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

D5.4

Bilingual Dictionary Extraction Tools

Dissemination Level:	Public
Delivery Date:	June 30th 2012
Status – Version:	Final v1.0
Author(s) and Affiliation:	Yvette Graham (DCU), Gregor Thurmair (Linguatec), Antonio Toral (DCU), Vera Aleksić (Linguatec)

Relevant Documents

D5.3	English-French and English-Greek parallel corpus for the Environment and Labour Legislation domains (M22)
D5.5	English-French and English-Greek bilingual dictionaries for the Environment and Labour Legislation domains
D7.4	Panacea Deliverable D7.4: Evaluation Report (Third cycle)

Table of Contents

1 Introduction

This document describes methods developed within task WP5.2. to automatically extract bilingual dictionaries from statistical machine translation (SMT) phrase tables, that were automatically generated from sentence aligned bilingual corpora (Deliverable 5.3).

The rest of the document is structured as follows. Firstly, in Section 2, the input parameters and output format of the bilingual dictionary extractors developed are defined. Secondly, in Section 3, the basic dictionary extraction tool developed at Linguatec for French to English is described. It uses standard SMT phrase-tables as input and outputs small precise dictionaries. Next, in Section 4, the advanced dictionary extraction tool developed at DCU is described. It takes a factored phrase table as input and outputs larger dictionaries with slightly lower precision. The evaluation for both tools will be presented in D7.4 Third Evaluation Cycle (to be delivered in month 34).

2 Parameters and Output Format

This section defines the mandatory parameters, Common Interface (CI) as defined in the PANACEA project, of the web services that provide dictionary extraction as well as the output format of the dictionaries produced by these services.

The mandatory parameters are the following:

- phrase_table. A file containing a phrase table.
- source_language. Source language in 2-char ISO code format.
- target_language. Target language in 2-char ISO code format.

The format chosen for the created dictionaries, in PANACEA terms the Travelling Object (TO), is a simple tab-based format, as such a format provides the required functionality for the data that is to be represented while ensures the efficient processing of such data, if it is to be handled by any subsequent tool.

Each entry of the dictionary takes one line. Each line contains four fields separated by tabs. The first is the term in the source language, the second the POS tag in the source language. Similarly, the third and fourth fields hold the term and POS, respectively, in the target language.

Such a format can be easily converted into a standard lexicon format, and a special converter will be supplied to bring the data into an LMF representation. Multiword representation is of particular interest here as most of the term candidates are multiword terms.

The services for the standard and the advanced tool can be found in the PANACEA registry:

- Standard tool: http://80.190.143.163/panaceaV2/services/LTPhr2Glo?wsdl, and a corresponding workflow in myExperiment
- Advanced tool: http://registry.elda.org/services/247

3 Basic Dictionary Extraction Tool: LT-P2G

The standard approach towards bilingual term extraction is a two-step procedure: first *identification* of term candidates in the source language, and then *mapping* of source to target term candidates. Usually the corpus data needs to be preprocessed, e.g. by applying lemmatisation / Part-of-Speech (POS) tagging (Caseli/Nunez, 2006).

The system presented here, called P2G (PhraseTable to Glossary), takes the opposite approach: it does mapping *first* (using state-of-the-art phrase aligners), and *then* it does extraction from the aligned phrases, by applying filters to the phrases. This approach follows the following considerations:

1. If a (monolingual) source language term candidate does not have a correspondence in the target language, it is unlikely that it is really a term. In turn, this means that if something is a term (i.e. a relevant concept) in a bilingual set-up, then it *must* show up in the alignment results, and the alignment can be used as a filter for term candidates.
2. The best available alignment tools produce translation tables which contain all possible term mappings (and beyond that many phrases which would not be considered as proper terms). So most of the correct term candidates *will* be represented in such translation tables.
3. As a result, the task consists in identifying 'good' term candidates from phrase table input. This is achieved by applying different *filters* to such input to extract the good terms.

The consequence is that *no preprocessing of corpus data is required* for the P2G tool; all information needed is either kept in the P2G tool (as language resources), or is derived from the input. Two input formats are supported:

- Phrase tables as produced by Moses (Koehn et al., 2007)
- Phrases aligned with AnymAlign (Lardilleux/Lepage, 2009; Lardilleux et al., 2012)

Moses output has proven in tests to have better alignment quality for the task at hand.

So the system expects as input: a phrase-aligned resource (like a phrase-table) and a source and target language mark-up. The basic tool supports the following languages as source and target: English (en), German (de), French (fr), Spanish (es), Italian (it) and Portuguese (pt).

3.1 Workflow of the Basic tool (P2G)

As mentioned earlier, the P2G approach is to apply filters on input records of aligned phrases. Among many other phrases, also term candidates must have been found; the task is to filter out those candidates. Formats of different alignment tools are supported as input for aligned phrases. Three filters are applied, as shown in Figure 1.

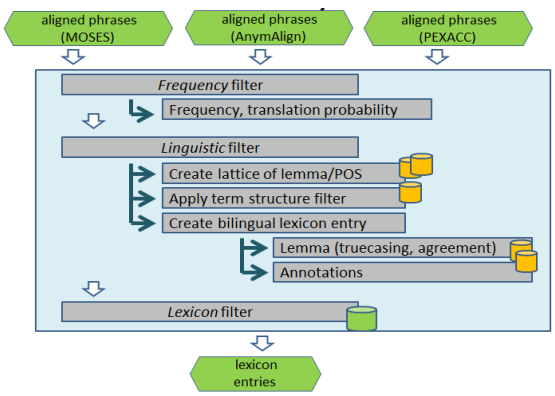

Figure 1: Operation flow of the P2G system

The filters are:

- A **Frequency filter**: Only phrases with a given frequency and / or translation probability are accepted as term candidates.
- A **Linguistic filter**: Only phrases which have certain linguistic properties are acceptable. If a candidate passes the linguistic filter, it is brought into the right lexicon form, in terms of lemma creation, assignment of annotations, etc.
- The **Lexicon filter** compares the lexicon entries just produced with a filter resource. This way, candidate entries can be removed which are already known, or are not wanted, or should not be part of the output for some other reason.

Further details are given in the following sections.

3.1.1 Frequency Filter

As the system does not create alignments itself (i.e. translation candidates), it must rely on the efficiency of the statistical alignment tools from which it receives the aligned candidates. The first step is therefore to identify the best translation proposals, in terms of recall (as many terms as possible) and precision (as good translations as possible).

Two factors influence the translation quality of the P2G tool: the selection of the alignment tool, and the selection of the thresholds for frequency and translation probability.

For the alignment tool, it can easily be seen that GIZA++ only is insufficient, as no multi-word entries are found, which form nearly 50% of a lexicon / term list, especially in narrow domains. So the focus was on phrase alignment tools, which also give superior quality in translation (Och and Ney, 2004). To create phrase alignment, two alignment methods were tried out[1]:

- Giza++ and **Moses** (Koehn et al., 2007), creating Phrase Tables. From the *LT_automotive* input data (cf. below), a phrase table with about 7.97 mio entries was built.
- Phrases as produced with **Anymalign (**Lardilleux/Lepage 2009**)**. Anymalign created about 3.14 mio word/phrase pairs from the same input data.

It soon turned out that if **frequency** is not considered, too much noise would be in the output. Therefore, frequency (on source and target side) is used and set to > 1.

For the **translation probability**, tests were done to find the optimal recall / precision combination.

The two alignment systems were compared, using different values for the translation probability. For evaluation, a random set of term candidates manually inspected[2], and the errors in alignment / translation were counted[3]. The results are given in Table1.

Tool	translation probability	no entries	errors
Moses	$p > 0.8$	12.000	5.54%
Moses	$0.6 < p < 0.8$	3.900	5.42%
Moses	$0.4 < p < 0.6$	20.000	55.11%
AnymAlign	$p > 0.7$	12.600	46.91%

[1] Input from PEXACC (Ion et al., 2011) for comparable corpora is also supported.
[2] Entries starting with the letters C, F, and S.
[3] There are always unclear cases among translations (e.g. transfers usable only in certain cases); they were not counted as errors. Errors are only clearly wrong translations; however a range of subjectivity remains.

AnymAlign	p > 0.8	10.900	47.56%

Table 1: Translation errors for different alignment methods and probabilities

It can be seen that the Moses alignment has much better quality, and is in the reach of being usable; AnymAlign error rates are approximately ten times higher. For AnymAlign, taking a higher threshold (0.8 instead of 0.7) does not improve alignment quality. Overall, Moses input with a threshold of 0.6 for P(f|e) seems to give the best results for term extraction, for this size of phrase tables[4], with an overall error rate of about 5.5%: it increases recall without reducing precision.

It should be noted that alignment errors result from external phrase alignment components, and are just 'inherited' by the current extraction system. However, they count in the overall workflow evaluation: Incorrect translation proposals lead to significantly higher human reviewing effort.

3.1.2 Linguistic Filter

Not all phrase aligned candidates which pass the frequency filter are linguistically meaningful. So only the ones which can be terms, or lexicon entries, are extracted[5]. Most such terms have an internal linguistic structure, described by a part-of-speech tag sequence. So the internal structure of the linguistic filter is:

- Create a word lattice for the input string, providing the different readings for each of the input words.
- Match the input lattice to the legal term patterns, on source and target side.
- Create a lexicon entry for candidates with a successful match on both source and target side, with proper lemma and its annotations.

a. Word lattice

First, each candidate input phrase is tokenized and normalized in spelling and casing[6]. Next, each token is lemmatised to find its base form and part-of-speech tag. Lemmatisation is basically done by lexicon lookup. Unknown words are handled by a POS-defaulting component; for German unknown words, a decomposer component is called to find a known head word. This procedure is documented in (Thurmair et al., 2012).

As tokens can have multiple readings, the result of this procedure is a word lattice consisting of the respective readings of each of the single words of a candidate. This procedure is language-specific, and is done on both source and target side.

b. Term Pattern matching

From the word lattice, all possible POS sequences are created, and compared to the legal term structure patterns. The patterns go significantly beyond the 'usual suspects'; they were collected as the result on an inspection of a large terminological database. For German, patterns for the structures are provided[7] as shown in Figure 2.

[4] However, this changes with the size of the phrase table, cf. section 5.5 below.

[5] As a consequence, there are phrases in the phrase table which are perfectly valid translations, however would never be found in a term bank.

[6] Normalisation in casing is problematic as it also lowercases proper names. However, *not* doing it would lead to significant errors due to the fact that phrase tables contain many capitalized non-propername words. The output would contain pseudo-doublets from capitalized and non-capitalized term proposals. Example: '*Financial debt' where* lowercased '*financial debt'* can also be found.

[7] Not covered: Proper nouns (*Lufthansa Service Center*), and terms containing conjunctions (*Facts and Figures*), as the backend MT system cannot cope with some of such structures.

Term ::= AdP? NoC (NoC | NP | PP)?
AdP ::= Ad | VbP
NP ::= Dt (AdP)? NoC
PP ::= (Ap Dt? AdP? NoC) | (ApPD AdP? NoC)

Figure 2: Term structure for German

The maximum length of such patterns is set to 6 members; longer terms are hardly ever found in term banks, and are even rarer in running texts.

The pattern filters are of course language-specific; e.g. in German and Greek, patterns must be foreseen which cover post-head NP's in genitive case, French and Spanish patterns cover both prenominal and postnominal adjectives, etc.

The matching strategy is a simple best-first approach, i.e. it returns the first match. It could be improved by sorting the multi-word patterns according to frequency, and/or giving weights to the different POS readings of an input word. However such extensions would only marginally affect the results, and would not avoid the most frequent errors of this filter (cf. the evaluation below, Section 4).

The pattern filter is applied to the candidates on both the source and target side, independently of each other, to be able to map a source language single word (e.g. a German compound) to a target language multi-word expression. If both side candidates pass the filter, then the sequence of readings corresponding to the matching patterns is given to the entry creation module.

c. Term and Lexicon Entry Creation

All entries which have passed the filter so far must be brought into a proper canonical form. The creation of lexicon entries for source and target consists of two parts:

- Creating proper **lemmata**. This is required for both term and lexicon use.
- Creating proper **lexicon entries**. This is relevant if the extracted terms are to be integrated into MT systems; such systems usually require certain annotations (at least part of speech information).

Lemma creation implies the creation of a canonical form for the entry. This has two aspects:

- **Truecasing** of all lemma parts: Proper names and German common nouns should be capitalized, the other forms lowercased.
- Production of the **canonical form** of the lemma.

  The *head* (or the term if it is a single word) is lemmatised, and the lemma is given as canonical form. In multiword entries, the head position is given in the pattern.

  The *modifiers* in a multiword entry are treated as follows:
  Head-modifying adjectives must be set into gender-number-agreement with their head (it '*cardiopatia coronarica'*, es '*cuestión política'*)[8]. Therefore the production of the lemma of multiword entries requires knowledge about the gender of the head. To provide this, a special component (gender defaulter) has been added to the system which consults an appropriate resource; depending on the gender of the noun, the adjective is inflected[9].
  The *post-head modifiers* of the multiword stay in their inflected form: de '*Oberfläche mit*

[8] In German, there are even two options, the weak inflection (<das> '*niedrige Zinsniveau'*) or the strong one (<ein> '*niedriges Zinsniveau'*).Both can be found in dictionaries; the strong inflection is more difficult as it requires knowledge of the head noun gender; unfortunately this is the form expected by the backend MT system.
[9] The system uses a static inflection resource for this.

speziellen Farbpigmenten', en '*surface with special color pigments*' would leave the PP untouched.

Based on these two principles, the multi-word lemma is composed[10]. It should be noted that the step of creating canonical forms can create duplicates (e.g. if a phrase table contains one entry for a singular and another one for a plural noun). Such duplicates must be eliminated before the final list is output.

Lexica go beyond term lists as their entries need **annotations**. The lexicon entries in P2G show the following annotations:

- All of them have a lemma, a part of speech, and a reading number, as these elements constitute an entry. In addition, they have annotations which depend on a feature called '*entrytype'*, with values '*singleword'*, '*compound'*, '*multiword'*.
- *Single word entries* are annotated with gender (in German) and inflection; this information is either taken from the lexicon, or defaulted.
- *Multiword entries* and compounds (i.e. the agglutinated German compounds) share the same entry structure; they provide: the head position, the sequence of lemmata, and the sequence of parts of speech of which the multiword consists. These annotations allow for a successful identification of multiword terms in texts.

Of course, the lexicon must contain much more information; however this goes beyond what the term extraction can contribute. In turn, the use which can be made of the provided annotations depends on the single backup MT systems and their import possibilities: Most systems can use (or even require) POS information, but e.g. not all multi-word term patterns are supported (e.g. terms containing conjunctions). Tests on transfers, like in (Caseli and Nunez, 2006), are not created, however.

The final output of the linguistic filter consists either of complete *lexical* entries (for MT import), or of *term* entries (for human lookup), depending an output format parameter.

3.1.3 Lexicon Filter

Before human post-editors select the entries which they really want to keep, a possibility has been created to remove unwanted term candidates. Such entries could be:

- Candidates which are already known; they need not be reviewed a second time.
- Candidates which do not belong to a specific domain (e.g. automotive); the filter then would be a general-domain lexicon, letting pass only narrow-domain words.
- Candidates which contain certain stopwords (like en '*large'*).
- Candidates which are known to be irrelevant.

The system offers the option to apply a filter which blocks this kind of entries. Users would provide the filter data themselves; only non-matching entries pass the lexicon filter.

The basic P2G tool is described in (Thurmair/Aleksić 2012).

[10] These heuristics for truecasing and for lemma creation leave room for errors, e.g. in cases where the prenominal adjective is in comparative form (de '*der frühere Präsident*' -> *'*der frühe Präsident*'), or in cases where the head should be in plural (en '*facts & figures*' -> *'*fact & figure*'). However, they show the best performance overall.

4 Advanced Dictionary Extraction Tool: DCU-P2G

The basic tool applies POS tagging to the phrases in the SMT phrases. The original context of these phrases is therefore unknown when POS tagging is applied. Using Moses (Koehn et al., 2007) in factored model mode, it is possible to run part-of-speech tagging on the training corpus before extracting phrases, instead of part-of-speech tagging SMT phrases, so that the preceding context of words is known, as this is used to estimate the probability of the tags and should result in higher accuracy tagging, especially for short phrases and single-word terms.

For this reason, when developing the advanced tool, POS tagging is applied prior to phrase extraction. In addition, the advanced tool is easily adapted to new language pairs. The tool was initially developed for French–English and has since been applied to Greek–English. All that is required for a new language pair is to define a regular expression for valid multi-word expression POS tag sequences for the new languages.

The web service created for the advanced automatic dictionary extraction tool can be accessed at http://www.cngl.ie/panacea-soaplab2-axis/. A screenshot is shown in Figure 3.

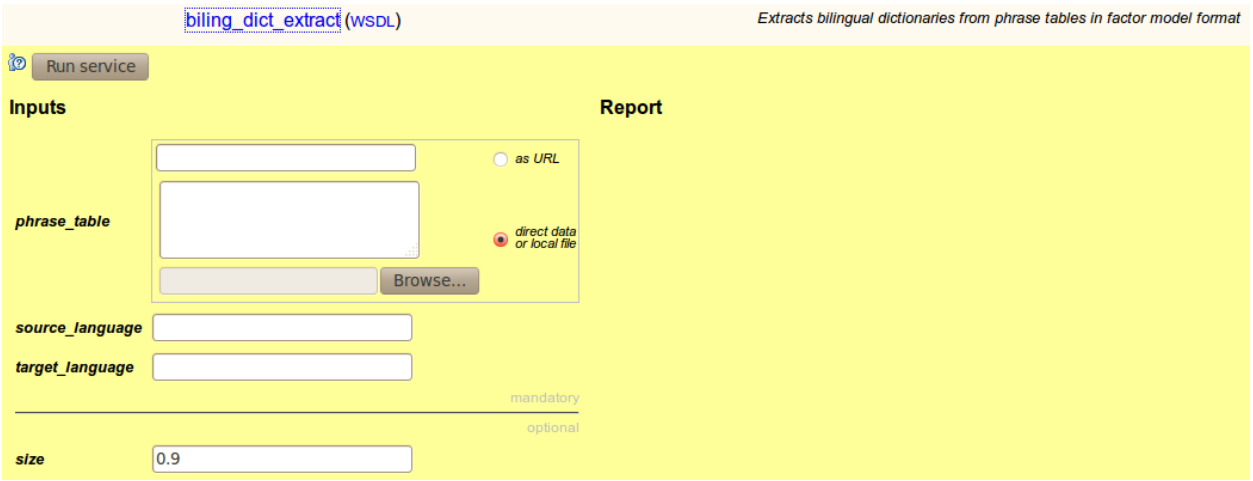

Figure 3: Interface for the web service of the advance dictionary extraction tool

4.1.1 Preparing Input

The advanced dictionary creation tool takes a factored SMT phrase table as input. The original bilingual corpus has been part-of-speech tagged, so that each word in the training data is now accompanied by its most likely part-of-speech tag and lemma form. Moses SMT training is then run in factored model mode, so that the resulting phrase-table not only contains surface-form phrases with features scores, in addition, each word in a given phrase retains its POS tag as well as its lemma form.

4.1.2 Part-of-Speech Tag Filtering

Language-specific POS tag filters are applied separately to each side of a given SMT phrase and phrases where the POS tag sequence of either the source side or the target side is not a valid POS sequence is filtered out. Valid POS tag sequences are defined for each language for single-word terms (consisting of a single part-of-speech tag) and multi-word terms. The following is a comprehensive list of POS tag sequences permitted by the POS tag filter for each language:[11]

Single-word Terms (all languages):

- noun

[11]Notation: ? Is used to denote 0 to many words with that part-of-speech tag. For example, a phrase will be allowed through the filter “adjective? noun” if it consists of 0 or more adjectives followed by a noun.

- verb
- adjective

French Multi-word Terms:

- adjective? noun noun? adjective?|verb?
- adjective? noun noun? adjective?|verb? preposion determiner? adjective?|verb? noun adjective?|verb?

English Multi-word Term:

- adjective? noun noun? Adjective?
- adjective? noun noun? adjective? prep det? adjective? noun noun?

Greek Mult-word Terms:

- aj|verb:mnppxx? no no?
- adjective|verb:mnppxx? nooun noun? as at? adjective|vbmnppxx? noun noun?

The tag filter is language-specific, esp. in the case of multiword terms: Noun – noun appositions are rather frequent in English, noun – NP structures are found in languages with Genitive case-markers (Greek, German), languages differ in pre- / post-nominal adjective position etc. However, these filters can be easily adapted to new languages.

4.1.3 Lemmatising the Head Word

By default, the tool converts the head word of both the source and target phrases to lemma form. For multi-word terms, there are some exceptions where it is not appropriate to lemmatise the head-word. Since Greek is so highly inflected, the tool does not attempt to lemmatise the head of any multi-word terms. For French, the tool does not lemmatise the head if the multi-word expression contains an adjective. To summarize, if either the Greek or French side of a phrase meets the following criteria, the headwords of both sides of the phrase are not lemmatised:

French:

- the phrase is a multi-word term containing an adjective

Greek:

- the phrase is a multi-word term

4.1.4 Feature Score Filtering

The SMT phrase table, given as input to the tool, includes the following feature scores for each phrase:

- inverse phrase translation probability $\varphi(f|e)$
- inverse lexical weighting $lex(f|e)$
- direct phrase translation probability $\varphi(e|f)$
- direct lexical weighting $lex(e|f)$

In order to estimate the quality of potential phrases that were permitted through the POS tag filter described in Section 3.1.2, we use a log-linear combination of the above feature scores for each phrase using default weights (each feature is given equal weight). An extension of this method would be to optimise the weights on a development set of terms to obtain better ranking of dictionary entries. However, due to time constraints we leave this to future work. Lower scoring phrases are filtered out.

4.1.5 Precision and Recall Trade-off

Inevitably, an automatically constructed dictionary will contain entries that a human would judge

as unsuitable for one reason or another. For example, the entry could be an incorrect translation, due to an error in word alignment during SMT training. There is a trade-off between how precise and how large the dictionaries produced by the tool will be. In the extreme case, we can allow in all phrases that pass through the POS filter into our dictionary, which is likely to give low precision but with very high recall, whereas if we only require a dictionary with as few as ten entries, the top ten scored entries are all likely to be correct giving us 100% precision but with very low recall.

We allow the user control over this trade-off by letting them specify what size dictionary they would like to produce. If they would like high precision and low recall, they should request a small dictionary. However, if precision is not their top priority but need a large dictionary, they can request a large dictionary with high recall. Since the size of the phrase table is not known until run-time, the requested size is specified as a percentage. For example, the user enters a *1,000* word entry phrase table with size *0.75*, the tool produces a dictionary with *750* entries.

5 Evaluation

Two kinds of evaluation were performed in the task 5.2:

- The basic P2G tool was evaluated for aligned phrases of all languages covered by it (en, fr, es, de, it, pt), using phrase tables of several sizes and formats.
- In addition, a Gold Standard evaluation was done for a small subset of the PANACEA domain tables, with a comparison of the approaches.

Details are given in the Evaluation report D7.4 of the PANACEA project; the following section gives just a short summary.

5.1 P2G Basic Tool Evaluation

As explained above, while there is no clear view which entries *should* be in the term list, there is agreement on which candidates should *not* be presented, and be considered as noise: It is *this* type of entry, which the term extract evaluation will focus on.

Several corpora were used for testing, related to several projects:

- The PANACEA corpora for environment, prepared by DCU: (*DCU_ENV*) and labour legislation (*DCU_LAB*)[12]
- Corpora in the Health and Safety domain, collected by Linguatec (*LT_H&S*) in different languages
- A corpus on automotive texts, collected by Linguatec (*LT_autom*.)
- The ACCURAT corpora for automotive, in two versions, prepared by DFKI: *DFKI_adapt* and *DFKI_lexacc*[13].

The size, languages treated, size of phrase tables created, and number of glossary entries extracted is given in Table 2.
From all corpus data sets, term candidates were extracted by the P2G system. From these candidates, term candidates were selected randomly. These candidates were evaluated manually by two evaluators. Overall, 99 K bilingual term candidates were extracted of which 17.2 K (17%) were manually evaluated.

Two kinds or errors are distinguished in the evaluation: *Translation* errors, i.e. the candidates are not translations of each other; and *P2G errors*, i.e. Lemma and annotation errors created by the P2G tool.
Table 2 shows the evaluation results. The average error rate of the complete P2G system is 9.26%, varying from 7.3 to 14.4%.

[12] cf. Mastropavlos / Papavassiliou. 2011.
[13] cf. ACCURAT Deliverable D4.2: Improved baseline SMT systems adjusted for narrow domain. 2012

Corpus	Lang.	PhrTab size	Gloss. size	Transl. error	P2G error	Total error
DCU_ENV	en-fr	400	2.8	5.2%	1.3%	7.8%
DCU_LAB	en-fr	800	4.5	4.9%	1.2%	7.3%
LT_H&S	fr-en	2.900	10.7	11.3%	1.3%	13.9%
LT_H&S	es-en	2.600	13.2	10.9%	0.4%	11.6%
LT_H&S	it-en	2.100	9.9	9.8%	2.3%	14.4%
LT_H&S	pt-en	600	4.4	12.7%	0.4%	13.5%
LT_autom.	de-en	7.970	15.7	5.7%	2.8%	10.3%
DFKI_adapt	de-en	85.000	23.2	1.5%	3.3%	8.0%
DFKI_lexacc	de-en	83.900	23.3	1.7%	3.1%	7.9%

Tab. 2: Evaluation results: Phrase Table size (K entries), size of extracted glossaries (K entries), error rates of translation, of P2D, and combined error rates

Translation errors: Translation errors vary from 1.5% to 12.7%, with 5.1% on average. They are produced by MOSES alignment, and are not accessible to the P2G tool; however, they increase the total error rate. Translation errors seem to correlate with the size of the phrase tables[14]: Larger phrase tables show a lower translation error rate for the extracted terms.

P2G errors: P2G errors vary from 0.4% to 3.3%, depending on the languages involved[15], with an average error rate of 2.1%. Many of these errors can be corrected by improvements of the backend components (dictionary, gender defaulters etc.), which would bring the P2G error rate down by an estimated 1%. The P2G errors do not depend on the size of the data; they are also language-dependent.

Total errors: As the output of the system is a bilingual lexicon, i.e. description of two source terms plus their translation, the error rates accumulate, so the overall error rate of the tool is two P2G error rates plus translation error rate; the total error rate is somewhat linear to the translation error rate. In total it is between 7.3% and 14.4%, which means that 8 entries out of 100 need to be corrected by human reviewers. This can be considered a reasonable result of a term extraction component.

Another observation is that the translation probability threshold for the frequency filter should be set depending on the size of the phrase table. To test this, the *DFKI_lexacc* data were split into packages depending on the translation probabilities. In each package, about 1000 entries were manually evaluated. The results show that the entry sets with a probability > 0.4 have basically the same error rate entry sets from 0.2 to 0.4 have a slightly increased error rate, and entries < 0.2 cannot be used. This means that recall can be improved dramatically by lowering the probability threshold for large phrase tables, with no or just minimal loss in precision, cf. Table 3. This result is also corroborated by the Gold Standard Evaluation below.

translation probability	no. entries retrieved	expected trans-lation error rate
P (f|e) > 0.4	67.664	2.25 %
P (f|e) > 0.2	109.418	3.53 %

Tab. 3: Recall improvement for large phrase tables (DFKI_lexacc)

As a result, the P2G term extraction tool can produce a 110 K bilingual glossary from phrase tables where 92 out of 100 entries are correct (7.7% total error rate[16]).

[14] DCU_ENV and DCU_LAB need to be considered in more detail.
[15] P2G supports the languages en de fr es it pt
[16] Two times the average P2G of 2.1% plus the translation error rate of 3.53%

5.2 Gold Standard Evaluation

In order to evaluate the tool and specifically to evaluate the effects of applying feature score filtering to the dictionary extraction tools, a gold standard of dictionary entries was manually created for each language pair and domain. The human evaluators were asked to annotate a random sample of dictionary entries produced by the tool that had not been filtered with feature scores. Results are included in the following section for Precision, Recall and F-score for each tool on the gold standard sets.

5.2.1 Basic P2G tool

Results for the standard dictionary extraction tools for French-English for the gold standard test sets for each domain are as follows:

Corpus	filter	precision	recall	F-score
fr-en ENV	only Freq > 1	89.38	21.82	35.07
	p(e|f) > 0.6 & Freq > 1	95.81	12.17	21.59
fr-en LAB	only Freq > 1	84.88	25.18	38.83
	p(e|f) > 0.6 & Freq > 1	96.25	11.96	21.27

Tab.4: Basic tool, taking only frequency or frequency plus translation probability

This evaluation showed that

- the component should only extract what can safely be used by human posteditors, i.e. precision should be close to human percision
- the translation probability is a significant factor in quality determination. Section 5.1.5 above shows that it depends on the size of the phrase tables, and can cautiously be lowered if the phrase table size increases.

5.2.2 Advanced Dictionary Extraction Tool

Bilingual dictionaries were automatically extracted using the advanced tool for French-English and Greek-English and the output dictionaries were compared with the maunally labeled gold standard. Results for the advanced dictionary extraction tool for French-English and Greek-English on the gold standard test sets are as follows:

Evaluation results of Advanced Method

Corpus	Lexicon Size	Precision	Recall	F-score
fr-en ENV	(min) 0.1	86.00	9.82	17.63
	(max) 1	84,60	100.00	91.66
fr-en LAB	(min) 0.1	80.00	9.35	16.75
	(max) 1	81.60	100.00	89.87
el-en ENV	(min) 0.1	76.00	10.53	18.49
	(max) 1	68.40	100.00	81.24
el-en LAB	(min) 0.1	74.51	10.05	17.72
	(max) 1	69.26	100.00	81.84

Table 5: Variation of lexicon size, changes in F-scores

The results on the gold standard show the trade-off between precision and recall for all language pairs and domains. As recall increases precision decreases, however much slower than the recall

increases. The f-score is included, so that a comparison can be made for the cases when precision and recall are equally important. For all language pairs, as the size of the dictionary increases, the f-score increases, showing that the drop in precision is less than the increase in recall as more phrases are allowed through the feature score filter.

5.3 Comparison of Basic and Advanced Tools

The basic tool has only been compared for French-English language pairs and achieves very high precision with very low recall, resulting in a low f-score for each domain when compared to that of the advanced tool. If a large dictionary is required, therefore the advanced tool is better. However, if small with high precision are needed the basic tool achievges higher precision, and this is probably due to the extra filter of phrases that ocurr only once in the corpus being filtered out by the tool, as this filter is not applied in by the advanced tool.

6 Conclusions and Future Work

6.1 Basic P2G tool

For the basic extraction tool, the following development requirements exist:

- extension to other languages; this includes the creation / adaptation of language resources for these languages.
  The P2G basic tool requires as resources (cf. the yellow boxes in Fig. 1 above): resources for normalisation and lemmatisation; little grammars for term filtering; POS recognition and inflection in cases of noun – adjective agreement. The extended tool only needs the filter expressions; lemmatisers etc. are already presupposed in the factorised input.
- improving recall without lowering precision. Current precision is close to human judgement; it needs to be seen if recall can be extended without losing acceptance by human post-editors. An option is to evaluate different settings of translation probabilities, as proposed by the advanced tool
- improving extraction quality by using better (monolingual) resources; esp. in English there are mistakes (like non-capitalisation of proper names) due to lexicon gaps
- creating complete lexicon entries (not just term candidates), by using defaulting techniques not just for POS and gender, but also for unknown lemmata, inflection classes and other annotations which real lexica consider to be obligatory
- integrating the bilingual lexicon component into a complete processing chain. Such a chain can be:
  - use monolingual lexicon analysis to improve the (monolingual) lexicon resources
  - use these resources to improve bilingual lexicons, extract complete lexicons from phrase tables; use stoplists to filter candidates which are already known
  - use the sentential contexts from which the phrase tables were built to extract transfer tests and transfer selection information (cf. deliverable D5.6).

6.2 Advanced Tool

The advanced bilingual dictionary extraction tool uses a log-linear combination of feature scores for SMT phrases to rank candidate term glossary entries. For this, only default weights are applied. A potential improvement of the method would optimize these weights on a development set to improve ranking of term glossaries, so that a higher level of precision could be reached without a drop in recall. Due to time constraints we leave this to future work.

7 Bibliography

Caseli, H., Nunes, M., 2006: Automatic induction of bilingual resources for machine translation: the ReTraTos project. Machine Translation 20,4

Daille, B., Morin, E., 2005: French-English Terminology Extraction from Comparable Corpora. Proc. IJCNLP 2005

Fung, P., McKeown, K., 1997: Finding Terminology Translations from Non-parallel Corpora. Proc. 5th Annual Workshop on Very Large Corpora (VCL 97), Hong Kong

Gamallo Otero, P., 2007: Learning Bilingual Lexicons from Comparable English and Spanish Corpora. Proc MT Translation Summit Copenhagen

Gamallo Otero, P., 2008: Evaluating Two Different Methods for the Task of Extracting Bilingual Lexicons from Comparable Corpora. Proc. LREC Workshop on Comparable Corpora, Marrakech

Ideue, M., Yamamoto, K., Utiyama, M., Sumita, E., 2011: A Comparison of Unsupervised Bilingual Term Extraction methods Using Phrase Tables. Proc. MT Summit XIII, Xiamen

Ion, R., Ceauşu, A., Irimia, E., 2011: An Expectation Maximization Algorithm for Textual Unit Alignment. Proc. 4th Workshop on Building and Using Comparable Corpora (BUCC), Portland, USA

Kit, Ch., 2002: Corpus Tools for Retrieving and Deriving Termhood Evidence, Proc. 5th East Asia Forum of Terminology

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, Evan Herbst, 2007 **Moses: Open Source Toolkit for Statistical Machine Translation**, Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.

Lardilleux, A., Lepage, Y., 2009: Sampling-based multilingual alignment. Intern. Conf. on Recent Advances in Natural Language Processing (RANLP 2009), Borovets, Bulgaria

Lardilleux, A., Yvon, F., Lepage, Y., 2012: Hierarchical Sub-Sentential alignment with AnymAlign. Proc EAMT Trento

Macken, L., Lefever, E., Hoste, V,.2008: Linguistically-based sub-sentential alignment for terminology extraction from a bilingual automotive corpus. Proc. 22nd COLING, Manchester

Mastropavlos, Nikos; Papavassiliou, Vassilis. (2011). Automatic Acquisition of Bilingual Language Resources. Proceedings of the 10th International Conference on Greek Linguistics. Komotini, Greece

Menezes, A., Richardson, St.D., 2001: A best-first alignment algorithm for automatic extraction of transfer mappings from bilingual corpora. Proc. ACL / DMMT

Morin, E., Prochasson, E., 2011: Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora. Proc. BUCC, Portland, Oregon, USA

Och, F., Ney, H., 2004: The Alignment Template Approach to Statistical Machine Translation. Computational Linguistics 30,4

Rapp, R., 1999: Automatic identification of word translations from unrelated English and German corpora. Proc. 37th ACL, College Park, Maryland

Robitaille, X., Sasaki, X., Tonoike, M., Sato, S., Utsuro, S., 2006: Compiling French-Japanese Terminologies from the Web. Proc. 11th EACL, Trento

Thurmair, Gr., 2003: Making Term Extraction Tools Usable. Proc. CLT, Dublin

Thurmair, Gr., Aleksić, V., Schwarz, Chr., 2012: Large-scale lexical analysis. Proc. LREC Istanbul

Thurmair, Gr., Aleksić, V., 2012: Creating Term and lexicon Entries from Phrase Tables, Proc. EAMT Trento

Vu, Th, Aw, A.T., Zhang M., 2008: Term Extraction Through Unithood And Termhood Unification Proc. IJCNLP 2008, Hyderabad, India

Weller, M., Gojun, A., Heid, U., Daille, B., Harastani, R., 2011: Simple methods for dealing with term variation and term alignment. Proc TIA 2011: 9th International Conference on Terminology and Artificial Intelligence, Paris, France

Wolf,P., Bernardi, U., Federmann, Chr., Hunsicker, S., 2011: From Statistical Term Extraction to Hybrid Machine Translation. Proc EAMT Leuven

Wong, W., Liu, W., Bennamoun, M., 2007: Determining Termhood for Learning Domain Ontologies using Domain Prevalence and Tendency, Proc. Sixth AusDM 2007), Gold Coast, Australia. CRPIT 70