# An analysis of translation divergence patterns using PanLex translation pairs

Francesca Gola

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2012

Committee:
Emily M. Bender
Scott Farrar

Program Authorized to Offer Degree:
Department of Linguistics, Computational Linguistics

University of Washington

**Abstract**

An analysis of translation divergence patterns using
PanLex translation pairs

Francesca Gola

Chair of the Supervisory Committee:
Associate Professor Emily M. Bender
Department of Linguistics

This analysis was performed to understand the patterns of translation divergences occurring in high and low frequency verbs, and to test the hypothesis that high frequency verbs are more prone to translation divergences than low frequency ones. Four types of divergences were considered: Thematic, Conflational, Categorial, and Structural (Dorr, 1990), with samples from three language pairs: Italian to French, Italian to English and English to Thai.
The analysis is also an evaluation of the possibility of using the online multilingual dictionary PanLex (Baldwin et al., 2010) to automatically derive transfer rules, as part of a larger effort to create a machine translation system based on customizable language-specific grammars for both source and target languages, using semantic representations in the format of Minimal Recursion Semantics, or MRS, (Copestake et al. 2005) as the input and output of the transfer stage.
Based on the samples analyzed, this evaluation suggests that manual transfer rules creation and tweaking of automatic rules would be most needed for high frequency verbs, while low frequency verbs seem likely to have a lower translation divergence error rate.

# TABLE OF CONTENTS

Page

## LIST OF FIGURES

# LIST OF TABLES

## ACKNOWLEDGMENTS

My sincerest thanks go to my adviser, Emily Bender, for her exceptional guidance and insight through all the stages of this work. Her tireless dedication, generosity, and patience made it all possible.

Thank you to Scott Farrar, for graciously accepting to help and for his precious comments on the draft.

Jonathan Pool made this research possible by providing all the PanLex material, and I am very grateful for that.

I am also grateful to Glenn Slayden, Nuttanart Facundes, Jiradech Kongthon and Natjaree Chutikul, whose help was essential to complete the Thai portion of this research.

Thank you to Nicola, for his quiet strength.

# Chapter 1
# INTRODUCTION

## *1.1 Overview*

Translating often results in text very different from the source language. Such divergences can be observed, for instance, when analyzing translations produced by current machine translation (MT) technologies: Rule-based (RBMT) and Corpus-based (statistical or example based MT).

Rule-based systems, such as previous generations of Systran[1] and Promt[2], rely on grammar rules and dictionaries, and are capable of achieving very high terminology consistency, especially in particular terminology domains. However, these systems are very labor intensive, as they require a considerable amount of manual work to create or validate rules. Statistical Machine Translation (SMT) systems, like Google Translate[3], on the other hand, are less labor intensive, and are trained on large quantities of text.

The example of the *black cat* (e.g. Birch, 2011) is often used to compare rule-based and statistical approaches. When translating *the black cat* in Italian, the RB method will have a rule prescribing that the adjective follows the noun, and will produce *il gatto nero*. Google Translate achieves the same result for this particular sentence, because this is a high frequency combination. However, when feeding a rather unusual combination, for instance *the magenta cat*, the result will be different. A RB system will still return a perfectly adequate result *il gatto magenta*, because it uses the same grammar rule. On the other hand, Google Translate, as of May

[1] http://www.systransoft.com/
[2] http://www.online-translator.com
[3] http://translate.google.com/#

16th 2012, translated it as *il magenta cat*, as it had not seen this segment in training, and therefore proceeded to translate it word-by-word.

The analysis presented here can be viewed as an evaluation of the possibility of using the online multilingual dictionary PanLex (Baldwin et al., 2010) to derive transfer rules in a system based on customizable language-specific grammars for both source and target languages, using semantic representations in the format of Minimal Recursion Semantics, or MRS, (Copestake et al. 2005) as the input and output of the transfer stage. In this type of system, the MT part is taken care of by transfer rules, which map between source language and target language MRSs, and which could be, in part, automatically acquired from bilingual dictionaries. This was done for the LOGON Machine Translation system, which initially used a small corpus of tourist brochures translated in English and Norwegian (Lønning et al., 2004).

When creating transfer rules automatically from bilingual dictionaries, the size of the dictionaries determines the coverage of the MT. For this reason, if transfer rules were created using a dictionary with an enormous amount of languages and translation pairs, such as PanLex (Baldwin et al., 2010), it would be possible to have a very broad coverage for many language pairs. The problem would come from the fact that naïve transfer rules, as created automatically from a bilingual dictionary, would incur translation divergences, which may require tuning to take care of. Since modifying transfer rules is very labor intensive, it would be desirable to know beforehand where the focus of such modification efforts should be. Analyzing the sentences created from verb translation pairs found in PanLex might help determine if a system could be made to work effectively with PanLex, by focusing the efforts only in the areas that need it.

Four types of divergences were considered: Thematic, Conflational, Categorial, and Structural. These divergences can sometimes be solved at MRS level, or by the naïve transfer rules created from the bilingual dictionary. Initial observation of some PanLex verb translation samples led us to formulate the hypothesis that low frequency verbs have less divergence than high frequency ones. If this is true, we can say we have identified a focus for rule tweaking – we will need to attribute less reliability to translations of high frequency verbs, and put in place more manual checks for them that help ensure better translation. Since low frequency verbs make up a larger proportion of the lexicon than the high frequency ones, this would also mean that a naïve PanLex transfer rules approach might work without too many problems.

This experiment was conducted with two sets of translation verb pairs extracted from PanLex: Italian to English and Italian to French. A third set, English to Thai, is different in that it starts from the English sentences obtained from the initial PanLex translations, and therefore was not directly extracted from PanLex. Starting from the verbs extracted, sentences were created in

Italian and then translated in each target language. The sentences were then analyzed for divergences, and the divergence patterns where finally examined for high and low frequency verbs. The results obtained for Italian to English and Italian to French seemed to identify similar patterns, and suggest that the hypothesis could be right. The results for the third language pair, English to Thai, were used to test if these results could be generalized to a language very different from the previous three. Thai, of the Tai-Kadai language family, historically had a limited amount of borrowing from English or other Indo-European languages, since the country was never colonized and no language was imposed, even though the trend has changed and English terminology has permeated the language, especially in the fields of business, advertisement and technology (Kapper 1992).

The fact that Thai has prepositions, as well as direct and indirect objects, and its SVO structure is similar to English, made it easy to compare its structures with English, and to compare the divergences analyzed for the other two language sets. The lack of determiners in Thai did not have any effect on this analysis.

### *1.2 Goals*

This analysis was done to try to establish a pattern in the type and prevalence of translation divergences occurring within certain categories of verbs, namely high and low frequency verbs, and to test the hypothesis that high frequency verbs are more prone to translation divergences than low frequency ones. The identification of such pattern might be useful when creating automatic transfer rules from a dictionary, since it could help determine which areas might be more error free, and which ones should be the focus of further development and checks, in order to minimize translation errors.

### *1.3 Thesis Overview*

In Chapter 2, a review of the relevant literature is presented, starting from the established definitions of translation divergences, to an overview of the PanLex dictionary used for this analysis, to current efforts to extract rules based on bilingual dictionaries, in order to provide the basis of the translation divergence system used and put this work in the context of efforts made in this field. In Chapter 3, the hypothesis at the basis of the analysis is stated, and detailed information is provided on how this study was set up for each of the three language pairs, while outlining the methodology used. In Chapter 4, the results of the verb analysis are classified and

evaluated by verb categories such as high and low frequency, and type of divergence, and compared with the initial hypothesis. Finally, Chapter 5 contains the conclusions of this study, the possible implications, and further developments.

# Chapter 2
# LITERATURE REVIEW

## *2.1 Overview*

The machine translation structure which served as inspiration for the analysis developed in this thesis is a system that starts with grammars mapping surface strings to semantic representations, or MRSs, (Copestake et al., 2005), as developed by the researchers of the LOGON project for English and Norwegian (Lønning et al., 2004). In order to translate between these two languages, the transfer component of the LOGON system was initially developed by creating manual transfer rules, and subsequently, by automatically creating naïve transfer rules (by analogy) from bilingual dictionaries. Acquiring rules automatically from dictionaries has the advantage of saving time, since creating these rules manually is very time consuming. Starting from this research set-up, the process of acquiring naïve transfer rules could be expanded to increase the coverage to numerous languages by using a massive multilingual dictionary available online, such as PanLex.

However, translation between two natural languages is not always straightforward, and several different types of translation divergences could occur (Dorr, 1994). To analyze the possible divergences, the first step would be to identify them and categorize them by type, as well as to detail how they occur and what POS they involve.

This would be the basis for finding out what types of divergences might be problematic when using MRS based naïve transfer rules to translate between two languages, and which ones are easy to deal with.

The following subsections detail previous research on which this thesis builds.

### *2.2 LOGON Machine Translation*

For the LOGON project, the researchers of Oslo, Bergen, and NTNU Trondheim universities (Lønning et al., 2004), chose a traditional semantic transfer-based approach as their starting point, as, in their view, statistical methods need to be integrated with linguistic methods to overcome long-term performance issues.

Their approach, starting from a symbolic foundation and augmenting it with probabilistic methods to help narrow down results, is summed up as (Lønning et al., 2004, p. 2):

1. In depth grammatical and semantic analysis of Norwegian, resulting in language-specific logical semantic representations.
2. Transfer of these representations into language-specific English representations.
3. Generation of English sentences from the English representation.

As to the format of semantic representation, the starting point is Minimal Recursion Semantics (Copestake et al., 2005). In particular, for English, the grammar used is the LinGO English Resource Grammar (Flickinger, 2000; Flickinger, 2011). For Norwegian, on the other hand, a LFG grammar was used, called NorGram (Dyvik, 1999). The analysis augments LFG with MRS representations, thus integrating the MRS framework with LFG.

The transfer component is a resource sensitive rewriting process over MRS structures, which takes after VerbMobil (Wahlster, 2000), with some additional elements.

During the pilot phase, sentences in English and Norwegian extracted from brochures about hiking in the backcountry were used.

All three stages of the system yield multiple results. An average of 30 target language MRSs were produced for each source language MRS. Probabilistic methods were then applied to rank the output. At the time of writing, lexical selection had not been tackled, but the expectation was that a growing coverage would produce more transfer of ambiguities and bad translations.

Fine-tuning of the transfer rules and refining of the output ranking were eventually performed via stochastic processes, which also contributed to managing ambiguity and improving robustness (Velldal, 2004).

### *2.3 LOGON: Extracting Transfer Rules from Bilingual Dictionaries*

Further developments of the LOGON project, including, critically, the introduction of fully automated lexical transfer rule acquisition, are described by Nygaard, Lønning, Nordgård, and Oepen (2006). The training data set used is composed of 5000 sentences, translated by professional translators, with 3 translations for each source sentence. In addition, the machine had access to a bilingual dictionary of 80,000 translation pairs from Norwegian to English. Finally, the generation component produces sentences from the MRSs, using English Resource Grammar (Flickinger, 2000; Flickinger, 2011). The use of the ERG combined with MRSs ensures the high quality of the translation. One thing to notice is that if a word is not found, no output is produced. Consequently, this system is highly dependent on the size of the dictionaries accessed.

As mentioned in 2.2, LOGON's approach to semantic transfer borrows from the main tenets of Verbmobil, with the addition of two new elements: Typing, which creates a hierarchy of transfer rules, and a chart-like treatment of transfer-level ambiguity (Nygaard et al., 2006).

The bottleneck in LOGON was the creation of transfer rules, a painstakingly manual exercise at first. To overcome this hurdle, the researchers adopted an approach which uses handcrafted transfer rules as initial templates, and, with the help of the bilingual dictionary, creates more rules by analogy. The mapping occurs in three stages. First, an MRS predicate in the source language is related to dictionary entries. Second, candidate translations are looked up in the dictionary and their internal structure is analyzed. Third, target language predicates matching each translation are located. This process draws on the Semantic Interfaces (SEMI) in the two languages involved (Nygaard et al., 2006).

### *2.4 PanLex*

PanLex is the world's largest multilingual sense-distinguished dictionary, which covered 1,353 language varieties and about 12 million expressions, as of 2010 (Baldwin et al., 2010). The PanLex project was started to contribute to long-term linguistic diversity, and with the goal of enabling the translation of any word from any language into any other language, and it was created by compiling and connecting thousands of free dictionaries and other knowledge sources available online.

Table 1: PanLex, current and projected growth (Baldwin et al., 2010)

| | Current (2010) | Goal |
|---|---|---|
| Resources | 766 | 10K |
| Language varieties | 1353 | 7000 |
| Expressions | 12M | 350M |
| Expression–meaning pairs | 27M | 1000M |
| Expression–expression pairs | 91M | 1000M |

To further expand PanLex beyond the translation pairs found in the source documents, the research team took the direction of automatic inference, for instance, if an attested translation of the Santiago del Estero Quichua word *unta* is found in Spanish, but not in Hungarian, and its translation in Spanish, *lleno*, can be translated in Hungarian as *tele*, it could be inferred that the Quichua word *unta* could be translated in Hungarian as *tele* (Baldwin et al., 2010). This is a complex task, since lexical ambiguity can easily lower the quality of the translations.

To perform this inference, a translation graph was created which combines all attested lemmatic translations among all languages. The graph (Mausam et al., 2009) is then populated with nodes and edges by extracting translation pairs from all translingual lexical resources available, and by inferring new edges between existing expressions in PanLex, as per the above example.

### *2.5 Translation Divergences*

Several types of translation mismatches and divergences stemming from different properties of text are identified and described by Dorr (1994). The categories of translation divergences analyzed by Dorr fall under the umbrella of lexical-semantic translation divergences, for which she identifies seven different types (Table 2). Other types of translation divergences or mismatches, briefly mentioned, but not considered in Dorr's analysis, are the ones based on purely syntactic information, as well as mismatches due to idioms, aspect information, world knowledge, etc. (Dorr, 1990). Dorr provides the examples below for each of the seven lexical-semantic divergences, using English, Spanish and German for her examples.

Table 2: Categories for Dorr's lexical-semantic translation divergence types (1994, p. 3)

**Thematic divergence:**

[eng] I like Mary <> [spa] María me gusta a mí (=Mary pleases me)

**Promotional divergence:**

[eng] John usually goes home <> [spa] Juan suele ir a casa (=John tends to go home)

**Demotional divergence:**

[eng] I like eating <> [ger] Ich esse gern (=I eat likingly)

**Structural divergence:**

[eng] John entered the house <> [spa] Juan entró en la casa (=John entered in the house)

**Conflational divergence:**

[eng] I stabbed John <> [spa] Yo le di puñaladas a Juan (=I gave knife-wounds to John)

**Categorial divergence:**

[eng] I am hungry <> [ger] Ich habe Hunger (=I have hunger)

**Lexical divergence:**

[eng] John broke into the room <> [spa] Juan forzó la entrada al cuarto (=John forced the entry to the room).

Thematic divergence happens when the arguments of a given predicate are realized differently in different languages. In the example above, the object in English *Maria* becomes the subject in Spanish. Dorr also identifies two more types of divergences, which in Dorr (1990) are described as two cases of thematic divergence: Promotional and Demotional. Both of these divergences are "head switching", and while the first happens when a modifier is "promoted" to a main verb in translation, the latter describes the case where the main verb is "demoted" to a modifier. Conflational divergence happens when different arguments are needed for a certain action. An example would be *kick* (someone), which in Italian becomes *dare calci* (=give kicks to someone). Categorial divergence happens when a predicate changes part of speech type from one language to the other. For instance, *I am thirsty*, where the adjective *thirsty* becomes *Ho sete* (=I have thirst) in Italian, an auxiliary with a nominal predicate argument.

Additionally, structural divergence happens when the object of the verb is of a different type, for instance the direct object in this English sentence *Enter the room*, and the prepositional object in Italian *Entrare nella stanza* (=enter in the room). Finally, Lexical divergence captures the situation where an event is lexically realized with two different verbs in different languages, for

instance, the English *break into* is realized as *forzar* (=force) in Spanish.

### *2.6 Examples of MRS and Transfer Rules*

Of the four divergences considered for this analysis, Structural and Categorial divergences are, in some cases, solved at MRS level. The Structural ones, in particular, can sometimes be solved since an NP or a PP argument might look the same at MRS level, as long as the preposition in PP is semantically null, such as in Figure 1[4], where the preposition *to* is not represented by a separate predication.

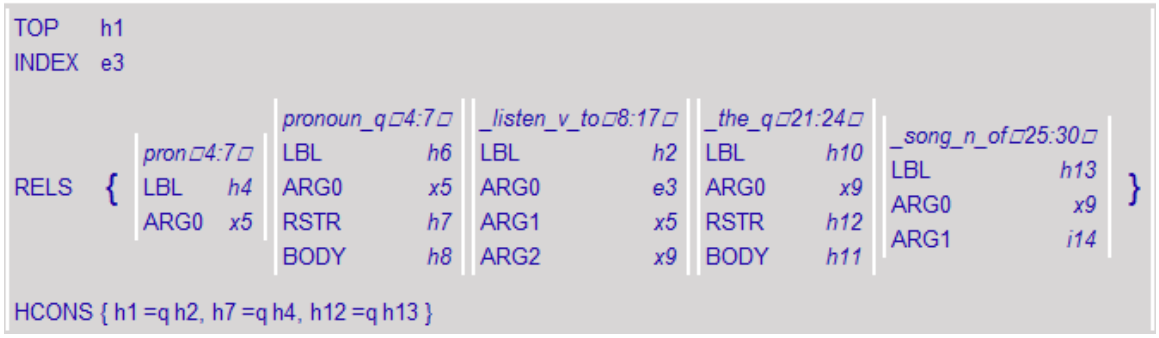


Figure 1: MRS of *Are you listening to the song?*

This is, however, not always the case, as in the English sentence *The salt had eaten into the paintwork*, where the preposition *into* carries semantic value and is not washed out in the MRS (Figure 2[5]).

[4] http://erg.delph-in.net/logon, check mrs, enter "Are you listening to the song?", and click Analyze.

[5] http://erg.delph-in.net/logon, check mrs, enter "The salt had eaten into the paintwork", and click Analyze.

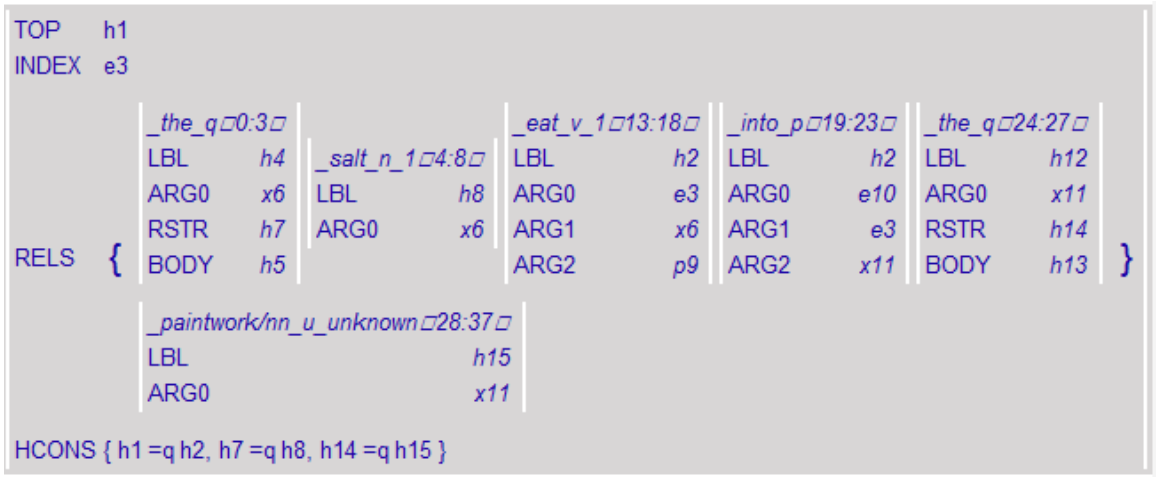


Figure 2: MRS of *The salt had eaten into the paintwork*

Categorial divergences might be problematic in some cases, since, at the level of MRS adjectives and verbs look very similar, but they are different from nouns. No examples of that case were found in this sample.

One last category that will be discussed in Chapter 4 are the English phrasal verbs, which were not classified as divergences, since the divergence is handled at MRS level, as seen in Figure 3[6] below.

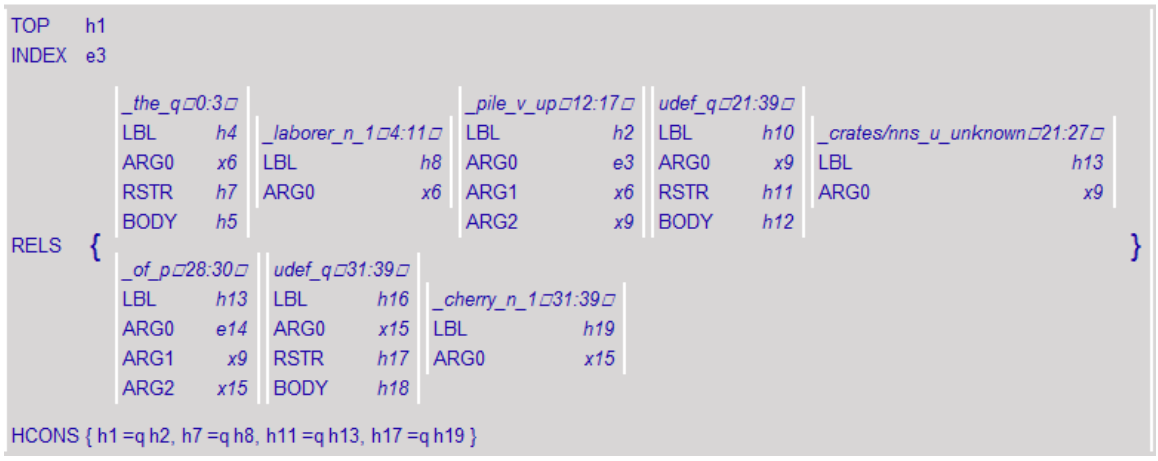


Figure 3: MRS of *The laborer piled up crates of cherries*

The MRSs are called by the naïve transfer rules, which rewrite parts of the input MRS. Successive application of transfer rules results in the final output (target language) MRS. An

[6] http://erg.delph-in.net/logon, check mrs, enter "The laborer piled up crates of cherries", and click Analyze.

example of a generic transfer rule for an intransitive verb is presented below (Lønning et al., 2004). In this example, since the verb is intransitive, we have just a subject, represented by the ARG1, and no objects. Each argument in the input verb entry maps to an argument in the output verb.

```
arg1_v_mtr := mrs_transfer_rule &
[ INPUT.RELS < [ LBL #h, ARG0 #e, ARG1 #x ] >,
  OUTPUT.RELS < [ LBL #h, ARG0 #e, ARG1 #x ] > ].
```

The transfer rule is applied below to the Italian intransitive verb *zampillare* (=to gush):

```
zampillare := arg1_v_mtr &
[ INPUT.RELS < [ PRED "_zampillare_v_rel" ] >,
  OUTPUT.RELS < [ PRED "_gush_v_rel" ] > ].
```

Conflational divergences are not handled by naïve transfer rules or MRS, and Thematic divergences are always problematic when using MRS based naïve transfer rules to translate between two languages, since a naïve rule would have no way of knowing that it would need to switch ARG1 with ARG2 (subject and object) between Input and Output.

### *2.7 Summary*

This chapter presented the relevant literature about a machine translation structure that starts with grammars mapping surface strings to semantic representations, or MRSs, (Copestake et al., 2005), as developed by the researchers of the LOGON project (Lønning et al., 2004). The transfer component of such system would be developed by automatically acquiring naïve transfer rules from bilingual dictionaries. Such process of acquiring naïve transfer rules could be expanded to increase the coverage to numerous languages by using a massive multilingual dictionary, such as PanLex. However, since translation between two languages often results in translation divergences (Dorr, 1990), an analysis of these divergences would be useful to understand how well the automatically acquired naïve transfer rules would work.

The following chapter describes the methodology followed to analyze and evaluate these divergences.

# Chapter 3
# METHODOLOGY

## *3.1 Overview*

In this chapter, the methodology used to set up the study is reviewed in detail. For each of the three language pairs involved, the differences in methodology and the reasons for these differences are considered. In addition, further information is provided on the reliability of the examples as well as a disclaimer for Thai, where the analysis was limited by the lack of direct knowledge of the language.

## *3.2 Hypothesis*

The initial hypothesis in this study is that low frequency verbs would have a lower rate of translation divergence compared to the high frequency ones. The purpose of the research is to test this hypothesis, and to validate the assumption that divergences occur more often in high frequency verbs, when transfer rules are automatically acquired from a bilingual dictionary and used for translation. The initial test involved relatively similar languages, such as Italian, English and French. However, to make the results more generalized, Thai, a language with very limited borrowing from the previous three, was added. Should the results of this research suggest the initial hypothesis is correct, subsequent confirmation in further research would lead to predictions about the likelihood of errors we might incur if we were to create PanLex-based transfer rules between two languages. Accordingly, low frequency verbs, which constitute the majority of the verbs, would be expected to be relatively error-free, whereas high frequency verbs would be expected to yield a higher error rate and require more work, such as tweaking the naïve transfer rules, or implementing additional rules to supplement the ones created automatically from the

dictionary.

### *3.3 Study Setup*

#### *3.3.1 Divergences Relevant to This Analysis*

Similar to Dorr's approach, all divergences that do not fall under the lexical-semantic umbrella were excluded from this analysis. Of the seven types of divergences identified by Dorr, the focus was restricted to four: Thematic, Conflational, Categorial, and Structural.

Table 3: List of divergences, with examples. (Dorr, 1990, 1994)

| | |
|---|---|
| **Conflational** | The translation of two or more words in one language into one word in another language, or incorporation of necessary participants (or arguments) of a given action (the manner argument, the likingly token, is incorporated into the main verb in English).<br><br>**Examples:**<br>[eng] I stabbed John <> [spa] Yo le di puñaladas a Juan. The English *stabbed* becomes *di puñaladas* (=give knife-wounds) in Spanish.<br>[eng] I like Mary <> [ger] Ich habe Mary gern. The English *like* becomes *have likingly* in German. |
| **Categorial** | Predicate changes part of speech type, or translation of words in one language into words that have a different part of speech in another language (e.g., predicate is adjectival in English and nominal in German).<br><br>**Examples:**<br>[eng] I am hungry <> [ger] Ich habe Hunger. The adjective *hungy* becomes the noun *Hunger* (=hunger) in German. |
| **Structural** | The realization of verb arguments in different syntactic configurations.<br><br>**Examples:**<br>[eng] I saw John <> [spa] Vi a Juan. The NP *John* is realized as a PP in |

Spanish *a Juan* (=to John).

[eng] John entered the house <> [spa] Juan entró en la casa. The NP *the house* is realized as a PP in Spanish *en la casa* (= in the house).

**Thematic**

1. Switching of arguments for a given predicate. For instance, the object in English is the subject in Spanish.
2. Promotion of a complement to the main verb and demotion of the main verb into an adjunct position.

**Examples:**

[eng] I like Mary <> [spa] María me gusta a mí. The subject is *I* in English, and *Maria* in Spanish.

[eng] John usually goes home <> [spa] Juan suele ir a casa. Promotion: *usually* becomes the main verb in Spanish (suele).

[eng] I like eating <> [eng] Ich esse gern. Demotion: *like* becomes an adverb in German (gern).

### *3.3.2 Selecting the Verbs*

The initial step was to choose Italian as a source language, and select 20 high-frequency, as well as 20 low-frequency verbs. To identify high and low frequency verbs in Italian, I used as a reference a website for teaching Italian as a second language[7]. The website contains a quite comprehensive list of 4,707 Italian verbs with grammatical information, including whether the verb is transitive or intransitive, personal or impersonal, regular or irregular, as well as its active and passive forms, and choice of auxiliary. The information used is a 4-point scale frequency classification provided for each verb. The Italian high frequency verbs were selected only from level-4 (high) verbs, while the low frequency ones were taken from level-2 (low) verbs only. No level-1 frequency verbs were found in the list, and it was decided to avoid all the level-3 frequency (medium) ones in order to maintain a certain separation between the two samples. A check on the relative frequency of the verbs selected was performed in the Italian language pages of google.it, which confirmed with reasonable approximation the frequencies provided in the verbs website (full results in Appedix A). The low frequency verbs range between 2,930 to 670,000, while the high frequency ones range between 3,810,000 to 78,700,000, for the infinitive

[7] http://www.saenaiulia.com

form of the verb. The checks in Appendix A were extended also to the past participle and the 3rd Present Indicative forms, and no outliers were evident. Some cells contain the notation "Names" for searches that returned brand names or homonym nouns. These counts were provided as approximate figures aimed at supporting the reliability of the frequencies found in saenaiulia.com, while noting the limitations of counting Google hits (Fletcher, 2011).

One more note about the selection process is that auxiliaries, Italian modal verbs, such as *potere* (can), *volere* (want), *dovere* (must/have to), as well as the verb *fare* (to do/make), were not used in this analysis.

It is also worth mentioning that high frequency verbs might include low frequency senses. All senses of a high frequency verb were counted in the high frequency class, regardless. Low frequency verbs, on the other hand, are necessarily low frequency in all of their senses, but these distinctions were not considered in the analysis.

The 40 Italian verbs thus selected were then extracted from the Italian to English dictionary lists of PanLex[8] (Baldwin et al., 2010). For each of the 40 verbs several translation pairs—in many cases up to hundreds—were found. To clean up the sample, the irrelevant translation pairs were excluded: for instance, book or movie titles; and then the remaining translations were subdivided into four categories (Table 4, below):

Table 4: List of verb categories with examples from Italian to English

| Structure | Italian > English |
|---|---|
| verb to verb | uscire > quit |
| verb to phrase | uscire > come out |
| phrase to phrase | invitare a uscire > invite out |
| phrase to verb | uscire dal guscio > hatch |

Since the number of translation pairs was still very high despite the initial filter, to make the sample more manageable, the number of verbs considered was restricted to a maximum of five occurrences for each of the 4 categories in Table 4. The maximum number of translations for each verb would be 20. Where more than five translations were available in PanLex for a given category (for instance, for the Italian verb *uscire*, category "verb to verb"), only five would be randomly selected from the sample found. Conversely, in cases where PanLex only provided fewer than 6 translations for a certain category, all of them would be selected. Because not all

[8] http://panlex.org/

verbs in PanLex yield the same number, or type, of translation pairs, the distribution of the examples differs for each verb.

Out of the 20 or fewer translation pairs found for each verb, not all pairs were usable. Pairs where a viable example could not be found, which were deemed wrong translations[9] (167 total in English, 138 in French), or where homonym nouns or noun phrases were provided instead of verbs, such as the case of the verb *piacere* used as a noun *con piacere* > *with pleasure*, were then excluded from the sample.

### *3.3.3 Creating the Examples*

Once the irrelevant translation pairs had been filtered out, the remaining pairs were analyzed, and sentences were created in Italian, which were then translated in the target language. To make sure none of the uses of a certain verb were being overlooked, dictionaries in the source and target language containing syntax information were used. The goal was to provide one example for each verb construction found in the dictionaries. This also meant that each of the 5 translation pairs initially selected could yield more than one possible sentence (Tables 5, and 6). The examples below for the Italian to English pair *mangiare* (=eat) include cases with and without divergences.

Table 5: Several possible sentences from each synset

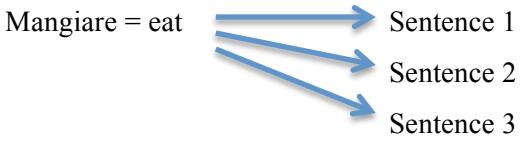


Table 6: Examples for the pair (mangiare > eat)

| **Italian sentence** | **English translation** | **Divergence** |
|---|---|---|
| (tr.) Mangio una mela. | (tr.) I'm eating an apple. | No |
| (tr.) La salsedine aveva mangiato la vernice. | (intr.) The salt had eaten into the paintwork. | Yes |
| (tr.) Mangio con gli amici. | (tr.) I eat with friends. | No |

[9] In order for these pairs to not create problems in a system making transfer rules out of PanLex entries, some means of filtering them out must be found.

When creating sentences, idioms were avoided, unless PanLex provided them as entries, such as the case of the French *giocare a carte scoperte/jouer cartes sur table* (=put one's card on the table). In that case, while sentences with PanLex idioms were created, they were not counted as divergences.

All the sentences were then examined to establish the occurrence of the four divergence phenomena described above. The incidence of the four divergences was analyzed by observing the structures used in the source and target languages, and the type of divergence was categorized. The specifics of each language pair are detailed in the following three subsections.

### *3.3.4 Italian to English*

For this language pair, English, Italian, as well as Italian to English dictionaries were used. The Italian sentences were reviewed by myself, as well as a second Italian native speaker, while the English translations were reviewed by several native English speakers, to ensure grammaticality and fluency. As mentioned above, each synset[10] often ended up with more than one translation pair, such as in the case of *mangiare* (=eat) in Table 6. The total number of sentences analyzed for this language pair was 45 low and 155 high frequency.

### *3.3.2 Italian to French*

A similar process was applied to the Italian to French language pairs. This time, the initial Italian verbs had already been selected, and the work thus started with the same high and low frequency verbs used for the English experiment. Based on these verbs, the translations pairs were extracted in the Italian > French PanLex verbs list. Two of the Italian low frequency verbs initially selected in the Italian > English lists were not available in the Italian > French ones (*acciambellare/coil* up and *accastellare/pile up*) and are therefore not included. For this reason, the French low frequency list started with 18 verbs. Note that the French translations found in PanLex for each verb were generally fewer than for English, so often all the verbs available in PanLex were used for the analysis, without having to randomly select five verbs per category. All French sentences were reviewed by a native speaker, to ensure grammaticality and fluency. The sentences were translated using various dictionaries[11] providing detailed syntax information.

[10] Throughout the text, a synset is defined as each translation pair found in PanLex.
[11] http://www.cnrtl.fr/definition/, http://www.larousse.com/en/dictionaries/french-monolingue, http://atilf.atilf.fr/tlf.htm

The total number of sentences analyzed for this language pair was 35 low and 150 high frequency.

### *3.3.3 English to Thai*

Thai was selected in order to have one language quite distant from the other three, which has a very limited amount of borrowing with them, to allow generalization of the observations by comparing languages that are closely related, as well as a dissimilar one. The study for Thai was structured differently, and is less systematic than for the previous two languages, due to this researcher's lack of direct knowledge of the language, and the relative difficulty of finding native speakers in the different phases of the project. Because it would be extremely hard to find people proficient in both Italian and Thai, English was used as the source language instead of Italian. Instead of extracting the verbs from PanLex, as was done for the other two language pairs, the English sentences previously translated were used as the starting point for the Thai translations. The reason is that while the painstaking process of checking all possible syntax structures for each language pair was possible for English and French, it could not have been done for Thai, and this solution was considered an acceptable compromise, since the source sentences were created and checked when working on the Italian to English translations. In addition, even though the translations were not done using PanLex pairs, a check was performed on 10 high frequency and 10 low frequency English verbs, which confirmed that most of the verbs used in this sample were available in the PanLex English to Thai bilingual lists. For the high frequency check, out of 10 verbs, in 7 instances the translator used a verb available in PanLex, in one case, no English<>Thai translation pair for that English verb was available in PanLex, and in the remaining two cases, the translator chose a different translation than the PanLex one. For the low frequency check, out of 10 verbs, in 6 instances the translator used a verb available in PanLex, and in the remaining 4 cases the translator chose a different translation than the PanLex one.

The translation into Thai was performed by a native speaker living in Thailand, with a very good knowledge of English. All of the low frequency (38 without repetitions), as well as the high frequency (156 without repetitions) English sentences were translated into Thai, for a total of 194 sentences.

The second difference in the methodology, also caused by the lack of direct knowledge of the language, was due to the inability to establish directly whether a divergence occurred in the translations. The method devised consisted in creating a series of questions (Table 7), which would cover the possible divergence cases. These questions were intended to be sent to a number

of Thai speaking linguists in Thailand, and then, by evaluating the responses received, it would have been possible to make informed guesses on whether a divergence occurred or not.

Table 7: Questions to evaluate the divergences between English and Thai

1. Are the subjects the same in English and Thai?
2. Are the objects the same in English and Thai?
3. Does the subject of one language become the object of the other?
4. Is the verb similar between English and Thai (same amount of words and similar meaning)? If not, can you explain what is different?
5. Are there words in English that do not appear in Thai, or words in Thai that do not appear in English? If yes, please list them by language.
6. Are different prepositions used in English and Thai? Can you tell me which ones?
7. Could you do a very, very literal translation in English of the Thai sentence, without using the original English text?

A slightly different solution was eventually adopted, to make it easier to double-check the results and to make guesses. A Thai speaker was consulted, and asked to consider each translation pair. For each sentence in Thai, some or most of the questions in Table 7 were then asked dynamically, depending on the findings. After a few tries, it was decided to start with question seven, and have the speaker provide a literal, oral translation of the Thai (or morpheme-by-morpheme gloss), without using the source English text. The following questions depended on how closely the literal translation mapped to the previously given one; if a clear divergence emerged (i.e., if the existing translation wasn't morpheme-by-morpheme isomorphic), the questions focused on pinning down the type of divergence, by probing similarities and dissimilarities in the syntax of the two languages. If a divergence was not reported by the speaker, more questions were asked (such as “What word is the subject in Thai?”, “What word is the object in Thai?”, “What kind of object is it?” and so on) to make sure that the English and Thai structures were indeed similar, and did not contain a divergence.

Based on the answers and the literal translation of the Thai into English, it was possible to determine with a moderate degree of confidence where a divergence occurred, and, in most cases, what kind of divergence it was. The total number of sentences analyzed for this language pair was 38 low, and 156 high frequency.

### *3.4 Summary*

In this chapter the methodology devised for this study was described. The initial step was the selection of the Italian verbs, both low and high frequency, used to extract translation pairs from PanLex. The translations thus extracted included up to four different categories, such as verb to verb, phrase to verb, verb to phrase and phrase to phrase. Based on these dictionary entries, sentences were created in Italian, which were then translated into English, and French. The sentences and their translations were then analyzed for divergences, as well as categorized in the four divergence categories adopted. Further details included how the Thai part of the analysis differs from the French and English, due to the lack of direct knowledge of the Thai language, and how the divergence information was extracted in an interview with a native speaker. The total number of examples produced between English and French combined is 80 low, and 305 high frequency, while the total number of examples from English that were translated into Thai is 38 low frequency, and 156 high frequency.

# Chapter 4
# RESULTS

## *4.1 Overview*

The purpose of the analysis laid out in Chapter 3 is to evaluate the occurrence and distribution of divergences, in order to identify a pattern. By comparing the hand-written translations in each language pair, and considering whether a MRS based naïve transfer rule would be sufficient to solve any divergences found, it is possible to estimate the frequency of the divergences within the sample and estimate how well these would be taken care of by naïve transfer rules created automatically from PanLex. The results are analyzed from the angle of the initial hypothesis, which postulates that low frequency verbs experience less divergence issues than high frequency ones.

Statistics and conclusions will be drawn from all the data gathered, such as what the prevalence of divergences observed in the sample used is, how these divergences are handled by SMT, and how a naïve transfer rule with MRS would handle them in comparison. Based on these considerations, a prediction could be put forward about how effectively PanLex could be used as a source of transfer rules, and what the need for modeling of translation divergences would be.

All the verbs referred to in the following subsections can be inspected from each of the 6 files available (one low and one high frequency sample file per language)[12]. The structure of the data in the English and French files is similar to the one illustrated in Table 8, and includes information about the initial Italian verb (*ID* column), as well as the *synset* (the bilingual pair extracted from PanLex from the initial verb), an *Italian* column which contains the original sentence created in Italian, as well as a *French* or *English* column with the manual translation. Finally, the *SMT* column contains the Google Translate version of the Italian sentence, as of May

[12] http://www.delph-in.net/matrix/gfra-thesis

7th 2012, for low frequency verbs, and May 8th for high frequency verbs. Another column, called *Naïve transfer rule*, is marked with a *no*, when a naïve transfer rule with MRS could not handle the divergence, and with *ok*, when the divergence can be solved by the naïve transfer rule and/or MRS. Finally, four more columns (not represented below) are marked with an *x* for the type of divergence detected: Conflational, Categorial, Thematic, and Structural.

Table 8: Italian to English: Example of data set-up

| ID | Synset | Italian | English | SMT Translation | Naïve transfer rule | Divergence |
|---|---|---|---|---|---|---|
| aerare | aerare=give an airing | Mi piace aerare la stanza ogni ora. | I like to give the room an airing every hour. | I like to ventilate the room every hour. | no | Yes |

The English to Thai data files have a simpler structure, as shown in Table 9 below. They do not have *ID*, nor *Synset* or *Italian* columns, since this analysis started from the English sentences, rather than PanLex. The *SMT* translation column was also excluded, because of the lack of direct control over the Thai translation, and the inability to assess the type of problems related to SMT translation. In this case as well, four more columns (not represented below) included information about the occurrence of Conflational, Categorial, Thematic, and Structural divergences.

Table 9: English to Thai: Example of data set-up.

| English | Thai | Literal Translation of Thai (when different) | Naïve Transfer Rule | Divergence |
|---|---|---|---|---|
| I watered the cattle. | ฉันให้น้ำฝูงวัวควาย | I gave water to (the) cattle | No | Yes |

### *4.2 Italian to English results*

#### *4.2.1 Italian to English Low Frequency Results*

A total of 57 bilingual synsets were extracted from PanLex from the 20 initial verbs. None of the synsets had to be excluded for the purpose of limiting the sample to a maximum of 20, since all synsets extracted per verb were under that number. A total of 12 synsets were considered wrong translations, and did not produce any sentences. As an example, of the two synsets for the

Italian *abbeverare* found in PanLex, *abbeverare* (=water), and *abbeverare* (=give to drink), one was considered wrong. The verb *abbeverare* describes the action of *giving water to animals*, and since the English *giving to drink* would not be correctly used to describe the action of *giving water to animals*, but only to humans, the second of the two pairs extracted was considered a wrong translation.

Three divergences were found in 45 sentences analyzed, none of which could be handled by MRS based naïve transfer rules. Of these three, one was only structural; the verb *ammantare* (=cover something with something), requires the preposition *di* (=of) in Italian, while in English, *with* is used in *he blanketed the bed with flowers*. The second sentence has a conflational divergence: *aerare* (=air out) becomes *give an airing*, as well as a structural one in *room,* which goes from a direct object to indirect.

In the remaining divergence, a Conflational one, *tosare qualcuno* (=clip someone, i.e. cut a human's or an animal's hair) becomes *crop someone's hair*. The Italian verb includes the meaning of *cutting hair*, while the English verb only means *cutting*, and the object is made explicit *child's hair*.

Of these divergences, Google Translate, as of May 7th 2012, translated correctly only one with a non-divergent synonym - *aerare* (=give an airing) as *ventilate*.

In this sample there was a special group of five verbs, which eventually became a category of its own, separated from the divergences, namely phrasal verbs. While these five pairs where the English side was constituted by a phrasal verb, for instance *accastellare/pile up*, were initially considered as showing Conflational divergences, it was eventually decided that they couldn't be justified as divergences. These verbs should be handled at lexical level, and shouldn't be problematic when handled via MRSs through transfer rules (Figure 3). The two tables below present a summary of the statistics discussed.

Table 10: Italian to English: Statistics for low frequency verbs

| # Italian verbs selected | # Bilingual pairs from PanLex | # Sentences<br># Wrong pairs | Divergences | Handled by Naïve Transfer Rule | SMT handled divergences |
|---|---|---|---|---|---|
| 20 | 57 | 45 sentences<br>12 wrong pairs | 3<br>5 phr. verbs | 0<br>5 phr. verbs | 1 |

Table 11: Italian to English: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---|---|---|---|
| 2 | | | 2 |

### *4.2.2 Italian to English High Frequency Results*

Twenty high frequency verbs were selected in Italian, and extracted from PanLex, to obtain this sample of 310 synsets. Unlike for the low frequency sample, the translations in PanLex were very numerous, so the sample had to be reduced, as mentioned in Chapter 3, to a maximum of 20 synsets per verb (maximum 5 synsets randomly chosen per each verb category: verb to verb, verb to phrase, phrase to verb and phrase to phrase). A total of 155 bilingual verb pairs were considered wrong translations and were for this reason excluded. Some of the "wrong" translations were nouns, for instance *con piacere* translated as *with pleasure*, but most of them were simply wrong translations of a certain verb, such as *mangiare leccando* (=to eat by licking) translated as *to try*. The total of divergences found in the 155 sentences analyzed amounts to 75.

Sixteen of these divergences were only structural, such as, for instance, the sentence *La salsedine aveva mangiato la vernice* (=The salt had eaten into the paintwork), where the preposition *into* cannot be solved at MRS level (Figure 2). An example of a structural divergence which is solved at MRS level is provided below, in Figure 4. In this case, "on" is semantically null and will therefore not be problematic when translating.

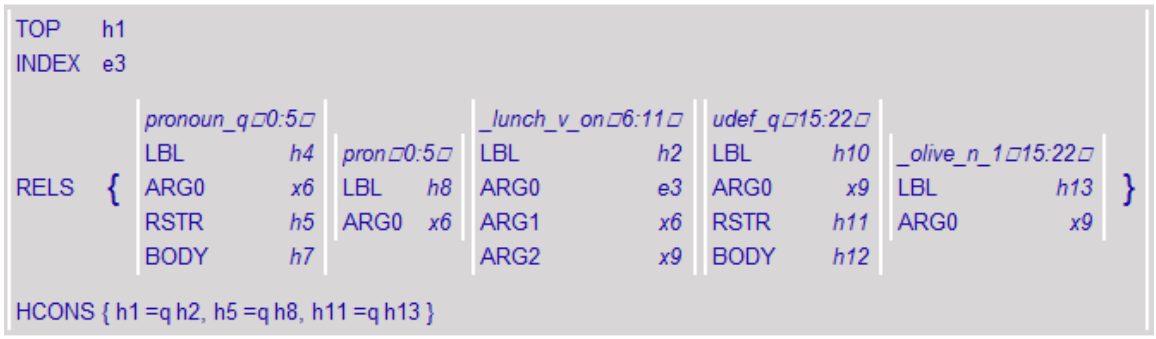


Figure 4: MRS of *Let's lunch on olives!*

The remaining 59 divergences, which constitute more than a third of the sentences (59/155) of this sample, are all non-structural, or a combination of structural with some other divergence. The total count for structural divergences is 42, while 49 were conflational, e.g. the pair *uscire dal guscio* (=get out of the shell) translated in English with the verb *hatch*, which could not be solved by a naïve transfer rule, since the syntactic structure of *uscire dal guscio* has the complement *dal guscio* (=from the egg) which the verb *hatch* doesn't have.

The Categorial category had 4 divergences, for instance the verb *finire* (= to finish, to end) translated in PanLex as *be finished*. A slightly difficult divergence to place was the sentence

*Sento freddo* (=I feel cold), translated in PanLex as *I'm cold*. Initially, this divergence seemed to be closer to a Categorial divergence more than any other, but by looking at the MRS of *I'm cold* (Figure 5) it appears that the divergence would rather be conflational, as the Italian *sento freddo* (=feel cold) seems to map to the English *cold*. This sentence would likely not be handled correctly by a naïve transfer rule:

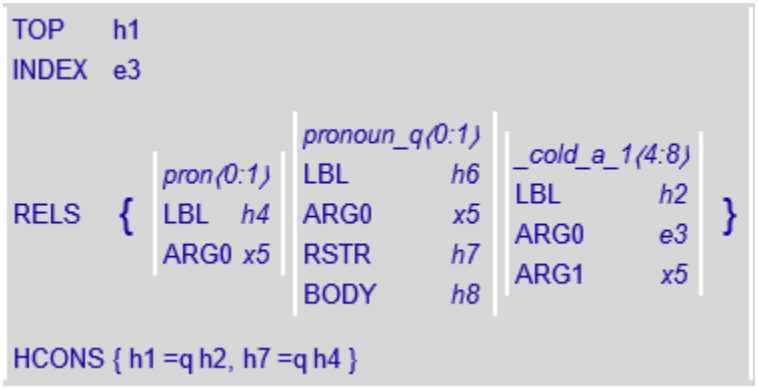


Figure 5: MRS of *I'm cold*

Finally, the Thematic divergence category had 7 divergent sentences. One example for this category is the sentence *Paolo piace a Maria*, were the verb *piacere* is translated in English with the verb *to like*: *Maria likes Paolo*. In this translation, the subject of the Italian, *Paolo* becomes the direct object in the English sentence, while the Italian predicate, *a Maria*, becomes the subject of the English sentence. This divergence would be problematic, since it could not be solved by the mapping from the surface string to the MRS, or by naïve transfer rules.

The complete syntactic units found in PanLex can provide an advantage when creating transfer rules automatically, because some of the divergences are taken care of by the equivalences found in the dictionary. In this sample, 5 divergences would be solved by naïve transfer rules and MRS representations.

One last thing to note is also that all the sentences containing idioms, as well as sentences with phrasal verbs in English, became categories of their own, and were not evaluated for divergences, as mentioned in section 3.3.3.

Google Translate, as of May 8th 2012, could handle 46 divergences, some of which were structural. Most of the divergences were solved by Google Translate via alternate non-divergent translations, e.g. *La mamma fa giocare i bambini*, which in the PanLex bilingual pair was *The mother entertains the children*, was solved by Google Translate as *The mother makes the children play*, which is the literal non divergent translation of the Italian sentence. The two tables below

present a summary of the statistics discussed above.

Table 12: Italian to English: Statistics for high frequency verbs

| # Italian verbs selected | # Bilingual pairs from PanLex | # Sentences<br># Wrong pairs | Divergences | Handled by Naïve Transfer Rule | SMT handled divergences |
|---|---|---|---|---|---|
| 20 | 310 | 155 sentences<br>155 wrong pairs | 75<br>11 Phr. verbs | 5<br>11 Phr. verbs | 46 |

Table 13: Italian to English: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---|---|---|---|
| 49 | 4 | 7 | 42 |

### *4.3 Italian to French results*

#### *4.3.1 Italian to French low frequency results*

Two of the initial 20 low frequency Italian verbs selected (*affastellare*/*pile up*, and *acciambellare/coil up*) could not be found in the Italian to French bilingual PanLex files. As a result, the initial verbs for this pair were 18, and yielded a total of 45 synsets in PanLex. Similarly as for the English low frequency sample, none of the synsets extracted had to be excluded for the purpose of limiting the sample to a maximum of 20, since all synset extracted per verb were below that number. However, a total of 10 bilingual verb pairs were considered wrong translations and were for this reason excluded from the evaluation. As an example, of the 8 translation pairs for the Italian verb *sorbire* (=to sip, to absorb) in PanLex, 5 were wrong translations, such as *sorbire* (=drink/sip) > *lecher* (=lick), *sorbire con rumore* (sip noisily) > *boire golument* (=drink greedily), and so on. In the first of the two mistranslations described, the two verbs are not synonyms: *sip* and *lick*. In the second one, the mistranslation concerns the adverbs, since the Italian *con rumore* (=noisily) is translated as *goulûment* (=greedily) in French.

Six divergences were found in 35 sentences analyzed. The evaluation of whether MRS would handle the divergences for French was less systematic, since an adequate MRS structure is not available for French. Based on the analysis performed, it appeared that none of the French divergences would be handled with MRS based transfer rules. Of the six divergences found, one was only structural; the verb *abiurare* (=abjure) has a direct object in Italian *una religione* (=a religion), while this particular translation, the French verb *renoncer*, takes an indirect object *renoncer à* (=renounce to). It is possible that *à* would be treated as semantically empty, and be

washed out in the mapping to MRS.

Of the remaining 5 divergent translations, one contained structural and conflational divergences, *abbeverare le vacche* (=water the cows), which becomes *donner à boire aux vaches* (=give water to the cows), and has a conflational divergence in *water* > *give water*, as well as a structural one in *cows* which goes from a direct object to indirect.

Another example of a conflational divergence is the case of *ammantare/couvrir d'un manteau* (=to cloak/cover with a cloak). While the occurrence of divergences is still low (6/35) for this sample, it is considerably higher than for the English low frequency sample.

Of these divergences, Google Translate, as of May 7th 2012, could not handle any of the conflational ones, while it could translate correctly the structural one. Google Translate again handled correctly the thematic divergence with a non-divergent synonym *ha alienato la sua famiglia* (=he alienated his family), which on PanLex was translated as *sa famille s'est détachée de lui* (=his family distanced itself from him), was translated as *il a aliéné sa famille* (=he alienated his family) by this system. Tables 14 and 15 below present the statistics discussed.

Table 14: Italian to French: Statistics for low frequency verbs

| # Italian verbs selected | # Bilingual pairs from PanLex | # Sentences<br># Wrong pairs | Divergences | Handled by Naïve Transfer Rule | SMT handled divergences |
|---|---|---|---|---|---|
| 18 | 45 | 35 sentences<br>10 wrong pairs | 6 | 0 | 2/6 |

Table 15: Italian to French: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---|---|---|---|
| 4 | | 1 | 2 |

### *4.3.2 Italian to French High Frequency Results*

Twenty high frequency verbs were initially selected in Italian, and then extracted from PanLex, to obtain the 278 synsets in this sample. The translations available in PanLex were much more numerous than the desired sample size, so the sample had to be reduced, as mentioned in Chapter 3, to a maximum of 20 synsets per verb; to get to this number, 5 synsets were randomly chosen per each verb category when necessary (verb to verb, verb to phrase, phrase to verb and phrase to phrase). A total of 128 bilingual verb pairs were considered wrong translations and were for this reason excluded from the evaluation. Some of the wrong translations were verbs

translated as nouns, for instance *parlare* (=to speak) translated as *langue parlée* (=spoken language), but most of them were simply wrong translations of a certain verb, such as *parlare* (=to speak) translated as *protester* (=to protest).

The total of divergences found in the 150 sentences analyzed amounts to 60. Five of these divergences were only structural. Since MRSs cannot be verified for French, it is difficult to guess whether prepositions could be treated as semantically null at MRS level. In this sample, the guesses about whether the divergence would be solved in this sample are mostly at naïve transfer rule level. For instance I conservatively guessed that the divergence in the sentence *Lo mangiava con gli occhi* (=he ate her with his eyes), which requires the preposition *con* (=with) in Italian, while is uses *des* in French, *des yeux* (=from/by his eyes), could not be solved by MRS, but this may not be the case.

The remaining 55 divergences, which constitute over a third of the sentences (55/150) of this high frequency sample, are all non-structural, or a combination of structural with some other divergence. The total count for structural divergences is 14, while 49 were conflational, e.g. the pair *studiare molto* (=study much) translated in French with the verb *bûcher*, in the sentence *Studiava sempre molto/Il bûchait tout le temps* (=He/she always studied a lot), which would probably not be solved by a verb to verb naïve transfer rule because of the lack of symmetry in the predications between Italian and French, but might be solved by a slightly different transfer rule, mapping the Italian verb + adverb to the simple verb of the French, which could be automatically derived from PanLex.

One issue with categorization was posed by non-reflexive to reflexive translation pairs, such as the verb *mangiare* (=to eat) translated in PanLex as *se nourrir* (=feeding oneself). Reflexive verbs don't entirely fit in any of these four divergence categorizes, but share some traits with some of them, for instance, they introduce a direct object, the reflexive pronoun, which might seem borderline thematic, since the subject becomes also the object in the target sentence. It was finally decided to consider them conflational, for instance, in the sentence *Amava mangiare le verdure* (he/she loved to eat vegetables), translated in French as *Elle aimait se nourrir de legumes*, *mangiare* (=to eat) was treated as if its meaning were expressed in French with two words *se nourrir* (=feed oneself). This particular sentence also has a structural divergence, since the direct object of the source language *verdure* becomes an indirect object in the target language *de legumes* (=with vegetables).

Of the 49 conflational divergences, 16 are reflexives. An interesting case is the sentence *Mangiamo avidamente le olive* where the Italian verb *mangiare avidamente* (=eat greedily), is translated in PanLex as *se goinfrer* (=stuff oneself). Based on our decision to consider reflexives

as conflational divergences, we would have here a double conflational divergence. The first one comes from treating the reflexive as a conflational divergence, as mentioned in the previous paragraph. The second one is due to the fact that the verb *goinfrer* incorporates the meaning of the Italian verb + adverb *mangiare avidamente*.

The last two categories, Categorial and Thematic have both 5 divergent sentences. One example for the Categorial is the sentence *Ha preso marito* (literally, he/she got a husband), were the verb *prendere marito* (=get a husband) is translated in French with *s'est mariée* (=got married), and where the predicate *husband* of the Italian changes part of speech in French. An example of Thematic divergence is the sentence *Mi piacciono le mie scarpe nuove* (literally, the syntax is: *My new shoes are pleasant to me*, but the meaning corresponds to: *I like my new shoes*), translated in Panlex as *J'aime mes nouvelles chaussures* (=I love my new shoes), where subject and object are switched. Another interesting Categorial case is the sentence *Parlò molto bene del suo insegnante* (=he/she spoke well of her teacher), translated as *Il dit beaucoup de bien de son enseignant* (=he said good things of his teacher), where the Italian modifier *molto bene* changes into a direct object in French.

The complete syntactic unit of the synset *partire da/quitter* (=leave from somewhere) helped solve the structural divergence, since the preposition *da* is incorporated in the information provided from PanLex. In this sample, only 1 divergence could be solved by naïve transfer rules and/or MRS representations.

One last thing to notice is also that all the eight sentences containing idioms, such as *non capire un corno/ne comprendre que dalle*, which literally means *do not understand anything* were not evaluated for divergences, as mentioned in Chapter 3.

Google Translate, as of May 8th 2012, could handle 15 divergences (2 of which were structural, but solved via alternate non-divergent translations, e.g. *Mangiava poco* (=he/she ate little), which in the PanLex bilingual pair was *Elle picorait*, became *Il mangeait peu* (he ate little). The two tables below present a summary of the statistics discussed above.

Table 16: Italian to French: Statistics for high frequency verbs

| # Italian verbs selected | # Bilingual pairs from PanLex | # Sentences<br># Wrong pairs | Divergences | Handled by Naïve Transfer Rule | SMT handled divergences |
|---|---|---|---|---|---|
| 20 | 278 | 150 sentences<br>128 wrong pairs | 60 | 1 | 15 |

Table 17: Italian to French: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---|---|---|---|
| 49 | 5 | 5 | 14 |

### *4.4 English to Thai Results*

#### *4.4.1 Limitations of This Analysis*

There are two main limitations to the Thai analysis. The first limitation concerns the type of sample used. In the Italian to English and Italian to French pairs, the basis for creating the examples were translation pairs extracted from PanLex. There were four categories of translation pairs found: verb to verb, verb to phrase, phrase to verb and phrase to phrase, but the initial structure used for the examples was a verb. However, a check performed by extracting the English to Thai translations from PanLex, and searching for matches between these and the verbs in this translation sample proved that 13/20 of the verbs in this sample were available in PanLex, 6/20 of the verbs were also available but a different verb was used, while only in 1 instance out of 20 the initial verb was not found.

Secondly, verb frequency may differ when translating from one language to another. Because high and low frequency distribution can change in translation, despite the fact that an effort was made to maintain the same register when translating, our initial composition of 20 high and 20 low frequency verbs in Italian may not have carried over to English or French. This is particularly important for the Thai analysis, since its starting point are the English translations that were created from the initial high and low frequency Italian sentences. The possibility of re-categorizing the English verbs into high and low frequency before evaluating this sample was considered, but was eventually dismissed in order to maintain a consistent verb group, and a certain continuity with the initial Italian sample selected.

A third cautionary note could be added to the limitations above. Due to the lack of direct knowledge, and the difficulty of finding available Thai native speakers with a linguistics background, the results cannot be considered as reliable as for the other two samples.

#### *4.4.2 English to Thai Low Frequency Results*

The initial English sentences translated into Thai numbered 38, after excluding sentences that were identical from the initial sample. The English literal retranslation from Thai was provided

only when the morpheme-by-morpheme translation between the two languages was divergent. Based on this, only 14 sentences were retranslated into English and are recorded to have a divergence. All the Thai sentences reviewed by the second Thai consultant, were considered adequate translations of the English source, for this low frequency sample (while some were found to be wrong in the high frequency sample). Of these 14 divergences, one could not be categorized in any of the four types, since it involves a subordinate dependent from the main verb. It is the case of the sentence *He forswore his bad habits*, translated in Thai as *เขาสาบานว่าจะเลิกนิสัยเลว* and literally retranslated back to English as *He swore that he will quit his bad habits*. Six of these divergences were Structural, but none of them was purely structural Fourteen conflational divergences were also found, of which 6 were both conflational and structural, for example *The neighbor's house shaded the garden*, translated in Thai as *บ้านของเพื่อนบ้านเป็นร่มเงาให้สวน*, and literally translated back to English as *Neighbor's house gave shadow to garden*. In this sentence, the conflational divergence is represented by *shadow/give shadow*, while the structural one *the garden/to the garden*. The categorization of these sentences was quite straightforward, based on the data available. The occurrence of divergences in this sample is much higher than for the low frequency in English and French, since almost half the sentences are divergent (14/38). Looking at the type of divergences, the incidence of Conflational divergences seem much higher than all other ones, and it may suggest a tendency to paraphrase the original English translation, for verbs that are unusual or uncommon. Of 10 verbs searched in PanLex from this sample, all were found in PanLex, even though in 4 cases the translator had used a different Thai translation than the one available in PanLex. Further investigation could be done to find out whether the 4 verbs in PanLex not used by the translator were also divergent, or if they were non-divergent verbs, which would have lowered the percentage of divergences in this sample.

None of the divergences found would seem to be able to be handled correctly by naïve transfer rules.

Table 18: English to Thai: Statistics for low frequency verbs

| Initial English sentences | Divergences | Handled by Naïve |
|---:|---:|---:|
| 38 | 14 | 0 |

Table 19: English to Thai: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---:|---:|---:|---:|
| 14 | | | 6 |

### *4.4.3 English to Thai High Frequency Results*

The initial English sentences translated into Thai were 156, after excluding sentences that were identical in the initial English sample. Seven translations were considered wrong by the second Thai consultant, when working on the morpheme-by-morpheme translation back to English, and were therefore excluded from the analysis. As a result, the sample of viable sentences decreased to 149. The English literal retranslation from Thai was provided only when the morpheme-by-morpheme translation between the two languages was divergent, and a total of 56 sentences were considered divergent. One divergence could not be categorized in any of the four types used so far, since it involves a subordinate dependent from the main verb. It is the case of the sentence *He drank himself unconscious*, translated in Thai as *เขาดื่มจนไม่รู้สึกตัว* and literally retranslated back to English as *He drank until not feel himself*. Seventeen of these divergences were only Structural. Of these, 3 were believed to be handled correctly by naïve transfer since the structure replicates a similar structure as the one in Figure 1. Five more structural divergences are classified as *maybe*, since not enough information is available to understand if they could be handled by MRSs. The total of Structural divergences both alone and in combination with other divergences is 29. Thirty-one Conflational divergences were also found, of which some were both Conflational and Categorial, e.g. *Paolo is idle*, whole Thai translation *เปาโลไม่ทำอะไร* is literally translated in English as *Paolo doesn't do anything*. Others were both Conflational and structural. The total of Categorial divergences is 10, while 3 are Thematic.

In this sample as well, all the sentences containing phrasal verbs in English, and no other divergence, became categories of their own, and were not considered divergent, as previously mentioned. For instance, the verb *set off* translated as *leave*, like in the sentence *I set off without warning anyone*. Other cases worth mentioning, which were not considered divergent, but had differences in the translation, which seemed to be a tense/aspect difference, for instance, *I decided*, in Thai *ฉันตัดสินใจแล้ว*, which was retranslated in English as *I decided already*.

The occurrence of divergences in this sample is lower than for the low frequency, but generally in line with the high frequency samples of English and French (56/149).

Table 20: English to Thai: Statistics for high frequency verbs

| Initial English sentences | Wrong translations | Divergences | Handled by Naïve |
|---|---|---|---|
| 156 | 7 | 56 | 3<br>5 maybe |

Table 21: English to Thai: Statistics by type of divergence

| Conflational | Categorial | Thematic | Structural |
|---|---|---|---|
| 31 | 10 | 3 | 29 |

### *4.5 Hypothesis Testing*

The statistical hypothesis test was run on the divergence results for each low frequency language pair sample, to understand their significance, and to help support the claim that low frequency verbs experience much lower divergence than high frequency ones. This is particularly helpful, since the sample of verbs and sentences analyzed is small.

For each of the three tests (one per language pair), the null hypothesis, $H_0$ is expressed below, and it describes the statement that the expected value of divergence in the normal population would be equal for the high and low frequency samples:

| | | | |
|---|---|---|---|
| Italian to English | $H_0$: $p = p_0$ | or p = 0.48 | (48.4%) |
| Italian to French | $H_0$: $p = p_0$ | or p = 0.4 | (40%) |
| English to Thai | $H_0$: $p = p_0$ | or p = 0.38 | (37.6%) |

The claim in this case, or $H_1$, is that low frequency divergence cannot be explained by the null hypothesis, and is, instead, lower.

| | |
|---|---|
| Italian to English | $H_1$: $p < p_0$ |
| Italian to French | $H_1$: $p < p_0$ |
| English to Thai | $H_1$: $p < p_0$ |

The test statistic z was calculated based on the values for each sample, using the claim and null hypothesis above, as well as the following values: "p̂" was identified as the occurrence of divergences in the low frequency samples, "n" as the total number of sentences in that sample, "p" as the occurrence of divergence in the high frequency sample (as per $H_0$). The formula used to find $z_{obs}$, and results are below:

$$\frac{\hat{p} - p0}{\sqrt{\frac{p0 - (1 - p0)}{n}}}$$

(1) Italian to English:

$\hat{p} = 0.067$ (3/45); n = 45; p = 0.48; q = 0.52

$z_{obs} = -5.55$

$Pr(z < z_{obs}) = Pr(z =< -5.55) = 0.0001$

(2) Italian to French:

$\hat{p} = 0.17$ (6/35); n = 35; p = 0.4; q = 0.6

$z_{obs} = -2.78$

$Pr(z < z_{obs}) = Pr(z =< -2.78) = 0.0027$

(3) English to Thai:

$\hat{p} = 0.37$ (14/38); n = 38; p = 0.38; q = 0.62

$z_{obs} = - 0.13$

$Pr(z < z_{obs}) = Pr(z =< 0.13) = 0.4483$

With a significance level of 0.01, both the English and the French p-values (0.0001 and 0.0027) are low enough to warrant rejection of the null hypothesis. The test results above confirm our suspicions that the low frequency results (1) and (2) cannot be explained within the null hypothesis. The only p-value which doesn't warrant rejection of the null hypothesis, as expected, is the one for Thai (3), which is not in line with the other two language pairs, as described in Results.

### *4.6 Summary*

From the results of the divergences observation across the three language pairs, it is manifest that the results for English and French are close to each other, and seem to be in line with our initial assumption. As seen in Table 22 below, the analysis of the Italian to English translations confirmed this hypothesis, with a 3/45 divergences found in the low frequency sample, and a 75/155 in the high frequency sample. The differences in the Italian to French verb sample were less striking between high and low, but still within the expected results: 6/35 for low frequency, and 60/150 for high frequency.

The results for Thai were not as expected, but there are some limitations to the Thai evaluation, explained in Chapter 4, which could have skewed the results for this language pair. While the English to Thai high frequency verbs are within the range of the Italian to French

sample, at 56/149, the low frequency sample was 14/38, with a much higher occurrence of divergences than the other low frequency samples.

Table 22: Divergences summary

| | Divergences | Conflational | Categorial | Thematic | Structural |
|---|---|---|---|---|---|
| ITA > ENG | | | | | |
| Low | **6.7%** (3/45) | 2 | | | 2 |
| High | **48.4%** (75/155) | 49 | 4 | 7 | 42 |
| ITA > FRA | | | | | |
| Low | **17.1%** (6/35) | 4 | | 1 | 2 |
| High | **40%** (60/150) | 49 | 5 | 5 | 14 |
| ENG > THA | | | | | |
| Low | **36.8%** (14/38) | 14 | | | 6 |
| High | **37.6%** (56/149) | 31 | 10 | 3 | 29 |

The prevalence of divergences for each category of synset is presented in Table 23. In the English and French high frequency samples, the highest percentage of divergence fell in the Phrase to verb category, but the rest of the percentages per each category were not consistent between these two languages.

Table 23: Divergences by synset category

| | Verb to Verb | Verb to Phrase | Phrase to Verb | Phrase to Phrase |
|---|---|---|---|---|
| ITA > ENG | | | | |
| Low | 66.7% (2/3) | 33.3% (1/3) | | |
| High | 24% (18/75) | 14.7% (11/75) | 44% (33/75) | 17.3% (13/75) |
| ITA > FRA | | | | |
| Low | 16.7% (1/6) | 83.3% (4/6) | | |
| High | 10% (6/60) | 31.6% (19/60) | 36.7% (22/60) | 21.7% (13/60) |

# Chapter 5
# CONCLUSION

## *5.1 Reconsidering the Hypothesis*

Our initial hypothesis stated that the occurrence of the four translation divergences analyzed would be more prevalent in high frequency than in low frequency verbs. The analysis of the Italian to English translations confirmed this hypothesis, with a very low 6.7% divergences found in the low frequency sample, but a very high 48.4% in the high frequency sample. The Italian to French verb evaluation showed less striking differences between high and low, but the same general pattern: 17.1% for the low frequency sample and 40% for the high frequency one.

The results for Thai were not as expected, but there were some limitations to the Thai evaluation, explained in Chapter 4. While the English to Thai high frequency verbs are close enough to the results for the other two language pairs, at 36.8%, the low frequency results for Thai, 37.6%, are higher than the high frequency ones, and seem to run counter to the hypothesis. Since almost all the divergences in the low frequency Thai sample are Conflational (14), we suspect that one possible reason is that the translation in Thai may have been paraphrased, or simplified, to explain rather than translate an uncommon and unfamiliar English verb. To prove that this were the case, a new set of translations would need to be done, from English to Thai, using PanLex pairs, and possibly by a professional translator, with clear directions to maintain a similar register and to use an uncommon word to translate the English one, if one that fits well were available.

## *5.2 Final Remarks and Future Developments*

We believe that this evaluation shows somewhat promising results, which may suggest that

manual transfer rules creation and tweaking of automatic rules should be focused on high frequency verbs, while low frequency verbs seems to have a very low translation divergence error rate. However, a few suggestions can be made to improve and continue this analysis, and make it more error-proof. For instance, it would be a good idea to analyze parallel results, and have the same source language for all language pairs. In this case, the same Italian verbs would have to be extracted from the Italian to Thai PanLex bilingual lists, and sentences then created from Italian to Thai. As mentioned, in the previous subsection, it would be a good idea to have the translations done by professional translators, possibly more than one, and give clear directions that the target text should match the register of the source language as much as possible.

Also, and more generally, the analysis could be extended to more verbs, and more languages, in order to make the results more generalized and reliable.

# BIBLIOGRAPHY

Timothy Baldwin, Jonathan Pool and Susan M. Colowick. 2010. PanLex and LEXTRACT: Translating all Words of all Languages of the World. Coling 2010: Demonstration Volume, pp. 37-40.

Alexandra Birch. Reordering Metrics for Statistical Machine Translation. 2011. PhD Thesis, University of Edinburgh.

Ann Copestake, Dan Flickinger, Ivan Sag and Carl Pollard. 2005. Minimal Recursion Semantics: An introduction. *Journal of Research on Language and Computation*, 3(2-3), pp. 281-332.

Bonnie J. Dorr. 1990. Solving Thematic Divergences in Machine Translation. *Proceedings of the 28th annual meeting on Association for Computational Linguistics*, University of Pittsburgh, Pittsburgh, PA, pages 127-134.

Bonnie J. Dorr. 1994. Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics*, 20:4, pages 597-633.

Helge Dyvik. 1999. The universality of f-structure. Discovery or stipulation? The case of modals. *Proceedings of the 4th International Lexical Functional Grammar Conference*. Manchester, UK.

William H. Fletcher. 2011. Corpus Analysis of the World Wide Web. Chapelle, Carol A, (Ed.). *Encyclopedia of Applied Linguistics*. Wiley-Blackwell.

Dan Flickinger. 2000. On Building a More Efficient Grammar by Exploiting Types. *Natural Language Engineering*, 6 (1), *Special Issue on Efficient Processing with HPSG*, pp. 15-28.

Dan Flickinger. 2011. Accuracy v. Robustness in Grammar Engineering. CSLI Publications, Stanford, CA, pp. 31-50.

James Kapper. 1992. English Borrowing in Thai as Reflected in Thai Journalistic Texts. Dooley, Robert A. and Marshall, David F., Eds. *Work Papers of the Summer Institute of Linguistics*, University of South Dakota Session, Volume 36.

J. T. Lønning, S. Oepen, D. Beermann, L. Hellan, J. Carroll, H. Dyvik, D. Flickinger, J. B. Johannessen, P. Meurer, T. Nordgård, V. Rosén, and E. Velldal. 2004. LOGON - a Norwegian MT effort. *Proceedings of the Workshop in Recent Advances in Scandinavian Machine Translation,* Uppsala, Sweden.

Mausam, Stephen Soderland, Oren Etzioni, Michael Skinner, Daniel S. Weld, and Jeff Bilmes. 2009. Compiling a massive, multilingual dictionary via probabilistic inference. ACL 2009.

L. Nygaard, J. T. Lønning, T. Nordgård, and S. Oepen. 2006. Using a Bi-Lingual Dictionary in Lexical Transfer. An Experience Report and Some Preliminary Findings. In Jan Tore Lønning & Stephan

Oepen (ed.), *Proceedings of the 11th Conference of the European Association for Machine Translation.* European Association for Machine Translation (EAMT), pages 233-238.

Erik Velldal, Stephan Oepen and Dan Flickinger. 2004. Paraphrasing treebanks for stochastic realization ranking. *Proceedings of 3rd Workshop on Treebanks and Linguistic Theories*. Tübingen, Germany.

Wolfgang Wahlster, (Ed.). 2000. Verbmobil. Foundations of speech-to-speech translation. Berlin, Germany: Springer.

# Appendix A

## GOOGLE HITS FREQUENCY CHECK

| **Low Frequency** | **Infinitive** | **Past Participle** | **Present Indicative 3rd Person** |
|---|---|---|---|
| abbeverare | 161,000 | 139,000 | 44,000 |
| abbrustolire | 427,000 | 102,000 | 16,500 |
| abbuiare | 13,500 | 4,100 | 23,800 |
| abiurare | 70,600 | 33,000 | 179,000 |
| accastellare | 2,930 | 2,110 | 40,500 |
| acciambellare | 3,750 | 21,300 | 11,600 |
| adombrare | 48,800 | 164,000 | 90,100 |
| adunare | 31,400 | 44,300 | 122,000 |
| aerare | 63,800 | 137,000 | Names |
| affastellare | 12,300 | 29,000 | 8,680 |
| agglomerare | 80,100 | 775,000 | 20,500 |
| alienare | 263,000 | 283,000 | Names |
| ammantare | 30,400 | 98,700 | 119,000 |
| ammollare | 670,000 | 236,000 | 60,400 |
| sorbire | 264,000 | 60,400 | 32,000 |
| tediare | 90,100 | 49,200 | 22,900 |
| tosare | 96,200 | 333,000 | Names |
| triplicare | 279,000 | 336,000 | 246,000 |
| zampillare | 45,000 | 11,800 | 74,600 |
| zappare | 377,000 | 28,000 | Names |

| High Frequency | Infinitive | Past Participle | Present Indicative 3rd Person |
|---|---|---|---|
| mangiare | 36,800,000 | 8,710,000 | 12,700,000 |
| parlare | 66,400,000 | 31,100,000 | Names |
| piacere | Names | 19,500,000 | 121,000,000 |
| leggere | 74,700,000 | Names | Names |
| uscire | 36,100,000 | 24,600,000 | 25,000,000 |
| amare | 16,800,000 | Names | Names |
| prendere | 74,200,000 | 74,400,000 | 42,200,000 |
| bere | 27,600,000 | 3,280,000 | 3,410,000 |
| finire | 28,800,000 | 32,100,000 | 24,700,000 |
| scrivere | 53,400,000 | 129,000,000 | 29,900,000 |
| capire | 67,500,000 | 48,200,000 | 14,600,000 |
| arrivare | 50,700,000 | 40,800,000 | 49,900,000 |
| partire | 78,700,000 | 34,900,000 | Names |
| dormire | 26,600,000 | 2,990,000 | 7,810,000 |
| cuocere | 3,810,000 | 11,100,000 | 571,000 |
| conoscere | 53,800,000 | 26,300,000 | 28,000,000 |
| sentire | 44,400,000 | 33,500,000 | 26,500,000 |
| vivere | 60,300,000 | 13,700,000 | Names |
| studiare | 19,200,000 | 9,070,000 | 6,100,000 |
| giocare | 43,500,000 | 20,900,000 | 29,100,000 |