Design of a hybrid high quality machine translation system

Kurt Eberle
Johanna Geiß
Mireia Ginestí-Rosell
Lingenio GmbH
Karlsruher Straße 10
69 126 Heidelberg, Germany

[k.eberle,j.geiss,m.ginesti-rosell]
@lingenio.de

Bogdan Babych
Anthony Hartley
Reinhard Rapp
Serge Sharoff
Martin Thomas
Centre for Translation Studies
University of Leeds
Leeds, LS2 9JT, UK

[B.Babych,A.Hartley,R.Rapp,
S.Sharoff,M.Thomas]@leeds.ac.uk

Abstract

This paper gives an overview of the ongoing FP7 project HyghTra (2010 – 2014). The HyghTra project is conducted in a partnership between academia and industry involving the University of Leeds and Lingenio GmbH (company). It adopts a hybrid and bootstrapping approach to the enhancement of MT quality by applying rule-based analysis and statistical evaluation techniques to both parallel and comparable corpora in order to extract linguistic information and enrich the lexical and syntactic resources of the underlying (rule-based) MT system that is used for analysing the corpora. The project places special emphasis on the extension of systems to new language pairs and corresponding rapid, automated creation of high quality resources. The techniques are fielded and evaluated within an existing commercial MT environment.

1 Motivation

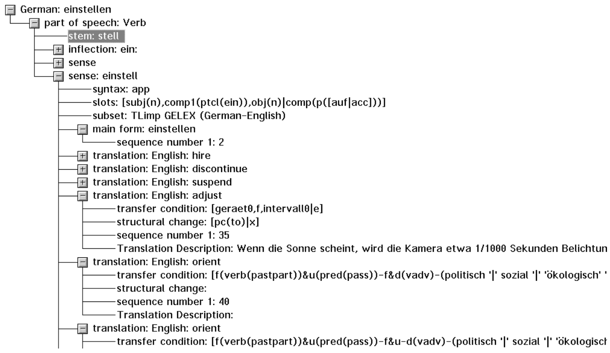
Statistical Machine Translation (SMT) has been around for about 20 years, and for roughly half of this time SMT and the 'traditional' *Rule-based Machine Translation* (RBMT) have been seen as competing paradigms. During the last decade however, there is a trend and growing interest in combining the two methodologies. In our approach these two approaches are viewed as complementary.

Advantages of SMT are low cost and robustness, but definite disadvantages of (pure) SMT are that it needs huge amounts of data, which for many language pairs are not available and are unlikely to become available in the future. Also, SMT tends to disregard important classificatory knowledge (such as morphosyntactic, categorical and lexical class features), which can be provided and used relatively easily within non-statistical representations.

On the other hand, advantages of RBMT are that its (grammar and lexical) rules and information are understandable by humans and can be exploited for a lot of applications outside of translation (dictionaries, text understanding, dialogue systems, etc.).

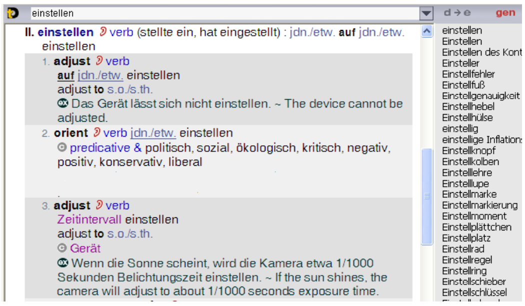
The slot grammar approach used in Lingenio systems (cf. McCord 1989, Eberle 2001) is a prime example of such linguistically rich representations that can be used for a number of different applications. Fig.1 shows this by a visualization of (an excerpt of) the entry for the ambiguous German verb *einstellen* in the database that underlies (a) the Lingenio translation products, where it links up with corresponding set of the transfer rules, and (b) Lingenio's dictionary product *TranslateDict*, which is primarily intended for human translators.

Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 101–112,
Avignon, France, April 23 - 27 2012. ©2012 Association for Computational Linguistics



```
German: einstellen
  part of speech: Verb
    stem: stell
    inflection: ein:
    sense
    sense: einstell
      syntax: app
      slots: [subj(n),comp1(ptcl(ein)),obj(n)|comp(p([auf|acc]))]
      subset: TLimp GELEX (German-English)
      main form: einstellen
        sequence number 1: 2
      translation: English: hire
      translation: English: discontinue
      translation: English: suspend
      translation: English: adjust
        transfer condition: [geraet0,f,intervall0|e]
        structural change: [pc(to)|x]
        sequence number 1: 35
        Translation Description: Wenn die Sonne scheint, wird die Kamera etwa 1/1000 Sekunden Belichtun
      translation: English: orient
        transfer condition: [f(verb(pastpart))&u(pred(pass))-f&d(vadv)-(politisch '|' sozial '|' 'ökologisch' '
        structural change:
        sequence number 1: 40
        Translation Description:
      translation: English: orient
        transfer condition: [f(verb(pastpart))&u(pred(pass))-f&u-d(vadv)-(politisch '|' sozial '|' 'ökologiscl
```

Fig 1 a) data base entry *einstellen*
('translation' represents links between SL and T entries)



II. einstellen verb (stellte ein, hat eingestellt) : jdn./etw. **auf** jdn./etw. einstellen

1. **adjust** verb
auf jdn./etw. einstellen
adjust **to** s.o./s.th.
ex Das Gerät lässt sich nicht einstellen. ~ The device cannot be adjusted.

2. **orient** verb jdn./etw. einstellen
predicative & politisch, sozial, ökologisch, kritisch, negativ, positiv, konservativ, liberal

3. **adjust** verb
Zeitintervall einstellen
adjust **to** s.o./s.th.
Gerät
ex Wenn die Sonne scheint, wird die Kamera etwa 1/1000 Sekunden Belichtungszeit einstellen. ~ If the sun shines, the camera will adjust to about 1/1000 seconds exposure time.

Fig 1 b) product entry *einstellen*

The obvious disadvantages of RBMT are high cost, weaknesses in dealing with incorrect input and in making correct choices with respect to ambiguous words, structures, and transfer equivalents.
SMT output is often surprisingly good with respect to short distance collocations, but often misses correct choices are missed in cases where selectional restrictions take effect on distant words. RBMT output is generally good if the parser assigns the correct analysis to a sentence and if the target words can be correctly chosen from the set of alternatives. However, in the presence of ambiguous words and structures, and where linguistic information is lacking, the decisions may be wrong.
Given the complementarity of SMT and RBMT and their very different strengths and weaknesses, we take a view that an optimized MT architecture must comprise elements of both paradigms. The key issue therefore lies in the identification of such elements and how to connect them to each other. We propose a specific type of hybrid translation – *hy*brid *high* quality *tra*nslation (HyghTra), where core RBMT systems are created and enhanced by a range of reliable statistical techniques.

2 Development Methodology

Many hybrid systems described in the literature have attempted to put some analytical abstraction on top of an SMT kernel.[1] In our view this is not the best option because, according to the underlying philosophy, SMT is linguistically ignorant at the beginning and learns all linguistic rules automatically from corpora. However, the extracted information is typically represented in huge data sets which are not readable by humans in a natural way. This means that this type of architecture does not easily provide interfaces for incorporating linguistic knowledge in a canonical and simple way.
Thus we approach the problem from the other end, , integrating information derived from corpora using statistical methods into RBMT systems. Provided the underlying RBMT systems are linguistically sound and sufficiently modular in structure, we believe this to have greater potential for generating high quality output.
We currently use and carry out the following work plan:

(I) Creation of MT systems
(with rule-based core MT information and statistical extension and training):
(a) We start out with declarative analysis and generation components of the considered languages, and with basic *bilingual dictionaries* connecting to one another the entries of relatively small vocabularies comprising the most frequent words of each language in a given translation pair (cf. Fig 1 a).
(b) Having completed this phase, we extend the dictionaries and train the analysis-, transfer- and generation-components of the rule-based core systems using monolingual and bilingual corpora.

[1] A prominent early example is Frederking and colleagues (Frederking & Nirenburg, 1994). For an overview of hybrid MT till the late nineties see Streiter et al. (1999). More recent approaches include Groves & Way (2006a, 2006b). Commercial implementations include *AppTek* (http://www.apptek.com) and *Language Weaver* (http://www.languageweaver.com). An ongoing MT important project investigating hybrid methods is EuroMatrixPlus (http://www.euromatrixplus.net/)

(II) Error detection and improvement cycle:
(a) We automatically discover the most frequent problematic *grammatical constructions* and multiword expressions for commercial RBMT and SMT systems using automatic construction-based evaluation as proposed in (Babych and Hartley, 2009) and develop a framework for fixing corresponding grammar rules and extending grammatical coverage of the systems in a semi-automatic way. This shortens development time for commercial MT and contributes to yielding significantly higher translation quality.

(III) Extension to other languages:
Structural similarity and translation by pivot languages is used to obtain extension to further languages:
High-quality translation between closely related languages (e.g., Russian and Ukrainian or Portuguese and Spanish) can be achieved with relatively simple resources (using linguistic similarity, but also homomorphism assumptions with respect to parallel text, if available), while greater efforts are put into ensuring better-quality translation between more distant languages (e.g. German and Russian). According to our prior research (Babych et al., 2007b) the pipeline between languages of different similarity results in improved translation quality for a larger number of language pairs (e.g., MT from Portuguese or Ukrainian into German is easier if there are high-quality analysis and transfer modules for Spanish and Russian into German (respectively). Of course, (III) draws heavily on the detailed analysis and MT systems that the industrial partner in HyghTra provides for a number of languages.

In the following sections we give more details of the work currently done with regard to (I) and with regard to parts of (II): the creation of a new MT system following the strategy sketched. We cannot go further into detail with (II) and (III) here, which will become a priority for future research.

3 Creation of a new system

Early pilot studies covering some aspects of the strategy described here (using information from pivot languages and similarity) showed promising results (Rapp, 1999; Rapp & Martín Vide, 2007; see also Koehn & Knight, 2002).

We expect that the proposed semi-automatic creation of a new MT system as sketched above will work best if one of the two languages involved is already 'known' by modules to which the system has access. Against the background of the pipeline approach mentioned above in (III), this means that we assume an analysis and translation system that continuously grows by 'learning' new languages where 'learning' is facilitated by information about the languages already 'known' and by exploiting similarity assumptions – and, of course, by being fed with information prepared and provided by the human 'companion' of the system.
From this perspective, we assume the following steps of extending the system (with work done by the 'companion' and work done by the system)

1. Acquire parallel and comparable corpora.
2. Define a core of the morphology of the new language and compile a basic dictionary for the most frequent words and translations. Morphological representations and features for new languages are derived both manually and automatically, as proposed in (Babych et al., 2012 (in preparation)).
3. Using established alignment technology (e.g. Giza++) and parallel corpora, generate a first extension of this dictionary.
4. Expand the dictionary of step 3 using comparable corpora as proposed in a study by Rapp (1999). This is applicable mainly to single word units.
5. Expand coverage of multiword-units using novel technology.
6. Cross-validate the new dictionary with respect to available ones by transitivity.
7. Integrate the new dictionary into the new MT system as developing from reusing components and adding new components as in 8.
8. Complete morphology and spell out declarative analysis and generation grammar for the new language.
9. Automatically evaluate the translations of the most frequent grammatical constructions and multiword expressions in a machine-translated corpus, prioritising support for these constructions with a type of risk-assessment framework proposed in Babych and Hartley (2008).
10. Extend support for high-priority constructions semi-automatically by mining correct

translations from parallel corpora.

11. Train and evaluate the new grammar and transfer of the new MT system using the new dictionary on the basis of available parallel corpora.

The following sections give an overview of the different steps.

Step 1: Acquire parallel and comparable corpora

As our parallel corpus, we use the Europarl. The size of the current version is up to 40 million words per language, and several of the languages we are currently considering are covered. Also, we make use of other parallel corpora such as the Canadian Hansards (Proceedings of the Canadian Parliament) for the English–French language pair. For non-EU Languages (mainly Russian), we intend to conduct a pilot study to establish the feasibility of retrieving parallel corpora from the web, a problem for which various approaches have been proposed (Resnik, 1999; Munteanu & Marcu, 2005; Wu & Fung, 2005).

In addition to the parallel corpora, we will need large monolingual corpora in the future (at least 200 million words) for each of the six languages. Here, we intend to use newspaper corpora supplemented with text collections downloadable from the web.

The corpora are stored in a database that allows for assigning analyses of different depth and nature to the sentences and for alignment between the sentences and their analyses. The architecture of this database and the corresponding analysis and evaluation frontend is described in (Eberle et al 2010, 2012). Section *Results* contains examples of such representations.

Step 2: Compile a basic dictionary for the most frequent words

A prerequisite of the suggested hybrid approach with rule-based kernel is to define morphological classifications for the new language(s). This is done exploiting similarities to the classifications as available for the existing languages. Currently, this has been carried out for Dutch (on the basis of German) and for Spanish (on the basis of French/other Romance languages). The most frequent words (the basic vocabulary of a language) are typically also the most ambiguous ones. Since the Lingenio systems are lexically driven transfer systems (cf. Eberle 2001), we define (a) structural conditions, which inform the choice of the possible target words (single words or multiword expressions) and (b)restructuring conditions, as necessary (cf. Fig 1 a: attributes *'transfer conditions'* and *'structural change'*). In order to ensure quality this must be done by human lexicographers and therefore costly for a large dictionary. However, we manually create only very small basic dictionaries and extend these (semi-automatically) step 3 and those which follow.

Some important morphosyntactic features of the language are derived from a monolingual corpus annotated with publicly available part-of-speech taggers and lemmatisers. However, these tools often do not explicitly represent linguistic features needed for the generation stage in RBMT. In (Babych et al., 2012) we propose a systematic approach to recovering such missing generation-oriented representations from grammar models and statistical combinatorial properties of annotated features.

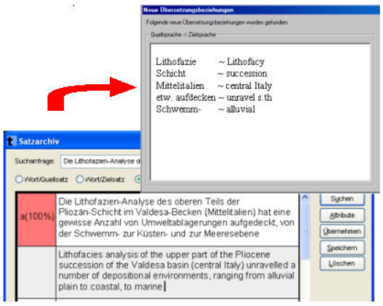
Step 3: Generating dictionary extensions from parallel corpora

Based on parallel corpora, dictionaries can be derived using established techniques of automatic sentence alignment and word alignment. For sentence alignment, the length-based Gale & Church aligner (1993) can be used, or – alternatively – Dan Melamed’s GSA-algorithm (Geometric Sentence Alignment; Melamed, 1999). For segmentation of text we use corresponding Lingenio-tools (unpublished).[2]

For word alignment Giza++ (Och & Ney, 2003) is the standard tool. Given a word alignment, the extraction of a (SMT) dictionary is relatively straightforward. With the exception of sentence segmentation, these algorithms are largely language independent and can be used for all of the languages that we consider. We did this for a number of language pairs on the basis of the

[2] If these cannot be applied because of lack of information about a language, we intend to use the algorithm by Kiss & Strunk (2006). An open-source implementation of parts of the Kiss & Strunk algorithm is available from Patrick Tschorn at http://www.denkselbst.de/sentrick/index.html.

Europarl-texts considered (as stored in our database). In order to optimize the results we use the dictionaries of step 1 as set of *cognates* (cf. Simard at al 1992, Gough & Way 2004), as well as other words easily obtainable from the internet that can be used for this purpose (like company names and other named entities with cross-language identity and terminology translations). Using the morphology component of the new language and the categorial information from the transfer relation, we compute the basic forms of the inflected words found. Later, we intend to further improve the accuracy of word alignment by exploiting chunk type syntactic information of the narrow context of the words (cf. Eberle & Rapp 2008). An early stage variant of this is already used in Lingenio products. The corresponding function AutoLearn<word> extracts new word relations on the basis of existing dictionaries and (partial) syntactic analyses. (Fig 2 gives an example).



Folgende neue Übersetzungsbeziehungen wurden gefunden:

Quellsprache -> Zielsprache

Lithofazie	~ Lithofacy
Schicht	~ succession
Mittelitalien	~ central Italy
etw. aufdecken	~ unravel s.th
Schwemm-	~ alluvial

Suchanfrage: Die Lithofazien-Analyse d

a(100%) Die Lithofazien-Analyse des oberen Teils der Pliozän-Schicht im Valdesa-Becken (Mittelitalien) hat eine gewisse Anzahl von Umweltablagerungen aufgedeckt, von der Schwemm- zur Küsten- und zur Meeresebene

Lithofacies analysis of the upper part of the Pliocene succession of the Valdesa basin (central Italy) unravelled a number of depositional environments, ranging from alluvial plain to coastal, to marine

Fig 2 AutoLearn<word>: new entries using transfer links and syntactic analysis

Given the relatively small size of the available parallel corpora, we expect that the automatically generated dictionaries will comprise about 20,000 entries each (This corresponds to first results on the basis of German↔English). This is far too small for a serious general purpose MT system. Note that, in comparison, the English↔German dictionary used in the current Lingenio MT product comprises more than 480,000 keywords and phrases.

Step 4: Expanding dictionaries using comparable corpora (word equations)

In order to expand the dictionaries using a set of monolingual comparable corpora, the basic approach pioneered by Fung & McKeown (1997) and Rapp (1995, 1999) is to be further developed and refined in the second phase of the project as to obtain a practical tool that can be used in an industrial context.

The basic assumption underlying the approach is that across languages there is a correlation between the co-occurrences of words that are translations of each other. If – for example – in a text of one language two words *A* and *B* co-occur more often than expected by chance, then in a text of another language those words that are translations of *A* and *B* should also co-occur more frequently than expected. It is further assumed that a small dictionary (as generated in step 2) is available at the beginning, and that the aim is to expand this basic lexicon. Using a corpus of the target language, first a co-occurrence matrix is computed whose rows are all word types occurring in the corpus and whose columns are all target words appearing in the basic lexicon. Next a word of the source language is considered whose translation is to be determined. Using the source-language corpus, a co-occurrence vector for this word is computed. Then all known words in this vector are translated to the target language. As the basic lexicon is small, only some of the translations are known. All unknown words are discarded from the vector and the vector positions are sorted in order to match the vectors of the target-language matrix. Using standard measures for vector similarity, the resulting vector is compared to all vectors in the co-occurrence matrix of the target language. The vector with the highest similarity is considered to be the translation of our source-language word.

From a previous pilot study (Rapp, 1999) it can be expected that this methodology achieves an accuracy in the order of 70%, which means that only a relatively modest amount of manual post-editing is required.

The automatically generated results are improved and the amount of post-editing is reduced by exploiting sense (disambiguation) information as available from the analysis component for the 'known' language of the new language pair.. Also we try to exploit categorial and underspecified syntactic information of the contexts of the words similar to what has been suggested for improving word alignment in the previous step (see also Fig.2). Also, as the frequent words are already covered by the basic lexicon (whose production from parallel corpora on the basis of a manually compiled kernel does not show an ambiguity problem of similar significance), and as experience shows that most low frequency words in a full-size lexicon tend to be unambiguous, the ambiguity problem is reduced further for the words investigated and extracted by this comparison method.

Step 5: Expanding dictionaries using comparable corpora (multiword units)

In order to account for technical terms, idioms, collocations, and typical short phrases, an important feature of an MT lexicon is a high coverage of multiword units. Very recent work conducted at the University of Leeds (Sharoff et al., 2006) shows that dictionary entries for such multiword units can be derived from comparable corpora if a dictionary of single words is available. It could even be shown that this methodology can be superior to deriving multiword-units from parallel corpora (Babych et al., 2007). This is a major breakthrough as comparable corpora are far easier to acquire than parallel corpora. It even opens up the possibility of building domain-specific dictionaries by using texts from different domains.

The outline of the algorithm is as follows:

- Extract collocations from a corpus of the source language (Smadja, 1993)
- To translate a collocation, look up all its words using any dictionary
- Generate all possible permutations (sequences) of the word translations
- Count the occurrence frequencies of these sequences in a corpus of the target language and test for significance
- Consider the most significant sequence to be the translation of the source language collocation

Of course, in later steps of the project, we will experiment on filtering these sequences by exploiting structural knowledge similarly to what was described in the two previous steps. This can be obtained on the basis of the declarative analysis component of the new language which is developed in parallel.

Step 6: Cross-validate dictionaries

The combination of the corpus-based methods for automatic dictionary generation as described in steps 3 to 5 will lead to high coverage dictionaries as the availability of very large monolingual corpora is no major problem for our languages. However, as all steps are error prone, it can be expected that a considerable number of dictionary entries (e.g. 50%) are not correct. To facilitate (but not eliminate) the manual verification of the dictionary, we will perform an automatic cross-check which utilizes the dictionaries' property of *transitivity*. What we mean by this is that if we have two dictionaries, one translating from language A to language B, the other from language B to language C, then we can also translate from language A to C by use of the intermediate language (or interlingua) B. That is, the property of transitivity, although having some limitations due to ambiguity problems, can be exploited to automatically generate a raw dictionary for A to C. Lingenio has some experience with this method having exploited it for extending and improving its English ↔ French dictionaries using French ↔ German and German ↔ English.

As the corpus-based approach (steps 3 to 5) allows us to also generate this type of dictionary via comparable corpora, we have two different ways to generate a dictionary for a particular language pair. This means that we can validate one with the other. Furthermore, with increasing number of language pairs created, there are more and more languages that can serve as interlingua or 'pivot': This, step by step, gives an increasing potential for mutual cross-validation.

Specific attention will be paid to automating as far as possible the creation of selectional restrictions to be assigned to the transfer relations of the new dictionaries in all steps of dictionary creation (2–6). We will try to do this on the basis of the analysis components as available for the languages considered: These are: a completely worked out analysis component for the 'old' language, a declarative (chunk parsing) component for the new one (compare the two following steps for this).

Step 7: Integrate dictionaries in existing machine translation systems

Lingenio has a relatively rich infrastructure for automatic importation of various kinds of lexical information into the database used by the analyses and translation systems. If necessary the information on hand (for instance from conventional dictionaries of publishing houses) is completed and normalized during or before importation. This may be executed completely automatically – by using the existing analyses components and resources respectively as databases – or interactively – by asking the lexicographer for additional information, if needed.

For example, there may be a list of multiword expressions to be imported into the database. In order to have available correct syntactic and semantic information for these expressions, they are analysed by the parser of the corresponding language. From the analysis found, the information necessary to describe the new lemma in the lexicon with respect to semantic type and syntactic structure is obtained. The same information is used to automatically create correct restructuring constraints for translation relations which use the new lemma as target. If the parser does not find a sound syntactic description, for example because some basic information or the expression is missing in the lexical database, the lexicographer is asked for the missing information or is handed over the expression to code it manually.

Using these tools importation of new lexical information, as provided in the previous steps, is considerably accelerated.

Step 8: Compile rule bases for new language pairs

Although experience clearly shows that construction and maintenance of the dictionaries is by far the most expensive task in (rule-based) Machine Translation, the grammars (analysis and generation) must of course be developed and maintained also. Lingenio has longstanding experience with the development of grammars, dictionaries and all other components of RBMT.

The used grammar formalism (*slot grammar*, cf. McCord 1991) is unification based and its structuring focuses on dependency, where phrases are analysed into heads and grammatical roles – so called (complement and adjunct) *slots*.

The grammar formalism and basic rule types are designed in a very general way in order to allow good portability from one language to another such that spelling out the declarative part of a grammar does not take very much time (2-4 person months approx. for relatively similar languages like Romance languages according to our experience). The portation of linguistic rules to new languages is also facilitated by the modular design with clearly defined interfaces that make it relatively straightforward to integrate information from corpora.

Given a parallel corpus as acquired in step 1, the following procedure defines grammar development:

1. Define a declarative grammar for the new language and train this grammar on the parallel -corpus according to the following steps:
2. Use a chunk parser for the grammar on the basis of an efficient part-of-speech tagger for the new language.
3. Combine the chunk analyses of the sentence, according to suggestions for packed syntactic structures (cf. Schiehlen 2001 and others) and underspecified representation structures respectively (cf. Eberle, 2004, and others), such that the result represents a disjunction of the possible analyses of the sentence.
4. Filter the alternatives of the representation by using mapping constraints between source and target sentence as can be computed from the lexical transfer relations and the structural analysis of the sentence. For instance, if we know, as in the example of the last section, that in the source sentence there is a relative clause with lexical elements A, B, . . . modifying a head H and that there are translations TH, TA, TB, . . . of H, A, B,. . . , in the target sentence which, among other possibilities, can be supposed to stand in a similar structural relation there, then we prefer this relation to the competing structural possibilities. (Fig. 3 in section *results* shows the corresponding selection for a German-Spanish example in the project database).
5. For each of the remaining structural possibilities of the thus revised underspecified representation, take its lexical material and underspecified structuring as a context for its successful firing. For instance, if the possibility is left that O is the direct object of VP, where VP is an underspecified verbal phrase and O an underspecified nominal phrase (i.e. where details of the substructuring are not spelled out), take the sentence as a reference for direct object complementation and O and VP as contexts which accept this complementation.

6. Develop more abstract conditions from the conditions learned according to (5) and integrate the different cases.
7. Tune the results using standard methods of corpus-based linguistics. Among other things this means: Distinguish between training and test corpora, adjust weights according to the results of test runs, etc.

The basic idea of the proposed learning procedure is similar to that used with respect to learning lexical transfer relations: Do not define the statistical model for the 'ignorant' state, where the surface items of the bilingual corpora are considered. Instead, define it for appropriate maximally abstract analyses of the sentences (which, of course, must be available automatically), because, then, much smaller sets of data will do. Here, the important question is: What is the most abstract level of representation that can be reached automatically and which shows reliable results? We think that it is the level of underspecified syntactic description as used in the procedure above.

The result of training the grammar is a set of rules which assign weights and contexts to each filler rule of the declarative grammar and thus allow to estimate how likely it is that a particular rule is applied in a particular context in comparison with other rules (Fig. 4 and 5 in section *results* give an overview of the relevance of grammar rules and their triggering conditions w.r.t. German).

We mentioned that the task of translating texts into each other does not presuppose that each ambiguity in a source sentence is resolved. On the contrary, translation should be *ambiguity preserving* (cf. Kay, Gawron & Norvig 1994, compare the example above). It is obvious that underspecified syntactic representations as suggested here are also especially suited for preserving ambiguities appropriately.

Step 9: Automatically evaluate translations of the most frequent grammatical constructions and multiword expressions in a machine-translated corpus

In a later work package of the project, we will run a large parallel corpus through available (competitive) MT engines, which will be enhanced by automatic dictionaries developed during the previous stages. On the source-language side of the corpus we will automatically generate lists of frequent multiword expressions (MWEs) and grammatical constructions using the methodology proposed in (Sharoff et al., 2006). For each of the identified MWEs and constructions we will generate a parallel concordance using open-source CSAR architecture developed by the Leeds team (Sharoff, 2006). The concordance will be generated by running queries to the sentence-aligned parallel corpora and will return lists of corresponding sentences from gold-standard human translations and corresponding sentences generated by MT. Each of these concordances will be automatically evaluated using standard MT evaluation metrics, such as BLEU. Under these settings parallel concordances will be used as standard MT evaluation corpora in an automated MT evaluation scenario.

Normally BLEU gives reliable results for MT corpora over 7000 words. However, in (Babych and Hartley, 2009; Babych and Hartley, 2008) we demonstrated that if the corpus is constructed in this controlled way, where evaluated fragments of sentences are selected as local contexts for specific multiword expressions or grammatical constructions, then BLEU scores have another "island of stability" for much smaller corpora, which now may consist of only five or more aligned concordance lines. This concordance-based evaluation scenario gives correct predictions of translation quality for the local context of each of the evaluated expressions.

The scores for the evaluated MWEs and constructions will be put in a risk-assessment framework, where we will balance the frequency of constructions and their translation quality. The top priority receive the most frequent expressions that are the most problematic ones for a particular MT engine, i.e., with queries with lowest BLEU scores for their concordances. This framework will allow MT developers to work down the priority list and correct or extend coverage for those constructions which will have the biggest impact on MT quality.

Step 10: Extend support for high-priority constructions semi-automatically by mining correct translations from parallel corpora

At this stage we will automate the procedure of correcting errors and extending coverage for problematic MWEs and grammatical constructions, identified in Step 9. For this we will exploit alignment between source-language sentences and gold-standard human translations. In the target human translations we will identify linguistically-motivated multiword expressions, e.g., using part-of-speech patterns or tf-idf distribution templates (Babych et al., 2007) and run standard alignment tools (e.g., GIZA++) for finding the most probable candidate MWEs that correspond to the problematic source-language expressions. Source and target MWEs paired in this way will form the basis for automatically-generated grammar rules. The rules will normally generalise several pairs of MWEs, and may be underspecified for certain lexical or morphological features. Later such rules will be manually checked and corrected by language specialists in MT development teams that work on specific translation directions.

This procedure will allow to speed up the grammar development procedure for large-scale MT projects and will focus on grammatical constructions with the highest impact on MT quality, establishing them as a top priority for MT developers. In HyghTra and with respect to the languages considered there, this procedure will be integrated into the grammar development and optimization of step 8, in particular it will be related to step 4 of the procedure sketched there. With regard to integration, we aim at an interleaved architecture in the long run.

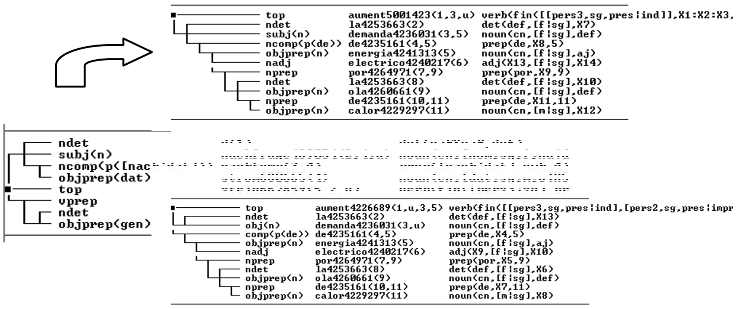
Step 11: Bootstrap the system

In Step 11, the new grammar and the transfer of the new MT system and the new dictionary may be mutually trained further using the steps before and applying the system to additional corpora.

4 Results

Declarative slot grammars for Dutch and Spanish have been developed using the patterns of German and French – where *declarative* means that there has been used no relevant semantic or other information in order to spell out weighting or filters for rule application -- the only constraint being morphosyntactic accessibility. The necessary morphological information has been adapted similarly from the corresponding model languages. The basic dictionaries have been compiled manually (Dutch) or extracted from a conventional electronic dictionary (*translateDict* Spanish).

For a subset of the Spanish corpus (reference sentences of the grammar, parts of the open source Leeds corpus (Sharoff, 2006), and Europarl), syntactic analyses have been computed and stored in the database. As the number of analyses grows extremely with the length of sentences, only relatively short sentences (up to 15 words) have been considered. These analyses are currently compared to the analyses of the German translations of the corresponding sentences (one translation per sentence), which are taken as a kind of 'gold' standard as the German analysis component (as part of the translation products) has proven to be sufficiently reliable. On the basis of the comparison a preference on the competitive analyses of the Spanish sentence is entailed and used for defining a statistical evaluation component for the Spanish grammar. Fig.3 shows the corresponding representations in the database for the sentence *Aumenta la demana de energía eléctrica por la ola de calor*[3] and its translation *die Nachfrage nach Strom steigt wegen der Hitzewelle/the demand for electricity increases because of the heat-wave.*



```
top            aument5001423(1,3,u) verb(fin([[pers3,sg,pres|ind]],X1:X2:X3,
ndet           la4253663(2)         det(def,[f|sg],X7)
subj(n)        demanda4236031(3,5)  noun(cn,[f|sg],def)
ncomp(p(de))   de4235161(4,5)       prep(de,X8,5)
objprep(n)     energia4241313(5)    noun(cn,[f|sg],aj)
nadj           electrico4240217(6)  adj(X13,[f|sg],X14)
nprep          por4264971(7,9)      prep(por,X9,9)
ndet           la4253663(8)         det(def,[f|sg],X10)
objprep(n)     ola4260661(9)        noun(cn,[f|sg],def)
nprep          de4235161(10,11)     prep(de,X11,11)
objprep(n)     calor4229297(11)     noun(cn,[m|sg],X12)

ndet
subj(n)
ncomp(p([nac
objprep(dat)
top
vprep
ndet
objprep(gen)

top            aument4226689(1,u,3,5) verb(fin([[pers3,sg,pres|ind],[pers2,sg,pres|impr
ndet           la4253663(2)           det(def,[f|sg],X13)
obj(n)         demanda4236031(3,u)    noun(cn,[f|sg],def)
comp(p(de))    de4235161(4,5)         prep(de,X4,5)
objprep(n)     energia4241313(5)      noun(cn,[f|sg],aj)
nadj           electrico4240217(6)    adj(X9,[f|sg],X10)
nprep          por4264971(7,9)        prep(por,X5,9)
ndet           la4253663(8)           det(def,[f|sg],X6)
objprep(n)     ola4260661(9)          noun(cn,[f|sg],def)
nprep          de4235161(10,11)       prep(de,X7,11)
objprep(n)     calor4229297(11)       noun(cn,[m|sg],X8)
```

Fig.3 *Selection of analyses via correspondences (prefer first Spanish analysis because of* subj*-congruity)*

The analyses are associated with the corresponding creation protocols, which are structured lists whose items describe, via the identifiers, which rule has been applied when and to what structures in the process of creating the analysis. From the selection of a best analysis for a sentence, we can entail the circumstances under which the application of particular rules are preferred. This has been carried

[3] Sentence taken from the online newspaper *El Día de Concepción del Uruguay*

out - not yet for the 'new' language Spanish, but for the 'known' language German, in order to obtain a measure about how correctly the existing grammar evaluation component can be replaced by the results of the corresponding statistical study.

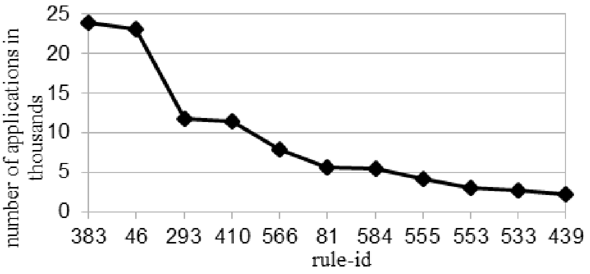


Fig.4 Frequency of applications of rules

cluster applications	similarity	feas mod	feas head
383, 384,..	0,86	sent, ...	emosentaffv,..
557,558,566,..	0,68	denselb,..	gebv, ...

Fig.5 Preliminary constraints related to grammar rule clusters

Fig.4 shows the distribution of rule usages within the training set of analyses (of approx.30.000 sentences). 390 different rules were used with a total of 133708 rule applications. The *subject* rule (383) and the noun determiner rule (46) the most used rules (35% of all applications). Fig 5. illustrates the preliminary results of a clustering algorithm where different rule applications are grouped into clusters and the key features of the head and modifier phrases for each cluster are extracted.

Currently, we try to determine further and tare the linguistic features and the weighting which models best the evaluation for German. (The gold standard that is used in this test is the set of analyses mentioned above). The investigations are not yet completed, but preliminary results on the basis of the morphosyntactic and semantic properties of the neighboring elements are promising. After consolidation, the findings will be transferred to Spanish on the basis of the selection procedure illustrated in Fig. 3. The next step of grammar training in the immediate future will consist of changing the focus to underspecified analyses as described in step 8

5 Conclusions

The project tries to make state-of-the-art statistical methods available for dictionary development and grammar development for a rule-based dominated industrial setting and to exploit such methods there.

With regard to SMT dictionary creation, it goes beyond the current state of the art as it also aims at developing and applying algorithms for the semi-automatic generation of bilingual dictionaries from unrelated monolingual (i.e., comparable) corpora of the source and the target language, instead of using relatively literally translated (i.e., parallel) texts only. Comparable corpora are far easier to obtain than parallel corpora. Therefore the approach offers a solution to the serious data acquisition bottleneck in SMT. This approach is also more cognitively plausible than previous suggestions on this topic, since human bilinguality is normally not based on memorizing parallel texts. Our suggestion models human capacity to translate texts using linguistic knowledge acquired from monolingual data, so it also exemplifies many more features of a truly self-learning MT system (shared also by a human translator).

In addition, the proposal suggests a new method for spelling out grammars and parsers for languages by splitting grammars into declarative kernels and trainable decision algorithms and by exploiting cross-linguistic knowledge for optimizing the results of the corresponding parsers.

For developing different components and dictionaries for the system a bootstrapping architecture is suggested that uses the acquired lexical information for training the grammar of the new language, which in turn uses the (underspecified) parser results for optimizing the lexical information in the corresponding translation dictionaries. We expect that the suggested methods significantly improve translation quality and reduce the costs of creating new language pairs for Machine Translation. The preliminary results obtained so far in the project appear promising.

6 Acknowledgments

This research is supported by a Marie Curie IAPP project taking place within the 7th European Community Framework Programme (Grant agreement no.: 251534)

7 References

Armstrong, S.; Kempen, M.; McKelvie, D.; Petitpierre, D.; Rapp, R.; Thompson, H. (1998). Multilingual Corpora for Cooperation. *Proceedings of the 1st International Conference on Linguistic Resources and Evaluation (LREC)*, Granada, Vol. 2, 975–980.

Babych, B., Hartley, A., Sharoff S.; Mudraya, O. (2007). Assisting Translators in Indirect Lexical Transfer. Proceedings of the 45th Annual Meeting of the ACL.

Babych, B., Anthony Hartley, & Serge Sharoff (2007b) Translating from under-resourced languages: comparing direct transfer against pivot translation. Proceedings of MT Summit XI, 10-14 September 2007, Copenhagen, Denmark, 29-35

Babych, B. & Hartley, A. (2008). Automated MT Evaluation for Error Analysis: Automatic Discovery of Potential Translation Errors for Multiword Expressions. ELRA Workshop on Evaluation "Looking into the Future of Evaluation: When automatic metrics meet task-based and performance-based approaches". Marrakech, Morocco 27 May 2008. *Proceedings of LREC'08.*

Babych, B. and Hartley, A. (2009). Automated error analysis for multiword expressions: using BLEU-type scores for automatic discovery of potential translation errors. Linguistica Antverpiensia, New Series (8/2009): Journal of translation and interpreting studies. Special Issue on Evaluation of Translation Technology.

Babych, B., Babych, S. and Eberle, K. (2012). Deriving generation-oriented MT resources from corpora: case study and evaluation of de/het classification for Dutch Noun (in preparation)

Baroni, M.; Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. Proceedings of LREC 2004.

Callison-Burch, C., Miles Osborne, & Philipp Koehn: Re-evaluating the role of BLEU in machine translation research. EACL-2006: 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 3-7, 2006; pp.249-256

Charniak, E.; Knight, K.; Yamada, K. (2003). Syntax-based language models for statistical machine translation". Proceedings of MT Summit IX.

Eberle, Kurt (2001). FUDR-based MT, head switching and the lexicon. *Proceedings of the the eighth Machine Translation Summit,* Santiage de Compostela.

Eberle, Kurt (2004). *Flat underspecified representation and its meaning for a fragment of German.* Habilitationsschrift, Universität Stuttgart.

Eberle, K.; Rapp, R. (2008). Rapid Construction of Explicative Dictionaries Using Hybrid Machine Translation. *In: Storrer, A.; Geyken, A.; Siebert, A.; Würzner, K._M (eds.) Text Resources and Lexical Knowledge: Selected Papers from the 9th Conference on Natural Language Processing KONVENS 2008. Berlin: Mouton de Gruyter..*

Eckart,K., Eberle, K.; Heid, U. (2010) An infrastructure for more reliable corpus analysis. *Proceedings of the Workshop on Web Services and Processing Pipelines in HLT of LREC-2010* , Valetta.

Eberle, K.; Eckart,K., Heid, U.,Haselbach, B. (2012) A tool/database interface for multi-level analyses. *Proceedings of LREC-2012* , Istanbul.

Frederking, R.; Nirenburg, S.; Farwell, D.; Helmreich, S.; Hovy, E.; Knight, K.; Beale, S.; Domashnev, C.; Attardo, D.; Grannes, D.; Brown, R. (1994). Integrated Translation from Multiple Sources within the Pangloss MARK II Machine Translation System. *Proceedings of Machine Translation of the Americas*, 73–80.

Frederking, Robert and Sergei Nirenburg (1994). Three heads are better than one. In: Proceedings of ANLP-94, Stuttgart, Germany.

Fung, P.; McKeown, K. (1997). Finding terminology translations from non-parallel corpora. *Proceedings of the 5th Annual Workshop on Very Large Corpora,* Hong Kong: August 1997, 192-202.

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

González, J.; Antonio L. Lagarda, José R. Navarro, Laura Eliodoro, Adrià Giménez, Francisco Casacuberta, Joan M. de Val and Ferran Fabregat (2004). SisHiTra: A Spanish-to-Catalan hybrid machine translation system. Berlin: Springer LNCS.

Gough, N., Way, A. (2004). Example-Based Controlled Translation. *Proceedings of the Ninth Workshop of the European Association for Machine Translation*, Valetta, Malta.

Groves, D. & Way, A. (2006b). Hybridity in MT: Experiments on the Europarl Corpus. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, Oslo, Norway, 115–124.

Groves, D.; Way, A. (2006a). Hybrid data-driven models of machine translation. *Machine Translation,* 19(3–4). Special Issue on Example-Based Machine Translation. 301–323.

Habash, N.; Dorr, B. (2002). Handling translation divergences: Combining statistical and symbolic techniques in generation-heavy machine translation. *Proceedings of AMTA-2002,* Tiburon, California, USA.

Kiss, T.; Strunk, J. (2006): Unsupervised multilingual sentence boundary detection. Computational Linguistics 32(4), 485–525.

Koehn, P. (2005). Europarl: *A Parallel Corpus for Statistical Machine Translation*. Proceedings of MT Summit X, Phuket, Thailand

Koehn, P.; Knight, K. (2002). Learning a translation lexicon from monolingual corpora. In: *Proceedings of ACL-02 Workshop on Unsupervised Lexical Acquisition*, Philadelphia PA.

Language Industry Monitor (1992). Statistical methods gaining ground. In: *Language Industry Monitor*, September/October 1992 issue.

McCord, M. (1989). A new version of the machine translation system LMT. *Journal of Literary and Linguistic Computing, 4*, 218–299.

McCord, M. (1991). The slot grammar system. *In: Wedekind, J., Rohrer, C.(eds): Unification in Grammar,* MIT-Press.

Melamed, I. Dan (1999). Bitext maps and aligment via pattern recognition. *Computational Linguistics*, 25(1), 107–130.

Munteanu, D.S.; Marcu, D. (2005). Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4), 477–504.

Och, F.J.; Ney, H. (2002). Discriminative trainig and maximum entropy models for statistical machine translation. Proceedings *of the Annual Meeting of the Association for Computational Linguistics*, Philadelphia, PA, 295–302.

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

Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. (2002). BLEU: A method for automatic evaluation of machine translation. In: *Proceedings of the 40th Annual Meeting of the ACL*, Philadelphia, PA, 311–318.

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

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics 1999, College Park, Maryland. 519–526.*

Rapp, R. (2004). A freely available automatically generated thesaurus of related words. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC)*, Lisbon, Vol. II, 395–398.

Rapp, R.; Martin Vide, C. (2007). Statistical machine translation without parallel corpora. In: Georg Rehm, Andreas Witt, Lothar Lemnitzer (eds.): *Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007.* Tübingen: Gunter Narr. 231–240

Resnik, R. (1999). Mining the web for bilingual text. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics.

Sato, S.; Nagao, M. (1990). Toward memory-based translation. *Proceedings of COLING 1990*, 247–252.

Schiehlen, M. (2001) Syntactic Underspecification. In: Special Research Area 340 – Final report, University of Stuttgart.

Sharoff, S. (2006) Open-source corpora: using the net to fish for linguistic data. In International Journal of Corpus Linguistics 11(4), 435–462.

Sharoff, S.; Babych, B.; Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. In: Proceedings of COLING/ACL 2006, 739–746.

Sharoff, S. (2006). A uniform interface to large-scale linguistic resources. In Proceedings of the Fifth Language Resources and Evaluation Conference, LREC-2006, Genoa.

Simard, M., Foster, G., Isabelle, P. (1992). Using Cognates to Align Sentences in Bilingual Corpora. P*roceeedings of the International Conference on Theoretical and Methodological Issues*, Montréal.

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1), 143–177.

Streiter, O., Carl, M., Haller, J. (eds)(1999). *Hybrid Approaches to Machine Translation*. IAI working papers 36.

Streiter, O.; Carl, M.; Iomdin, L.L.: 2000, A Virtual Translation Machine for Hybrid Machine Translation'. In: *Proceedings of the Dialogue'2000 International Seminar in Computational Linguistics and Applications*. Tarusa, Russia.

Streiter, O.; Iomdin, L.L. (2000). Learning Lessons from Bilingual Corpora: Benefits for Machine Translation. International Journal of Corpus Linguistics, 5(2), 199–230.

Thurmair, G. (2005). Hybrid architectures for machine translation systems. *Language Resources and Evaluation,* 39 (1), 91–108.

Thurmair, G. (2006). Using corpus information to improve MT quality. *Proceedings of the LR4Trans-III Workshop*, LREC, Genova.

Thurmair, G. (2007) Automatic evaluation in MT system production. MT Summit XI Workshop: Automatic procedures in MT evaluation, 11 September 2007, Copenhagen, Denmark,

Veronis, Jean (2006). *Technologies du Langue. Actualités – Comentaires – Réflexions. Translation. Systran or Reverso?* http://aixtal.blogspot.com/2006/01/translation-systran-or-reverso.html

Wu, D., Fung, P. (2005). Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. *Second International Joint Conference on Natural Language Processing (IJCNLP-2005).* Jeju, Korea.