser (LT-Parser) for this purpose, and add a tree-comparison and extraction component.
3) For conceptual tests, a classifier similar to the one of the topic identifier is required, so this technology will be adapted to the task at hand.

Such tools will be adapted to the PANACEA task 5.3 in a first implementation phase.

6 Resource description

This section provides description of data resources (project deliverables and internal resources) to be produced in WP5.

6.1 Alignment

6.1.1 Types of parallel corpora

The following types of parallel corpora (with respect to different types of alignment) will be created for the purposes of other tasks in WP5 (WP5.2 Bilingual dictionary induction and WP5.3 Transfer grammar induction) as well as other WPs (WP7 Evaluation of components integration and produced resources, WP8 Evaluation of industrial environments, see D7.1 and D8.1 for details):

1) *Sentence-aligned parallel corpora* as the deliverable D5.3: sententially aligned texts, cleaned and prepared for training an SMT system.
2) *Word-aligned parallel corpora* as an internal package: word-aligned corpora from D5.3 ready for translation table extraction.
3) *Chunked-aligned parallel corpora* as an internal package: chunk-aligned corpora from D5.3 ready for translation table extraction.
4) *Subtree-aligned parallel corpora* as an internal package: subtree-aligned corpora from D5.3 ready for translation table extraction.

All the produced resources will be provided in the format described in Section 6.1 “Travelling object. Corpus and data format” of D3.1.

6.1.2 Domains and languages

The above types of corpora will be provided at least for these language pairs and domains: The general domain corpora will be provided for English-German, English—Greek, and English—French. Corpora for the automotive domain will be provided for English—German, corpora for the environment and legal domains will be provided for English—Greek and English—French. The news domain corpora will be used as a fall-back option if PANACEA is less successful in acquiring corpora from other domains. An overview of the language-pair/domain distribution is presented in Table 6.1.1:

language pair/domain	*general*	*automotive*	*environment*	*legal*	*news*
English—German	√	√			?
English—Greek	√		√	√	?
English—French	√		√	√	?

Table 6.1.1: Language pairs and domains of parallel corpora to be provided by WP5.1

Four parallel corpus types will be provided for each language pair/domain combination, which makes a total of 32 corpora to be produced. The domain-specific corpora may be produced in multiple versions (with improving quality and size from version to version) depending on the output of WP4 and depending on advances in parallel sentence extraction from comparable corpora achieved during the course of the project.

6.2 Bilingual dictionary induction

For the representation of the bilingual dictionary, formats exist, but most of them are idiosyncratic to specific MT systems. So it seems to be advisable to first define the type of information which a bilingual dictionary should contain, and then describe their representation.

Bilingual dictionaries usually contain the following information items:

- source language lemma (can be single or MWE)
- target language lemma (can be single or MWE)
- source language part-of-speech
- target language part-of-speech
- (reading)

The reading annotation would be needed in cases of entries which are identical in source and target lemma and POS, but differ in meaning as in:

en *Barcelona* (*ProperNoun*)	de *Barcelona* (*ProperNoun*)	// the city
en *Barcelona* (*ProperNoun*)	de *Barcelona* (*ProperNoun*)	// the province
en *cell* (*Noun*)	de *Zelle* (*Noun*)	// prison
en *cell* (*Noun*)	de *Zelle* (*Noun*)	// battery

However, it is usually not represented; the only 'surrogate' which sometimes is coded is a domain tag. Accordingly, a standard bilingual dictionary consists of a lemma and POS tag in both source and target languages.

The following bilingual dictionaries will be created for the purposes of task 3 in WP5.3 (transfer grammar induction) as well as for other WPs:

- A dictionary covering the automotive domain for the language pair English—German. This dictionary is expected to hold between 5 and 10 thousand lemmas. The evaluation will be carried out in WP8 and the results reported in D8.2 (t36).
- Dictionaries covering the domains of legislation and environment for the language pairs English—French and English—Greek. Each of these dictionaries is expected to hold circa 100,000 lemmas. These dictionaries will be provided in D5.5, while their evaluation will be carried out in WP7.

6.3 Transfer grammar induction

As explained above, the transfer tests are stored as annotations to bilingual lexicon entries. They are required if several translations exist for a given source lemma.

It can be seen that transfer entries are not independent of each other in cases where several translations exist: The different translations will be distinguished by tests and actions; tests may be applied in a specific order (cf. Section 7.3 below). The representation must therefore provide annotations which help the disambiguation process. Such annotations include:

- alignment of the lemma parts (important in case of multiwords)
- probability of the translation (frequency of this translation, related to the frequency of all possible translations)[12]
- sequence of tests (to be used in cases where the translation should be selected at transfer time)
- conditions of transfer selection actions following a transfer selection (covering both source-target actions (e.g. mapping of prepositions) and target actions (e.g. setting some number of gender values)

In terms of formalism, there is no proposal yet which covers all these required annotations. Proposals in the framework of LMF (e.g. Francopoulo et al., 2009) do not seem to support all requirements for transfer entries yet; they only foresee tests, and they seem to assume bidirectional transfers. Also, proposals like OLIF (Lieske et al., 2001; www.olif.net) do not cover all aspects required by transfer dictionaries.

The proposed format extends the definition of the Stat-XFER project (Lavie, 2008) where transfer rules contain similar elements, cf. Figure 6.3.1. It is pragmatic and intended to explicit enough to convert the transfer dictionaries into any emerging standard of transfer entries. The core is given in Figure 6.3.2.

The features for the SL-Test would be something like '*domain = sports*' (for topic tests), '*hasDirectObject = TRUE*' (for grammatical tests), '*concepts = cluster12*' (reference to a cluster of conceptual contexts). The details will be defined later.

```
{NP1,2}
;;SL: $MLH ADWMH
;;TL: A RED DRESS
;;Score:2
NP1::NP1 [NP1 ADJ] -> [ADJ NP1]
(
 (X2::Y1)
 (X1::Y2)
 ((X1 def) = -)
 ((X1 status) =c absolute)
 ((X1 num) = (X2 num))
 ((X1 gen) = (X2 gen))
 (X0 = X1)
)
```

```
{NP1,3}
;;SL: H $MLWT H ADWMWT
;;TL: THE RED DRESSES
;;Score:4
NP1::NP1 [NP1 "H" ADJ] -> [ADJ NP1]
(
 (X3::Y1)
 (X1::Y2)
 ((X1 def) = +)
 ((X1 status) =c absolute)
 ((X1 num) = (X3 num))
 ((X1 gen) = (X3 gen))
 (X0 = X1)
)
```

Figure 6.3.1: Examples of transfers (from Lavie, 2008);[13]
more examples in Figure 4.3.1 and 4.3.2 above.

12 There should be support for cases where bilingual dictionaries are used to improve/replace translation tables (cf. Koehn 2010). This feature is also important in architectures where transfer decisions are delayed until generation takes place.

13 Although this example goes beyond the lexical level, it shows some of the main elements.

```
TranferEntry    ::=   PackageID  SL-Lemma  SL-POS  Transfer+
PackageID       ::=   <integer>
SL-Lemma        ::=   <string>
SL-POS          ::=   <a legal POS of the source language>
Transfer        ::=   TransferID TL-Lemma TL-POS Alignment Probability
                      Testsequence SL-Test+ SL-TL-Action+ TL-Action+
TransferID      ::=   <integer>
TL-Lemma        :=    <string>
TL-POS          ::=   <a legal POS of the target language>

Probability     ::=   <double>
Alignment       ::=   (SLposition  '::'  Tlposition)+
SLposition      ::=   <integer>
TLposition      ::=   <integer>

Testsequence    ::=   integer

SL-Test         ::=   No  SLconstituent feature  '='  value
SL-TL-Action    ::=   No  SLconstituent feature  '='  TLconstituent feature
TL-Action       ::=   No  TLconstituent feature  '='  value
No              ::=   <integer>                          // test or action number
SLposition      ::=   <integer>                          // SL-wordnumber
TLposition      ::=   <integer>                          // TL-wordnumber
SLconstituent   ::=   'SL'  <integer>                    // the reference in SL
TLconstituent   ::=   'TL'  <integer>                    // the reference in TL
feature         ::=   <a legal feature name>
value           ::=   <a legal value name>
```

Figure 6.3.2 Format of annotated bilingual dictionary.

7 Solution path and work plan

This section presents the envisaged solution path and details the work plan and strategy to be followed during the lifetime of the WP. It also establishes which available tools will be integrated in the factory and which tools will have to be developed. The evaluation details on each of the tasks tackled are found in Deliverable 7.1.

7.1 Alignment

Here, we describe the general strategy applied in WP5.1, the setup identifying required resources, and the main steps of the solution path.

7.1.1 Strategy

The following strategy will be applied in order to produce the resources described in Section 6.1 of this report. It will consist of a preprocessing step relevant for all types of resources, a sentential alignment step specific for parallel corpora and comparable corpora, and a sub-sentential alignment step specific for each type of sub-sentential alignment. The general solution path for WP5.1 is given in Figure 7.1.1. External resources and tasks are in blue. The parsing step is optional, because it is not necessarily required for subtree alignment.

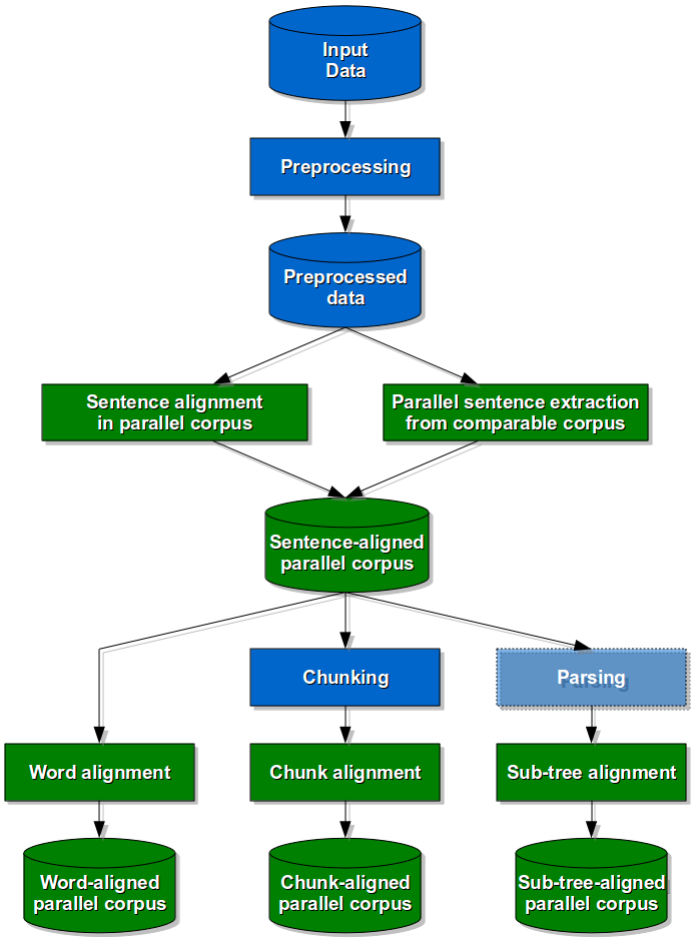

Figure 7.1.1: WP5.1 Solution path.

7.1.2 Setup

WP5.1 will require two types of bilingual resources:

1) *general-domain corpora* for language pairs as specified in Section 6.1 of this report. The publicly available Europarl corpus (Koehn, 2005) will be used for this purpose. It consists of several parallel corpora which include all project language pairs, it has a sufficient size of about 1—2 million sentence pairs (40—50 million words) for each language pair, and can be assumed to contain general domain texts (it is extracted from the proceedings of the European Parliament).
2) *domain-specific corpora* for language pairs and domains as specified in Section 6.1 of this report. These resources will be delivered by WP4 in a form of document-aligned comparable corpora described in Section 6 of D4.1. These corpora should contain enough material to extract a parallel corpus of 10—20 thousand sentence pairs (250—500 thousand words) for each language pair and domain.

7.1.3 Preprocessing

The first step of this task is external and will be performed within WP4. All parallel and comparable corpora used in WP5.1 will be tokenized, sentence-segmented, and analysed and disambiguated on morphological level (each token will be provided with a POS tag and lemma).

7.1.4 Parallel sentence alignment and extraction from comparable corpora

This step will be specific with respect to the character of the input data. If it is parallel data, standard sentence alignment tools will be applied (e.g. Hunalign which is widely used and known to perform with the state-of-the-art results (Varga et al., 2005). For comparable data, sentence extraction techniques will be employed. Since there are no tools for this specific task available, we will have to either modify one of the sentence aligners (described in Section 5.1.1) to be applicable on comparable data or develop a new tool from scratch. This tool will be integrated to the factory as a webservice.

7.1.5 Sub-sentential alignment

Sub-sentential alignment will be specific for each type of aligned parallel corpora to be produced. Giza++ will be used for word alignment. Currently, it is the most widely used tool for this task in MT and produces high quality alignments for translation phrase extraction (Ma, 2009). We will also consider BerkeleyAligner as an alternative to Giza++ for integration into the platform because as opposed to Giza++, once trained, it can be applied to single sentences. The OpenMaTrEx chunk aligner will be employed for chunk alignment, and Subtree Aligner will be used for subtree alignment. Both of these tools have been developed at Dublin City University (DCU) and we will take advantage of their developers to assist with integration of the tools into the platform. All these aligners will be integrated in t22 as deliverable D5.2.

7.1.6 Additional experiments

After a comparable corpus is produced by WP4, a series of experiments with the sentence alignment tools from Section 5.1.1 will be carried out in order to evaluate their ability to be used for extraction of parallel sentences from comparable corpora. Based on this evaluation, either one of the tools will be selected (and modified if needed) and integrated into the platform, or else a new tool will be developed.

An additional set of experiments will be performed in the area of MT domain adaptation. The basic approach takes a union of general-domain and domain-specific data and use it for training an SMT system. Other approaches are based on system combination rather than on data combination. In both areas, we have already performed some experiments and further research in this area will be carried out during the later stages of the project:

1) For the Workshop on Statistical Machine Translation 2010 translation task and system combination task we have developed MT systems based on system combination which outperformed most of the other systems. In manual and automatic evaluation, our systems were the best performing ones in English-Spanish translation task, and English-to-Czech and English-to-French system combination tasks (Penkale et al., 2010; Du et al., 2010).
2) In cooperation with Trinity College Dublin in June 2010, we have developed a system for automatic translation of Tweeter messages related to football World Cup 2010 (http://myisle.org/worldcup2010). A general domain MT system was adapted to the domain of football by careful selection of training data.

7.1.7 Testing

All testing of parallel tools and resources in WP5.1 will be done extrinsically by MT systems, as described in D7.1. If needed, manual analysis of results will be carried out, too.

7.2 Bilingual dictionary induction

The current task builds on top of the alignments produced in WP5.1 applying techniques which address the issues of precision, thus reducing the amount of human intervention needed on the resulting dictionaries.

7.2.1 Strategy

The bilingual dictionaries that can be automatically derived from the alignments produced in WP5.1 should be suitable for an SMT system, but probably do not have the required precision for more general uses. Therefore, additional processing which emphasises precision is necessary. In addition, there might be terms whose low frequency does not justify their inclusion in a dictionary.

7.2.2 Setup

This task will require two types of sub-sententially aligned data, both to be provided by WP5.1:

1) *Word-aligned data* for dictionaries of single words and non-compositional MWEs.
2) *Chunk-aligned and subtree-aligned data* for dictionaries of compositional MWEs.

7.2.3 Methodology

The basic idea is then to exploit (i) confidence measures provided by the alignment algorithms and (ii) frequencies of the aligned terms in the input corpora. Based on these, we will experiment with different filtering techniques in order to derive dictionaries which have higher quality and thus require less amount of human intervention. Finally, the entries of the dictionaries will be linguistically annotated using the PANACEA annotation tools (described in D4.1).

Regarding MWEs, two lines will be followed in this WP. First, we will apply word packing on top of the word-aligned corpora in order to derive correspondences for non-compositional MWEs. Second, the chunk-aligned and tree-aligned corpora will be filtered in order to increase the precision of the aligned elements. By doing this, we will have better quality correspondences for compositional MWEs.

WP6 deals with MWEs too but from a monolingual perspective. Initially no overlap between the treatment of MWEs in both WPs is foreseen as the languages covered (Italian in WP6 and English, French and Greek in WP5) differ. However, in case the MWE monolingual component is applied to English (see D6.1), there would be an interaction that might lead to an improvement of the proposed methods. In any case, both tasks will stay in touch in order to identify possible interactions that might benefit each other.

7.2.4 Testing

The dictionaries induced will be evaluated intrinsically (see section 3.2. of D7.1 for details). Besides, these dictionaries are the input of WP5.3, so the evaluation carried out there will extrinsically evaluate the quality of the dictionaries.

Furthermore, the corpora used for the induction (for each domain and language pair) will be evaluated with respect to the their suitability to induce dictionaries by computing their translation fertility and alignment density. These figures will allow to derive more meaningful conclusions from the quality of the dictionaries produced.

7.3 Transfer grammar induction

In Section 3.3. we reported that three very common types of transfer selection will be exploited:

1) *domain tag marking*, i.e. we want to know whether competing translations can be disambiguated by setting a subject area tag;
2) *grammatical analysis*, i.e. we try to disambiguate based on grammatical context of a given translation candidate;
3) *conceptual analysis*, i.e. we want to know whether a translation can be selected along the lines of collocational analysis.

The three approaches need different sets of tools.

7.3.1 Strategy

The strategy will consist of a preparatory phase, relevant for all approaches, and specific strategies for the three research directions. These strategies follow a supervised learning approach; we pre-define the topic taxonomy, the syntactic patterns and templates the system should search for, and the conceptual clusters.

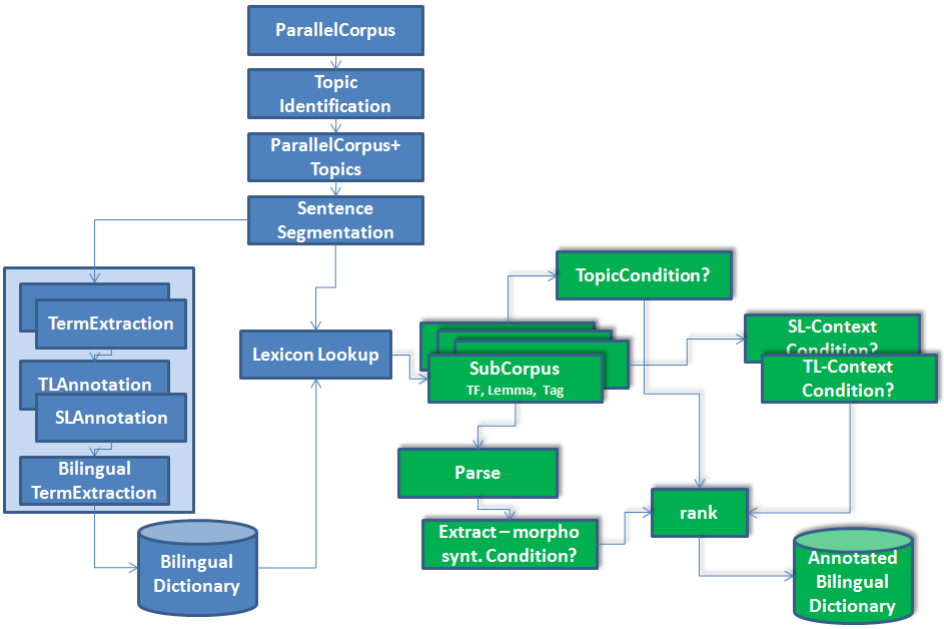

Figure 7.3.1: WP5.3 Solution path.

The general solution path is given in Figure 7.3.1. Two 'external' tasks are required as preprocessing (the blue parts of the figure): collection of sentence-aligned bilingual corpora, and creation of dictionaries. The PANACEA tasks 5.1 and 5.2 will create such resources. For the time being, an existing bilingual dictionary will be used instead of the PANACEA tool, to be able to start the developments.

The green parts of the figure show the specific developments for transfer selection, all starting from subcorpora created by splitting the corpus for a given SL term into parts related to the respective translations.

7.3.2 Setup

The setup requires mainly two kinds of resources: corpus data and dictionary data.

A. Corpus Data

Work in this task needs a sentence-aligned parallel corpus. Some of the following data will be used: Europarl v5 (Koehn, 2005), JRC-Acquis (Steinberger et al., 2006), EMEA (Tiedeman,

2009), MultiUN (Eisele and Chen, 2010). The language pairs are German—English. For the different subtasks in WP 5.3, subcorpora will be created, described below. Parts of these subcorpora will be kept aside for later tuning and testing.

B. Dictionary Data

Work on transfers will use a dictionary provided by the Linguatec resources, containing more than 300 thousand lemmata. In a production environment, dictionary data would be collected by running term extraction tools (either monolingual and bilingual extraction, or all in one using GIZA++), and annotating the entries (lemmata) using the PANACEA annotation tools. For WP 5.3, however, it is assumed that such a dictionary has already been created, and an already existing bilingual dictionary for German—English is used. Basically the interest is in entries with 1:N transfers.

Also, in a production environment, only unknown entries would have to be processed in the way described in the current task.

7.3.3 Preprocessing

The corpus data must be preprocessed.

A. Linguistic preprocessing

Preprocessing consists of tokenisation, lemmatisation, and POS-tagging. Each sentence finally is supposed to be in a relevant format described in D3.1. After lemmatisation, tokenisation will be modified to take multiword entries into account.

B. Word candidate production

In an additional column, for each content word all possible translations as found in the bilingual dictionary are stored, together with their POS information. Based on this information, all content words with more than one translation are collected in a special sub-dictionary; they are the candidates of the following investigations. In case too many candidates are found a selection based on frequency will be used.

C. Language and topic Iidentification

Then, for each document, each paragraph, and each sentence of the corpus, language and topic are determined using a language and topic identifier. This tool uses a taxonomy which distinguishes about 40 different classes, following the topic hierarchy used in the Linguatec translation system.

The result is annotated in the header tag of the respective element (<doc>, <p>, <s>, respectively), thus enabling the system to have subtopics for paragraph or sentences in a larger topic. If no topic can be assigned, no attribute is set, and the topic must be inherited from the next large tag. As a result, every sentence has a topic assignment. As the quality of the topic assignment is an essential factor in this strategy, manual evaluation of some of the assignments

is foreseen. The SL terms will be indexed to create subcorpora for each SL terms and each of its translations. Indexing will be done both on the paragraph and on sentence level.

7.3.4 Topic test processing

For each translation of the SL candidate word, a TL-specific subcorpus is produced, and for each subcorpus, its context (in terms of topic annotation) is inspected to decide if it can be disambiguated using the topic annotation from the topic assignments of the other translations of the candidate term. For each SL term, correlations between translation equivalents and topic assignments will be calculated; a measure will be computed as to how reliably a translation equivalent can be predicted from a certain topic setting.

There will be translations with broad coverage (occurring in many topics), and maybe some with restricted coverage, occurring in only one, or rather few, topics. The more specific translations will be marked with domain tags. Testing consists of enlarging the data sets for the domain-tagged translations found, and verifying whether it really only is selected if the annotated domain is found.

7.3.5 Grammatical testing

This approach consists of searching syntax trees for the pre-specified set of transfer-relevant phenomena, checking whether some of them can be assigned to some subcorpus, and differentiating it from the other subcorpora. The envisaged solution path is the following:

1) Create a subcorpus for each of the translations of a given SL candidate word. Remove long and complex sentences.
2) 2.Parse every sentence containing a translation candidate, create an analysis tree. To do this, the Linguatec parser will be used.
3) Inspect the analysis tree for the phenomena mentioned in the list of grammatical phenomena to be investigated as transfer disambiguation candidates. These phenomena will be described in the form of underspecified tree structures; each of them will be matched against the input tree. Extract the respective configurations, annotate the translation candidates accordingly.
4) Identify, for all members of a TL-related subcorpus, if sufficiently symetrical grammatical contexts exist which would justify the use of such contexts for translation disambiguation.
5) Compare the different TL-related corpora to find whether the disambiguation criterion found is valid for all (or sufficiently many) contexts; readjust the contexts, and redefine the disambiguation criteria.
6) Validate the disambiguation criteria found by enlarging the context to yet unseen examples.

In the case where transfer actions can be considered, a similar procedure would be implemented on the target language side, to search for common grammatical properties of the sub-corpus of the translations of a given source term. Candidates of interest are argument switching, and preposition determination.

7.3.6 Conceptual context determination

This approach consists of creating vectors of concepts which are able to be used for disambiguation. It is like WSD where the senses to be distinguished are predefined based on TL-concepts, and sense-related sub-corpora are already available. The analysis steps are the following:

1) For each TL-related corpus, create a vector of concepts per paragraph describing the context of the candidate. Concepts are weighted, based on frequency and on their semantic distance to the translation candidate.
2) Each term is then weighted according to its relevance for a given vector. The best measure for this particular setup will be determined (mutual information, TF-IDF etc.). One of the issues is to decide whether the reduction of feature vector dimension is necessary in this setup.
3) The resulting vectors will be used as conceptual contexts for testing the transfer options, i.e. the dictionary will contain links to such vectors as transfer tests.

7.3.7 Order of tests, entry packages

The analysis of transfer tests so far has shown that several criteria will have to be used to support disambiguation of a translation candidate. Such criteria use different knowledge sources and work on different levels of representation. In a given context, it may be the case that one translation is disambiguated by a domain tag, while another is disambiguated by a syntactic constraint. This fact raises the question of the order of the tests. The order of tests therefore is a significant issue.

The need to define the order of tests has a severe consequence for the whole dictionary structure, as it introduces contextual information, i.e. the dictionary entries are no longer independent of each other. Instead, the test order information creates groups of bilingual entries, i.e. entry packages, defined by source lemma and source part-of-speech, which have multiple translations; among those, a connection exists in the sequence of tests. This fact has two consequences:

1) Dictionary entries are not reversible any more; i.e. dictionaries with transfer tests are directional. It is a simplification to believe that dictionaries can be used bidirectionally.
2) Adding new entries to a dictionary affects the existing entries. If an entry is added to a package, then the whole package needs to be balanced anew: tests may have to be modified, and the sequence of tests must be adapted.[14]

Several heuristics for ordering tests have been applied so far:

14 Of course this is also true for data-driven MT where appropriate training data would have to be added to the corpus, and the whole translation table would have to be rebuilt.

- Manual ordering, for each package. While this may turn out to be the most fine-grained way, it means that a package has to be manually re-ordered whenever a new entry is added. This increases dictionary maintenance cost.
- Ordering according to the number of tests, assuming that transfers with a higher number of tests are more specific than transfers with fewer tests; accordingly the more specific entries would be taken first.
- Ordering according to the kind of tests: for instance, at first, all grammatical tests are tried, and then all domain tag tests are executed. This leads to the consequence that a transfer with a syntactic test is selected even if it has a 'wrong' domain tag, i.e. the domain tag is overwritten. End-users who code domain tags may not be aware of this fact.
- Ordering according to the specificity of a test: for instance, all tests containing lexical material are tried first (In METAL such tests were called 'hard tests', as opposed to 'soft tests' without lexical material), assuming that if a test contains a specific lemma it must be a multiword representation: if English 'kick' has a direct-object test for 'lemma=bucket', then this is a strong argument in favour of executing this test first.
- As there will always be cases in which the test sequence turns out to produce improper results, it remains to be explored how a good sequence, mirroring human intuition, might look like. Due to limited resources, PANACEA will not do any research in the area of test ordering, but it is clear that corpus-based approaches can definitely help in fine-tuning test sequences.

7.3.8 Testing

Testing the component will include the following questions:

- How well suited is each of the different transfer test strategies (topic test, grammatical tests, conceptual context tests)? This will be tested by running translations with the single strategies.
- Does one strategy include the others? It could turn out that the conceptual context strategy subsumes the topic tests, as both use conceptual contexts.
- How are grammatical tests and conceptual tests related?
- How are contextual disambiguation strategies of source and target contexts related?
- How important is the ordering of tests, based on these results?

These questions will be answered using some selected lexical entries of different POS, in both language directions, as test data. Existing MT dictionaries may be used as a reference.

8 Bibliography

Abney, S. (1991). Parsing by chunks. In Principle-Based Parsing, pages 257–278. Kluwer Academic Publishers, Dordrecht.

Agirre, E., Màrquez, L., Wicentowski, R. (2009). Computational semantic analysis of language: SemEval-2007 and Beyond. In Language Resources and Evaluation. 43,2, p.97 104.

Alshawi, H., Carter, D., Rayner, M., Gambäck, B. (1991). Translation by Quasi Logical Form transfer. Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.

Andersen, G. (2008). Quantifying domain-specificity: the occurrence of financial terms in a general corpus. In Synaps 21, 2008: Festskrift for Magnar Brekke. Norges Handelshøyskole. p. 37-52

Argyle A., Shen L., Stenchikova S., and Melamed D. I. (2004.) Geometric Mapping and Alignment (GMA) tool. http://nlp.cs.nyu.edu/GMA/

Bangalore S., Haffner, P., Kanthak, S. (2007). Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. In Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL-07), Prague, Czech Republic, pp.152–159.

Blunsom, P. and Cohn, T. (2006). Discriminative word alignment with Conditional Random Fields. In Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, pages 65–72, Sydney, Australia.

Brown, P. F., Della-Pietra, S. A., Della-Pietra, V. J., and Mercer, R. L. (1993). The mathematics of Statistical Machine Translation: Parameter estimation. Computational Linguistics, 19(2):263–311.

Brown, R. (2001). Transfer-Rule Induction for Example-Based Translation. Proceedings of the MT Summit VIII.

Brown, R. (2003). Clustered Transfer-Rule Induction for Example-Based Translation. In Carl,M., Way, A.: Recent advances in example-based machine translation (Kluwer).

Cabré Castellvi, M.T., Estopá Bagot, R., Vivaldi Palatresi, J. (2001). Automatic term detection: A review of current systems. In Bourigault, D., Jacquemin, Chr., L'Homme, M.-Cl. (eds.), Recent Advances in Computational Terminology.

Cabré, M.T. (1999). Terminology: theory, methods, and applications. John Benjamins.

Capuat. M., Wu. D. (2005). Word sense disambiguation vs. statistical machine translation. In Proceedings of the 43rd Annual Meeting of the ACL.

Carbonell, J., Probst, K., Peterson, E., Monson, Chr., Lavie, A., Brown, R., Levon, L. (2002) Automatic Rule Learning for Resource-Limited MT. In Proceedings of the 5th Conference of the Association for Machine Translation of the Americas.

Carl, M. (2007). METIS-II: the German to English MT system. In Proceedings of Machine Translation Summit XI. Copenhagen, Denmark.

Carl, M. (2008). Using log-linear models for tuning machine translation output. In Proceedings of the sixth international conference on language resources and evaluation (LREC 2008). Marrakech, Morocco.

Carl, M., Schmidt, P., Schütz, J. (2005). Reversible template-based shake & bake generation. In MT summit X Workshop: Second workshop on example-based machine translation. Phuket, Thailand, pp 17-26.

Carpuat, M., Wu, D. (2005). Evaluating the word sense disambiguation performance of statistical machine translation. In Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP-05), Jeju Island, Republic of Korea, pp.120-125.

Carpuat, M., Wu, D. (2007). Context-dependent phrasal translation lexicons for statistical machine translation, Proceedings of MT Summit XI, Copenhagen, Denmark, pp.73-80.

Caseli, H.M., Nunez, M.V, Forcada, M. (2008). Automatic induction of bilingual resources from aligned parallel corpora: application to shallow-transfer machine translation. In Machine Translation, 20:227–245.

Chan, Y.S., Ng, H.T., Chang, D. (2007). Word sense disambiguation improves statistical machine translation. In Proceedings of the 45th Annual meeting of the Association for Computational Linguistics (ACL-2007), Prague, Czech Republic. pp.33-40.

Chen, J., Palmer, M.S. (2009). Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. In Language Resources and Evaluation, 43,2, 181-208.

Copestake, A., Briscoe, T., Vossen, P., Ageno, A., Castellon, I., Ribas, F., Rigau, G., Rodriquez H., and Samiotou A. (1994). Acquisition of lexical translation relations from MRDs. Machine Translation, 3(3–4):183–219.

Damle, D., Uren, V. (2005). Extracting significant words from corpora for ontology extraction. In Proceedings of the 3rd International Conference on Knowledge capture.

Dandapat S., Forcada, M. L., Groves D., Penkale, S., Tinsley J., Way, A. (2010). OpenMaTrEx: A Free/Open-Source Marker-Driven Example-Based Machine Translation System. In Proceedings of IceTAL - 7th International Conference on Natural Language Processing, Reykjavík.

Dang, H.T., Palmer, M. (2002). Combining Contextual Features for Word Sense Disambiguation. Proceedings SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation. Philadelphia, ACL.

Danielsson, P., Ridings, D. (1997). Practical presentation of a vanilla aligner. Technical report, Sprakbanken, Institutionen for svenska spraket, Göteborgs Universitet.

DeNero, J., Klein D. (2007). Tailoring Word Alignments to Syntactic Machine Translation. In proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

Drouin, P. (2006). Termhood, Quantifying the Relevance of a Candidate Term. In Proceedings of the 15th European Symposium on Languages for Special Purposes.

Du, J., Pecina, P., Way, A. (2010). An Augmented Three-Pass System Combination Framework: DCU Combination System for WMT 2010. In Proceedings of the Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden.

Eisele, A., Chen, Yu (2010). MultiUN: A Multilingual Corpus from United Nation Documents. In Proceedings of the seventh international conference on Language Resources and Evaluation.

Eisner, J. (2003). Learning Non-Isomorphic Tree Mappings for Machine Translation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics. Companion Volume, pp. 205–208. Sapporo, Japan.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M., Soria, C. (2009). Multilingual resources for NLP in the lexical markup framework (LMF). In Language Resources and Evaluation 43, 57-70.

Frank, A., Hoffmann, Chr., Strobel, M. (2004). On Gender Issues in Machine Translation. University of Bremen.

Frantzi, K., Ananiadou, S., Tsuji, J. (1999). Classifying Technical terms. Proceedings of the ICCC/IFIP Conference.

Frantzi, K.T., Ananiadu, S. (2003). The C-Value/NC-Value Domain Independent method for Multi-Word Term Extraction. In Journal of Natural Language Processing, 6,3.

Fung, P. (1995). Compiling bilingual lexicon entries from a non-parallel english-chinese corpus. In Proceedings of the 14th Annual Meeting of Very Large Corpora, pp. 173–183.

Fung, P. (1998). A Statistical View of Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. Proceedings of the 3rd Conference of the Association for Machine Translation of the Americas.

Fung, P., Church, K.W. (1994). K-vec: a new approach for aligning parallel texts. In Proceedings of the 15th conference on Computational linguistics.

Gale, W.A., Church, K.W. (1991). Identifying word correspondences in parallel texts. In the Fourth DARPA Workshop on Speech and Natural Language, pages 152–157.

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

Giménez, J., Màrquez, L. (2007). Context-aware discriminative phrase selection for statistical machine translation. In Proceedings of the 45th Annual meeting of the Association for computational Linguistics (ACL-2007): 2nd Workshop on Statistical Machine Translation, Prague, Czech Republic, pp.159-166.

Giménez, J., Màrquez, L. (2008). Discriminative Phrase Selection for Statistical Machine Translation. In C. Goutte, N. Cancedda, M. Dymetman and G. Foster (eds.) Learning Machine Translation. NIPS Workshop Series. MIT Press.

Goller, Chr., Löning, J., Will, Th., Wolff, W. (2000). Automatic document classification: A thorough evaluation of various methods. Proceedings of the 7th International Symposium für Informationswissenschaft.

Gough N., Way, A. (2004). "Robust large-scale EBMT with marker-based segmentation," in Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation, Baltimore, Maryland, pp. 95–104.

Green, T. (1979). "The necessity of syntax markers. two experiments with artificial languages," Journal of Verbal Learning and Behavior, vol. 18, pp. 481–496.

Haghighi, A., Blitzer, J., DeNero, J., Klein, D. (2009). Better Word Alignments with Supervised ITG Models. In Proceedings of the Joint conference of the 47th annual meeting of the Association for Computational Linguistics and 4th International Joint conference on natural language processing of the AFNLP.

Hannemann, Gr., Ambati, V., Clark, J.H., Parlikar, A., Lavie, A. (2009). An Improved Statistical Transfer System for French-English Machine Translation. In Proceedings of the 4th Workshop on Statistical Machine Translation.

Hannemann, Gr., Huber, E., Agarwal, A., Ambati, V., Parlikar, A., Peterson, E., Lavie, A. (2008). Statistical Transfer Systems for French-English and German-English Machine Translation. In Proceedings of the 3rd Workshop on Statistical MT.

Haque, R., Naskar, S.K., Ma, Y., Way, A. (2009a). Using Supertags as Source Language Context in SMT. In Proceedings of the 13th EAMT Conference, Barcelona, Spain, pp. 234-241.

Haque, R., Naskar, S.K., Bosch, A. van den, Way, A. (2009b). Dependency Relations as Source Context in Phrase-Based SMT. In Proceedings of PACLIC 23: the 23rd Pacific Asia Conference on Language, Information and Computation, Hong Kong, pp. 170-179.

Harper, M.P., Wang, W., White, C.M. (2001). Approaches for Learning Constraint Dependency Grammar from Corpora. In Proceedings of the Workshop on Grammar and NLP.

Hasan S., Ganitkevitch, J., Ney, H., Andrés-Ferrer, J. (2008). Triplet lexicon models for statistical machine translation. Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-2008), Honolulu, Hawaii, USA, pp.372-381.

Helmreich, S., Guthrie, L., Wilks, Y. (1993). The use of machine readable dictionaries in the Pangloss project. In AAAI Spring Symposium on Building Lexicons for Machine Translation.

Hiemstra, D. (1998). Multilingual domain modeling in Twenty-One: automatic creation of a bi-directional translation lexicon from a parallel corpus, In: Peter-Arno Coppen, Hans van Halteren and Lisanne Teunissen (eds.) Proceedings of the eightth CLIN meeting, pages 41-58.

Hoste, V., Daelemans, W., Hendrickx, I., van den Bosch, A. (2002). Dutch Word Sense Disambiguation: Optimising the Localness of Context. In Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation. Philadelphia, ACL.

Ide, N., Erjavec, T., Tufis, D. (2002). Sense Discrimination with Parallel Corpora. In Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation, Philadelphia, ACL.

Imamura, K. (2000). Hierarchical Phrase Alignment Harmonized with Parsing. In Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium (NLPRS'01), pp. 377–384. Tokyo, Japan.

Innselset, K., Kristiansen, M., Øvsthus, K. (2008). Looking back to move forward, Challenges related to deceitful parallel texts and slippery terms. In Synaps 21, 2008: Festskrift for Magnar Brekke. Norges Handelshøyskole. p. 72-89.

insertions, and reversalsPDF". Soviet Physics Doklady 10: 707–10.

Jassem, K., Graliński, F., Krynicki, Gr. (2000). POLENG – Adjusting a Rule-based Polish-English Machine Translation System by Means of Corpus Analysis. In Proceedings of the 5th European Association for Machine Translation Conference.

Jellinghaus, M. (2007). Automatic Acquisition of Semantic Transfer Rules for Machine Translation. Dipl.A. Univ. Saarland.

Kaji, H., Kida Y., Morimoto Y. (1992) Learning Translation Templates from Bilingual Text. In Proceedings of the 15th International Conference on Computational Linguistics (CoLing'92), ed. Christian Boitet. vol. 2, pp. 672–678. Nantes, France.

Kaplan, R. M., Bresnan, J. (1982). Lexical-Functional Grammar: A Formal System for Grammatical Representation. In The Mental Representation of Grammatical Relations, ed. Joan Bresnan, pp. 173–281. Cambridge, MA: The MIT Press.

Karlgren, J., Sahlgren, M. (2005). Automatic bilingual lexicon acquisition using random indexing of parallel corpora. Journal of Natural Language Engineering, Special Issue on Parallel Texts, 11(3):327–341.

Kay, M., Röscheisen, M. 1988. Text-Translation Alignment, Technical report. Xerox Palo Alto Research Center (Martin Kay and Martin Röscheisen (1988). Text-Translation Alignment. Technical Report, Xerox Palo Alto Research Center. Published in Computational Linguistics, MIT Press, March 1993.

Kikui, G. (1999). Resolving Translation Ambiguity using Non-parallel Bilingual Corpora. In Proceedings of the ACL Workshop on Unsupervised Learning in NLP.

Kit, C. (2002). Corpus Tools for Retrieving and Deriving Termhood Evidence. In Proceedings of the 5th East Asia Forum of Terminology.

Klein. D., Toutanova, Kr., Tolga Ilhan, H., Kamvar, S.D., Manning, Chr. D. (2002). Combining Heterogenous Classifiers for Word-Sense Disambiguation. In Proceedings of the ACL-02 workshop on Word sense disambiguation.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. Proceedings of the 10th Machine Translation Summit.

Koehn, P. (2010). Statistical Machine Translation. Cambridge University Press.

Korhonen, A. (2002). Subcategorization acquisition. Ph.D. thesis, University of Cambridge.

Lavie, A. (2008). Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation. In Proceedings of CICLing-2008.

Lavie, A., Parlikar, A., Ambati, V. (2008). Syntax-driven Learning of Sub-sentential Translation Equivalents and Translation Rules from Parsed Parallel Corpora. In Proceedings of the Second ACL Workshop on Syntax and Structure in Statistical Translation (SSST-2),

Lavoie, B., White, M., Korelsky, T. (2002). Learning Domain-Specific Transfer Rules: An Experiment with Korean to English Translation. In Proceedings of the 19th International Conference on Computational Linguistics.

Lee, H.A., Kim, G.C. (2002). Translation Selection through Source Word Sense Disambiguation and Target Word Selection. In Proceedings of the 19th International Conference on Computational Linguistics.

Leusch, G., Ueffing, N., Ney, H. (2006). "CDER: Efficient MT evaluation using block movements," in Proceedings of the 11 th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, pp. 241–248.

Levenshtein, V.I. (1966). "Binary codes capable of correcting deletions,

Liang, P., Taskar, B., Klein, D. (2006). Alignment by Agreement. In Proceedings of the Conference on Human Language Technology and Annual Meeting of the North American Chapter of the Association of Computational Linguistics.

Lieske, Chr., McCormick, S., Thurmair, Gr. (2001). The Open Lexicon Interchange Format (OLIF), In Proceedings of the MT Summit VIII.

Liu, Y., Liu, Q., and Lin, S. (2005). Log-linear models for word alignment. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 459–466, Ann Arbor, MI.

Martínez, D., Agirre, E., Màrquez, L. (2002). Syntactic features for high precision Word Sense Disambiguation. In Proceedings of the 19th International Conference on Computational Linguistics.

Matsumoto, Y., Ishimoto, H., Utsuro T. (1993). Structural Matching of Parallel Texts. In 31st Annual Meeting of the Association for Computational Linguistics (ACL'93), pp. 23–30. Columbus, OH.

Mausam, Soderland, S., Etzioni, O., Weld, D.S., Skinner, M., Bilmes, J. A. (2009). Compiling a Massive, Multilingual Dictionary via Probabilistic Inference. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.

Ma, Y. (2009). Constrained Word Alignment Models for Statistical Machine Translation. PhD thesis. Dublin City University.

Melamed, D.I. (1996). A Geometric Approach to Mapping Bitext Correspondence Proceedings of the First Conference on Empirical Methods in Natural Language Processing (EMNLP), Philadelphia, PA.

Melamed, D.I. (1997). A word-to-word model of translational equivalence. Proceedings of the 35th annual meeting of the Association for Computational Linguistics.

Menezes, A., Richardson, S. D. (2003). A Best-first Alignment Algorithm for Automatic Extraction of Transfer Mappings from Bilingual Corpora. In Recent Advances in Example-Based Machine Translation , eds. Michael Carl and Andy Way, chap. 15, pp. 421–442. Vol. 21 of Text, Speech and Language Technology. Dordrecht.

Menezes, A., Richardson, St. (2001). A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. In Proceedings of the Workshop on data-driven Machine Translation, ACL 2001, Toulouse.

Meyers, A., Yangarber, R., Grishman, R., Macleod, C., Moreno-Sandoval, A. (1998). Deriving Transfer Rules from Dominance-Preserving Alignments. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics (CoLing-ACL'98). vol. 2, pp. 843–847. Montreal, QC, Canada.

Mihάltz, M. (2005). Towards a hybrid approach to word sense disambiguation in Machine Translation. In Proceedings of RANLP.

Molina, A. and Pla, F. (2002). Shallow parsing using specialized hmm. Journal of Machine Learning Research.

Moore, R.C. (2002). Fast and Accurate Sentence Alignment of Bilingual Corpora, Springer-Verlag.

Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 81–88, Vancouver, BC.

Munteanu, D., Marcu D. (2005). Improving Machine Translation Performance by Exploiting Comparable Corpora. Computational Linguistics, 31 (4), pp. 477-504.

Nasr, A., Rambow, O., Palmer, M., Rosenzweig, J. (1997). Enriching Lexical Transfer With Cross-Linguistic Semantic Features or How to Do Interlingua without Interlingua. In Proceedings of the MT Summit.

Neff, M., McCord, M. (1990). Acquiring lexical data from machine-readable dictionary resources for machine translation. In 3rd Intl Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language.

Och, F.J., Ney, H. (2003). "A Systematic Comparison of Various Statistical Alignment Models", Computational Linguistics, volume 29, number 1, pp. 19-51.

Otero, P.G. (2007). Learning Bilingual Lexicons from Comparable English and Spanish Corpora. In Proceedings of the MT Translation Summit.

Otero, P.G. (2008). Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. In Proceedings of the LREC Workshop on Comparable Corpora.

Penkale, S., Haque, R., Dandapat, S., Banerjee, P., Srivastava, A. K., Du, J., Pecina, P., Naskar, S. K., Forcada, M.L. , Way, A. (2010). MATREX: The DCU MT System for WMT 2010. In Proceedings of the Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden.

Preiss, J., Briscoe, T., Korhonen, A. (2007). A System for Large-Scale Acquisition of Verbal, Nominal and Adjectival Subcategorization Frames from Corpora. In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics.

Probst, K. (2005). Learning Transfer Rules for Machine Translation with Limited Data. PhD thesis, Carnegie Mellon University.

Probst, K., Carbonell, J., Levin, L. (2002). Semi-automatic learning of transfer rules for machine translation of low-density languages. In Proceedings of ESSLI.

Ramshaw, L.A., Marcus, M.P. (1995). Text chunking using transformation-based learning. In Workshop on Very Large Corpora, pages 82–94.

Rapp, R. (1995). Identifying word translations in non-parallel texts. Proceedings of the 33rd Meeting of the Association for Computational Linguistics. Cambridge, MA, 1995, 320-322.

Rayson, P., Garside, R. (2000). Comparing Corpora using frequency profiling. In Proceedings of the Workshop on Comparing Corpora, ACL.

Resnik, P., Yarowsky, D. (1997). A perspective on Word Sense Disambiguation Methods and their Evaluation. In Proceedings of the ACL SIGLEX workshop on tagging text with lexical semantics: Why, what,and how.

Rosell, M.G. (2010). Documentation of the Open-Source Shallow-Transfer Machine Translation Platform Apertium. Report University of Alicante.

Sánchez-Martínez, F. (2008). Using unsupervised corpus-based methods to build rule-based machine translation systems. PhD thesis, Departament de Llenguatges i Sistemes Infomàtics, Universitat d'Alacant, Spain.

Sánchez-Martínez, F., Forcada, M.L. (2007). Automatic induction of shallow-transfer rules for open-source machine translation. In Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation.

Sánchez-Martínez, F., Forcada, M.L. (2009a). Inferring Shallow-Transfer Machine Translation Rules from Small Parallel Corpora. In Journal of Artificial Intelligence Research 34, 605-635.

Sánchez-Martínez, F., Forcada, M.L., Way, A. (2009). Hybrid Rule-Based – Example-Based MT: Feeding Apertium with Sub-sentential Translation Units. In Proceedings of the 3rd International Workshop on Example-Based Machine Translation.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34,1, p.1–47.

Sikuan W., Yanquan, Z. (2009). "Chinese-English Chunk Alignment Based on Anchor Chunk," Intelligent Information Technology Application Workshops, International Symposium on, pp. 398-401, 2009 Third International Symposium on Intelligent Information Technology Application Workshops.

Skut, W., Brants, T. (1998). Chunk tagger: statistical recognition of noun phrases. In ESSLLI-1998 Workshop on Automated Acquisition of Syntax and Parsing.

Smadja, F., McKeown, K.R., Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach., Computational Linguistics, 22(1):1–38.

Smith, J., Quirk, C., Toutanova, K. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In proceedings of Human Language Technologies:

The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Los Angeles, California.

Snover M., Dorr, B., Scwartz, R., Makhoul, J., Micciula, L. 2006. A study of translation editrate with targeted human annotation. In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA 2006), Cambridge, MA, pp. 223-231.

Specia, L., Das Graças Volpe Nunez, M., Castello Branco, R.G., Stevenson, M. (2006). Multilingual versus Monolingual WSD. In Proceedings of the Workshop Making Sense of Sense.

Steinberger, R., Pouliquen,B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th International Conference on Language Resources and Evaluation.

Stroppa N., Bosch, A. van den, Way, A. 2007. Exploiting Source Similarity for SMT using Context-Informed Features. In Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-2007), Skövde, Sweden, pp.231-240

Stroppa, N., Groves, D., Way, A., Sarasola, K. (2006). Example-based machine translation of the Basque language. in Proceedings of the 7th Conference of the Association for Machine Translation of the Americas. Cambridge, Massachusetts, pp. 232–241.

Stroppa, N., Way, A. (2006). MaTrEx: DCU machine translation system for IWSLT 2006. In Proceedings of the International Workshop on Spoken Language Translation, pages 31-36.

Taskar, B., Simon, L.-J., Dan, K. (2005). A discriminative matching approach to word alignment. In Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 73–80, Vancouver, BC, Canada.

TEI: Text Encoding Initiative, www.tei-c.org

Thurmair, Gr (1990). Complex Lexical Transfer in METAL. Proceedings of the third International Conference On Theoretical and. Methodological Issues in Machine Translation.

Thurmair, Gr. (2003). Making Term Extraction Tools Usable. In Proceedings of the Joint Conference combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop.

Thurmair, Gr. (2006). Using Corpus Information to Improve MT Quality. In Proceedings of the Workshop LR4Trans-III, LREC.

Thurmair, Gr. (2010). Proposal for corpus representation in PANACEA. PANACEA internal report.

Tiedemann, J. (2003). Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing, Doctoral Thesis, Studia Linguistica Upsaliensia 1.

Tiedemann, J. (2009). News from OPUS – A collection of Multilingual Parallel Corpora with Tools and Interfaces. In N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pages 237-248, John Benjamins, Amsterdam/Philadelphia.

TMX: TMX 1.4b Specification, OSCAR Recommendation (2005). http://www.lisa.org/fileadmin/standards/tmx1.4/tmx.htm

Tsang, V., Stevenson, S., Merlo, P. (2002). Crosslinguistic Transfer in Automatic Verb Classification. In roceedings of the 19th International Conference on Computational Linguistics.

Tufiş, D., Ion, R., Ide, N. (2004). Fine-graned word-sense disambiguation Based on Parallel Cvorpora, Word Alignment, Word Clustering and Aligned WordNets. In Proceedings of the 20th International Conference on Computational Linguistics.

Varga, D., Németh, L., Halácsy, P., Kornai, A., Trón, V., Nagy, V. (2005). Parallel corpora for medium density languages. In Proceedings of RANLP 2005, pages 590-596.

Varma, N. (2002). Identifying Word Translations in Parallel Corpora Using Measures of Association. MsC thesis, University of Minnesota.

Vickrey, D., Biewald, L., Teyssier, M., Koller, D. (2005). Word-Sense Disambiguation for Machine Translation. In Proceedings of the Joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.

Vu, T., Aw, A. T., Zhang, M. (2007). Term Extraction Through Unithood and Termhood Unification. In Proceedings of the 3rd International Joint Conference on Natural Langauge Processing.

Winiwarter, W. (2004a). Automatic Acquisition of Transfer Rules from Translation Examples. In Proceedings of the EsTAL Conference.

Winiwarter, W. (2004b). Incremental Learning of Transfer Rules for Customized Machine Translation. In Proceedings of the 15th Conference on Applications of Declarative Programming and Knowledge Management.

Wong, W., Liu, W., Bennamoun, M. (2007). Determining termhood for learning domain ontologies using domain prevalence and tendency. In Proceedings of the 6th Australian Conference on Data mining and analytics.

Wu, D. (2000). Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars. In Parallel Text Processing: Alignment and Use of Translation Corpora, ed. Jean Veronis, chap. 7, pp. 139–167. Dordrecht: Kluwer. ___

Wu, D., Xia, X. (1995), Large-scale automatic extraction of an english-chinese translation lexicon., Machine Translation, 9:285–313.

Yarowsky, D. (1995). Unsupervised Word sense disambiguation rivaling supervised methods. In Proceedings of the 33rd Meeting of the Association for Computational Linguistics.

Yarowsky, D. (2000). Hierarchical decision lists for word sense disambiguation. In Computers and the Humanities 34,1-2.

Yuan D., Gildea, D., Palmer, M. (2003). An Algorithm for Word-Level Alignment of Parallel Dependency Trees. In Proceedings of the MT Summit IX, pp. 95–101. New Orleans, LA.

Yu, K., Tsujii, J. (2009). Bilingual Dictionary Extraction from Wikipedia. Proceedings of Machine Translation Summit XII.

Zhang, T., Damereau, F., Johnson, D. (2002). Text chunking base on a generalization of winnow. Journal of Machine Learning Research.

Zhechev, V. (2009). Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System. School of Computing, Dublin City University: Ph.D. Thesis. Dublin, Ireland.

Zhechev, V. (2009). Unsupervised Generation of Parallel Treebanks through Sub-Tree Alignment. The Prague Bulletin of Mathematical Linguistics, 91: 89–98.

Zhechev, V., Way, A. (2008). Automatic Generation of Parallel Treebanks. In Proceedings of the 22nd International Conference on Computational Linguistics (CoLing '08), pp. 1105–1112. Manchester, UK.

A. Tool Documentation Forms

A. 1 Hunalign

Administrative information	
* Tool name	hunalign
Short name	
* Short description	aligns bilingual text on the sentence level
* Organization name	
* Latest versionand release date	1.0 – september 2009
Tool web page	http://mokk.bme.hu/resources/hunalign
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, V. Nagy (2005). Parallel corpora for medium density languages. In Proceedings of the RANLP 2005, pages 590-596.
Relevant project(s)	Hunglish
* License and availability	LGPL 2.1
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	ISO-8859-1 and UTF-8
* Character encoding (output)	ISO-8859-1 and UTF-8 (depending on input)
* Format (input)	- Bilingual corpus (two plain text files with one sentence per line) - Bilingual dictionary (optional, newline-separated dictionary items. An item consists of a target language phrase and a source language phrase) - Reference alignment (optional, for evaluation purposes)
* Format (output)	- Numeric ladder format - Text format
* Compatibility of the input and/or output data with national/international standards/common practices	Unknown
* Language resources required for the operation of the application	None (optionally a bilingual dictionary)

* Operating system	OS independent
* Implementation language	C++ (and awk, bash, python scripts)
* Other software requirements	None
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	400 sentences/sec (500 sentence corpus on an Intel core 2 duo E8400)
Evaluation information	
* Methodology and reference data	Data: Manual alignment of the Hungarian version of Orwell's 1984
* Results	Without bilingual dictionary: 97.93%P, 97.80% R With bilingual dictionary: 99.34% P, 99.34% R

Sample data

Input

English

```
6      he be a sad day a day of grey unrest of discontent
7      he gently moving air seem to be celebrate the loss of some gay thing
with a soft tender elegy
```

Hungarian

```
15     szomorú szürke nap volt ez a nyugtalanság elégedetlenség
16     a szellõ finom fuvallat lágy s gyengéd elégia mintha       valami vidám
dolog búcsúztat
```

Output

Hungarian-English. Numeric ladder:

```
15      6       1.55833
16      7       1.34143
```

Hungarian-English. Text:

```
szomorú szürke nap volt ez a nyugtalanság elégedetlenség   he be a sad day
a day of grey unrest of discontent     1.55833

a szellõ finom fuvallat lágy s gyengéd elégia mintha valami vidám dolog
búcsúztat    the gently moving air seem to be celebrate the loss of some gay
thing with a soft tender elegy   1.34143
```

A.2 Geometric Mapping and Alignment

Administrative information	
* Tool name	Geometric Mapping and Alignment
Short name	GMA
* Short description	The GMA software package implements the Smooth Injective Map Recognizer (SIMR) algorithm for mapping bitext correspondence and the Geometric Segment Alignment (GSA) post-processor for converting general bitext maps to monotonic segment alignments.
* Organization name	
* Latest versionand release date	2.1 – September 2004
Tool web page	http://nlp.cs.nyu.edu/GMA/
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	Ali Argyle, Luke Shen, Svetlana Stenchikova, and I. Dan Melamed (2004.) Geometric Mapping and Alignment (GMA) tool.
Relevant project(s)	
* License and availability	GPL
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	ISO-8859-1 and UTF-8
* Character encoding (output)	ISO-8859-1 and UTF-8 (depending on input)
* Format (input)	- bilingual corpus/ bitext (two plain text files with one sentence per line)
* Format (output)	GMA outputs aligned blocks, one per line. The two sides of each aligned block are separated by the five ASCII characters " <=> "
* Compatibility of the input and/or output data with national/international standards/common practices	Unknown
* Language resources required for the operation of the application	None
* Operating system	OS independent

* Implementation language	Java
* Other software requirements	None
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	
Evaluation information	
* Methodology and reference data	
* Results	

Sample data

Output

```
1 <=> 1
2 <=> 2,3
3,4,5 <=> omitted
6 <=> 4
omitted <=> 5
```

Segment A1 is aligned with segment B1; segment A2 is aligned with segments B2 and B3; segments A3, A4, and A5 are not aligned with anything; segment A6 is aligned with segment B4; and segment B5 is not aligned with anything.

A.3 Bilingual Sentence Aligner

Administrative information	
* Tool name	BSA
Short name	Bilingual Sentence Aligner
* Short description	An algorithm for finding which sentences do translate one-for-one in a parallel bilingual corpus.
* Organization name	
* Latest versionand release date	1.0 – May 2003
Tool web page	http://research.microsoft.com/
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	
Relevant project(s)	
* License and availability	MSR-SSLA
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	UTF-8
* Character encoding (output)	UTF-8
* Format (input)	- bilingual corpus/bitex (one sentence per line)
* Format (output)	- bitext format
* Compatibility of the input and/or output data with national/international standards/common practices	Unknown
* Language resources required for the operation of the application	none
* Operating system	OS independent
* Implementation language	Perl
* Other software requirements	None
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	

Evaluation information	
* Methodology and reference data	
* Results	

A.4 Giza++

Administrative information	
* Tool name	GIZA++
Short name	
* Short description	a statistical machine translation toolkit that is used to train IBM Models 1-5 and an HMM model.
* Organization name	CLSP/JHU and RWTH Aachen
* Latest versionand release date	1.0.3 – march 2009
Tool web page	http://code.google.com/p/giza-pp/
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	Franz Josef Och, Hermann Ney. "A Systematic Comparison of Various Statistical Alignment Models", *Computational Linguistics*, volume 29, number 1, pp. 19-51 March 2003.
Relevant project(s)	
* License and availability	GPL
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	UTF-8
* Character encoding (output)	UTF-8
* Format (input)	GIZA++ input format: vocabulary files, bitext files and dictionary (optional) or plain text (tokenised and lowercased one sentence per line, e.g. by Moses scripts), which is converted to GIZA++ input format by the utility plain2snt.out
* Format (output)	GIZA++ output format: alignment file
* Compatibility of the input and/or output data with national/international standards/common practices	Both the input and the output are compatible with Moses
* Language resources required for the operation of the application	Parallel corpus tokenised and lowercased Bilingual dictionary (optional)
* Operating system	OS independent
* Implementation language	C++
* Other software requirements	mkcls

URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	100 sentences/sec (44k sentence corpus on an Intel core 2 duo E8400)
Evaluation information	
* Methodology and reference data	2,500 sentences of Czech-English parallel corpus manually aligned on the word level. David Mareček. Improving Word Alignment Using Alignment of Deep Structures. TSD 2009.
* Results	Intersection: 95.8 P, 79.0 R, 13.2 AER Grow-diag-final: 71.5 P, 92.0 R, 20.3 AER Union: 68.5 P, 93.2 R, 22.1 AER

Sample data

Input

English

```
nothing could be further from the truth .
as a result , pakistan was rewarded with american financial assistance and
arms .
```

French

```
on ne saurait être plus loin de la vérité .
le pakistan a donc été récompensé par l' assistance et les armes des états-
unis .
```

Output

French—English

```
on ne saurait être plus loin de la vérité .
NULL ({ }) nothing ({ 1 2 }) could ({ 3 }) be ({ 4 }) further ({ 5 6 }) from
({ 7 }) the ({ 8 }) truth ({ 9 }) . ({ 10 })
le pakistan a donc été récompensé par l' assistance et les armes des états-
unis .
NULL ({ 13 }) as ({ }) a ({ }) result ({ }) , ({ }) pakistan ({ 1 2 }) was ({
3 }) rewarded ({ 4 5 6 }) with ({ 7 }) american ({ }) financial ({ })
assistance ({ 8 9 }) and ({ 10 }) arms ({ 11 12 14 }) . ({ 15 })
```

English—French

```
nothing could be further from the truth .
```

```
NULL ({ }) on ({ }) ne ({ }) saurait ({ 1 2 }) être ({ 3 }) plus ({ }) loin ({
4 5 }) de ({ }) la ({ 6 }) vérité ({ 7 }) . ({ 8 })
as a result , pakistan was rewarded with american financial assistance and
arms .
NULL ({ 4 }) le ({ }) pakistan ({ 5 }) a ({ }) donc ({ }) été ({ 6 })
récompensé ({ 1 2 3 7 8 9 10 }) par ({ }) l' ({ }) assistance ({ 11 }) et ({
12 }) les ({ }) armes ({ 13 }) des ({ }) états-unis ({ }) . ({ 14 })
```

A.5 Berkeley Aligner

Administrative information	
* Tool name	berkeleyaligner
Short name	
* Short description	The BerkeleyAligner is a word alignment software package that implements recent innovations in unsupervised word alignment.
* Organization name	University of California, Berkeley
* Latest versionand release date	2.1 - 28.09.2009
Tool web page	http://code.google.com/p/berkeleyaligner/
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	Tailoring Word Alignments to Syntactic Machine Translation: John DeNero and Dan Klein. In proceedings of ACL, 2007
Relevant project(s)	
* License and availability	GPL v2
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	Unicode (UTF-8)
* Character encoding (output)	Unicode (UTF-8)
* Format (input)	Tokenised (optionally lowercased) parallel corpus, plain text one sentence per line Optionally parsed trees of the sentences (Berkeley Parser format)
* Format (output)	- GIZA++ output format: alignment file - Pairs of alignment correspondences (numbers) - Alignment tables
* Compatibility of the input and/or output data with national/international standards/common practices	Compatible with GIZA++, and therefore with Moses and Marclator
* Language resources required for the operation of the application	Optionally a parser (to train a syntactic HMM alignment model), e.g. Berkeley Parser
* Operating system	OS independent

* Implementation language	Java
* Other software requirements	none
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	55 sent/sec (2830 sentences corpus on an Intel core 2 duo E8400), 12.5 sentences/sec for the syntactic model
Evaluation information	
* Methodology and reference data	English-French Hansards data from the NAACL 2003 Shared Task
* Results	Classic: 93.9% P, 93% R, 6.5% AER Syntactic: 95.2% P, 91.5% R, 6.4% AER Baseline (GIZA++): 96% P, 86.1% R, 8.6% AER

Sample data

Input

French

```
monsieur le orateur , ma question se adresse à le ministre chargé de les
transports .
```

English

```
mr. speaker , my question is directed to the minister of transport .
```

Syntax for English (optional)

```
(S (NP (NNP mr.) (NNP speaker)) (, ,) (NP (PRP$ my) (NN question)) (VP (VBZ
is) (VP (VBN directed) (PP (TO to) (NP (NP (DT the) (NNP minister)) (PP (IN
of) (NN transport)))))) (. .))
```

Output

French—English **(**GIZA++ like)

```
# sentence pair (9) source length 13 target length 16 alignment score : 0
monsieur le orateur , ma question se adresse à le ministre chargé de les
transports .
NULL ({ 2 12 13 }) mr. ({ 1 }) speaker ({ 3 }) , ({ 4 }) my ({ 5 }) question
({ 6 }) is ({ 7 }) directed ({ 8 }) to ({ 9 }) the ({ 10 }) minister ({ 11 })
of ({ 14 }) transport ({ 15 }) . ({ 16 })
```

Pairs of alignment correspondences (numbers)

0-0 14-11 10-9 2-1 3-2 4-3 5-4 6-5 7-6 8-7 9-8 13-10 15-12

A.6 OpenMaTrEx

Administrative information	
* Tool name	OpenMaTrEx
Short name	
* Short description	OpenMaTrEx is a free/open-source (FOS) example-based machine translation (EBMT) system based on the marker hypothesis. It comprises a marker-driven chunker, a collection of chunk aligners, and two engines: one based on the simple proof-of-concept monotone recombinator (previously released as Marclator) and a Moses-based decoder
* Organization name	Dublin City University
* Latest versionand release date	0.7 - 28.04.2010
Tool web page	http://openmatrex.org/
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	- Stroppa, N., D. Groves, A. Way, and K. Sarasola. 2006. Example-based machine translation of the Basque language. In Proceedings of AMTA 2006, pages 232-241. - Stroppa, N. and A. Way. 2006. MaTrEx: DCU machine translation system for IWSLT 2006. In Proceedings of the International Workshop on Spoken Language Translation, pages 31-36.
Relevant project(s)	
* License and availability	GPL v3
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	Any supported by Java, e.g. Unicode (UTF-8)
* Character encoding (output)	Any supported by Java, e.g. Unicode (UTF-8)
* Format (input)	Tokenised and lowercased parallel corpus, plain text one sentence per line
* Format (output)	Tokenised and lowercased parallel corpus, SGML wrapping plain text sentences
* Compatibility of the input and/or output data with national/international standards/common practices	Compatible with GIZA++ and Moses

* Language resources required for the operation of the application	Marker files for the source and target languages. The current version includes marker files for Catalan, Czech, English, Spanish, French, Italian and Portuguese
* Operating system	GNU/Linux, MacOS
* Implementation language	Java (and some python scripts)
* Other software requirements	GIZA++ Moses scripts Moses decoder (optional, for MaTrEx mode)
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	Training: 35 sent/sec (90k sentences corpus on an Intel core 2 duo E8400) Decoding: 25 sent/sec (200 sentences on an Intel core 2 duo E8400)
Evaluation information	
* Methodology and reference data	
* Results	

Sample data

Input

French

```
toutefois la commission a , elle aussi , d' évidentes responsabilités en la
matière .

je ne puis donc que soutenir et recommander le précieux travail réalisé par le
rapporteur .
```

Output

English

```
<seg id="5">
<sol num="1" prob="1.0">however the commission has of obvious responsibilities
in this area </sol>

<seg id="8">
<sol num="1" prob="1.0">i can therefore that support and recommend the
valuable work done by the rapporteur </sol>
```

A.7 Subtree Aligner

Administrative information	
* Tool name	Subtree Aligner
Short name	
* Short description	Automatically generates parallel treebanks from parallel corpora. It has two stable methods to do so: - tree-to-tree (requires PoS tagged, constituency annotated text) - string-to-string (works on plain or PoS annotated text) - string-to-tree and tree-to-string (combinations of the above)
* Organization name	NCLT, DCU
* Latest versionand release date	2.8.6 – March 2009
Tool web page	http://www.ventsislavzhechev.eu/Home/Software/Software.html
* Contact person in the context of PANACEA (i.e. person responsible for integration of the tool in the factory)	Pavel Pecina Antonio Toral
* Contact person's email	ppecina@computing.dcu.ie atoral@computing.dcu.ie
* Technical report or publication relevant to the application	Zhechev, Ventsislav. 2010. Automatic Generation of Parallel Treebanks. An Efficient Unsupervised System: Lambert Academic Publishing. ISBN 978-3-8383-2795-2
Relevant project(s)	ATTEMPT project
* License and availability	GPL
Descriptive information	
* Languages covered	Language independent
* Character encoding (input)	UTF-8
* Character encoding (output)	UTF-8
* Format (input)	- source-to-target and target-to-source word alignment probabilities (Moses format) - source-to-target phrase alignment probabilities (Moses format, optional) - input corpus (aligned phrase-based-parsed sentences in bracketed format for the tree-to-tree module; for the string-to-string, string-to-tree and tree-to-string modules, the string-based side can be instead plain text or PoS annotated)

* Format (output)	Described in module documentation. Plain text with data blocks separated by empty lines. Each data block contains the following data in the given order: a line with the source sentence in phrase-based bracketed format, a line with the target in phrase-based bracketed format, a line with space separated integer IDs, each pair of IDs representing an alignment between a source and a target node.
* Compatibility of the input and/or output data with national/international standards/common practices	Input compatible with Moses (alignments) Output compatible with DOT grammar extraction tool. Due to its plain text nature, easily convertible to other formats.
* Language resources required for the operation of the application	- word alignment probabilities (e.g. from GIZA++) - phrase alignment probabilities (e.g. from Moses, optional) - bilingual parallel corpus (optionally PoS tagged and/or parsed using a phrase-based parser)
* Operating system	Any system with POSIX Unix-compatible tools. Tested on Linux and MacOS X
* Implementation language	C++
* Other software requirements	Boost libraries (tokenizer and regex for string-based modules), word alignment (e.g. GIZA++)
URL providing access to the tool as a web service	
URL providing guidelines on accessing the tool as a web service	
Hardware requirements	
* Processing speed	Depends on sentence length, the amount of word- and phrase-alignment data and on whether phrase-alignment data is used at all. When using the parallel implementation of the system with only word-alignment data based on 60k EN–DE sentence pairs, the tool processes 50 sentence pairs per second on average on an Intel Core 2 Duo E8400 with 3GB RAM.
Evaluation information	
* Methodology and reference data	Intrinsic: compare treebank produced by tree-to-tree method to the handcrafted HomeCentre treebank (810 EN–FR sentence pairs)

	Extrinsic: DOT system trained using the manual treebank vs. the same system trained with an automatically derived treebank
* Results	Intrinsic: Presicion 61.79% Recall 78.49%

Sample data

Input

English—French

```
( ROOT ( S ( NP ( D the ) ( NPzero ( N ink ) ( N cartridges ) ) ) ( VPcop (
Vcop are ) ( ADV running ) ( A low ) ) ) ( PERIOD . ) )
( LISTITEM ( S[decl] ( NPdet ( D les ) ( NPpp ( N cartouches ) ( PP ( P de ) (
N encre ) ) ) ) ( VPaux[pass] ( AUXpass sont ) ( ADV presque ) ( V épuisées )
) ) ( PERIOD . ) )
```

Output

English—French tree-to-tree:

```
(ROOT-1 (S-2 (NP-3 (D-4 the)(NPzero-5 (N-6 ink)(N-7 cartridges)))(VPcop-8
(Vcop-9 are)(ADV-11 running)(A-12 low)))(PERIOD-13 .))
(LISTITEM-1 (S[decl]-2 (NPdet-3 (D-4 les)(NPpp-5 (N-6 cartouches)(PP-7
(P-8 de)(N-9 encre))))(VPaux[pass]-10 (AUXpass-11 sont)(ADV-13
presque)(V-14 épuisées)))(PERIOD-15 .))
1 1 2 2 3 3 4 4 5 5 6 9 7 6 8 10 9 11 11 13 12 14 13 15
```

English—French string-to-string:

```
(X-100000 (D-1 the)(X-27 (X-9 (N-2 ink)(N-3 cartridges))(X-22 (X-17 (Vcop-4
are)(X-12 (ADV-5 running)(A-6 low)))(PERIOD-7 .))))
(X-36 (X-22 (X-9 (D-1 les)(N-2 cartouches))(P-3 de)(N-4 encre))(X-26 (X-20
(AUXpass-5 sont)(X-14 (ADV-6 presque)(V-7 épuisées)))(PERIOD-8 .)))
2 4 3 9 4 5 5 6 6 7 7 8 9 22 12 14 17 20 22 26 27 36
```