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

D5.6

Transfer Selection Support

Dissemination Level:	Public
Delivery Date:	June 15th 2012
Status – Version:	Final v1.0
Author(s) and Affiliation:	Gregor Thurmair, Vera Aleksić (Linguatec)

Relevant Documents

Panacea Deliverable D5.4: English-French and English-Greek bilingual dictionaries for the Environment and Labour Legislation domains

Panacea Deliverable D5.5: English-French and English-Greek bilingual dictionaries for the Environment and Labour Legislation domains

Panacea Deliverable D5.7: Sample of Transfer Entries produced

Panacea Deliverable 7.4: Evaluation Report (Third cycle)

Table of Contents

1. Introduction

The objective of the task 5.3, and the tool LT-XFR created to meet this challenge, is to find automatic means for transfer selection: If a source word has several translations then the right one for a given context must be found. This problem becomes the more important the larger the dictionary is, and occurs much more frequently than a case where there is no translation at all.

Of course, the problem is only relevant for 1:n transfers; if a source term has exactly one translation then there is no selection problem.

The basis of the investigation is a transfer dictionary; this is a bilingual and directed resource. The terminology used in the following section is:

- an *entry* (or *transfer*) is a combination of a source and a target term (defined by: <source-lemma, source-part-of-speech, target-lemma, target-part-of-speech>)
- a *package* is a set of entries with by a common source side (defined by: <source-lemma, source-part-of-speech>). A package consists of at least one entry; in the present case, only packages containing more than one entry are of interest. Packages differ depending on language direction, therefore bilingual lexicons are directed.

The context of the investigation is knowledge-driven (rule-based) MT. As for SMT, it should be noted that transfer selection fully depends on the presence of the translations in the training corpus; this fact makes transfer selection in SMT very much domain (or even trainings-text) dependent. It will be shown that even in large training sets, many entries do not occur at all.

Another difference is that the means to disambiguate transfers is on the *source* side, whereas in SMT it is on the *target* side (the target LM selects the best from a set of transfer options coming from the phrase table). So SMTs can better react to local contexts, for the cases where the transfers are in the training set.

The original title of the package was 'transfer rule creation', following the paradigm of rule-based MT. However, it turned out that the only 'rule' applied in this work is to look up the conceptual context, and the task is to provide significant context clues; therefore the title of the deliverable was selected broader than just focusing on the rule aspect.

1.1 Types of Transfer

Another distinction which is relevant here is the type of transfer to be considered. We distinguish between the following transfer types:

- **structural** transfer is a change in the target which is independent of the lexical material involved. Example: complex prenominal adjectives in German must be represented in English as relative clauses
- **lexical** transfer is dependent of the lexicon entries involved. Several cases exist here:
    - **Simple** lexical transfer is just a replacement of a word by its translation: (en) '*incineration*' -> (de) '*Einäscherung*'
    - **Complex** lexical transfer takes additional information to disambiguate, and performs tests to find the correct transfer.
        - *Local* transfer considers features on the *node* which must be transferred (e.g. number: (de) '*Schuld*' -> (en) '*guilt*' if singular but -> (en) '*debt*' if plural.
        - *Contextual* transfer inspects the context of the node to be transferred. This can be done on several levels: Lexical context [Frye 2012], syntactic

context (like transitivity) [Thurmair 1990], semantic context (e.g. (en) '*eat*' -> (de) '*essen*' for humans, -> (de) '*fressen*' for animals), pragmatic contexts (domain features, locale etc.) and others.

The present work deals with *conceptual* context, which is a form of lexical transfer based on lexical context, however not related to specific syntactic structures. It looks at concepts surrounding a translation candidate, and determines its transfer depending on such concepts. E.g. (en) '*interest*' -> (de) '*Zins*' in context of '*money*', '*pay*', '*loan*' etc. but (en) '*interest*' -> (de) '*Interesse*' in other contexts like '*sports*', '*activity*', '*research*' etc. The challenge is to find such contexts in an automatic way, using parallel corpus data.

1.3 Related Work

1. There are approaches of **word sense** disambiguation which use bilingual material [e.g. Agirre/Edmonds, ed., 2006]. However, word senses and translations do not go parallel; polysemous words like (de) '*Zelle*' transfer all their meanings into the target (en) '*cell*'. The goal of the current approach is not to disambiguate word senses but to find the best transfers[1].

2. There is significant work on automatic creation of transfer rules, recently cf. [Tyers et al. 2012]. However, they look at **close contexts** of the transfer candidates (windows of trigrams to pentagrams); however such windows are rather small, and do not always contain the relevant information for disambiguation; and there will be significant overhead in the rules once the lexicon gets bigger.

3. A similar approach of disambiguation of source language contexts was presented in [Thurmair 2005], called 'neural transfer' there. Only **monolingual** corpora were used there, disambiguation of contexts for translation candidates was done by manual annotation of training data, and the lookup context was extended from sentences to paragraphs; but very high accuracy could be reported. The current approach does automatic context disambiguation from parallel corpora, and uses only sentential contexts.

4. There are approaches to do disambiguation at the **target side**, not at the source side. This is the current paradigm in SMT [Koehn 2010], and also tried in METIS-II [Carl et al., 2008]. This approach must carry all possible transfers of all source words into the target, and then try to disambiguate there. This creates a massive overhead, which could be reduced by using some source-language information.

1.4 Approach

The approach taken here tries to model human intuition which, looking at the context words of a term, is able to determine how it should be translated. As this intuition works quite successfully for humans, it is tried to identify such conceptual context, based on parallel corpora.

The task takes two resources:

- a lexicon containing possible transfers of a given word; such a lexicon can e.g. result from a bilingual term extraction component as described in [Thurmair/Aleksić 2012], from legacy systems, or from other available data.
- a parallel corpus which allows to identify contexts for certain translations

It produces a resource (a corpus-based add-on to a transfer lexicon) which can be queried at

[1] A similar approach towards transfer can be found in [Brown et al. 1991], but they use just one contextual 'informant'.

runtime as an additional source of information. This resource is a static resource, logically independent of the MT system and can be used for both 'deep' and 'shallow' MT.

The task is executed in the following way:

- Take a bilingual dictionary, and identify the packages they contain; these packages are the target objects of the disambiguation effort.
- For all source and target lemmata in the packages, index the bilingual corpus for the sentences in which they both occur (on source and target side)
- For all translations of each source entry of each package, create subcorpora consisting of the sentence pairs containing the source lemma and the target lemma of this entry. This step will subdivide the monolingual source and target corpora into subsets of parallel sentences in which the source term has the same translation.
- Try to identify significant co-occurrences in the source language subcorpus which are specific for this translation. The goal is to be able to determine a cluster of source language words which indicates a certain transfer selection.

The result of the task will be a resource which, for each translation in a package, gives a vector of contexts which trigger the translation in question. At runtime, this resource will be queried, by matching the context of the source language candidates with all possible translations, and selecting the best matching cluster and its related translation.

The test would consist creating a test set containing sentence contexts with 'right' (reference) translations for the test terms, and in analysing these sentences and their context and comparing the transfer proposals with the transfers used by the reference.

The tool is called 'LT-Xfr', and has been developed here for German-to-English language direction.

2. The Lexicon

For the investigations, a dictionary was taken as it is used for human lookup but modified for machine processing. The LinguaDict lexicon [http://www.linguatecapps.com/linguadict] was selected in order to extend the coverage of transfers beyond normal MT lexicons, and have a realistic size of a bilingual lexicon. Compared to MT lexicons, lexicons for human lookup contain much more transfers, and give clues how to select the best transfer in a given situation. The challenge is to find such transfer disambiguation clues by corpus analysis.

	no. entries	no. SL terms	no. transfers/term
de->en	213.200	144.900	1.47
en->de	213.200	136.100	1.57

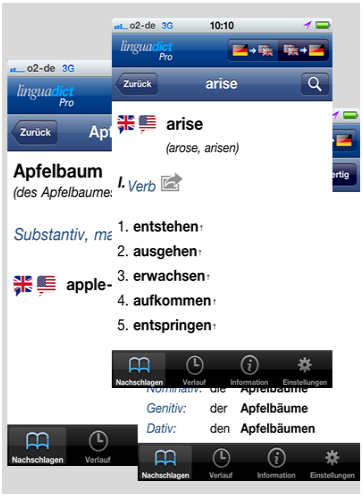
Tab. 2-1: Size of LinguaDict

As the transfer selection is directed, i.e. specific for a given language direction, the tests were made for German -> English.

2.1 The LinguaDict Lexicon

The LinguaDict lexicon in its German-English version is a state-of-the-art lookup resource; it is available both for online and offline (on mobile devices) lookup. It is a bilingual and directed lexicon; the language directions are built on the fly from a common data base.

It consists of single and multiword entries, and offers part-of-speech, gender and inflection information, and a pronunciation for most entries. Examples are given in Fig. 2-2.



en lemma	en POS	de lemma	de POS	en lemma	en POS	de lemma	de POS
constructional	Ad	konstruktiv	Ad	consular invoice	NoC	Konsulatsfaktura	NoC
constructional	Ad	Konstruktions-	NoC	consular invoice	NoC	Konsularrechnung	NoC
constructional unit	NoC	konstruktive Einheit	NoC	consulate	NoC	Konsulat	NoC
constructive	Ad	konstruktiv	Ad	consulate general	NoC	Generalkonsulat	NoC
constructively	Av	konstruktiv	Ad	consulship	NoC	Konsulat	NoC
constructivism	NoC	Konstruktivismus	NoC	consult	NoC	Beratung	NoC
constructor	NoC	Erbauer	NoC	consult	Vb	beraten	Vb
constructor	NoC	Erbauerin	NoC	consult	Vbi	richten	Vbr
constructor	NoC	Konstruktor	NoC	consult	Vbt	nachschlagen	Vbt
constructor's championship	NoC	Konstrukteurswertung	NoC	consult	Vbt	heranziehen	Vbt
construe	Vbt	analysieren	Vbt	consult	Vbt	hinzuziehen	Vbt
consubstantiation	NoC	Konsubstantiation	NoC	consult	Vbt	konsultieren	Vbt
consuetude	NoC	Gewohnheit	NoC	consult	Vbt	wenden	Vbt
consul	NoC	Konsul	NoC	consult	Vbt	besprechen	Vbt
consul general	NoC	Generalkonsul	NoC	consult	Vbt	einholen	Vbt
consul general	NoC	Generalkonsulin	NoC	consult	Vbt	zuziehen	Vbt
consular	Ad	konsularisch	Ad	consult	Vbt	sehen	Vbt
consular department	NoC	Konsularabteilung	NoC	consultancy	NoC	Beratung	NoC
				consultancy contract	NoC	Beratervertrag	NoC

Fig. 2-2: Example of LinguaDict entries: GUI (left), entry examples (right); a typical example for the transfer selection problem here is the entry for '*consult*'

Overall, the lexicon contains 212,000 entries (plus about 1000 entries without transfer (links for strong verbs etc.).

2.2 Lexicon Preparation

2.2.1 Preparatory Steps

The lexicon was prepared in the following way:

- All packages with only a single transfer were removed. For these packages, the problem of transfer selection does not exist. After this, 104.200 entries remained.
- All function word entries were removed, as they need a different type of transfer selection, and are much more interwoven with the MT system internals
- The treatment of all entries containing multiwords on either source or target side was postponed, to reduce the initial complexity of the task. They need to be integrated later.
- All entries containing part-of-speech changes were removed, as this is syntactic information: (de-adj) '*sicher*' -> (en-adj) '*secure*' and (en-adv) '*securely*'. Entries of this kind are in lexicons if the adverb formation is not completely regular. However, to determine the right part-of-speech selection, syntactic analysis is needed, which goes beyond the scope of the current Lt-Xfr work, and cannot easily be modelled by an approach for conceptual transfer

After these operations, 27.000 packages with 71.400 entries remained for the investigation. Table 2-3 gives the details on the lexicon used for the following analysis.

part of speech	no. packages	no. entries	no. transfers / entry
adjectives	6,900	18,200	2,83
nouns	15,600	35,400	2,27
verbs	4,500	17,800	3,26
total	27,000	71,400	2,63

Tab. 2-3: Packages in the lexicon

2.2.2 Lexicon Inspection

A short investigation of the lexicon entries reveals that conceptual transfer will never have full coverage, and a multitude of transfer selection strategies is required to do proper transfer, as many transfers will not be able to be disambiguated on a purely conceptual level:

- locale: (de) '*geschmack*' -> '*flavor*' (en-us) / '*flavour*' (en-uk));
- spelling: (en) '*adaptable*' ->: (de-old) '*anpaßbar*' and (de-new) '*anpassbar*'
- register: (en) '*anglophobe*' -> (de-lit) '*anglophob*' and (de-coll) '*englandfeindlich*';
  (en) '*adiposity*' -> (de-lit) '*Adipositas*' and (de-coll) '*Verfettung*'
- topic: (en) '*case*' -> (de-legal) '*Fall*' and (de-mechan) '*Gehäuse*'

The lexicon just provides the different alternatives in such cases; it is the task of the global system to identify which one to select. This is often done by user settings (locale, topic etc.), or automatic tools like topic identification, register and spelling selection etc. must be run.

As many of the cases just presented could be considered to be synonyms on a semantic level, these aspects are not considered in the following analysis; the focus here is on transfer selection based on conceptual contexts[2].

[2] It could have been an option to normalise such varaints before cluster building; this could have resulted in better clusters.

3. The Corpus

For the remaining packages of the lexicon, an automatic contextual disambiguation is tried. To do this, a parallel corpus is used. The goal is to find conceptual contexts in the corpus which allow the disambiguation of translation alternatives.

3.1 Corpus Collection

The corpus used for the LT-Xfr experiments consists of parallel sentences collected from different domains; details are given in Tab. 3-1:

Domain	no sentences
automotive	47,485
dgt	530,760
europarl	1,739,154
health&safety	57,155
jrc-acquis	1,239,731
e-books	82,635
statmt_dev	15,134
statmt_news	136,227
total	3,848,281

Tab. 3-1: Parallel corpus used (sentences)

Overall, 3.8 mio parallel sentences German-English were used for the experiment.

3.2 Corpus Processing

The corpus data were processed in the following way:

Step 1: Format conversion

All corpus sentences were converted into the PANACEA TO format: Text converted into UTF8, <s> tags were inserted with unique sentence-ids and language attribute. Errors in the original sentence segmentation were *not* corrected, as the sentences were already parallelised, and the sentence alignment could have been lost.

Step 2: Lemmatisation and tagging

All sentences of the corpus underwent lexical analysis, i.e. they were tokenised and lemmatised as described in [Thurmair et al. 2012]. Cases of homography, as far as related to content words, were disambiguated using a simple tagger. This step produced the <textform, lemma, POS> triples to work with later on.

Step 3: Monolingual Indexing

Each <lemma, POS> pair of the corpus which also occurred in the lexicon test set was

indexed (lemma -> sentence ids), and its frequency was computed[3]. The result were two index files (one for German, one for English), containing lemmata pointing to sentence ids.

Step 4: Bilingual indexing

The two index files were merged, such that: for each package: for each transfer, all common sentence-id's were collected into a bilingual index file; an example is given in fig. 3-2.

erläuterung_|_explanation	2	automotive-2685,automotive-2686
schloss_|_castle	1	automotive-33298
individualität_|_individuality	0	
befund_|_finding	0	
mischung_|_blend	2	automotive-2613,automotive-2614
wäschetrockner_|_tumble-drier	0	
vorgänger_|_predecessor	3	automotive-11700,automotive-21322,automotive-33984
stabilisierung_|_consolidation	0	
kasse_|_checkout	0	
straße_|_road	12	automotive-2524,automotive-7396,automotive-16426,automotive-27523,automotive-28250,automotive-28740,automotive-30381,automotive-37056,automotive-37122,automotive-38936,automotive-40052,automotive-42763
leidenschaft_|_passion	3	automotive-1763,automotive-45451,automotive-47047
aufwertung_|_revaluation	0	
stichwort_|_headword	0	
schiene_|_track	1	automotive-703
feind_|_enemy	0	
aufseher_|_warden	0	
stück_|_thing	0	
verleihung_|_renting	0	
tagesordnung_|_agenda	23	automotive-6690,automotive-11394,automotive-11684,automotive-17380,automotive-17381,automotive-17382,automotive-17383,automotive-17384,automotive-21465,automotive-28602,automotive-28603,automotive-28604,automotive-28605,automotive-28606,automotive-28607,automotive-32495,automotive-34783,automotive-38536,automotive-38537,automotive-38538,automotive-39174,automotive-39176,automotive-46898
echtheit_|_authenticity	0	
führung_|_captaincy	0	

Fig. 3-2: Example of bilingual index file (automotive corpus). many entries have no sentence in common, i.e. there is no evidence for this transfer in the corpus.

For many entries, no bilingual sentences could be found, mainly because there were no correspondences in the corpus. These entries had to be eliminated.

3.3 Subcorpus creation

From the bilingual index file, subcorpora were created in the following way:

Step 1: Corpus collection

For each transfer of a package, the common subset of sentence ids was identified, and the respective sentences were collected. As an output, one file per package was built, containing:

- For each package: the number of transfers for this package, and the sum of all sentences used therein
- For each transfer: the sentences in which it occurred (i.e. for the sentence-IDs oin fig.3-2, the real sentences were fetched.

[3] Indexing was not done on textforms but on lemmata.

This operation left 2.96 mio sentences which contained relevant lemmata.

Step 2: Word alignment

In order to avoid accidental co-occurrence of a SL-TL pair, the subcorpora were filtered using the criterion of word alignment: Only SL-TL word pairs which could be word-aligned were kept in the data[4]. For word alignment, GIZA++ was used. All sentence pair candidates which could not be word-aligned were removed from the subcorpora.

This operation removed another 280K sentences from the text base, leaving 2.68 mio sentences for the following steps. It would be worth looking at the difference; it could result either from real accidental co-occurrences, or from word alignment errors.

More importantly, this step also removed entries, and whole packages, for which no word alignment could be found, either because they did not co-occur in any sentence pair, or because they could not be word-aligned.

Table 3-3 shows the remaining data sets.

part of speech	original packages	after bilingual indexing	after word alignment
adjectives	6,900	4,670	1,240
nouns	15,600	11,360	3,690
verbs	4,500	3,930	1,680
Total	27,000	19,960	6,610

Tab. 3-3: Data sets (packages) available at the beginning, after bilingual indexing, and after word alignment.

It can be seen that only 6.600 packages out of 27.000 could be used for the experiment. So, even in a large parallel corpus, for only 25% of the entries, parallel data can be provided to try contextual transfer selection. As a consequence, additional means of transfer selection must be provided for a working system, beyond parallel-corpus-driven automatic extraction.

Step 3: Classification

The resulting subcorpora were classified according to the data which were available for disambiguation:

- class 1: all transfers of a package have at least five sentences where this transfer is used
- class 2: each transfer of a package has at least one sentence in which it occurs
- class 3: a package contains one or several transfers with no occurrence in any sentence pair.

Of the remaining subcorpora (one per package), only one third shows more than five sentences per transfer. Nouns are slightly better represented than verbs and adjectives, cf. Tab. 3-4

	adj	noun	vrb[5]	total
packages	1244	3693	1677	6614

[4] This was possible as the XFRlexicon contains only single word transfers.

[5] Note that verb statistics are not fully accurate as verbs with separated prefixes (‚*kommt … an*') are not correctly lemmatised

class 1	301	24.2%	1370	37.1%	447	26.6%	2118	32.0%
class 2	667	53.6%	1789	48.4%	859	51.2%	3315	50.1%
class 3	276	22.2%	534	14.5%	371	22.2%	1181	17.9%

Tab. 3-4: Distribution of corpus data for the different packages: Assignment to classes

From these subcorpora, about 1000 sentences of class 1 were subtracted to be used as a test set; the rest was used for training.

4. Creation of the Lt-Xfr Lexicons

The analysis of the package coverage showed that sufficiently many contexts would only be available for one third of the translation entries resulting from subcorpus collection. To provide disambiguation means for the other entries, *additional* information had to be provided.

Therefore a strategy was adopted which is based on *two* kinds of information:

- conceptual context clusters, as the original approach suggested. These data are collected in a *conceptual* lexicon (ConcLex);
- a translation based on frequency information as a fallback: In case no cluster is available, different probability measures are used for transfer selection. The probabilities are collected in a *probability* lexicon (ProbLex).

Both lexicons are consulted at runtime, in sequential order.

4.1 Conceptual Lexicon

The conceptual lexicon is created by analysing the subcorpus attached to each entry for co-occurrences: All lemmata of all sentences of the subcorpus are compared, and the ones with the best co-occurrence score are taken. Experiments to restrict the resulting clusters to a certain size, or to use a threshold, showed that the data sparsity requires to basically leave *all* candidates in the clusters.

Also, lemmata co-occurring with *several* transfers of a package were not eliminated but left in the clusters, as they could still help to disambiguate from other translation candidates, and would leave many transfers with very few contexts.

In addition, experiments to include the *distance* of the co-occurring word to the lemma weight as an additional precision measure have been postponed; their influence would only be marginal[6].

The output of the component which builds the conceptual lexicon is the lexicon itself (ConcLex). It gives, for each source language package defined by <lemma, part-of-speech>, a list of translations, consisting of: the translation, its part-of-speech, and an optional cluster of variable size, consisting of pairs of <sourcelanguage-lemma, weight>, the weight giving the strength of the co-occurrence.

Such a cluster can be matched to the context lemmata of an input sentence at runtime, and their similarity can be computed.

An example is given in fig. 4-1.

In case a translation has no example sentences, the conceptual cluster for this translation must be left empty. To be able to still include them in the transfer selection, a fallback strategy was implemented using probabilities.

[6] An exception may be certain prepositions indicating strict subcategorisation; to be investigated.

füllung	No	fill	No	mittelwert|0.8, abweichung|0.8, jeweilig|0.6, zulässig|0.4, maximal|0.4, gemäß|0.4, füllung|0.4, last|0.4, gerät|0.4, ernähren|0.2, jahr|0.2, afrikaner|0.2, reichen|0.2, benötigen|0.2, ethanol|0.2, suv-tank|0.2, menge|0.2, getreide|0.2, (1)|0.2, klasse|0.2
füllung	No	stuffing	No	
füllung	No	stopping	No	
füllung	No	filling	No	
fürsorge	No	care	No	kind|0.28, mensch|0.16, schutz|0.11, behandlung|0.10, medizinisch|0.14, neu|0.06, brauchen|0.08, bereich|0.06, europäisch|0.05, aids|0.05, art|0.09, dienstleistung|0.05, sozial|0.06, leben|0.08, familie|0.06, bildung|0.05, frau|0.09, zugang|0.06, liebe|0.06, erhalt|0.05
fürsorge	No	welfare	No	sozial|0.47, öffentlich|0.31, gefahr|0.26, gut|0.15, mensch|0.15, europäisch|0.15, finden|0.15, ausgabe|0.10, sozialwissenschaft|0.10, investition|0.10, pflege|0.10, anzahl|0.10, weg|0.10, meinen|0.10, geben|0.10, rahmen|0.10, gemeinsam|0.10, verfahren|0.10, handeln|0.10, psychologisch|0.10
fürsprecher	No	intercessor	No	
fürsprecher	No	booster	No	entwickeln|1.0, markt|1.0, stark|1.0, usa|1.0, verwerten|1.0, erhalt|1.0, chance|1.0, unternehmen|1.0, mittlere|1.0, klein|1.0, machen|1.0, zugänglich|1.0, kosten|1.0, geringfügig|1.0, information|2.0, fungieren|1.0, customer|1.0, launching|1.0, staat|1.0
fürsprecher	No	advocate	No	herr|0.35, gut|0.23, präsident|0.17, uribe|0.11, woche|0.11, verwenden|0.11, direkt|0.11, persönlich|0.11, unterstützung|0.11, wichtig|0.11, us-parlamentarier|0.11, bemühen|0.11, zweifellos|0.11, rating-agentur|0.11, kommissar|0.11, bericht|0.11, ganz|0.11, groß|0.11, weg|0.05, dritte|0.05
fürst	No	prince	No	vermeiden|0.33, machiavelli|0.33, hass|0.33, metternich|0.33, krieg|0.16, napoleonisch|0.16, heilige|0.16, allianz|0.16, schöpfung|0.16, österreicher|0.16, umstand|0.16, machen|0.16, deutlich|0.16, sorgsam|0.16, klar|0.16, darstellen|0.16, begegnen|0.16, person|0.16, weise|0.16, konstruktiv|0.16
fürst	No	ruler	No	land|1.0, eigen|1.0, feind|1.0, volk|1.0, vernachlässigung|1.0, machtgier|1.0, machen|1.0, amt|1.0, kehren|1.0, schah|1.0, autoritär|1.0
fütterung	No	feeding	No	
fütterung	No	feed	No	bestimmt|0.34, mischfuttermittel|0.26, tier|0.21, mischung|0.21, futtermittel-ausgangserzeugnis|0.24, zusatzstoff|0.21, ergänzungsfuttermittel|0.21, tierernährung|0.24, verwenden|0.12, rückstand|0.12, gemäß|0.07, artikel|0.07, erzeugnis|0.07, groß|0.07, hoch|0.09, tierpflege|0.07, melken|0.07, futterbereitung|0.07, ölkuchen|0.07, fest|0.07
gabel	No	fork	No	messer|0.43, essen|0.26, schieben|0.16, mund|0.1, zerdrücken|0.1, legen|0.1, tisch|0.1, zinken|0.1, soup|0.06, nehmen|0.06, kochen|0.06, schneiden|0.06, führen|0.06, passiersieb|0.06, gemüse|0.06, hoch|0.06, erbse|0.06, halten|0.06, hand|0.06, zeit|0.06
gabel	No	cradle	No	knallen|1.0, hörer|1.0

Fig. 4-1: Example of conceptual lexicon

4.2 Probability Lexicon

In case the conceptual transfer does not lead to a result (and this case is rather likely given the amount of transfers without any context because there were no sentences), a fallback strategy is created, which consists in computing a translation probability score.

Previous experiments in the creation of the LinguaDict lexicon have shown that simply using the (target monolingual) corpus frequency of a translation is not the best option: We want to know how often the target lemma occurs *as translation of a given source lemma*. Otherwise target lemmata which are very frequent (like '*be'* or '*have'*) disturb the transfer selection.

Also, a relevant factor is *for how many words* a given target lemma is a translation: If a target lemma has high frequency as translation of only one source word, then this is much more important than if the frequency results from the fact that it is the translation of e.g. five source words.

Therefore a more complex approach than simple target corpus frequency was taken: The translation probability consists of *three* scores, differing in the reference used to compute them. These scores are:

- **Package probability**: probability of a given translation related to the other translations *of this package*. Number of sentences for the given translation DIV total number of sentences in this package. (0 for all transfers without a sentence in the subcorpus);
- **Target probability**: probability of a given translation related to *other source terms* (i.e. for how many SL lemmas is this a possible transfer?) (0 for all terms which are in no package)
- **Corpus probability**: probability that this translation is used at all in the *target language*. Number of occurrences of TL lemma DIV number of lemmata in the total TL corpus.

Querying of the probability lexicon is done sequentially, i.e. if a score is zero then the next 'weaker' score is taken. Finally, a corpus probability is nearly always available[7].

[7] Only in this particular setup. Note that many entries of LinguaDict do *not* have any reference in the corpus.

The format of the probability lexicon is a tuple of <source lemma, source-pos, tl-lemma, tl-pos, package prob, target-prob, corpus-prob>. An example is given in fig. 4-2.

fülle	No	abundance	No	0.21951219512195122	8.586079335909694E-6	2.2571946085091414E-6
füllung	No	fill	No	1.0	1.3415748962358896E-6	6.17606865221863E-6
füllung	No	stuffing	No	0.0	0.0	0.0
füllung	No	stopping	No	0.0	0.0	0.0
füllung	No	filling	No	0.0	0.0	1.6328641848789535E-7
fürsorge	No	care	No	0.8207547169811321	3.6195690700444304E-4	8.769441181144026E-5
fürsorge	No	welfare	No	0.1792452830188679	3.702746713611055E-5	4.705530377483525E-5
fürsprecher	No	intercessor	No	0.0	0.0	9.605083440464432E-9
fürsprecher	No	booster	No	0.05555555555555555	2.6831497924717793E-7	7.972219255585478E-7
fürsprecher	No	advocate	No	0.9444444444444444	3.622252219836902E-5	2.2120507163389588E-5
fürst	No	prince	No	0.8571428571428571	1.6098898754830676E-6	2.1611437741044973E-6
fürst	No	ruler	No	0.14285714285714285	8.854394315156871E-6	4.293472297887601E-6
fütterung	No	feeding	No	0.0	0.0	0.0
fütterung	No	feed	No	1.0	4.454028655503154E-5	5.629539404456204E-5
gabel	No	fork	No	0.967741935483871	8.049449377415339E-6	1.2486608472603762E-6
gabel	No	cradle	No	0.03225806451612903	2.6831497924717793E-7	1.3831320154268783E-6
galgen	No	gallows	No	0.8888888888888888	2.1465198339774235E-6	1.7289150192835978E-7

Fig. 4-2: Example of the probabilistic lexicon[8]

These two lexicons (ConcLex and ProbLex) are used for the test and evaluation. The challenge is to determine the transfer of a source-lemma based on the context in which it occurs.

[8] Words without even a corpus probability could be due to lemmatisation / tagging errors

5. Test and Evaluation

The transfer selection component is tested by determining the transfer of a test lemma in a given sentence context, and comparing it with the one of a reference translation. In the best case, all translations proposed by the Lt-Xfr component are identical with the transfers selected in the reference translations.

As the LinguaDict lexicon contains many near translations, which can hardly be distinguished on the basis of conceptual transfer, a special evaluation procedure was adopted, consisting of three ranks instead of a binary decision:

- **Rank 1**: the translation proposed by the system is *identical* to the one in the test reference sentence
- **Rank 2**: the proposed translation close / *synonym* to the one in the test reference sentence. This was decided to be the case if
  - the proposed translation belongs to the same WordNet synset as the reference
  - the proposed translation is orthographically similar to the reference (like: '*electric'* vs. '*electrical'*, '*agglutinating'* vs. ‚*agglutinative*', '*dialogue'* (UK) vs. '*dialog'* (US) etc.
- **Rank 3**: the two translations are (still) different.

Evaluation would allow rank1 and rank2, and reject rank3 results.

Based on the three ranks, a simple scoring system is used (rank1 = 1, rank2 = 2, rank3 = 3) to compute an overall score: The lower the score the closer the translation is to the reference.

5.1 Test data

5.1.1 Test corpus

The test corpus was taken from the subcorpora used for the research (cf. section 3.3 above). 1044 sentences were extracted, containing transfers for nouns, verbs, and adjectives.

5.1.2 Resources for ranking

For ranking (esp. rank2: similarity), two additional resources were produced:

- an indexed version of WordNet V3, whereby for a given input lemma a list of possible synonyms was retrieved (i.e. the synset lemmata[9]).
- a resource for orthographic similarity. For all parts of speech, a resource was used which unifies US and UK spelling (This list contains about 4,700 entries). For adjectives, additional patterns were considered, like 'adj + -ed' ('*abstract'* vs. '*abstracted'*), 'adj-ic + al' ('*acoustic'* vs. '*acoustical'*) etc.

The test frame applies pattern matching for the strings, and simple lookup for the differences in locale.

5.1.3 Test frame

It was not possible with the available resources to integrate the Lt-Xfr component into a complete MT system. Therefore a special test system was written which has a translation candidate (source lemma) and a sentence context as an input, and returns the 'best matching'

[9] As the test lexicon contains only single words, also only the single words of the synsets were taken.

transfer (target lemma). This return lemma can be compared to the reference translation, and ranked: In case they are not identical, it can be checked if they are both in the same WordNet synset, or are orthographically similar (rank 2).

5.2 Test systems

Two test systems were built:

- one with the full component (called *Lt-Xfr* below), with all options produced, and both the conceptual and the probability lexicon
- one with only the fallback (called *Lt-Xfr-frq* below), using the probability lexicon but not the conceptual lexicon; this is relevant in cases where no conceptual context information would be available.

For comparison, the test sentences were also given as input to several available MT systems, both with statistical and rule-based architecture. Their translations of the test lemmata were extracted, and also ranked according to the three ranks chosen (also using the synset and the orthographic similarity).

5.3 Test results

First, the output of the two Lt-Xfr systems was evaluated against the reference translation (absolute evaluation), and then it was compared to the output of the other MT systems (comparative evaluation). Results are shown in Table 5-1.

5.2.1 Absolute Evaluation

For this evaluation, the test sentences were analysed with the LT-Xfr frame, and the resulting transfer was compared to the reference translation. As explained, this procedure was done for two system variants:

- One which takes both conceptual and probability lexicon (Lt-Xfr)
- One which searches transfers only based on probability information (Lt-Xfr-frq)

It can be seen that 60% of the test terms are correctly translated (rank 1), and if WordNet and string-similarity synonyms are taken into account, then 75% of the test sentences return a correct transfer. The values are kind of similar for all parts-of-speech, with verbs doing a bit better than the other parts of speech.

As a result, if a random selection of transfers is assumed as a baseline (with about 41% correctness), then the Lt-Xfr improves over the baseline by absolute 34%, and relative 83%; improvement is most significant for verbs (with more than 100% relative). For the fallback system (only frequency-based), the improvement is still 25.6% absolute, and 61.6% relative.

5.2.2 Comparative Evaluation

In order to have an impression how the result is compared to the state of the art, the test sentences were translated with several available MT systems, to have an impression how useful they would be. The systems selected for comparison were one SMT and four RMT systems. The test sentences were translated, and the translations for the test words were identified and compared to the reference translation. Like for the absolute evaluation, total (rank1) and partial (rank2) identity were computed, as well as the overall scores. Tab. 5-1 shows the evaluation result.

It can be seen that the LT-Xfr system clearly shows the best performance of all systems in all categories. It has much better scores than all RMT systems, and also better scores than Google. It is absolute 20% better than the least-performing MT system, and still 7% better

than the best-performing one. Even the fallback frequency-based (LT-Xfr-freq) version outperforms all RMT systems, and is better than Google in three of six categories (Verbs1, Verbs/1+2, Adj/1+2).

		Google		RMT1		RMT2		RMT3		RMT4		LtXFR		LtXFR-frq	
		sent	in %	sent	in %	sent	in %	sent	in %	sent	in %	sent	in %	sent	in %
Nouns	694														
	rank1	384	55,3	279	40,2	307	44,2	263	37,9	264	38,0	425	61,2	342	49,3
	rank2	97	14,0	122	17,6	119	17,1	121	17,4	123	17,7	97	14,0	121	17,4
	rank3	213	30,7	293	42,2	268	38,6	310	44,7	307	44,2	172	24,8	231	33,3
	rank1+2	481	69,3	401	57,8	426	61,4	384	55,3	387	55,8	522	75,2	463	66,7
Adjectives	145														
	rank1	77	53,1	62	42,8	58	40,0	58	40,0	54	37,2	85	58,6	72	49,7
	rank2	16	11,0	19	13,1	21	14,5	29	20,0	23	15,9	19	13,1	24	16,6
	rank3	52	35,9	64	44,1	66	45,5	58	40,0	68	46,9	41	28,3	49	33,8
	rank1+2	93	64,1	81	55,9	79	54,5	87	60,0	77	53,1	104	71,7	96	66,2
Verbs	204														
	rank1	97	47,5	92	45,1	91	44,6	69	33,8	79	38,7	125	61,3	103	50,5
	rank2	38	18,6	37	18,1	43	21,1	54	26,5	52	25,5	37	18,1	36	17,6
	rank3	69	33,8	75	36,8	70	34,3	81	39,7	73	35,8	42	20,6	65	31,9
	rank1+2	135	66,2	129	63,2	134	65,7	123	60,3	131	64,2	162	79,4	139	68,1
Total	1043														
	rank1	558	53,5	433	41,5	456	43,7	390	37,4	397	38,1	635	60,9	517	49,6
	rank2	151	14,5	178	17,1	183	17,5	204	19,6	198	19,0	153	14,7	181	17,4
	rank3	334	32,0	432	41,4	404	38,7	449	43,0	448	43,0	255	24,4	345	33,1
	rank1+2	709	68,0	611	58,6	639	61,3	594	57,0	595	57,0	788	75,6	698	66,9
scores															
	Nouns	1,75		2,02		1,94		2,07		2,06		1,64		1,84	
	Verbs	1,86		1,92		1,90		2,06		1,97		1,59		1,81	
	Adj's	1,83		2,01		2,06		2,00		2,10		1,70		1,84	
	Total	1,81		1,98		1,97		2,04		2,04		1,64		1,83	

Tab. 5-1: Evaluation results, compared to the reference. Number sentences, ranks (sentences, percentage), per part of speech, total, and score, for all systems.

However the result shows that significant improvement in transfer selection can be achieved with the techniques used by LT-Xfr, compared to the state-of-the-art of MT systems.

More detailed information on the evaluation is given in PANACEA deliverable D7-4.

6. PANACEA Integration

It was not possible, due to the lack of resources, to provide a full workflow how to build the two lexicons for additional language directions:

- input data, consisting of large bilingual transfer lexicons, and of large parallel corpora, must be provided
- the tools in the processing chain must be streamlined, and brought into a better sequence; exploratory steps can be skipped.

However, to demonstrate the scope of the tool, the test frame was made available as a web service in the PANACEA registry.

6.1 Formats

The service is available as a web service in the PANACEA registry, called:

http://80.190.143.163/panaceaV2/services/LTXfr?wsdl

Parameters are:

- source language (only ‘de’ is supported)
- target language (only ‘en’ is supported)
- text (a string containing an URL pointing to the input file)

The text must be an UTF8 file, and have a 3-column layout:

- source language lemma (in normalised form, incl. lowercasing); only single word lemmata are supported so far
- source lemma part-of-speech (No for noun, Vb for verb, Ad for adjectives)
- context sentence (containing the lemma to be translated)

The format of the resulting file is the same, with an additional column inserted which contains the translation which the system proposed for this particular sentence context.

An example is given in fig. 6-1.

input.xlsx

	A	B	C	D	E	F	G
1	blatt	No	In unserer grünen Anlage sind schon alle Blätter abgefallen.				
2	gang	No	Dort steht ein Haus mit einem schmalen Gang und zwei kleinen Zimmern.				
3	paar	No	Dieses wurde neulich von einem jungen Paar als ihr Sommerhaus bezogen.				
4	laster	No	Die Möbel wurden auf zwei Lastern gebracht, welche schwer beladen waren.				
5	gang	No	Gestern hatte das Pärchen zum Abendessen ein Menü mit fünf Gängen.				
6	gericht	No	Es enthielt Pfeffer, Salz und andere Gewürze.				
7	linse	No	Als Erstes wurden				
8	mandel	No	Als Nachtisch hatte				
9	bremse	No	Das Einzige, was di				
10	bremse	No	Eine Bremse flog s				
11	pony	No	Eine Bremse flog s				
12	ader	No	Da ihr Mann eine k				
13	geschick	No	Da ihr Mann eine k				
14	blatt	No	Das Photo soll am				
15	absatz	No	Die Frau sah gut au				
16	ausschnitt	No	Die Frau sah gut au				
17	messe	No	Das Paar kommt so				
18	gericht	No	Danach gehen sie				
19							
20							
21							
22							
23							
24							
25							
26							

output.xlsx

	A	B	C	D
1	blatt	No	sheet	In unserer grünen Anlage sind schon alle Blätter abgefallen.
2	gang	No	aisle	Dort steht ein Haus mit einem schmalen Gang und zwei kleinen Zimmern.
3	paar	No	couple	Dieses wurde neulich von einem jungen Paar als ihr Sommerhaus bezogen.
4	laster	No	truck	Die Möbel wurden auf zwei Lastern gebracht, welche schwer beladen waren.
5	gang	No	gear	Gestern hatte das Pärchen zum Abendessen ein Menü mit fünf Gängen.
6	gericht	No	court	Es enthielt Pfeffer, Salz und andere Gewürze.
7	linse	No	lentil	Als Erstes wurden gekochte Linsen aufgetischt.
8	mandel	No		Als Nachtisch hatten sie ein Gericht mit Mandeln.
9	bremse	No		Das Einzige, was die jungen Leute störte, war das laute Fliegen und Summen der Bremsen, Mücken und Hornissen.
10	bremse	No		Eine Bremse flog sogar ins Gesicht der jungen Frau, dann flog sie in ihre Haare und dort blieb sie auf dem Pony neben der Augenbraue stehen.
11	pony	No	fringe	Eine Bremse flog sogar ins Gesicht der jungen Frau, dann flog sie in ihre Haare und dort blieb sie auf dem Pony neben der Augenbraue stehen.
12	ader	No	conductor	Da ihr Mann eine künstlerische Ader und sehr viel Geschick hat, machte er schnell eine Aufnahme.
13	geschick	No	skill	Da ihr Mann eine künstlerische Ader und sehr viel Geschick hat, machte er schnell eine Aufnahme.
14	blatt	No	sheet	Das Photo soll am nächsten Sonntag im örtlichen Blatt erscheinen.
15	absatz	No	market	Die Frau sah gut aus, mit ihrem kurzen Rock, den hohen Absätzen und dem rosafarbenen Pulli mit dem unwiderstehlichen Ausschnitt.
16	ausschnitt	No	part	Die Frau sah gut aus, mit ihrem kurzen Rock, den hohen Absätzen und dem rosafarbenen Pulli mit dem unwiderstehlichen Ausschnitt.
17	messe	No	mass	Das Paar kommt sonntags immer in die Kirche zur katholischen Messe.
18	gericht	No	food	Danach gehen sie mit dem Anwalt und dem Richter ins Gericht.

Fig. 6-1: Example of input and output of the LT-Xfr lookup service

7. Assessment

7.1 Relevance

The transfer strategy presented here is just one of possible transfer strategies; others are transfer selection based on external information (topic, locale etc., which are passed to the transfer selection component by external features), or based on morphosyntactic content.

The approach presented here shows the following features:

- It fits to the architecture of rule-based system inasmuch as it provides transfer selection on the source side, not on the target side, and controls the transfer selection strategies for such systems.
- It can be used as additional information source, as it provides a static resource which can easily be linked to a system: Most MT systems have an internal structure for their transfer packages, like a sequence of tests, and there is always a 'default' translation in cases where all tests fail. This could be the place where the current resource could successfully be used, i.e. for cases where no system relevant information is available.
- As it relates to conceptual contexts, and is not linked to a particular syntactic structure or configuration, it is more robust than current selection strategies, which usually fail in cases where the required syntactic structure is not built (e.g. due to a parse failure). So it could be used as a fallback in cases where the analysis component returns improper results
- For the same reason, the approach is independent of the specific system structure, the type of analysis results, syntactic structures etc.; it can support shallow MT systems just as well as all kinds of deep RMT.

It simply leads additional information into the transfer selection process which is not used up to now.

7.2 Quality

The quality of the component crucially depends on the quality of the match between the text context and the clusters of the conceptual lexicon.

1. One option to improve the matching is to extend the context from sentences to paragraphs; this step has been taken in [Thurmair 2005] and improves transfer quality to a level of 96% accuracy. However, most of the parallel data available today are aligned on sentence level, not on paragraph level, so such an approach would be difficult to train.

2. Another option is to review the clusters. A look at the current clusters shows that there are lemmata which would be considered to be irrelevant for transfer selection, and that other lemmata are missing which would be expected here.

```
blatt   No      leaf    No
zweig|0.28      clémentine|0.28 de|0.28
corse|0.28      bringen|0.28    patentieren|0.28
meer|0.28       mitte|0.28      zypern|0.28
insel|0.28      aphrodite|0.28  goldgrün|0.28
entfernt|0.14   fruchtholz|0.14 belichten|0.14
schlecht|0.14   befindlich|0.14 ausdünnen|0.14
tragend|0.14    frucht|0.14
```

Fig. 7-1: Cluster for *blatt -> leaf*

In fig. 7-1, the translation of (de) '*blatt*' -> (en) '*leaf*' (as opposed to '*sheet*' or '*newspaper*')

would be corroborated by '*zweig'*, '*frucht'*, and also missing terms like '*ast'* or '*blüte'*, whereas '*patentieren'*, '*zypern'* etc. would not really contribute to the disambiguation of this reading.

Additional missing concepts could be collected by doing monolingual correlation analysis, and add lemmata which are highly correlated with the terms of a given seed cluster. Such a strategy could provide good additional terms

3. Clusters suffer from data sparsity, and the more so the less frequent the translations are: Many transfers in the conceptual lexicon simply have no conceptual context information at all. If there is no seed cluster there is no monolingual extension either.

4. Therefore, to improve quality, an option must be foreseen to have the conceptual lexicon edited by human coders: They should be able to add / remove terms to improve the cluster accuracy, and adapt the transfer to specific types of texts, contexts, or other needs. Human editing would require a review of the current scoring mechanism, to be changed e.g. into a simple three-level score (very relevant – relevant – somewhat relevant), and the lookup would have to be adapted accordingly.

5. Transfer selection on the target side, as done in SMT, has only one advantage: In cases of idiomatic expressions, created by an idiosyncratic combination of two target words, can easier be solved on the target side. However, such expressions can easily be added to the transfer lexicon (as they have to be added to the training data on the other side); they would not even create ambiguities in transfer selection.

7.3 Extensions

To stabilise the results of the current investigation, the following items must be considered:

1. The analysis used only a subset of the lexicon; multiword entries and entries with changes in the part of speech were not considered.

- Adding multiwords requires more sophistication in the step where word alignment is required; GIZA++ would not be the appropriate tool here anymore, and full MOSES phrase alignment may be necessary.
- As for part-of-speech changes, the most frequent case in German->English is that adjectives used adverbially must be translated as adverbs. Care has been taken that the part of speech is given to the lookup as one of the input parameters; this can help in transfer selection.

2. The processing chain needs to be stabilised to be able to run the component with other lexicon and corpus data.

3. The coverage must be extended to other language directions. This would require large sentence-aligned bilingual corpora, a bilingual lexicon created e.g. with the Lt-P2G tool[10] from such a corpus, and language resources for analysis, esp. lemmatiser and tagger information.

Optional, for ranking and evaluation, a WordNet and a ortho-similarity resource could be used for the target language.

4. The cluster building itself could be improved by collapsing transfers which are clearly synonyms, or variants of each other (e.g. locale), before the analysis rather than afterwards in a step of ranking. This would provide more data for such readings in clustering.

[10] see task 5.2 of PANACEA, or [Thurmair/Aleksić 2012]

8 Citations

- Agirre, E., Edmonds, Ph., eds., 2006: Word Sense Disambiguation. Springer
- Brown, P., Della Pietra, St., Della Pietra, V., Mercer, R., 1991: Word-sense disambiguation using statistical methods. Proc. 29th ACL
- Carl, M., Melero, M., Badia, T., Vandeghinste, V., Dirix, P., Schuurman, i., Markantonatou, St., Sofianopoulos, S., Vassiliou, M., Yannoutsou, O:, 2008: METIS-II: low resource machine translation. in: Machine Translation 22, 1-2, p.67-99
- Koehn, Ph., 2010: Statistical Machine Translation. Cambridge Univ. Press
- Santos, D., 2000: The translation network, A model for a fine-grained description of translaitons. In: Véronis, ed.: Parallel Text Processing. Kluwer.
- Thurmair, Gr., 1990: Complex lexical transfer in METAL. Proc. 3rd TMI Conf., Austin, Tx
- Thurmair, Gr., 2005: Improving Machine Translation Quality. Proc. MT Summit X, Phuket
- Thurmair, Gr., Aleksić, V., 2012: Creating term and lexicon entries from phrase tables. Proc. EAMT, Trento
- Thurmair, Gr., Aleksić, V., Schwarz, Chr., 2012: Large scale lexical analysis. Proc. LREC, Istanbul
- Tyers, F.M., Sánchez-Mártinez, F., Forcada, M.L., 2012: Flexible finite-state lexical selection for rule-based machine translation: Proc EAMT Trento