Doktori Disszertáció

The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography

Quantifying Translational Equivalence

Héja Enikő

2015

Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

Doktori Disszertáció

Héja Enikő

The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography

Quantifying Translational Equivalence

Nyelvtudományi Doktori Iskola
Dr. Bárdosi Vilmos CSc, egyetemi tanár, a Doktori Iskola vezetője
Elméleti Nyelvészeti Doktori Program
Dr. Bánréti Zoltán CSc, egyetemi docens, a program vezetője

A bizottság tagjai:

A bizottság elnöke:
Dr. Kiefer Ferenc MHAS, professor emeritus

Hivatalosan felkért bírálók:
Dr. Kornai András DSc, egyetemi tanár
Dr. Prószéky Gábor DSc, egyetemi tanár

A bizottság további tagjai:
Dr. Gyuris Beáta PhD, a bizottság titkára
Dr. Hunyadi László DSc
Dr. Rebrus Péter PhD (póttag)
Dr. Komlósy András CSc (póttag)

Témavezető:
Dr. Váradi Tamás PhD, tudományos főmunkatárs

Budapest, 2015

ADATLAP
a doktori értekezés nyilvánosságra hozatalához

I A doktori értekezés adatai

- A szerző neve: Héja Enikő
- MTMT-azonosító: 10018657
- A doktori értekezés címe és alcíme: The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence
- DOI-azonosító: 10.15476/ELTE.2014.094
- A doktori iskola neve: Nyelvtudományi Doktori Iskola
- A doktori iskolán belüli doktori program neve: Elméleti Nyelvészeti Doktori Program
- A témavezető neve és tudományos fokozata: Dr. Váradi Tamás, PhD
- A témavezető munkahelye: MTA, Nyelvtudományi Intézet

II Nyilatkozatok

1. . A doktori értekezés szerzőjeként

   a) hozzájárulok, hogy a doktori fokozat megszerzését követően a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az ELTE Digitális Intézményi Tudástárban. Felhatalmazom az ELTE BTK Doktori és Tudományszervezési Hivatal ügyintézőjét, Manhercz Mónikát, hogy az értekezést és a téziseket feltöltse az ELTE Digitális Intézményi Tudástárba, és ennek során kitöltse a feltöltéshez szükséges nyilatkozatokat.

   b) kérem, hogy a mellékelt kérelemben részletezett szabadalmi, illetőleg oltalmi bejelentés közzétételéig a doktori értekezést ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;

   c) kérem, hogy a nemzetbiztonsági okból minősített adatot tartalmazó doktori értekezést a minősítés időtartama alatt ne bocsássák nyilvánosságra az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban;

   d) kérem, hogy a mű kiadására vonatkozó mellékelt kiadó szerződésre tekintettel a doktori értekezést a könyv megjelenéséig ne bocsássák nyilvánosságra az Egyetemi Könyvtárban, és az ELTE Digitális Intézményi Tudástárban csak a könyv bibliográfiai adatait tegyék közzé. Ha a könyv a fokozatszerzést követőn egy évig nem jelenik meg, hozzájárulok, hogy a doktori értekezésem és a tézisek nyilvánosságra kerüljenek az Egyetemi Könyvtárban és az ELTE Digitális Intézményi Tudástárban.

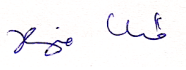
2. A doktori értekezés szerzőjeként kijelentem, hogy

   a) az ELTE Digitális Intézményi Tudástárba feltöltendő doktori értekezés és a tézisek saját eredeti, önálló szellemi munkám és legjobb tudomásom szerint nem sértem vele senki szerzői jogait;

b) a doktori értekezés és a tézisek nyomtatott változatai és az elektronikus adathordozón benyújtott tartalmak (szöveg és ábrák) mindenben megegyeznek.

3. A doktori értekezés szerzőjeként hozzájárulok a doktori értekezés és a tézisek szövegének Plágiumkereső adatbázisba helyezéséhez és plágiumellenőrző vizsgálatok lefuttatásához.

Kelt: Budapest, 2015. június 1.



a doktori értekezés szerzőjének aláírása

In order to say what a meaning *is*, we may first ask what a meaning *does*, and then find something that does that.

——— David Lewis (1970)

In order to say what a translation relation *is*, we may first ask what the translation relation *does*, and then find something that does that.

——— Based on David Lewis (1970)

Abstract

Creating bilingual dictionaries is very demanding in terms of human effort. From the compilation of the headword list to finding correct and relevant translations and providing information on how to use them in idiomatic target language sentences, lexicographic knowledge and experience are required. Although mono- and bilingual corpora are increasingly present in the process, human effort is still in the center and makes the creation of dictionaries a labor-intensive endeavor. This is particularly true when it comes to lesser-used language pairs, where finding competent bilingual speakers is already a challenge. The scope of the present thesis is to investigate to what extent state of the art language technology methods and resources can be exploited to help lexicographers construct bilingual dictionaries. As the main result of our research, we propose and evaluate a complete methodology to automatically produce "proto-dictionaries" for lesser-used language pairs, and show that not only our method presents economical advantages by reducing the amount of human effort needed, but it also addresses several methodological difficulties or inconsistencies with success. Moreover, a customizable dictionary query interface is presented whose features make the proto-dictionaries useful not only for lexicographers, but also for end users: Language learners as well as advanced users, such as professional translators.

The methodology we propose relies on parallel corpora as a resource and exploits word alignment to produce translation candidates. We suggest that conditional probabilities, as estimated via the word alignment process, can be conceived as modelling translation relation between SL and TL units. We argue that, from a theoretical viewpoint,

conditional probability is very well fit to *quantify translation relation* because of its ability to capture the *gradual, asymmetrical* nature of it. Moreover, the method implicitly produces a partition over the senses of the SL lemma by assigning different translation candidates to different contexts. This implicit feature addresses an omnipresent problem in lexicographic work: The subjective nature of word sense definitions and distinctions. We show via different word sense disambiguation experiments that the more we rely on contextual information, the less sense characterizations are prone to subjectivity.

The second part of the thesis deals with the methodology of creating a proto-dictionary from a list of translation candidates obtained from the corpus. Proto-dictionaries result from a filtering of candidates based on three parameters: SL and TL lemma frequencies and the conditional probabilities. These parameters were set via a series of experiments and evaluations, at the end of which a Hungarian-Slovenian and a Hungarian-Lithuanian proto-dictionary were available. On top of that, we confirmed that our method allows to easily produce the *reversed* dictionaries, which gave us the Slovenian-Hungarian and the Lithuanian-Hungarian resources as well. Contextual information is also provided in proto-dictionaries to illustrate the current sense and inform users on how to use the candidate lemma in a sentence. A dictionary query interface makes it possible for the user to adapt the proto-dictionary to their specific needs.

Although alignment units correspond to words by default, the method allows to create links between larger, eventually syntactically related expressions. The first results in this direction are presented and discussed in the thesis.

Since we are dealing with lesser-resourced languages, the scarce availability of parallel corpora results in a loss in the coverage of the vocabulary. This is the most important limitation of our methodology, for which we propose two solutions. First, we suggest that certain parameter settings can improve coverage without significant loss in

terms of precision. Second, we propose a method to extend the vocabulary by automatically detecting semantically similar clusters of words. The further elaboration of these topics constitutes our main direction for future work.

Acknowledgements

My heartfelt thanks go

...to my supervisor, Tamás Váradi, for accepting to supervise my work, for giving me the opportunity to develop my ideas under excellent working conditions, to attend many conferences and trainings throughout the years and for his valuable remarks.

...to András Kornai and to Gábor Prószéky for their effort to evaluate my work, as well as to Ferenc Kiefer, Beáta Gyuris, László Hunyadi, András Komlósy and Péter Rebrus for participating in my thesis committee.

...to the EFNILEX project partners: to František Čermák, John Simpson and Jolanta Zabartskaitė for their thorough and insightful comments, which have contributed to the outcome of this work.

...to Rūta Marcinkevičienė and Andrius Utka for making available the morphologically annotated and disambiguated Lithuanian texts of the Lithuanian National Corpus and that of the Lithuanian-English Parallel corpus. We are also grateful to Iván Mittelholcz for collecting even more Lithuanian texts by contacting publishers.

...to Bence Sárossy for gathering Slovenian texts from authors and translators and for the evaluation of the Hungarian-Slovenian proto-dictionary.

...to Beatrix Tölgyesi and Justina Lukaseviciute for the evaluation of the Hungarian-Lithuanian proto-dictionary; to Piroska Lendvai and to Annemieke Hoorntje for the evaluation of the French-Dutch proto-dictionary.

...to Eric Villemonte de la Clergerie for parsing the French part of the Dutch Parallel Corpus.

...to Bálint Sass for the extraction of verb centered constructions from the Dutch Parallel Corpus.

...to Miklós Rédei, who first drew my attention to the interesting notion of probability.

...to the outstanding teachers and linguists I was taught by at the Dept. of Theoretical Linguistics, ELTE, especially to Péter Rebrus and to András Komlósy. I am also very grateful to Kinga Gárdai, who always helped to resolve the omnipresent operative issues.

...to my colleagues at the Dept. of Language Technology, HAS, in particular to Judit Kuti, Csaba Oravecz and Bálint Sass. I benefited a lot from our collaboration.

...to Guillaume Jacquet for introducing me to the world of unsupervised meaning disambiguation. A considerable part of the research presented here was inspired by his former investigations.

...to Ágnes Sándor for providing insightful comments on the present thesis and on life in general. Our discussions during our afternoon walks were a great inspiration to me.

...to Dávid Takács, who greatly contributed to the work described in the present thesis. The program used in the automatic detection of near-synonym adjective classes was implemented by him. He also contributed to the interpretation of the results. He took part in the automatic extraction of parallel verbal structures from the Dutch-French parallel corpus, as well. Moreover, he designed and implemented the Dictionary Query System. Throughout our collaboration he turned out to be an excellent programmer with valuable linguistic insights. His friendship meant a lot to me, too.

...to Kata Gábor, who deeply influenced my research interest. Having the opportunity to work with her was the greatest motivation and inspiration for me during the past decade. Her sensibility towards

research issues, her ability to see the details and her professional commitment were indispensable in formulating new ideas, which show up in this dissertation, as well. I am also grateful to her for spending considerable amount of time and effort to render the thread of thoughts of the present thesis more coherent.

...to my friends, especially to Ági, Biga, Gergő and Linda.

...to my entire family: to my mother, Orsi, who was always there when I needed her. To my brothers, Dávid and Herki, who were nearly always there, when I needed them. To my daughters, Kamilla and Veronika, for their premature ability to sleep through the nights, which, without a doubt, was essential in the completion of the present dissertation. To my husband, Zoli, who supported everything.

...to my father, Gábor, for his desire to discover the hidden simplicity behind the complexity of phenomena. He showed me how interesting the world really is and what a good place it is to be. This dissertation is dedicated to him.

Contents

List of Figures

LIST OF FIGURES

LIST OF FIGURES

List of Tables

LIST OF TABLES

1

Introduction

1.1 Motivation

The oldest known bilingual dictionaries were Sumerian–Akkadian word lists dated back roughly 2300 BCE[1]. Although the history of bilingual lexicography spans at least over 4300 years and bilingual dictionaries form an integral part of our everyday life, no extensive consensus has been reached so far with regard to some of the fundamental notions of lexicography, such as meaning and translation relation.

Nevertheless, in spite of the obscurity of these notions, enormous amount of effort and time are invested into the production of new dictionaries and to keep up-to-date old ones. Compilation of bilingual dictionaries requires widespread expertise, including the appropriate knowledge of the relevant languages and of lexicography alike.

The motivation of this PhD thesis is twofold: Our basic objective is to find a *cost-effective method* which is able to facilitate lexicographers' work and to investigate *to what extent it meets the expectations put forward by meta-lexicography toward the editing principles of a dictionary.*

The increasing amount of language data available in electronic format and the appearance of new data processing techniques open up new perspectives even for

[1]See, for instance, `http://www.historyofinformation.com/expanded.php?id=2456`

such a widely studied field as lexicography. Therefore, the scope of this thesis is to explore to what extent language technology methods are able to facilitate the creation of bilingual dictionaries so that it correspond to the above requirements.

1.2 Research goals

In parallel with the increasing body of research in the field of machine translation, much attention has been paid to the automatic compilation of bilingual lexicons. There are several strategies to obtain translation pairs to create new bilingual lexicons or to extend existing ones. Nevertheless, these techniques were invented prevalently for the purpose of machine translation or for other natural language processing tasks, thus, somewhat surprisingly, current lexicographic practice does not reflect the achievements of these related fields. As far as we know, there are a few research projects (e.g. Lindemann, 2013) investigating the potential in automatic bilingual dictionary compilation, but there are no larger scale projects exploiting these techniques we are aware of. The aim of the present dissertation is to explore to what degree these relatively novel procedures are compatible with the usual practice and underlying theories of lexicography.

The main contribution to the research field The most important contribution of the present thesis to natural language processing and to multilingual computational lexicography is that THE AUTOMATIC LEARNING OF TRANSLATION PAIRS ON THE BASIS OF PARALLEL CORPORA USING CONDITIONAL PROBABILITIES is not only a cost-effective way of generating bilingual lexicons, but it is able to clarify the notion of translation in terms of quantifiable data, moreover, this conception of translation is compatible with the usual interpretation of translation in bilingual lexicography.

Conditional probability as translation relation The use of conditional probability as translation relation is usually motivated by its performance in the field of NLP, that is, there is no real lexicographical or translational insights behind the selection of this very measure to extract translation pairs from bilingual texts. On the other hand, bilingual lexicography and translatology came

up with a handful of expectations concerning translation relation resulting in a fragmented and somewhat obscure conception of it. Consequently, the novelty of the present dissertation with regard to translation relation comes from connecting the two research fields: It is claimed that conceiving translation relation as conditional probability is able to merge the various types of translation relation by turning it into a quantifiable notion. Therefore, by conceiving of translation relation as conditional probability we are able to grasp essential insights of bilingual lexicography.

Methodological notes It is important to note that the title of the dissertation might be somewhat misleading as it may imply that various language technology methods are empirically tested and the results are compared in some dictionary building tasks. Albeit—not surprisingly—this empirical approach seems to be prevalent in the natural language processing literature, we have decided to follow a more deductive methodology here. Accordingly, we have laid greater emphasis on the clarification of the basic notions in bilingual lexicography, such as meaning and translation relation. Thus, by the end of the first part of the dissertation it will be confirmed that the notion of conditional probability is suited to define translational relation and not merely because it yields better results but because it fits into the conceptual framework of translation in the context of bilingual dictionaries.

Therefore, the present PhD thesis is made up of two main parts: In the first part we start out from a general definition of dictionaries, and give a theoretic overview on the fundamental notions of bilingual lexicography (ch. 2-4). The second part puts the findings achieved by then into practice and investigates the pros and cons of the selected technique (ch. 5-8).

General definition of bilingual dictionaries Our investigations start out from the most general definition of bilingual dictionaries: Let us suppose that a dictionary is a relation in the mathematical sense. More precisely, let A be the source language (SL) vocabulary and B the target language (TL) vocabulary. In this case, the dictionary is $\rho \subseteq A \times B$, i. e. it is a set of ordered pairs (a, b) so that $a \in A$ and $b \in B$.

Encoding and decoding dictionaries Nevertheless, this general definition of bilingual dictionaries is excessively oversimplifying, as it is unable to reflect essential divergencies in their editing principles, which are obviously affected by the purpose of the dictionaries. From a user perspective, *encoding* (active) and *decoding* (passive) dictionaries should be distinguished. Encoding dictionaries provide speakers of the SL with information on how to express themselves in a foreign language. As opposed to encoding dictionaries, decoding dictionaries help the speakers of the TL to understand a foreign language, which is the SL in this case.

The ultimate questions On the basis of these definitions the present dissertation seeks answers to the following questions:

(1) What type of entities constitute A and B? How could these entities be characterized?

(2) Which expectations should ρ meet to be able to serve as translation relation?

As throughout this thesis we concentrate on the creation of encoding dictionaries[1], we also have to consider which additional requirements such a dictionary has to suffice, namely:

(3) Does the creation of encoding dictionaries impose additional constraints on A and B?

(4) Does the creation of encoding dictionaries impose additional constraints on ρ translation relation?

Proto-dictionaries After the clarification of the basic issues, the theory was put into practice and a method was selected, which corresponds to the answers to the above questions. Thus, dictionaries were built for two less-resourced language pairs (Hungarian-Lithuanian and Hungarian-Slovenian). We also created dictionaries for two additional language pairs: Hungarian-English and Dutch-French.

[1]This choice is motivated by the observation that encoding dictionaries are more difficult to compile and they can be converted into decoding dictionaries.

These resources were designed primarily for lexicographers' to facilitate their work when creating bilingual dictionaries. The automatically generated bilingual resources will be referred to as *proto-dictionaries* henceforward. Note that proto-dictionaries were created fully automatically, thus, some suitable heuristics were introduced to filter the results. Three parameters were used for that purpose, the SL and TL lemma frequencies, and the corresponding conditional probabilities. We investigated to what extent proto-dictionaries meet the expectations put forward in the first part of the thesis and address the relating difficulties.

Coverage One essential issue that should be overcome concerns the coverage of the resulting dictionaries. Note that this problem emerges principally in the case of the less-resourced language pairs, which are the focus of the present research.

Multi-word expressions (MWEs) Though MWEs play an important part in the case of encoding bilingual dictionaries, which primarily aim at enabling the user to produce idiomatically correct TL text, the selected technique does not treat MWEs in itself. Thus, we should investigate whether the original method could be complemented with a module that recognizes MWE translation pairs.

Dictionary Query System (DQS) Finally, we intended to make proto-dictionaries readily available even for end-users. For doing so, a Dictionary Query System was designed and implemented that is able to take advantage of the novelties of proto-dictionaries and compensate for their shortcomings.

In Section 1.3 the theses are listed, they are indicated with roman numerals.

1.3 Theses

1.3.1 General result

(I) Although word alignment techniques on parallel corpora are widely used for the purpose of machine translation, until recently they have been hardly—if at all—used in lexicographic projects. The main finding of this thesis is that THE AUTOMATIC LEARNING OF TRANSLATION PAIRS ON THE

BASIS OF PARALLEL CORPORA USING CONDITIONAL PROBABILITIES is particularly apt for lexicographic purposes for *theoretical*, *practical* and *economical* reasons.

1.3.2 Theoretical results

(II) From a theoretical perspective it is claimed that *the automatic learning of translation pairs on the basis of parallel corpora using conditional probabilities* has certain benefits both over traditional and corpus-based lexicography, inasmuch it provides answers to some questions inherently present in both methodologies. The proposed method is able to define some of the fundamental notions of lexicography in terms of quantifiable corpus data.

(III) Albeit the general view in lexicography takes form-meaning pairs as the atomic building blocks of dictionaries, it is argued that, if the proposed method is used, word forms (in the sense of lemmata) may serve as the basic units for bilingual encoding dictionaries. That is, in this case we do not have to address the rather difficult problem of how to characterize meanings of word forms, as it falls back to the problem of how to characterize mere word forms in a bilingual dictionary.

Based on the literature it was found that the notion of translation relation is inhomogeneous inasmuch several sub-types of it may be distinguished and, at the same time, translation relation tends to be asymmetric and gradual.

(IV) It was investigated how the symmetry of translation relation can be interpreted. We found that if $\rho \subseteq A \times B$, i.e. if ρ is a relation in the mathematical sense, then the symmetry of ρ translation relation is best interpreted as ρ being an invertible function mapping from the SL vocabulary A to the TL vocabulary B. It was also found that this definition is consistent with the cases when "symmetric translation relation" is exploited in practical lexicography, such as when designing reversible dictionaries or when applying the hub-and-spoke model.

(V) We accept that the translation relation is best to think of as a gradual notion, and we propose that we should be able to compare two possible translations of a given SL expression and to select the better one. That is, *translation relation should be quantifiable*. The strength of the translation equivalence could be measured by the number of contexts in which a TL expression appears as translation. This claim corresponds to the fact that perfect translation equivalents are defined as translation pairs that are interchangeable in every context, while on the other end of the scale, contextual translations appear only in a rather constrained set of contexts. Accordingly, we do not consider perfect translational equivalence (cognitive equivalence) and contextual equivalence separate types of translational equivalence, instead, we propose that they are the two ends of the very same scale.

(VI) In our view, instead of mathematical relation, ρ translation relation should be conceived of as *conditional probability*, $P(b|a)$, which gives an estimation of how many times the occurrences of $a \in A$ are translated as an occurrence of $b \in B$ on the basis of sentence aligned parallel corpora. We claim that conditional probability is a suitable mathematical construction to represent and to quantify over translation relation for multiple reasons. First, as opposed to the binary notion of mathematical relation, conditional probability is able to reflect the gradual nature of translation relation. Secondly, conditional probability captures the fact that translation relation tends to be asymmetric, as well. Thirdly, this mathematical construction is also able to reflect that translation relation is symmetric in the case of perfect translational equivalence.

Since our basic objective is to facilitate the creation of encoding dictionaries, we also investigated whether such dictionaries impose additional constraints on the traditional lexicographic methodology, i.e. when bilingual dictionaries are the result of translating some SL form-meaning pair list (these resources will be referred to as SL sense-inventories henceforward).

By definition, in an encoding environment we are aware of the meaning of the SL expression and want to find the contextually best translation for it, i.e. the one that produces an idiomatically correct translation when put into TL contexts. That is, the SL sense-inventory should enable lexicographers to anchor the form-meaning pairs in the SL sense-inventory to contexts with high confidence.

(VII) We found that a suitable SL sense-inventory should exhibit certain properties to enable the annotators to achieve high agreement on the annotation task. Namely,

(i) First, it has to comprise abundant contextual information that enable annotators to select the appropriate meaning on the basis of explicit distributional information.

(ii) Each SL headword in the sense-inventory should be characterized in a way that each occurrence of the given headword could be clearly assigned to a unique meaning. That is, there is no such occurrence that may be assigned to two different meanings.

(iii) It is also presupposed that meanings in the SL sense-inventory are non-overlapping entities.

That is, if our presuppositions hold, a suitable SL sense-inventory for a high-quality encoding dictionary should be characterized in a way that *the various meanings of a word form, i.e. lemma, create a partition in the mathematical sense over the occurrences of that word form.*

(VIII) Unfortunately, such neatly characterized data-base usually is not available. Therefore, in the absence of such a sense-inventory an alternative way of creating high-quality encoding dictionaries should be found. Another alternative is that we disregard word senses and try to retrieve translations by creating a partition of TL word forms directly over a given SL word form. The conditional probability $P(b_i|a)$ creates a partition of occurrences of the possible translations b_i on the set of occurrences of the SL word form a. Moreover, translation pairs of the form a-b_i are linked on the basis of their natural contexts. Thus, conceiving of translation relation

as conditional probability turned out to be suitable to create high quality encoding dictionaries.

Although automatic estimation of conditional probabilities on the basis of parallel corpora provides answers to some of the questions of bilingual lexicography, one serious difficulty is raised by the scarce availability of parallel texts, which has a serious effect on the size of the resulting proto-dictionaries. Note that this problem emerges principally in the case of less-resourced language pairs, which play a central part in our research. Proto-dictionaries are the results of filtering based on three parameters: SL and TL lemma frequencies and the conditional probabilities.

(IX) We found that a cascaded filtering technique significantly increases the coverage of the resulting proto-dictionaries: In the case of more frequent lemmata even lower values of conditional probabilities may yield correct translations. Hence, fine-tuning the parameters results in bigger proto-dictionaries.

1.3.3 Practical results

The main practical finding of this thesis is that a suitable dictionary query system (DQS) is capable of rendering proto-dictionaries a useful resource for not only lexicographers but for end-users, too. Therefore, a DQS was designed and implemented that displays some novel features compared to traditional dictionaries. The practical results of the present thesis concern the novelties of the dictionary query system.

(X) (a) The most important novelty of DQS is that it is *customizable*. That is, the users can select the sub-part of the proto-dictionary that suits most their needs. We think that various parameter settings match well different user needs. The scope of various users may span from novice language learners to professional translators. On the one end of the scale, keeping the most frequently occurring translation pairs

results in low-coverage but high-precision proto-dictionaries, which are appropriate for novice language learners. On the other end of the scale, selecting more relaxed parameters generates a proto-dictionary with a greater coverage but with a lower precision. Such a proto-dictionary may suit the needs of professional translators who may be interested in special uses of words. At the same time, they can easily catch wrong translations, therefore, low precision does not pose a problem for them. Thus, the customizability feature of the Dictionary Query System supports various user scenarios.

(b) Representing translation relation as conditional probability makes it possible to rank translations according to how likely they are. Presenting translation candidates in such a way is an obvious advance compared to the usual ordering techniques applied in bilingual dictionaries.

(c) As opposed to traditional dictionaries, the DQS gives a hint on the scope of usability of the translation based on some very simple heuristics. To know whether the TL word may show up in a more restricted or a more general set of contexts than the SL word is essential in the case of encoding dictionaries.

1.3.4 Economical results

From an economical perspective it is claimed that the proposed method facilitates the cost-effective generation of bilingual dictionaries for lesser-used languages.

(XI) Owing to the data-driven nature of the proposed technique, the amount of human effort needed to compile bilingual dictionaries is significantly decreased. The extraction method and the DQS are language-independent, thus, only the language dependent resources and tools need to be collected again when preparing dictionaries for new language-pairs. Once the required resources and tools are collected, the generation of the reversed dictionary is a straightforward process.

1.4 Framework

EFNIL The scope of this thesis is the work that was accomplished in the framework of the project EFNILEX between 2008 and 2012. EFNILEX is a lexicographical project launched by the European Federation of National Institutions for Language (EFNIL). The mission of the organization is formulated on its website[1] as follows:

> All the member states of the European Union have institutions whose role includes monitoring the official language or languages of their country, advising on language use, or developing language policy.
>
> The European Federation of National Institutions for Language provides a forum for these institutions to exchange information about their work and to gather and publish information about language use and language policy within the European Union.
>
> In addition, the Federation encourages the study of the official European languages and a coordinated approach towards mother-tongue and foreign-language learning, as a means of promoting linguistic and cultural diversity within the European Union.

EFNILEX The official website of the project is available at `http://www.efnil.org/projects/efnilex`. As described in the project home page[2]:

> The European Union wishes to contribute to policies aimed at the preservation and strengthening of the multilingualism of Europe and the plurilingualism of its citizens. This goal implies that as many languages as possible should be:
>
> (i) used in as many domains, functions and situations as possible;
>
> (ii) involved in cross-border European and global communication and information exchange, e.g. through the internet;

[1] `http://www.efnil.org/`
[2] http://www.efnil.org/projects/efnilex/description-of-efnilex

(iii) learned and used by as many users as possible, both native and non-native speakers.

The above objectives imply that special attention had to be paid to lesser used language pairs, where—due to low demand—dictionaries of appropriate size and quality are hardly available. The reason for this is that the creation of such dictionaries does not pay off for publishers. The targeted size of the dictionaries is between 15,000 and 25,000 entries covering every-day language vocabulary.

Regarding the limitations of material background, our primary objective was to decrease the amount of lexicographic labour needed to compile a bilingual dictionary. Nevertheless, in the framework of the present research we had had no opportunity to collaborate with lexicographers in the long-run and test their real needs. Thus, we did not have an exact idea of how much time it would take to convert the automatically generated resources into full-fledged dictionaries. While this is an obvious shortcoming of the project so far, we tried to compensate for the deficiency and come up with bilingual resources that are not only useful for lexicographers but for end-users without any post-editing phase, too.

1.5 Structure of the thesis

Chapter 2 Starting out from the hypothesis that a bilingual dictionary can be conceived of as a relation $\rho \subseteq A \times B$, in Section 2 we focus on the various approaches to the characterization of A, as the properties of A greatly influence the quality of the resulting dictionary. Three different methodologies are distinguished according to their relation to corpus data. First, traditional lexicography is considered, which is prevalently based on the linguistic intuition of lexicographers. Then, two corpus-oriented approaches are discussed: Corpus-based and corpus-driven lexicography, yielding the conclusion that high-quality monolingual sense-inventories are based on corpus data. The related theoretical considerations are discussed, as well.

Chapter 3 Chapter 3 focuses on the process of translation. Beside translation, linking, an alternative dictionary building method is introduced. In the

case of linking the target sense-inventory is built independently from the source sense-inventory. In the next step, the relevant elements of the source and target sense-inventories are linked resulting in a bilingual dictionary. The properties of ρ translation relation are investigated both in the case of translation and in the case of linking. It was found that albeit the notion of translation relation is inhomogeneous, it tends to be asymmetric in some sense and gradual. Nevertheless, in the case of linking, translation relation is expected to be symmetric, lexicographically speaking, which means that it should be an invertible function. Hub-and-spoke model, which is also covered in this chapter, raises the same expectation toward translation relation. By the end of Chapter 3 our expectations toward an automatically attained translation relation will be formulated and the possible approaches to compiling a bilingual dictionary are characterized. This is basically a summary of Chapters 2 and 3.

Chapter 4 In Chapter 4 the additional requirements are investigated that an encoding dictionary impose on vocabularies A and B and ρ translation relation. It is argued that high-quality encoding dictionaries are either based on neatly characterized SL sense-inventories, or we may disregard word senses and try to retrieve translations by creating a partition of TL word forms directly over a given SL word form. In Chapter 4 it will be also shown that the notion of conditional probability estimated on the basis of parallel corpus corresponds to the second approach.

Chapter 5 Once the main direction was selected, the special techniques should have be given a closer look so that the most appropriate methods could be selected for our purposes. Accordingly, Chapter 5 focuses on various sentence and word alignment techniques and discusses their pros and cons. As a result, we have decided to use Hunalign (Varga et al., 2005) to align sentences, and GIZA++ (Och and Ney, 2003) to extract translation pairs from the sentence aligned parallel corpora.

Chapter 6 In the next step the selected alignment techniques were applied to create Hungarian-Lithuanian (and vv.) and Hungarian-Slovenian (and vv.) proto-dictionaries. These are proof-of-concept experiments aiming to confirm the viability of the proposed approach and to explore the related difficulties. One of

the main difficulties, which should be addressed, is that the word alignment algorithm does not handle MWEs in itself. Both parallel corpora were converted into XML-format with uniform morphosyntactic annotation so that MWE extraction and alignment could be handled alike in the case of all language pairs.

Chapter 7 Chapter 7 explores to what extent the suggested method is able to handle MWEs through the alignment of verbal structures. Verbal structures are extracted using an algorithm described in Sass (2011), then the extracted verbal structures were merged in both sides of the parallel corpora. The alignment algorithm treated the MWEs as if they were one-token expressions in the rest of the workflow.

Chapter 8 Finally, a dictionary query system was designed and implemented that is able to compensate for the drawbacks of the selected method and to extend its advantages even further. Although no user case study has been performed, according to our expectations a proper query system is able to render the automatically generated resources useful for not only lexicographers but for end-users, too. The automatically generated online dictionaries are available at `http://efnilex.efnil.org`.

Chapter 9 Chapter 9 summarizes the results and determines the future research directions.

2

Compiling the Headword List

> [...] identifying and describing word senses is a major part of what lexicographers are expected to do. However, there is little agreement about what word senses are (or even whether they exist). Lexicographers are therefore in the position of having to describe something whose nature is not at all clear.
>
> ——— Atkins and Rundell (2008, p. 264)

2.1 Introduction

In the previous chapter the dictionary was defined as a relation $\rho \subseteq A \times B$, where A and B was defined as the SL and TL vocabulary, respectively. The present chapter focuses on the main characteristics of the SL vocabulary A in various lexicographic projects.

The dictionary building process ordinarily begins with the construction or selection of the SL vocabulary A. As the quality of the monolingual SL word lists greatly influence the quality of the resulting dictionaries (cf. Atkins and Rundell, 2008), the editing principles behind the SL word lists and the related methodological issues will be given a closer look here.

In Section 2.2 the traditional dictionary building process will be presented. In Section 2.3 the main building principles of the SL database will be clarified. Their relation to natural language data serves as the basis for classification. *Traditional*, *corpus-based* and *corpus-driven* approaches will be distinguished.

2.2 The Dictionary Building Process

2.2.1 The task

Form-meaning pairs The task of writing a bilingual dictionary might be conceived of as assigning the relevant language units of the TL to the relevant language units of the SL. As for the nature of these language units, they ideally can be characterized as form-meaning pairs. It is rather easy to realize that it is impossible to translate word forms without meaning, that is, meaning is inherently present during the translation process. Hence, the atomic linguistic units in a dictionary should be form-meaning pairs.

Lexical units (LUs) Form-meaning pairs are referred to as lexical units henceforward. As Atkins and Rundell (2008) assert:

> A headword in one of its senses is a lexical unit (or LU) [...]. LUs are the core building blocks of dictionary entries. (p. 162-163)

Thus, the dictionary building process involves:

(1) Including SL LUs into the dictionary Comprising not only bare word forms, the SL word list has to be characterized so that the corresponding meanings be attached to the relevant word forms. To the extent to which natural language data is relied on during the characterization process one can distinguish three different paradigms in lexicography viz., traditional, corpus-based and corpus-driven approaches.

(2) Finding the most appropriate TL LU(s) for every SL LU More frequently TL LUs are the results of translation of SL LUs, but occasionally

> TL LUs are characterized independently from SL LUs. In this case the task of translation can be conceived of as to find the most appropriate pairings between SL and TL LUs. This latter process is referred to as *linking*[1].

(1) will be elaborated more in Section 2.3, while (2) will be discussed in Chapter 3 in more detail. In fact, these two basic steps constitute two slightly interrelated dimensions of dictionary building (cf. Figure 3.6).

The best translation Either created by means of translation or by means of linking, dictionaries should comprise the best TL LUs for an SL LU. However, this task is not at all straightforward. As Atkins and Rundell (2008) notes:

> The perfect translation – where an SL word exactly matches a TL word – is rare in general language, except for the names of objects in the real world (natural kind terms, artefacts, places, etc.) (p. 467).

Since perfect translations are restricted to certain semantic domains, the next best translations should be included into the dictionaries, where perfect translations are not available. This in turn raises the question of how the next best translation can be defined and retrieved[2]. This topic will be elaborated more in Section 3.

Encoding dictionaries Encoding or active dictionaries are dictionaries that provide speakers of the SL with information on how to express themselves in a foreign language. In this case not only the best translation has to be found but relevant contextual information has to be also provided to give hints on how a TL expression should be used correctly. Thus, in the case of encoding dictionaries a further phase must be added to the dictionary building process:

(3) Providing relevant contextual TL information to help the SL speaker to find the best TL expression for the given situation.

[1] Linking, an alternative bilingual dictionary building method, will be introduced more closely in Section 3.3.

[2] A terminological note is in order here: The term perfect translation will be used for translations that suit every TL context. Next best translations are those that are not perfect but are the best ones among the available translations. Throughout this thesis we will usually use the term "best translation" referring either to the perfect translation, if there is one, or the next best translation in the other cases.

Decoding dictionaries As opposed to encoding dictionaries, decoding dictionaries help the speakers of the TL to understand a foreign language (SL). According to Melčuk (2006, p. 232-233) encoding dictionaries are also suitable for decoding, as encoding requires much more linguistic annotation than decoding. The underlying reason is that context might offer useful information when understanding, while text production requires exact knowledge on the appropriate expression conveying the right meaning and fitting the context at the same time.

2.2.2 The building process

According to Atkins and Rundell (2008) the process of building a bilingual dictionary is threefold:

Analysis stage A relevant headword list of the source language has to be compiled. An inherent part of this stage is making decisions on which alternative senses are to be included in the SL side of the dictionary. The exploitation of existing monolingual dictionaries, wordnets or monolingual corpora might facilitate the compilation of such a headword list. When starting out from monolingual corpora the production of a headword list is referred to as the *analysis stage*.

However, throughout this thesis we use this term in a wider sense: It always denotes the stage of compiling the SL headword list no matter what kind of resource is relied on, be it a monolingual explanatory dictionary, a wordnet or a monolingual corpus.

Transfer stage During this stage the linguistic units making up the headword list are translated into the TL. However, it is important to keep in mind, especially during the creation of an encoding dictionary, that the translated LUs will be used in discourse. Thus, the safest translation has to be obtained, i.e. the one that fits the most TL contexts, and possibly ranked at the first place in the relevant entry.

Synthesis stage In the *synthesis stage* of the dictionary building process the final entry will be produced through transforming the translated database records into a series of finished entries for a specific bilingual dictionary.

Figure 2.1 illustrates the workflow of dictionary building, where the SL LUs are characterized on the basis of corpus texts. Alternatively, instead of corpora other initial sense inventories might be relied on, too.

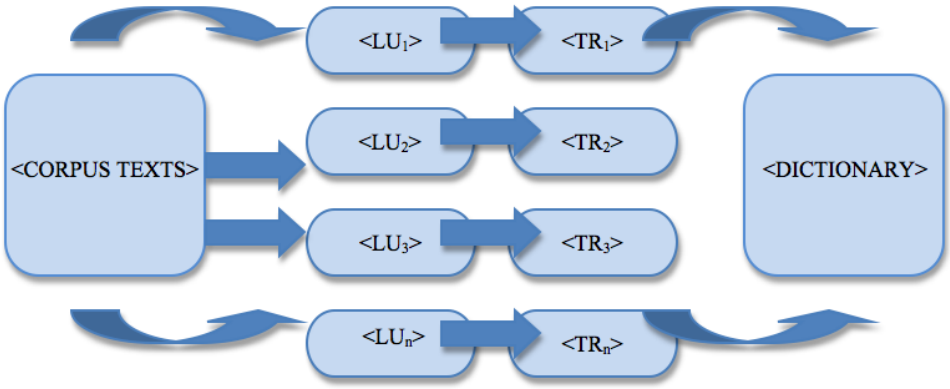


Figure 2.1: The workflow of dictionary building - At the first stage the source language LUs are characterised – analysis stage. In the next phase these LUs are translated – translations stage. Finally, the translated LUs (TRs) are compiled so that they could be included in a dictionary – synthesis stage.

Less resourced languages and the analysis stage Especially in the case of lesser used languages—where the resources are rather limited—starting out from a source language explanatory dictionary seems to be a reasonable choice. For instance, the Lithuanian-Hungarian Dictionary (Bojtár, 2007) relied on the Lithuanian Explanatory Dictionary (Keinys et al., 1993) as the source language headword list. As the initial monolingual database determines both the SL and the TL vocabulary of the bilingual dictionary, therefore, it is essential to be aware of the main properties of the initial source language database.

In what follows, an overview of such monolingual databases is given, classifying them on the basis of their relation to corpus data.

2.3 Sense Inventories and Language Data

As Jezek and Hanks (2010) claim:

> [Monolingual explanatory] Dictionaries describe the vocabulary of a language. For any given word, a good dictionary tells its readers the ways in which that word typically contributes to the meaning of an utterance, the ways in which it combines with other words, the types of text that it tends to occur in, and so on. Clearly it is desirable that this account is reliable. A reliable dictionary is one whose generalizations about word behaviour approximate closely to the ways in which people normally use (and understand) language. (p. 587)

Taking this citation as our starting point in Subsection 2.3 we seek answers to the following questions:

(1) How reliability is guaranteed by each of these approaches?

(2) To what extent do they describe the possible combination patterns with other words?

(3) How a word is supposed to contribute to the meaning of a sentence?

As we will see soon, each lexicographical methodology has its own—sometimes unarticulated—presumptions about the questions above.

2.3.1 Traditional lexicography

For the sake of reliability lexicographers have to take advantage of some sort of linguistic evidence when compiling a dictionary. In the case of traditional dictionaries they tend to make use of their own mental lexicon, that is, of their linguistic intuition. However, one major problem with such an approach is that it might easily lead to an unbalanced description of the relevant linguistic phenomena, even if lexicographers strive to include all possible expressions of a language and all possible uses of those expressions into the sense-inventory.

According to Atkins and Rundell (2008):

> Dictionaries generally divide polysemous words into numbered senses. A conventional dictionary entry consists of a list of 'neatly separated, consecutively numbered lexical meanings' (Geeraerts, 2006, p. 198). [...] This convention rests on two (unarticulated) assumptions:
>
> 1. There is a sort of Platonic inventory of senses 'out there' (so if the dictionary says word W has N senses, it can't possibly have $N-1$ or $N+2$ senses)
>
> 2. Each sense is mutually exclusive and has clear boundaries (so if a specific occurrence of a word is assigned to sense A, it cannot also belong to sense B. (p. 271-272)

If the assumptions above were right, i.e. if meanings would be completely intersubjective[1], non-overlapping entities with clear-cut boundaries, the agreement among native speakers should be high when asked to select the right meaning for a word in context. Nevertheless, more experiments will demonstrate in Chapter 4 that the above assumptions pose real problems for human annotators.

In addition, in traditional lexicons contextual information usually do not play a great role entailing the fact that apart from proverbs and some collocations, possible combination patterns with other words are only poorly characterized, usually by providing only the part-of-speech category of the relevant word.

Because monolingual explanatory dictionaries serve as a basis for several bilingual dictionaries, especially in the case of lesser used languages, in Subsection 2.3.1.1 we focus on sense inventories of this type.

[1]For the present thesis we presume that there is a subjective—intersubjective scale from an epistemological perspective. In this framework, 'objective' means the same as 'intersubjective', except for the fact that the former implies an ontological standpoint, as well. Accordingly, the term 'objective' presupposes a common ontological basis, which is considered to be independent from the observer.

2.3.1.1 Monolingual explanatory dictionaries

The established practice of lexicography in the pre-corpus era was citation collecting[1], which yielded the consequence that rare words and uncommon senses were over-represented in the dictionaries.

As Hanks (2010) puts it:

> Citation readers collect citations for unusual words like triskaidekaphobia 'irrational fear of the number 13' and for unusual senses. Computers, on the other hand, do not exercise judgement. (p. 586)

Such a methodology is suitable for dictionaries that aim at providing users with explanations on uncommon words. And indeed, a native speaker typically looks up less common words or senses. True for monolingual dictionary use but not for bilingual dictionaries! This fact is hard to reconcile with the ordinary usage of bilingual dictionaries where the more common words and senses are looked up.

2.3.1.2 Wordnets

In our days the most influental monolingual semantic databases are wordnets. The first wordnet, the Princeton WordNet (PWN) was the result of the WordNet Project beginning in 1986 (Fellbaum, 1998). Since then, PWN continues to serve as the basis of wordnets for other languages (Vossen, 2004). Princeton WordNet is a manually constructed vast database, which was built to reflect the mental lexicon in a systematic way.

The particular need for such a resource emerged from both theoretical and practical considerations. From a theoretical perspective, the role of the lexicon grew bigger and bigger in the description of natural language: Instead of being merely a dustbin of linguistically uninteresting idiosyncrasies, it began to be thought of as an inherently organized repository of lexemes. At the same time, various NLP

[1]Cf. Hanks (2010) "During the past 150 years or so, the Oxford Reading Programme has been devoted to reading texts and collecting citations for the words used in them. During its heyday in the late 19th and early 20th century it involved many volunteer readers." (p. 586)

applications, such as information extraction or machine translation required a neatly organized lexicon of appropriate size, too.

Linguistic hierarchy Accordingly, there are two interrelated striking differences between wordnets and traditional monolingual explanatory dictionaries. First, as opposed to explanatory dictionaries, WordNet is inherently an electronic database. Secondly, the elements of the database are assigned an inner structure. According to Vossen (2004):

> The wordnets are seen as linguistic ontologies rather than ontologies for making inferences only. They are 'wordnets' in the true sense of the word and therefore capture valuable information about conceptualisations that are lexicalised in a language: what is the available fund of words and expressions in a language, and what words and expressions can substitute each other (Cruse, 1986).

Synsets The basic elements of the hierarchy are synonymy sets (synsets). Synsets consist of words which have the same meanings at least in certain contexts. Thus, more word forms may belong to the same synset (symonymy), while the same word form may belong to more synsets (polysemy)[1]. Beside words, wordnets also might include multi-word expressions, such as phrasal verbs or collocations. Relevant example sentences are also provided. Synonyms to be included in PWN were found on the basis of traditional dictionaries of synonymy, such us Roget's International Thesaurus (Chapman, 1977), A Basic Dictionary of Synonyms and Antonyms (Urdang, 1983) and Urdang's revision of The Synonym Finder (LaRoche and Urdang, 1981).

Besides, corpus data[2] was also considered, but merely as resource of information on lemma frequency counts. Figure 2.2 depicts a synset from Princeton WordNet 3.0.

Relations The most important relations between PWN synsets are the following:

[1]Note that although the central element of wordnets are synsets instead of headwords, as it is usual in lexicography, we had no reason to neglect wordnets here, since in both cases the ultimate building blocks are LUs.

[2]The Brown corpus (Francis and Kučera, 1979) was relied on.

introduce, present, acquaint (ID) (Definition: cause to come to know personally)
"permit me to acquaint you with my son"; "introduce the new neighbors to the community"

Figure 2.2: A synset of Princeton WordNet 3.0

Hyperonymy The generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y .

Hyponymy The specific term used to designate a member of a class. X is a hyponym of Y if X is a (kind of) Y .

Meronymy The name of a constituent part of, the substance of, or a member of something. X is a meronym of Y if X is a part of Y .

Troponymy A verb expressing a specific manner elaboration of another verb. X is a troponym of Y if to X is to Y in some manner.

Entailment A verb X entails Y if X cannot be done unless Y is, or has been, done.

The WordNet is divided into four separate subparts according to part-of-speeches: The nominal, verbal, adjectival and adverbial semantic nets were built separately.

Objection According to Hanks and Pustejovsky (2005) many of WordNet's senses are indistinguishable from one another by any criterion—syntactic, syntagmatic, or semantic—other than the fact that they happen to have been placed in different synsets. They underpin this statement with the synset *write* claiming that the 10 different senses belonging to different synsets do not represent separate meanings in the reality, rather they are different facets of the same sense. Their objection corresponds to the result of our experiment described in 4.2.2.4.

2.3.1.3 Remarks on traditional lexicography

The Frege Principle In traditional lexicography words and their meanings are the basic building blocks of language. Sentences are derived from these meaningful units in a compositional, thus, predictable way. Thus, traditional

lexicography—along with several contemporary linguistic theories (generative grammars, formal semantics)—builds strongly upon the Frege principle (cf. Janssen, 1996 and Gendler Szabó, 2013), which asserts that the *meaning of a complex expression is fully determined by its structure and the meaning of its constituents.*

Productivity In traditional lexicography part-of-speech categories tend to be considered as ultimate generalizations on word usage. Therefore, the structure of the sentences is supposed to be determined predominantly by part-of-speech information, thus, part-of-speech categories are mostly used to give hints to the user on how to combine words into sentences. In accordance with generative grammars, the grammatical rules operate on part-of-speech categories.

Meaning of LUs In this framework meanings are considered to be independent of the rules that determine how the LUs of a language may be combined. This view corresponds to the principle of the independence of syntax and semantics put forward in Syntactic Structures (Chomsky, 1957):

> grammar is autonomous and independent of meaning, and that probabilistic models give no particular insight into some of the basic problems of syntactic structure (p. 17).

From a theoretical perspective, Chomsky's principle may be interpreted as the declaration of the insignificance of lexicon in the description of grammar: The main objective of grammar is to invent or describe rules that operate on the elements of the lexicon.

Note, that the other end of the scale is Harris' distributional hypothesis (Harris, 1954), which states that words that are used and occur in the same contexts tend to purport similar meanings. This means that the meaning of the words do influence where they may appear within a sentence, that is, the arguments of syntactic rules are more restricted: Instead of part-of-speeches they are determined by certain semantic classes. As we will see in Section 2.3.2 and 2.3.3, the distributional hypothesis plays an important role in corpus-based and corpus-driven approaches and entails a more restricted notion of productivity.

Presuppositions From this it follows that traditional lexicography presumes the following:

(1) The basic blocks of language are word form–meaning pairs. Which means, that

   (i) Word forms do have meanings

   (ii) These meanings are fairly stable across different contexts

(2) Word–meaning pairs are stored in the lexicon and can be assessed by means of introspection.

From (1) it follows that contextual information does not play a great role in traditional lexicography. While (2) has two questionable implications:

(i) Everyone has a strong belief that they know exactly the meanings of the word.

(ii) This knowledge is largely alike across the members of a language community. That is, meanings are objective or at least highly intersubjective entities.

In Chapter 4 the latter two assertions will be investigated and it will be proven that both are false.

Philosophical notes Accordingly, traditional lexicography takes its root from the philosophical tradition *rationalism*. Rationalism goes back to Plato, who claimed that ideas are objective entities 'out there'. This and the stipulated notion of the main idea (the idea of 'truth' or 'good') guarantees that ideas appear largely alike for each human being. Ideas turn out to be innate entities in Descartes' philosophy. In his epistemological system intersubjective equivalency of ideas across human beings is guaranteed by God: The existence of a loving God ensures that our clear and distinct ideas correspond to the reality, therefore, the clear and distinct ideas are the same for everyone. The same thread of thought re-emerges in the theories of Chomsky, since in this case the innate conceptual structure is determined by some human-specific biological necessities.

Interestingly enough, although corpus linguists claim to be empirisits, whose methodology is grounded in linguistic data, as we will shortly see, corpus-based bilingual lexicography might retain elements of rationalism, namely when determining translational equivalency.

According to Adamska-Sałaciak (2010):

> The only exception is cognitive equivalence, whose identification by skilled bilinguals is characterised by a high degree of intersubjective agreement, which may culminate in its objectification. (p. 400)

2.3.2 Corpus-based lexicography

Preliminaries Even if lexicographers exhibit profound expertise in their field, the reliance on merely human intuition might easily lead to an unbalanced description of the relevant linguistic phenomena. With the appearance of electronically available corpora an alternative approach emerged offering great amount of language data to support lexicographers' work.

The role of context Moreover, in accordance with Harris' distributional hypothesis (cf. page 25), through providing an essential source of distributional information, they can contribute to the characterization of prevalent word senses. This naturally entails a different view on meanings: Instead of being pre-existent Platonic entities, meanings are grasped through usage, relying on the contexts in which a word may occur. Thus, context and usage plays a much more important role in corpus-based lexicography, than before it.

As Hanks (2010) notes:

> Words have meanings—or rather, strictly speaking, they have the potential to make meanings when put into context—and they are associated with particular sets of syntagmatic patterns, which can be discovered through painstaking corpus analysis. (p. 581)

In the present section some widely known corpus-based lexicographic projects are described.

2.3.2.1 COBUILD

In 1983 John Sinclair and his colleagues started working on the first edition of the COBUILD dictionary. They discovered that many of the generalizations made in pre-corpus dictionaries, though plausible, were not quite right. The corpus they used is the continously growing Bank of English (650 million words nowadays, part of the Collins corpus). As the project's home page[1] indicates:

> When the first Collins cobuild Dictionary of English was published in 1987, it revolutionized dictionaries for learners, completely changed approaches to dictionary-writing, and led to a new generation of corpus-driven dictionaries and reference materials for English language learners.

Frequency information, for example, allowed them to rank senses by importance and usefulness to the learner (the most common meaning should be put first); and the corpus highlights collocates, information which had only been sketchily covered in previous dictionaries. Under Sinclair's guidance, his team also developed a full-sentence defining style, which not only gave the user the sense of a word, but showed that word in grammatical context.

According to Carter (1998) the innovations of the first COBUILD Dictionary include:

(1) Citations are examples of real English and do not involve made-up examples.

(2) Linguistic and stylistic differences between written and spoken usage, and British-English and American-English are stored separately.

(3) Relative frequencies of occurrences are indicated.

(4) Senses of polysemious words are ordered based on their frequencies in the corpus.

(5) Information on the main colligational and collocational properties of a word is also provided.

[1] http://www.mycobuild.com/about-john-sinclair.aspx

2.3.2.2 Explanatory combinatorial dictionaries (ECD)

ECD was proposed in the late 1960s by Žolkovskij and Mel'čuk. It also includes many of Apresjan's ideas. Only a few ECDs are currently available in print. See Mel'cuk and Zolkovskij (1984) for Russian and Mel'čuk et al. (1984), Mel'čuk et al. (1992), Mel'čuk et al. (1996) and Polguère (2000) for French. A dictionary of Spanish collocations—Diccionario de colocaciones del español—has been also developed (Ramos, 2005). Melčuk (2006) characterizes ECDs as follows:

> In a nutshell, the ECD is an active phrasal dictionary, based on the semantics of the LUs treated and stressing their restricted cooccurrence; its unit of description is a Lexical Unit, that is, roughly, a word or a set phrase taken in a particular sense (rather than a polysemous word, as in all current dictionaries). (p. 242.)

The ECD is based on the Meaning-Text Theory (MTT), where the ECD constitutes an integral part of the semantic module.

As opposed to generative grammars, in MTT the lexicon is claimed to be superior to the grammar itself in the course of linguistic description. According to Melčuk (2006):

> Most current linguistic theories view a linguistic description of a language as a grammar; a lexicon is taken to be an indispensable, but somehow less interesting annex to this grammar, where all the idiosyncrasies and irregularities that cannot be successfully covered by the grammar are stored. By contrast, MTT considers the lexicon as the central, pivotal component of a linguistic description; the grammar is no more than a set of generalization over the lexicon, secondary to it. (p. 228.)

Accordingly, ECDs are conceived of as theoretical rather than practical or conventional dictionaries: An ECD purports to store all the lexical knowledge shared by speakers of a given language within a clearly stated theoretical linguistic framework. As opposed to it, conventional dictionaries are normally not consistent with

a particular linguistic theory (cf. the significance of FrameNet for corpus-based lexicography on page 33).

An ECD is a formalized dictionary—a lexical database—which lays great emphasis on *explicitness* and *consistency*:

Expiliciteness

> The ECD's explicitness means that nothing should be left to the user's intuition or logical abilities; nothing should be communicated through analogy or examples; everything has to be stated in an overt and precise way. To achieve this, the lexicographer is obliged to use a pre-established and well-developed lexicographic metalanguage. (Melčuk, 2006, p. 229)

Consistency

> The goal of consistency in the ECD has two implications: first, similar LUs should be described in a similar way, so that the degree of semantic relatedness of two LUs is paralleled by the degree of the similarity of their entries; and second, different aspects of one LU, i.e. its semantic, syntactic, and lexical cooccurrence properties, should be described in conformity with each other. (Melčuk, 2006, p. 230)

ECDs has five linguistic, distinctive features:

Encoding dictionaries ECDs are encoding or active dictionaries, providing users with ample linguistic data on the usage of expressions in various contexts and situations. The theoretical basis of this standpoint is that according to Melčuk (2006) the speaking process is more linguistic than the understanding process, since it requires less extralinguistic knowledge and common sense, which has to be carefully separated from the linguistic competence in the course of dictionary compilation. A strong correlate of this perspective is that collocations and idioms form an important part of the dictionaries.

Semantic basis The ECD's semantic orientation is based on the theoretical conviction that natural language is primarily a tool for expressing meanings, so that semantic considerations underlie everything else in language.

Co-occurrence as the main target As encoding dictionaries should focus both on meanings and possible contexts, in ECDs all collocations of an LU are included in the entry of that LU. Thus, ECDs have to pay close attention to the restricted combinability of LUs.

The ECDs describe all LUs of a language together and in a similar way Unlike the common practice in lexicography, where idioms are subsumed under lexeme LUs, ECDs treats *idioms* as separate headwords[1]. This editing principle is a consequence of the observation that the meaning of idioms is unpredictable, thus they are LUs on their own right.

Each entry of the ECD describes one LU In the ECD, each LU has its own lexical entry, and each lexical entry corresponds to one LU. All relevant lexicographic information is, strictly speaking, attached to an individual LU. It ensures the internal coherence of lexical entries

2.3.2.3 Levin verb classes

Preliminaries Although Levin's work is not based on corpus data—instead, it considers data present in the linguistic scientific literature—moreover, it concentrates merely on verbs, we have decided to discuss it here. The main reason for this is that Levin's approach moves from intuitive lexical semantics toward a more corpus-based methodology. In accordance with Levin (1993), traditional argument realization theories (eg. Komlósy, 1992) also assume that verbal meaning accounts for the syntactic realization of its arguments in the verbal complement structure, i.e. arguments are realized as subjects, objects or obliques. These theories also presume that verbal arguments can be described with a *predefined*, *universal* and *finite set* of semantic roles (cf. Levin and Hovav, 2005). That is, several traditional argument realization theories have rather similar assumptions

[1]Collocations are presented only in the entry of their bases.

about meaning to that of traditional lexicography. Namely, verbal meanings are intersubjective Platonic entities 'out there'. The main contribution of Levin—along with the lines of corpus-based methodology—is that she starts out from observable surface syntactic behavior instead of verbal meaning.

Levin verb classes Levin (1993) takes the hypothesis that verbal meaning determines its syntactic behavior as her starting point, thus, surface syntactic behavior is informative of verbal meaning. Based on this assumption she classified over 3,000 English verbs according to their alternation behavior. The main contribution to argument realization theories (along with Pustejovsky, 1995) is that it refuses to rely on predefined semantic elements when representing verbal meaning. Instead, Levin seeks verbal *meaning components—meta-predicates*—that account for verbal behavior based on surface distributional data. The theory presumes that verbs participating in the same syntactic alternations share some common meta-predicates, too. The meta-predicates can be thought of as meaning components that at least partially describe the meaning of the given verb. Meta-predicates are then responsible for the syntactic behavior of the given verb class. Therefore, the inventory of meta-predicates can be explored in a given language by the investigation of the complement structure alternations of the verbs. Hence, it follows that this theory does not rely on a predefined finite set of meta-predicates, rather it justifies the existence of a given meta-predicate through observable syntactic behavior. Therefore, verbal meaning is grasped through intersubjective language data.

However, though the methodology of Levin (1993) seems to be rather sound, Hanks and Pustejovsky (2005) raised some objections against it:

Objections First, they claim that the classification of English verbs is based on Levin's intuition supported by the intuitions of other academics who have written about the same verbs. Instead of intuition, the classification should be based on corpus data, since word behavior is observable, thus, it is able to guarantee at least some level of intersubjectivity.

As they assert:

> Many of Levin's assertions about the behaviour (and sometimes also the meaning) of particular verbs in her verb classes are idiosyncratic or simply wrong. Our findings accord with those of Baker and Ruppenhofer (2002), that when compared with actual usage, Levin's comments about diathesis alternations for verb classes apply to some but not all members of the classes. This is a pervasive problem in the second half of the book.

The second problem concerns coverage. Although Levin discusses approximately 3000 English verbs, her classification does not comprise all of the major verbs (e.g. *specialize*, *specify*, *spell*, *spend*, *spoil*).

A further issue according to Hanks and Pustejovsky (2005) is that the major senses of the verbs are not included.

2.3.2.4 FrameNet

Preliminaries According to Atkins and Rundell (2008), although corpus-based lexicography is held to be more objective then the traditional intuition-based approaches, the problem of intersubjectivity arises again, when looking at the wealth of concordances a corpus might offer. In the absence of a clearly defined and coherent guide it is rather difficult to write entries that comprise all the relevant linguistic facts of an LU in a consistent way. Obviously, such a guide has to rest on a sound and detailed linguistic theory that is able to provide us with a full-fledged description of language and has to be explicit on how to select the lexicographically relevant linguistic facts. According to them:

> For many excellent lexicographers this underlying theory is never made explicit: their intuition tells them what's worth saying about the headword, once they've scrutinized the corpus evidence. (p. 150)

FrameNet Because of its underlying linguistic theory, Frame Semantics (Fillmore, 2005), FrameNet is claimed to be of considerable importance to professional lexicographers. In what follows, we give a short overview of FrameNet from a lexicographic point of view based on Atkins and Rundell (2008).

2. COMPILING THE HEADWORD LIST

Fillmore (1994) claims that:

> The proper way to describe a word is to identify the grammatical constructions in which it participates and to characterize all of the obligatory and optional types of companions (complements, modifiers, adjuncts, etc.) which the word can have in such constructions, in so far as the occurrence of such accompanying elements is dependent in some way on the meaning of the word being described.

In accordance with this, Frame Semantics describes words, their various meanings, and how these are combined with others to form the utterances and sentences of a language. Its aim is to analyse and record, for each sense of a word or phrase, the full range of its semantic and syntactic relations. To do this, they have devised a suite of codes denoting semantic roles ('frame elements') and grammatical relationships, which allow them to document in detail the corpus contexts in which a word is found.

Frames and frame elements A semantic frame is a schematic representation of a situation type (e.g. speaking, eating, etc.) together with a list of the typical participants, props, and concepts that are to be found in such a situation. These are the semantic roles, or *frame elements*. Frame semantics describes the meanings of words and phrases (lexical units) in terms of the frame to which they belong and the contexts in which these LUs are found. The context, in a frame semantics analysis, is normally the phrase or clause, and maximally the sentence, in which the target word appears in corpus data. Frame Semantics starts out from the hypothesis that successful communication presumes shared interpretations of what is said or written. These interpretations are principally common semantic frames that are evoked in our minds by the words and phrases we use.

Example Someone says: *"Jo asked her brother to help her"*. In our own personal experience the situation in which someone makes a request normally contains certain elements. The vocabulary and syntax of its context let us identify the LU *ask* in that sentence as belonging to the REQUEST frame. Now, based on this frame, we expect to find in the sentence an LU referring to someone who is doing the requesting (*Jo*) another LU denoting someone who is being

asked (*her brother*), and a phrase describing what that person is asked to do (*to help her*). Our knowledge of English leads us to interpret the subject of the verb as the 'requester', its object as the person being asked, and its infinitival complement as the requested action. These elements of the REQUEST frame are used to describe the behaviour of the other words in that frame, too: For instance, verbs such as order, appeal, command, suggest, beg, and nouns such as order, appeal, command, suggestion, and of course request itself.

Valency description First, the frame is defined, and the frame elements are named and described. Essential frame elements are parts of the complement structure of the verb. There are also 'peripheral' frame elements that are common to whole sets of frames (e.g. LOCATION or FREQUENCY). Since peripheral frame elements are expressed in a way describable by productive rules, they are not inherent parts of the valency description, therefore, they are not of primary importance here.

In the next step the set of words are identified that may evoke the given frame. Then, for each LU a list of corpus sentences is extracted. Finally, each frame element in each corpus sentence is annotated with their phrase boundaries, with their phrase types (e.g noun phrase, adjectival phrase, etc.) and with their syntactic function. Syntactic functions assigned by verbs are External Argument[1], Object and Dependent. Other part-of-speech categories assign syntactic functions, as well.

Hierarchy FrameNet is basically a hierarchy of interrelated frames. There are several types of inheritance, such us *Inheritance*, *Subframe*, *Causative of*, *Inchoative of* and *Using*. We do not discuss the exact nature of these relations here. However, it is worth noting that according to Ruppenhofer et al. (2010) the main advantage of building a systematic FrameNet hierarchy is that it enables paraphrasing, both for humans and for computers.

> In many ways, paraphrasing is at the core of what we intend FrameNet to facilitate. [...] Translation is paraphrasing with the limitation that

[1] Among others, External Argument includes subjects or any other constituent that controls the subject of a target verb.

> all the resulting paraphrase must be in the target language. This requires FrameNet-style data for both source and target language; this currently limits such efforts to English, German, Japanese, and Spanish.

Lexicographic aspect As Atkins and Rundell (2008) notes, there are several papers concerning the practical application of Frame Semantics to lexicography (e.g. Atkins, 1994, Fillmore and Atkins, 1998, Atkins et al., 2003a and Atkins et al., 2003b).

The underlying linguistic theory, Frame Semantics, ensures that the valency description include the most important facts that the lexicographer needs to be aware of when writing the dictionary entry. Moreover, it renders the corpus analysis more coherent and objective, which in turn can guarantee that no necessary linguistic fact is overlooked.

FrameNet vs. Levin verb classes According to Baker and Ruppenhofer (2002) the FrameNet project (Fillmore and Baker, 2001) is able to overcome some of the shortcomings of Levin's classification. In FrameNet words (not only verbs but nouns and adjectives, too) are grouped together on the basis of their underlying conceptual structures and their distributions are derived from corpus evidence. This entails that

> verbs grouped together in FrameNet (FN) might be semantically similar but have different (or no) alternations, and that verbs which share the same alternation might be represented in two different semantic frames.

That is,

> The FrameNet project is producing a lexicon with roughly comparable coverage of verbs, but with much more detail concerning the semantics and syntax of their arguments, more semantically consistent categories, and a richer set of relations among them.

Beside its merits, the corpus-based methodology of FrameNet has to face some difficulties, too. Hanks and Pustejovsky (2005) raised the following objections:

Objections: Although it exploits corpus data, FrameNet runs the risk of accidental omissions, as it relies on the intuitions of researchers. In accordance with this, some frames overlap to the point of being indistinguishable, while others are only partially populated. In some cases only minor or rare senses are included.

Besides, the workflow proceeds frame-by-frame and not word-by-word which entails that no word can be considered complete until all frames are finished.

2.3.2.5 Corpus Pattern Analysis

Corpus Pattern Analysis is an approach developed by Hanks (cf. Hanks and Pustejovsky, 2005) with the objective of producing a verbal database comprising verbs—The Pattern Dictionary of English Verbs[1] (PDEV)—without the methodological issues present in WordNet, FrameNet or in Levin's verb classes. The underlying linguistic theory is the Theory of Norms and Exploitations (TNE) (Hanks and Pustejovsky, 2005).

TNE As described in Cinkova and Hanks (2010), TNE relates prototypical meaning concepts to prototypes of phraseology (i.e. linguistic usage), as found in a large corpus. Corpus analysis shows that there are not only prototypical uses of words (i.e. normal and conventional uses: *Norms*) but also perfectly well-formed and well-motivated utterances that, in one way or another, deviate from the prototypical patterns. These are mostly creative innovations, but they include also domain-specific patterns. These patterns are called *exploitations*. An exploitation is an utterance that can be related to a corresponding phraseological norm.

The project aims to anchor word's meanings to their use based on corpus data. For that purpose the relationship between meanings and patterns of usage is explored. In fact, patterns are thought of as collocations that are characterized by *semantic types* and *lexical sets*.

[1] `http://nlp.fi.muni.cz/projects/cpa`

Semantic type Semantic types are context-independent cognitive concepts reflecting common sense knowledge, such as HUMAN, INSTITUTION, ANIMAL, EVENT, etc. Therefore, semantic types are characterized regardless of the verb which is the head of the given structure. Context-dependent semantic restrictions, i.e. those imposed by the verb, are described in terms of *semantic roles*. The semantic roles are mapped onto specific semantic types. Collocations that have a distinctive semantic feature in common are grouped together according to their semantic type.

Lexical set If there is no semantic type available, words are grouped together into lexical sets according to their syntagmatic behavior. In this case, the lexical set is specified extensionally, by simply enumerating typical members. Here, it remains an open question whether the relevant lexical items can be unified into a semantic type by means of common semantic features. Thus, the difference between lexical types and lexical sets is the way in which they are defined: Lexical types are defined intensionally, utilizing a salient common sense property, while lexical sets are rather defined by means of listing all the members of the corresponding set.

Verbal patterns in PDEV are characterized by relying both on lexical types and on lexical sets.

Shallow ontology An important by-product of CPA is a shallow ontology[1]. The significance of the ontology is that it is based on linguistic knowledge as opposed to conceptual knowledge (cf. WordNet). The semantic types are stated in a finite inventory, which constitutes a shallow ontology of about 200 items. Reflecting language data instead of scientific considerations, the CPA shallow ontology is intrinsically unbalanced.

Example Probably because of cultural reasons, there are many verbs (*bark*, *saddle*, etc.) that require dogs or horses as any of their collocates. Therefore, DOG and HORSE are parts of the ontology, while there are no semantic types for other species of animals, such as jackals or hyenas.

CPA and FrameNet According to Hanks and Pustejovsky (2005):

[1] `http://nlp.fi.muni.cz/projekty/cpa/public_onto.html`

> If CPA succeeds in its objective of analysing all the normal uses of each verb, it will complement FrameNet neatly in this respect. FrameNet offers a very full and detailed semantic analysis of each frame; CPA offers a contrastive analysis of the senses of each word. When a CPA entry for a given verb is finished, it has, by definition, completed analysis of all normal uses of that verb.

2.3.2.6 Referencie Bestand Nederlands

Preliminaries The relevance of 'Reference Database of Dutch' (RBN) for the present dissertation is given by the fact that this Dutch lexical database was designed primarily to support the construction of bilingual dictionaries, as described in the paper of van der Vliet (2007). One striking feature of the database is that it can be re-used to build dictionaries for various language pairs, on top of that, Dutch could play the role of both the source and the target language. These later issues will be elaborated more in Sections 3.3.2.1 and 3.3.3.2.

Description of tha database RBN consists of a definition and additional properties for each of its 45.000 entries. The properties might be syntactical, morphological, graphemic, semantic (countability, semantic type, systematic polysemy, synonyms), pragmatic and combinatorical (distributional). During the construction of the database great emphasis was laid on the explicit and systematic specification of properties. The explicit and systematic description of meanings was ensured by the following:

(1) They relied on a 38 million-word corpus when differentiating between meanings.

(2) They aimed at listing only the basic meanings, other senses were derived from the base meanings, for instances by means of systematic polysemy.

(3) Nouns were divided into 11 different semantic classes.

2.3.2.7 Lexical profiling: Sketch Engine

Preliminaries Without lexical profiling, lexicographers make use of their own intuition when selecting the relevant linguistic facts to be included in a dictionary even if a wealth of empirical data is available. As Kilgarriff and Kosem (2012) notes:

> Most of the first COBUILD dictionary was produced from a corpus of eight million words. Several of the leading English dictionaries of the 1990s were produced using the British National Corpus (BNC), of 100 million words. Current lexicographic projects we are involved in use corpora of around a billion words—though this is still less than one hundredth of one percent of the English language text available on the Web.

Therefore, in this section we give an overview of one of the leading lexical profiling tools.

Lexical profiling As corpus-size grew bigger and bigger, scrutinizing concordance turned out to be a suboptimal approach when tailoring headwords in a dictionary. As opposed to exploiting concordances directly, lexical profiling tools render possible the efficient evaluation of large amount of linguistic data with a reasonable amount of effort. A *lexical profile* is a sort of statistical summary of a word which provides lexicographers with the salient facts about the way a word most typically combines with other words.

Sketch Engine Sketch Engine is a web-based program which takes as its input a corpus of any language with an appropriate level of linguistic annotation. It generates word sketches on the basis of input data, which serve then as starting point when analyzing complex headwords (cf. Atkins and Rundell, 2008).

Language analysis functions Among many others, Sketch-Engine's most important language-analysis functions are the following:

The Concordancer It displays all occurrences from the corpus for a given query. The program is very powerful with a wide variety of query types and many different ways of displaying and organising the results.

The Word Sketch program It is a lexical profiling tool, which provides a corpus-based summary of a word's grammatical and collocational behaviour.

Word sketch A word sketch is a sort of lexical profile produced automatically by Sketch Engine. According to Kilgarriff et al. (2004) word sketches are

> one-page automatic, corpus-based summaries of a word's grammatical and collocational behavior. (p. 105)

Moreover, some additional features are also included to facilitate the characterization of headwords (Kilgarriff and Kosem, 2012):

Thesaurus The automatically generated thesaurus provides lexicographers with near-synonyms[1]. Nearest neighbours are calculated for a node word on the basis of their shared collocates (for more details see Subsection 2.3.3.2).

Sketchdiffs Sketch differences compare word sketches for two words, showing the collocations that they have in common and those they do not. For example, comparing *attractive* and *handsome* sketchdiff yields the information that *particularly*, and *extremely* are more typical modifiers of *attractive*; *strikingly* and *devastatingly* are more typical of *handsome*, while *truly* or *exceptionally* show similar salience with both *handsome* and *attractive*.

Good dictionary examples (GDEX) Good dictionary examples are hard to find in corpora, since several different characteristics have to be taken into account. According to Kilgarriff and Kosem (2012) readability and informativeness are important factors. Therefore, a GDEX is of ideal length, possibly comprises only alphabetical characters[2], is made up of frequent words and contains words that are frequently found in the vicinity of the key expression (and, therefore, are probably collocates). In the face of the difficulty of finding such ideal examples, lexicographers, in practice tend to

[1] And in some cases antonyms.
[2] Beside punctuation marks, of course.

invent, rather than find GDEXs, even in the presence of abundant language data. Unfortunately, this runs the risk of failing to provide a natural context for the expression being illustrated. Sketch Engine's GDEX module attempts to automatically sort the sentences according to how likely they are to be good dictionary examples. This module relies on both readability heuristics (e.g. sentence length, word frequency, etc.) and informativeness heuristics (e.g. including typical collocates).

2.3.2.8 Remarks on corpus-based lexicography

The great number of relevant projects is in accordance with the observation of Hanks (2010):

> At the core of lexicography, therefore, lies the corpus. (p. 597)

Accordingly, in our days it is widely accepted in the lexicographer community that high-quality dictionaries are based on corpora. The main reason for this is that linguistic data decreases the role of human intuition during lexicographic process.

Meaning of LUs In corpus-based lexicography the meaning of words is characterized on the basis of corpus data. This methodology is supported by the famous observation attributed to Firth (1957), who claimed:

> You shall know a word by the company it keeps.

From a corpus-based perspective, words' meanings can be grasped by relying on Harris' distributional hypothesis (Harris, 1954), which asserts that words that are used and occur in the same contexts tend to purport similar meanings. This hypothesis is compatible with two different interpretations of meaning:

(1) Possible contexts greatly influence the possible meanings of a word. Therefore, the basic building blocks of language are not words with their meanings, as in the case of intuition-based dictionaries, but *words with their meaning potentials* (cf. Hanks, 2010, p. 27).

(2) Instead of meaning potentials words do have meanings, as in the case of intuition-based lexicography. But as opposed to the traditional approach, the meanings of the words greatly influence where they may appear within a sentence. This equals to say that the arguments of syntactic rules are more restricted than in the case of traditional lexicography: Instead of part-of-speech categories they are made up of narrower semantic classes.

Example To give an example for both interpretations, the Hungarian words *nagy* (great) and *komoly* (serious) have the same meaning in certain contexts[1]. In other contexts *nagy* means the same as *szép* (nice)[2]. However, the synonymy holds only in certain contexts: *nagy*, *komoly* and *szép* in the phrase *nagy/komoly/szép fiú* (great/serious/nice boy) refer to different properties of the boy.

In the "meaning-potential" interpretation *szép* has a potential to mean both that something is greater-than-average and an aesthetic category. According to the alternative view, *szép* is a polysemious word that shows up with different meanings in different contexts.

Constrained productivity The second interpretation of meaning entails a view of constrained productivity: In this perspective, part-of-speech categories are not considered as the ultimate generalizations on word usage. It does not suffice any more to rely primarily on part-of-speech information when describing the grammar, presuming that every member of a part-of-speech category behaves more or less in the same way. Instead, a more subtle description of the behavior of words is needed, giving greater importance to the meaning component, since the semantic properties of words determine the ways in which they can be combined. Therefore, instead of part-of-speeches, grammatical rules operate on more restricted semantic word classes.

Increased role of lexicon A correlate of restricted productivity is a shift in theoretical work: As opposed to classical generative grammars, the corresponding theories (MTT, FrameNet, CPA, etc.) consider the lexicon the central element of

[1]For instance, before the Hungarian noun *baj* (trouble).
[2]For instance, before the Hungarian noun *siker* (success).

grammatical description. As we saw earlier on page 29, this theoretical standpoint is explicitly formulated by Melčuk (2006):

> Most current linguistic theories view a linguistic description of a language as a grammar; a lexicon is taken to be an indispensable, but somehow less interesting annex to this grammar, where all the idiosyncrasies and irregularities that cannot be successfully covered by the grammar are stored. By contrast, MTT considers the lexicon as the central, pivotal component of a linguistic description; the grammar is no more than a set of generalization over the lexicon, secondary to it. (p. 228.)

Therefore, in a corpus-based framework, lexicography i.e. the exhaustive description of the lexicon, and linguistics may be thought of as closely related disciplines.

Presuppositions Consequently, corpus-based lexicography presumes the following:

(1) The meaning of words' is highly dependent on the contexts in which they occur. This view on meanings is compatible with (at least) two interpretations:

   (i) Word forms do not have meanings in themselves, but they have *meaning potentials*.

   (ii) Words tend to be *highly polysemious* that show up with different meanings in different contexts.

(2) In either case, the various meanings can change significantly across different contexts.

(3) Because of the increased role of contexts in the description of meanings, introspection in itself is not enough to be able to list the relevant headwords of a dictionary and to provide a sufficient characterization of it. Therefore,

   (i) The lexicographic intuition should be underpinned by corpus evidence.

   (ii) Moreover, a sound linguistic theory is needed to draw lexicographers' attention to the lexicographically relevant facts (cf. page 33).

(iii) In addition, for the same purpose, the exploitation of lexical profiling tools turned out to be indispensable, too, as corpus size increased (cf. page 40).

Philosophical notes The philosophical background of corpus-based lexicography is *empiricism*. The simplest formulation of empiricism is that humans do not possess innate ideas. The metaphor of *tabula rasa* (blank sheet) of Locke (1841) depicts this insight. Such an approach raises the question of how the intersubjectivity of meanings can be guaranteed. In this perspective, intersubjectivity is guaranteed by usage. The members of a language community should use words with the same meaning to achieve the chief end of language, communication. According to Wittgenstein (1979):

> Words and chess pieces are analogous; knowing how to use a word is like knowing how to move a chess piece. [...] The meaning of a word is to be defined by the rules for its use, not by the feeling that attaches to the words.
>
> "How is the word used?" and "What is the grammar of the word?" I shall take as being the same question.

As the set of contexts in which an expression may appear reflects the use of language, the view of Wittgenstein might be reconciled with what Firth (1957) said.

Drawbacks One serious practical disadvantage of the corpus-based approach is that manufacturing a full-fledged sense-inventory is rather tedious requiring so much resource that is usually unavailable in the case of less resourced languages.

However, as it was discussed earlier (cf. page 33), the wealth of language data may raise additional problems for lexicographers when compiling monolingual sense-inventories. Namely, lexicographers should be able to select the relevant linguistic facts and describe them in a consistent way, which becomes more and more difficult as the corpus size increases. We saw that lexical profiling tools and linguistic theories may help in selecting the relevant facets of language data and describe them in a consistent way.

The next section discusses corpus-driven lexicography and some related research directions. As in our days the exploitation of corpus-driven methods is not the common practice in the field of lexicography, beside a few larger scale projects we present some insightful research ideas, as well. In our view, corpus-driven approaches have immense potential both for monolingual and for bilingual lexicography.

2.3.3 Corpus-driven lexicography

Corpus-based vs. corpus-driven approaches Although corpus-based and corpus-driven approaches are both based on the exploration of corpus data, there is a fundamental difference between them with regard to the role of observable data in formulating theories. According to Tognini-Bonelli (2001):

> ...in contrast with the corpus-based approach, the 'corpus-driven' approach where the corpus is used beyond the selection of examples to support or quantify a pre-existing theoretical category. Here the theoretical statement can only be formulated in the presence of corpus evidence and is fully accountable to it. This approach, it is argued, brings about a qualitative change in the description of language and shakes some major assumptions underlying traditional linguistics. (p. 11)

In corpus-driven lexicography the concept of meaning is the same as in the case of corpus-based approaches. That is, both methodologies grasp meaning primarily through the observation of language data, what makes the difference is the technique of observation.

Unsupervised learning techniques Corpus-driven approaches require a methodological shift in research. As we saw in Subsection 2.3.2.7, the size of available corpora continuously have been grown bigger and bigger. Analyzing such amount of data requires new techniques. One such approach is unsupervised learning which aims at finding hidden structures in unlabeled data. Relying on unlabeled data yields the advantage of eliminating unarticulated theoretical assumptions present in the labeling itself. Thus, these techniques decrease the role of human

intuition even further, but intuition is still not wholly excluded. Human intuition comes into play when selecting the investigated phenomenon, coming up with a representation set up, fine-tuning the parameters and throughout the evaluation (cf. Saldanha, 2009).

Thus, in what follows, five research directions will be described, which in one way or another intend to characterize an LU. The first method (Section 2.3.3.1) concentrates solely on verbal patterns. The second and third techniques (Sections 2.3.3.2 and 2.3.3.3) aim at building distributional thesauri completely automatically, while Sections 2.3.3.5 and 2.3.3.4 present two approaches, the ultimate goal of which is to create bilingual dictionaries in an unsupervised way.

2.3.3.1 Unsupervised extraction of verb frames

Objective This method, described in Sass (2009) and Sass (2011) in more detail (cf. Chapter 7) aims at the unsupervised extraction of verb-centered constructions (VCCs) from corpora. Thus, it does not intend to build a sense-inventory comprising all the relevant LUs of the SL, instead—similar to Levin's work and CPA—it concentrates on verbs. This technique treats VCCs with various structures alike in order to build a database comprising all relevant VCCs for any language. Here, various structure means that the algorithm is able to detect the complement structure of a verb, irrespective of its syntactic properties (e.g. the number of constituents, the syntactic function of the constituents or the order of the constituents). Moreover, the method is able to determine if a head content word is inherently part of the VCC forming a multiword verb (e.g. *take into consideration*).

The significance of the algorithm lies in its capability to deal with multiword verbs and their valences simultaneously. At the same time the method is claimed to be language independent.

Workflow After the detection of noun phrases and their head elements the corpus is converted into a sequence of clauses assuming that every clause contains one and only one verb. In the following step frequent frames are counted in a

cumulative way: One randomly selected longest subframe inherits the counts of small frequency subframes.

VCCs and meaning However, as opposed to Levin, the technique does not aim at accounting for verbal meanings, instead, this approach strives to list all the salient constructions in which a verb may appear. Unlike CPA the extraction method does not give hints on the arguments semantic types and semantic roles. Instead, it lists all the salient collocates, even if they belong to the same semantic class. In the next sections we discuss some research directions whose objective is to grasp meanings on the basis of distributional data.

2.3.3.2 Synonymy detection: Sketch Engine's thesaurus

Thesaurus In general usage, a thesaurus is a reference work that lists words grouped together according to similarity of meaning containing synonyms and sometimes antonyms. Thus, a thesaurus may be thought of as a repository of SL LUs, where the meaning of each SL word form is characterized by the list of similar words.

The automatic generation of thesauri The automatic generation of thesauri is a quite widespread technique in the NLP community to build monolingual sense-inventories. In accordance with Harris' distributional hypothesis (Harris, 1954), for each word, the words that share the most contexts based on some suitable statistics are the best candidates to be synonyms.

In what follows, Sketch Engine's thesaurus will be briefly introduced based on Rychlý and Kilgarriff (2007). According to them[1], the automatic creation of thesauri is made up of three stages:

(1) Setting up a corpus,

[1]The novelty of the method is that it computes thesaurus on the basis of large corpora, for thesauri generally improve in accuracy with corpus size. We do not want to delve into technical details here: The basic idea is that they do not compare all word pairs, only those word pairs that do have something in common. This allows them to create thesauri from 1B corpora in under 2 hours.

(2) Identifying contexts for each word,

(3) Identifying which words share contexts.

For instance, the target word *objective* has the following words similar in meaning ranked according to descending similarity: *aim*, *target*, *strategy*, *point*, *principle*, *task*, etc.[1]

Thesauri and meaning In this framework, a similarity measure is calculated, indicating the meaning distance between two words. In our view, such an approach is more compatible with Hanks (2010) conception of meaning, i.e. meaning potential, discussed on page 42. Namely, it grasps the meaning potential of the keyword in terms of semantically similar words, but it does not provide us with a detailed list of possible contexts, where these similar words might show up.

Thesauri and multiple meanings Thus, in our opinion, Sketch engine's thesaurus does not account for polysemy in itself. That is, albeit usually more than one semantically similar words are listed for a given keyword, the relation of these semantically similar words to each other is not indicated, e.g. whether they belong to the same semantic class or not. For instance, *bottle* is said to be similar both to *glass* and to *bag*, but it is not indicated whether these two word forms are similar to the very same sense of the keyword or they are related to two distinct senses.

The synonymy-detection in the next section concentrates on how the polysemy of a word may be grasped by means on distributional data.

2.3.3.3 Synonymy detection: Near-synonyms for adjectives

Preliminaries The present experiment takes a step further than Sketch Engine's thesaurus and tries to determine whether synonymy relation holds among the words that are in similarity relation to a given keyword. The basic objective of this thread of research is to build a system that is able to select the proper

[1] `http://www.sketchengine.co.uk/documentation/wiki/Website/Features$#$Distributionalthesaurus`

meaning of adjectives in contexts on the basis of a sense-inventory extracted automatically from a monolingual corpus[1]. Hence, *the experiment aims at building a database that is able to capture polysemy and homonymy associated to a word form on the basis of possible contexts.* Note that although the present investigation is only a proof-of-concept experiment, it presents insightful and illuminating ideas that play a great role in the present thesis (cf. Section 4.2.2.5). Therefore, we have decided to give a detailed overview of it, which may seem somewhat disproportionate.

Just as all corpus-orientated approaches, this technique strongly builds upon Zellig Harris' distributional hypothesis (Harris, 1954), which states that words that are used and occur in the same contexts tend to purport similar meanings. Hence, contexts are investigated to derive meanings on the basis of them. Starting out from the distributional hypothesis the definition of synonymy may be formulated as follows:

Synonyms and near-synonyms Among many others[2], according to Ploux and Victorri (1998), two lexical units are *synonyms* iff every occurrence of the one lexical unit can be substituted with an occurrence of the other lexical unit *in every context* so that the meanings of the corresponding utterances *never* change significantly[3].

On the other hand, they also define the notion of *near-synonymy.* Accordingly, two lexical units are *near-synonyms* iff every occurrence of the one lexical unit can be substituted with an occurrence of the other lexical unit *in a certain set of contexts* so that the meaning of utterance does not change significantly[4].

[1]This work has been accomplished in tight collaboration with Dávid Takács: He implemented the algorithm and took part in the interpretation of the results. (cf. Héja and Takács, 2010).

[2]This defintion goes back to Leibniz's *salva veritate* principle, which states that *"Two terms are the same (eadem) if one can be substituted for the other without altering the truth of any statement (salva veritate).*"

[3]Deux unités lexicales sont en relation de synonymie pure si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans tout environnement sans modifier notablement le sens de l'énoncé dans lequel elle se trouve.

[4]Deux unités lexicales sont en relation de synonymie si toute occurrence de l'une peut être remplacée par une occurrence de l'autre dans un certain nombre d'environnements sans modifier notablement le sens de l'énoncé dans lequel elle se trouve.

Near-synonymy and multiple meanings The notion of near-synonymy will be used to capture multiple meanings. Here, we suppose that near-synonyms represent one meaning of a given word-form W. That is, a word form W may belong to various near-synonym classes, thus these various near-synonymy classes are able to capture the polysemy and homonymy belonging to W.

Multiple meanings and graphs Now, near-synonyms are represented by sub-graphs exhibiting special properties. The underlying idea is that graphs are capable of representing a system of near-synonyms: Since every special sub-graph corresponds to a near-synonymy class, i.e. to a meaning, multiple sub-graphs of the same graph are able to represent that the given word-form W belongs to more near-snyonymy classes, that is, it conveys multiple meanings.

Method We followed the method described in Ah-Pine and Jacquet (2009). However, as opposed to them, instead of the unsupervised creation of a lexical database suitable for named entity disambiguation, we intended to investigate how this method is applicable to disambiguate semantically more intricate word classes, such us adjectives. Because systematic polysemy is inherently present in the class of proper names, the different semantic classes of the denotated entities usually can be clearly told apart. In this proof-of-concept experiment we investigated whether their method is applicable to word classes with possibly more overlapping meanings, such as adjectives. Figure 2.3 depicts a graph each edge of which represents a sub-meaning of the highly polysemous Hungarian word *nagy* (great, big, etc.).

The applied technique is made up of the following steps:

(1) Selecting the input corpus.

(2) Detecting adjectives and their relevant contexts.

(3) Constructing the distributional space of the adjectives based on the corpus.

(4) Computing pairwaise similarities between the adjectives.

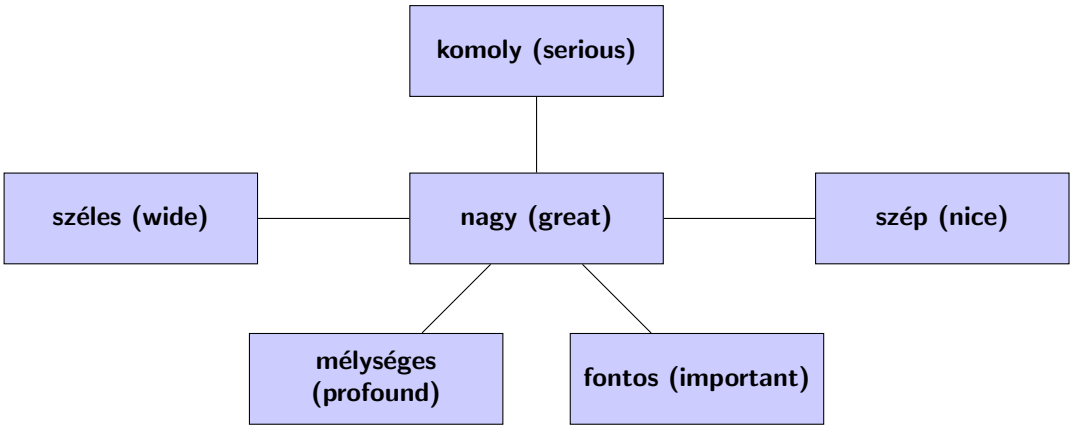


Figure 2.3: Subgraphs representing polysemous meanings of the Hungarian word *nagy*

(5) Computing the relevant sub-graphs of adjectives on the basis of the similarity matrix.

(6) Retrieving the relevant contexts for each adjective-clique.

Description of the steps

Input data The input corpus was a sub-corpus of the Hungarian National Corpus (Váradi, 2002) consisting of 1,877,661 tokens.

Detection of adjectives and their relevant contexts The corpus contained part-of-speech annotation. Nouns were used as relevant contexts for adjectives: The corpus contained 592,321 nouns, 203,685 adjectives and 143,682 adjective-noun pairs.

Constructing the distributional space The distributional space of the adjectives was constructed on the basis of the corpus. Let A denote the set of the adjectives and N denote the set of the nouns. The distributional space is given by the D adjective-noun matrix, where $1 \leq i \leq |A|$ and $1 \leq j \leq |N|$. The matrix element D_{ij} is computed by estimating the corresponding con-

ditional probability:

$$P_{MLE}(N_j|A_i) = \frac{count(N_j, A_i)}{count(A_i)} \tag{2.3.1}$$

Threshold However, sufficient amount of data should be considered, when estimating the probabilities, thus, throughout our experiment we confined ourselves to the investigation of adjectives and nouns that occur more than 50 times.

Smoothing Unfortunately, the maximum likelihood estimation of each D_{ij} leads to sparse data, resulting in too many zero elements in matrix D. To handle this problem we relied on the Jelinek-Mercer smoothing that allows to distribute the probability mass found by the maximum likelihood estimation to contexts with zero occurrences.

$$P_S(N_j|A_i) = (1-\lambda)P(N_j|A_i) + \lambda P(N_j|CORP) \tag{2.3.2}$$

where $0 \leq \lambda \leq 1$[1]

Computing pairwaise similarities between the adjectives Now, we are interested in the similarity of the probability distributions of adjectives. That is, the similarity of A_i is represented as $< D_{i1}, D_{i2}, D_{i3}, ..., D_{ik} >$ and A_j is represented as $< D_{j1}, D_{j2}, D_{j3}, ..., D_{jk} >$ vectors has to be calculated. Cross entropy (CE) was used for that purpose:

$$CE(A_i, A_j) = -\sum_{k=1}^{n} P_S(N_k|A_i) log P_S(N_k|A_j) \tag{2.3.3}$$

Thus, the $A \times N$-matrix was sent into an $A \times A$ matrix.

Computing the relevant sub-graphs of adjectives In this step two questions have to be answered. First, we have to decide what counts as a sub-

[1]The main insight behind smoothing is to give an estimation to unseen events. In the present experiment λ was set to 0.1. Thus, according to the formulae, the relative frequency of N_j in the corpus is given a relatively low weight, whereas, N_j relative to the adjectives is considered to be important.

graph representing a (sub)meaning. Secondly, these sub-graphs have to be calculated on the basis of the similarity matrix somehow.

Cliques: Sub-graphs representing a meaning Following Ah-Pine and Jacquet (2009) we have decided to use cliques as sub-graphs representing meanings. A clique is a subset of vertices in an undirected graph such that every two vertices in the subset are connected by an edge. Or to put it in an other way, a completely connected sub-part of the graphs is a clique.

Computing adjective cliques from the similarity matrix In this phase adjectival cliques are derived from the asymmetric $A \times A$ matrix. For that purpose the similarity matrix is converted into a binary and symmetric matrix such that each A is a node and each similarity between two As is an edge.

(i) The $A \times A$ matrix is symmetrized through selecting the greater CE value.

(ii) Then a certain threshold is introduced to map the values of the matrix elements to 0 or 1.

(iii) Adjectival cliques are searched from the resulting symmetric and binary $A \times A$ matrix.

Retrieving the relevant contexts for each adjective-clique A_i and A_j may belong to the same clique only if they have some contexts N_ks in common. For more than two A_is their common contexts will be the intersection of their pairwise common contexts N_ks. This intersection is rather straightforward to retrieve: Thus, it was not stored separately.

Examples Since this thread of research is in an early stage, detailed evaluation was not performed. However, as the example presented below shows, the method enables the detection of interchangeable adjectives: In accordance with the definition of near-synonymy, the substitutability is constrained to certain contexts. These contexts are also automatically provided. Interestingly, without the explicitly given contexts, some of the examples are rather counter-intuitive even for

native speakers. For instance, for most Hungarian speakers it may be surprising that *nagy* (big) and *komoly* (serious) can be used as synonyms. However, this is exactly the case:

NAGY KOMOLY (big serious) baj (trouble), beruházás (investment), eredmény (result), erőfeszítés (effort), feladat (task), gond (trouble), igény (demand), kihívás (challenge), kár (impairment), lehetőség (opportunity), munka (work), probléma (problem), pénz (money), segítség (help), siker (success), terv (plan)

NAGY SZÉLES (big wide) választék (assortment)

NAGY SZÉP (big nice) eredmény (result), siker (success), teljesítmény (achievement)

NAGY MÉLYSÉGES (big profound) bánat (sorrow), fájdalom (pain)

NAGY FONTOS (big important) alak (figure), alkotás (piece of work, artifact), esemény (event), teljesítmény (achievement), áttörés (breakthrough),

Conclusion Obviously, this is only a proof-of-concept experiment and definitively needs much more effort to produce a reliable database, but from our perspective the initial results are rather insightful. The different meanings of *"nagy"*—characterized by other adjective(s) that may appear in the same set of contexts—are told apart solely on the basis of the following nouns. Consequently, these adjectives are interchangeable before certain nouns.

As we will see in Chapter 4, such a database might facilitate the creation of encoding dictionaries. However, as this is only the initial phase of research, this technique was not exploited when generating bilingual dictionaries. Yet, we think that this thread of research is worth pursuing for it has the potential of increasing coverage of dictionaries when combined with word alignment (cf. Section 9.3).

2.3.3.4 Bilingual lexicography: Detection of translation pairs in monolingual corpora

Preliminaries In Section 2.3.3.4 and 2.3.3.5 two *bilingual* extraction methods will be considered, only partially fitting in the framework of the present discussion, which concentrates primarily on monolingual sense-inventories. However, for the sake of completeness we have decided to shortly describe these two related research methods here. The research described in 2.3.3.4 is also in line with our starting hypothesis, i.e. that bilingual dictionaries are usually based on monolingual ones. As opposed to it, the

research only briefly mentioned in 2.3.3.5, does not admit this presupposition.

Objective Mikolov et al. (2013) aim at generating bilingual dictionaries on the basis of two independent monolingual corpora and a small seed dictionary. For doing so, they use distributed representations of words and phrases to infer missing dictionary entries.

Method The method is composed of two steps. First, the monolingual models of the corresponding languages are built using large amount of texts. Two models were used for that purpose: The CBOW model and the Skip-gram model. The objective of the CBOW model is to combine the representations of surrounding words to predict the word in the middle, while the Skip-gram model learns word vector representations that are good at predicting its context in the same sentence.

Since semantically similar words tend to occur in similar contexts, closely related words have similar vector representations, e.g., *school* and *university*, *lake* and *river*. More interestingly, as Mikolov et al. (2013) claim, the vectors capture relationships between concepts, as well, via linear operations. For example, $vector(France) - vector(Paris)$ is similar to $vector(Italy) - vector(Rome)$.

Next, a small bilingual dictionary was used to learn a linear projection between the languages. Any word that has been seen in the monolingual corpora can be translated by projecting its vector representation from the source language space to the target language space. That is, the relationship between vector spaces that represent these two languages can be captured by a linear mapping, namely, by a combination of rotation and scaling.

Evaluation English-Spanish translations with high confidence values yield a precision of 75% in a first-best evaluation setup, as Mikolov et al. (2013) report.

Discussion In the course of this experiment the vectors generated for English may be conceived of as the SL sense-inventory. Nevertheless, since one word-form W is represented by a single vector, the vectors do not account for polysemies, inherently present in the language.

2.3.3.5 Bilingual lexicography: Lexicon extraction from parallel corpora

Recall that the present discussion is based on the presupposition that bilingual dictionaries are based on monolingual sense-inventories that have great impact on the quality of the resulting bilingual dictionary. In the case of automatic lexicon extraction from parallel corpora the SL sense inventory need not be previously characterized, rather it emerges as a result of finding the right translations for the SL word forms.

As the major part of this thesis is devoted to bilingual lexicon extraction from parallel corpora and the related issues, we do not want to delve into the details here, this approach was mentioned to describe an additional way of characterizing the SL sense-inventory.

2.3.3.6 Remarks on corpus-driven lexicography

Corpus-based and corpus-driven approaches With regard to the basic assumptions, corpus-driven lexicography is rather similar to corpus-based lexicography. Starting from Harris (1954) distributional hypothesis, according to which *words that are used and occur in the same contexts tend to purport similar meanings*, it investigates contexts to derive meanings on the basis of them. Nevertheless, data plays a greater role than in corpus-based lexicography, at least as far as corpus-driven techniques are suitable to handle greater amount of data than corpus-based approaches.

Unsupervised learning techniques The aim of unsupervised learning techniques is to learn hidden structures from unlabeled data. Thus, unsupervised techniques eliminate unarticulated theoretical presumptions present in the labeling itself. However, human intuition cannot be completely excluded: It comes into play again when selecting the investigated phenomenon, coming up with a representation set up, fine-tuning the parameters and throughout the evaluation, as well.

2.4 Conclusion

The dictionary building process We started out from the traditional dictionary building process which comprises three phases. In the analysis stage the SL headwords are selected and characterized to be included in the dictionary. In the transfer phase

the resulting SL LUs are translated into the TL. Finally, in the synthesis stage the final bilingual entry is completed.

Monolingual sense-inventories This chapter focuses on the analysis phase by giving an overview of some existing monolingual sense-inventories classified according to their relation to natural language data. These sense-inventories may serve as a possible basis for translation and greatly influence the quality of the resulting bilingual dictionary. Based on their relation to corpus data, traditional, corpus-based and corpus-driven approaches were distinguished.

Traditional approaches In traditional lexicography words and their meanings are the basic building blocks of language. Sentences are derived from these meaningful units in a compositional, thus, predictable way. Traditional lexicography presumes the following:

(1) The basic blocks of language are word form–meaning pairs. Which means, that

   (i) Word forms do have meanings

   (ii) These meanings are fairly stable across different contexts

(2) Word–meaning pairs are stored in the inner lexicon of lexicographers and can be assessed by means of introspection.

From (1) it follows that contextual information does not play a great role in traditional lexicography. While (2) has two questionable implications:

(i) Everyone has a strong belief that they know exactly the meanings of a word.

(ii) This knowledge is largely alike across the members of a language community. That is, meanings are objective or at least highly intersubjective entities.

In Chapter 4 the latter two assertions will be investigated and it will be proven that both are false.

Corpus-based approaches As the great number of the corresponding projects indicate, in our days it is widely accepted in the lexicographer community that high-quality dictionaries are based on corpora. The main reason for this is that linguistic data decreases the role of human intuition during lexicographic process, thus rendering the description of the investigated phenomena more intersubjective.

As a first conception of meaning Hanks (2010) view was accepted, according to which, in corpus-based lexicography the basic building blocks of language are words with their meaning potentials. This is primarily due to the fact that possible contexts greatly influence the possible meanings of a word.

Such a concept of meaning entails a more restricted view on productivity and thus compositionality: It does not suffice any more to rely primarily on part-of-speech information, when indicating how to use a word grammatically. Instead, a more subtle description of words' behavior is needed by carefully observing the set of contexts in which the word in question may appear. Such a methodology requires that lexicographers are provided with abundant language data.

And if the contexts can be neatly characterized, the meaning of a word-form can be given in terms of near-synonyms, which are interchangeable in a certain set of contexts. Corpus-based lexicography presumes the following:

(1) The meaning of words' is highly dependent on the contexts in which they occur. This view on meanings is compatible with (at least) two interpretations:

   (i) Word forms do not have meanings in themselves, but they have meaning potentials.

   (ii) Words tend to be highly polysemious that show up with different meanings in different contexts.

(2) In either case, the various meanings can change significantly across different contexts.

(3) Because of the increased role of contexts in the description of meanings, introspection in itself does not provide a proper ground for the sufficient characterization of LUs. Therefore,

   (a) The lexicographic intuition should be underpinned by corpus evidence.

   (b) Moreover, a sound linguistic theory is needed to draw lexicographers' attention to the lexicographically relevant facts (cf. page 33).

   (c) In addition, for the same purpose, the exploitation of lexical profiling tools turned out to be indispensable, too, as corpus size increased (cf. page 40).

Drawbacks One serious practical disadvantage is that manufacturing a full-fledged sense-inventory is rather tedious requiring so much resource that is usually unavailable in the case of less resourced languages.

However, the wealth of language data may raise additional problems for lexicographers when compiling monolingual sense-inventories. Namely, lexicographers should be able to select the relevant linguistic facts and describe them in a consistent way. This task in turn is becoming more and more difficult as the corpus size increases. We saw that lexical profiling tools and linguistic theories may help in selecting the relevant facets of language data and describe them in a consistent way.

Corpus-driven approaches With regard to the basic assumptions corpus-driven and corpus-based approaches are rather similar: Corpus-driven approaches investigate contexts to derive meanings on the basis of them, as well. Probably, data plays a greater role, than in corpus-based lexicography, at least as far as corpus-driven techniques are suitable to handle greater amount of data than corpus-based approaches.

This is primarily due to the fact that the methodology has changed and unsupervised learning techniques became widely used for natural language processing tasks. Unsupervised techniques are able to eliminate unarticulated theoretical presumptions present in the labeling itself, since they are designed to learn hidden structurte from unlabeled data. Unfortunately, human intuition comes into play again when selecting the investigated phenomena, coming up with a representation set up, fine-tuning the parameters and throughout the evaluation, as well.

3

The Translation Phase

> Irrespective of the current status of equivalence in translatology, lexicography can definitely learn something from the decades of discussion [...]. Foremost among the conclusions reached is the impossibility of formulating a single universally valid definition of equivalence.
>
> ——— Adamska-Sałaciak (2010, p. 403)

3.1 Introduction

Translation Recall the workflow of dictionary building. As Figure 3.1 indicates, after characterizing the SL sense-inventory, the resulting senses should be translated into the target language. Thus, in this framework, the process of building a bilingual dictionary can be decomposed into three steps: In the first stage the source language LUs are characterized (analysis stage). In the next phase these LUs are translated (translations stage). Finally, the translated LUs (TRs) are compiled so that they could be included in a dictionary.

Linking However, Figure 3.2 depicts an alternative approach to the creation of bilingual dictionaries. In the first step two monolingual sense-inventories are characterized independently and in the next step instead of translating the SL monolingual database, the corresponding items of the SL and TL language databases are linked.

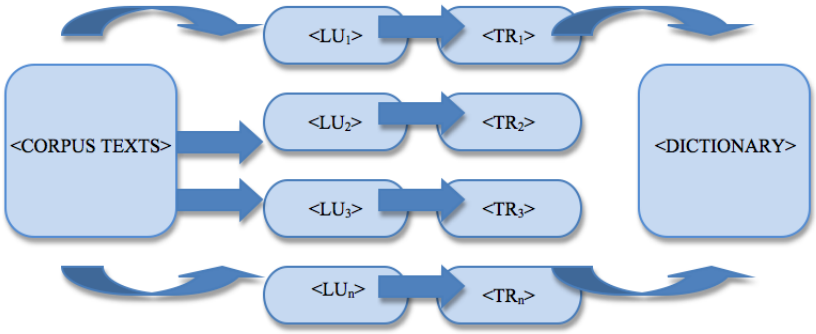


Figure 3.1: Dictionary building: Translation

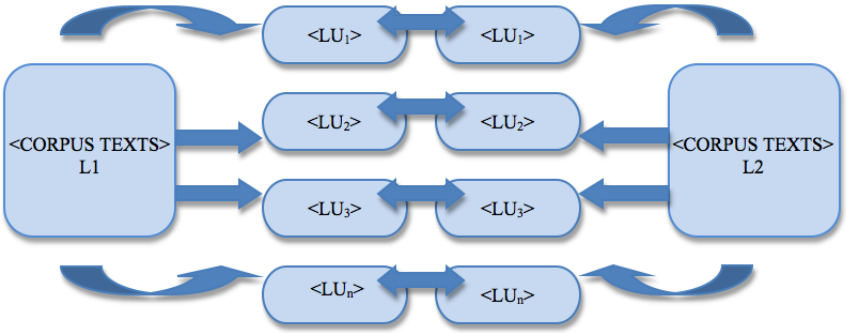


Figure 3.2: Dictionary building: Linking

In either case, translation relation has to hold between the corresponding SL and TL LUs. Chapter 3 focuses on the characteristics of translation relation.

Is translation relation a relation? Note that *translation relation* is a technical term in the field of bilingual lexicography. In fact, translation relation is not necessarily a relation in the mathematical sense. For the present chapter we presume that translation relation ρ is a mathematical relation: $\rho \subseteq A \times B$, where A is the SL vocabulary and B is the TL vocabulary. By the end of the chapter we will see what properties should the translation relation ideally exhibit and in Chapter 4 we will see which mathematical construction corresponds to these expectations even better. For that purpose let us consider the properties of translation relation more closely.

3.2 Translation Equivalency: The Best Translations

Interlingual synonymy Ideally, the relation of interlingual synonymy should hold between translation pairs. As an extension of the definition of monolingual synonymy (cf. page 50), interlingual synonymy can be conceived of as complete interchangeability, that is, the SL expression and the TL expression may occur exactly in the same set of contexts. Unfortunately, a perfect interlingual synonymy is only occasionally available. As Atkins and Rundell (2008) puts it :

> [...] pure synonymy is rare across languages, except for the names of concrete objects which the two cultures share. (p. 134-135)

Translation equivalency Thus, instead of interlingual synonyms, translation equivalents should be assigned to SL headwords, which can be thought of as the "best" equivalents. In the following sections the main types of translation equivalents will be examined based on Adamska-Sałaciak (2010). In her paper she focuses on the following questions:

(1) What is the nature of the relationship between an SL headword and its corresponding TL equivalent (e.g. identity, interchangeability, similarity)?

(2) Is equivalence a unitary concept or should different types thereof be recognised?

(3) Is equivalence 'discovered' (does it exist prior to being established by the lexicographer) or is it 'created' by the lexicographer's act?

We also add a further question that needs to be answered:

(4) Is there a methodology that is able to ensure that the best translation be included in the dictionary?

3.2.1 Relation between SL and TL headwords (Q1):

3.2.1.1 Translation relation is closeness

Based on the definition of interlingual synonymy we presuppose for the present discussion that translational identity and translational interchangeability are used in the same sense. As for the relation between SL and corresponding TL headwords there is a common agreement among lexicologists and lexicographers, namely, that

> the relationship is definitely weaker than identity: merely (maximum) closeness. (Adamska-Sałaciak, 2010, p. 392)

3.2.1.2 Arguments against interchangeability

According to her the reasons for this are at least twofold:

(1) Interchangeability and stylistic value First, interchangeability also presumes that equivalents have not only the same lexical meaning (if there is such thing), but also the same stylistic value. This expectation is especially important in the case of encoding dictionaries to be able to produce 'a smooth translation'. According to Adamska-Sałaciak (2010) to find such perfect equivalents is impossible in most cases due to inherent divergencies in the structures and vocabularies of languages.

(2) Interchangeability and symmetry Secondly, interchangeability or identity relations should be symmetric. However, as many author noted (cf. 3.3.2) dictionaries tend to be asymmetrical, that is, in many cases the TL expression is either more general or more specific than the SL expression. For instance, as she puts:

> the equivalent of both boyhood and girlhood in German is Kindheit (an instance of convergence), but if we start from the German word and look for its equivalent in English, we are likely to think of childhood, not boyhood and/or girlhood. (p. 392)

Figure 3.3 depicts the above example: The continuous arrows are representing the translation from English to German, while the dashed arrow stands for the reversed translation.

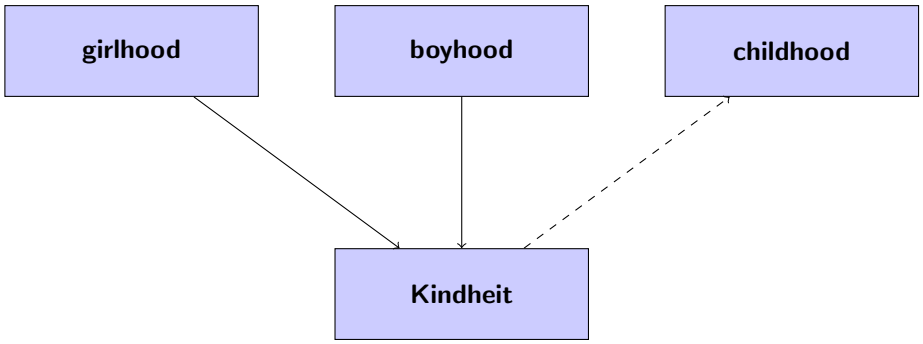


Figure 3.3: Translations of English LUs to German and back

3.2.1.3 Discussion of the arguments

The first argument The first argument is basically the same as that of Atkins and Rundell (2008, p. 134-135), which states that translation synonymy is rare across languages (cf. p. 63).

The second argument The second argument considering asymmetry needs to be given a closer look. How can we interpret the statement that *translation relation, as opposed to interchangeability (or identity), is asymmetric*?

Translation relation as relation First, translation relation ρ cannot be asymmetric (neither symmetric nor antisymmetric) relation in the mathematical sense. These properties are defined as follows:

$$Symmetric \qquad \forall a, b \in X, a\rho b \Rightarrow b\rho a \tag{3.2.1}$$

$$Asymmetric \qquad \forall a, b \in X, a\rho b \Rightarrow \neg(b\rho a) \tag{3.2.2}$$

$$Antisymmetric \qquad \forall a, b \in X, a\rho b \wedge b\rho a \Rightarrow b = a \tag{3.2.3}$$

As the definitions indicate, these relations are mappings from the elements of set X (domain) to the elements of the very same set X (range). Therefore, symmetry, asymmetry and antisymmetry hold only for homogenous relations. Here, both the domain vocabulary A and the range vocabulary B are made up of form-meaning pairs, therefore, the SL vocabulary A and the TL vocabulary B cannot coincide. Thus, the requirement of homogeneity definitively does not hold for bilingual dictionaries. That is, translation relation is neither symmetric, nor asymmetric, nor antisymmetric in the mathematical sense.

Example Figure 3.4 depicts a small part of a Hungarian-English dictionary to illustrate the possible connections between the elements of the SL vocabulary A and the TL vocabulary B.

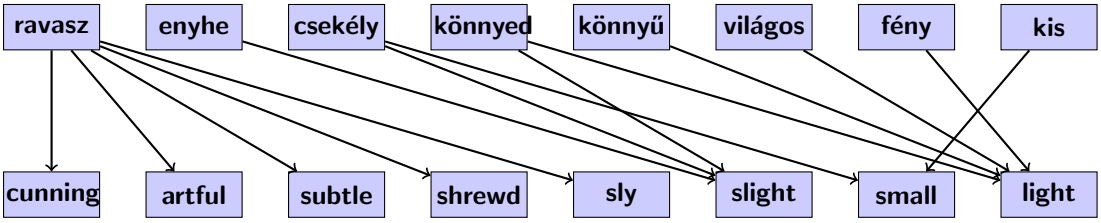


Figure 3.4: Mapping between the SL vocabulary A and the TL vocabulary B

Translation relation as mapping As Figure 3.4 represents, a dictionary as a mapping between A and B exhibits the following properties:

(1) Every $a \in A$ has at least one translation $b \in B$.

(2) Conversely, every $b \in B$ is the translation of at least one $a \in A$.

(3) The dictionary contains one-to-many mappings from A to B.

(4) The dictionary contains many-to-one mappings from A to B.

Thus, as Figure 3.4 indicates, the mapping might be extremely intricate, allowing both many-to-one and one-to-many mappings between the elements of A and B vocabularies.

Symmetric translation relation is an invertible function Recall the example of Adamska-Sałaciak (2010, p. 392) illustrating that translation is an asymmetric relation. The example, depicted in Figure 3.3, explicitly states that translational asymmetry means that

(1) In many cases the TL expression is either more general or more specific than the SL expression

(2) In some cases an SL word $a \in A$ the translation of which is the TL word $b \in B$ is not included in the TL' vocabulary A' in the reversed dictionary (e.g. $boyhood/girlhood \Rightarrow Kindheit \Rightarrow childhood$).

From (1) it follows that a symmetric translation relation ρ does not allow either one-to-many or many-to-one mappings. Whereas, (2) implies that for every $a \in A$, a should be part of the reversed dictionary, that is, be accessible as the reversed

translation of $\rho(a) = b$. Accordingly, $\rho^{-1}(\rho(a)) = a$, for every $a \in A$. These two requirements are satisfied iff the translation relation ρ is an invertible function.

Thus, on the basis of Adamska-Sałaciak (2010) if the translation relation ρ is symmetric in the lexicographical sense, then it can be thought of as an invertible function, whereas, if asymmetric, it can be conceived of either as a non-invertible function or a relation which is not a function.

Closeness and symmetry Recall the argument of Adamska-Sałaciak (2010), according to which *translation relation is best to interpret as closeness, since translation relation, as opposed to interchangeability (or identity), is asymmetric.* The intuitive conception of symmetry does not provide a strong argument against the identity interpretation of translation relation, as we do not have reason to consider closeness less symmetric than identity, since intuitively, if a is close to b, than b is also close to a. The same holds for similarity, too.

Closeness and quantifiability On top of that, considering translation relation as closeness, raises the question of quantifiability. If translation relation is closeness, we might ask how close are $a \in A$ and $b \in B$. Is $b \in B$ closer to $a \in A$ than $b' \in B$? That is, is b' a better translation of a than b? Therefore, $b \in B$ and $b' \in B$ should be comparable entities. In our view, a quantifiable translation relation is able to provide us with an ideal basis for comparison.

"Asymmetry" and quantifiability Consequently, on the one hand, since the translation relation is usually "asymmetric", it is best to conceived of as a non-invertible function or a mapping, which is not a function. On the other hand, interpreting translation relation as closeness implies that translation relation should be quantifiable, so that it could serve as a measure for comparing translations. This second expectation leaves us with the question whether there is a mathematical construction that is quantifiable and is able to reflect the "asymmetry" of the translation relation at the same time.

But before focusing on this question in Section 4.3.2, let us elaborate on the notion of translation relation. In the next section we will consider the types of translation relation that are distinguished in lexicography.

3.2.2 Types of translation equivalency (Q2):

The classification in Adamska-Sałaciak (2010) is discussed here, but as she notes, there is a wide agreement in the literature on the categories itself, albeit various authors tend to use different names for the various translation relations.

Cognitive equivalence (direct translation) Although cognitive equivalence[1] was originally conceived of as cross-lingual synonymy[2]—i.e. perfect interchangeability—and, thus, was the expected type of equivalence in a traditional bilingual dictionary, this notion has been extended to include more common types of equivalents, too. According to Adamska-Sałaciak (2010) cognitive equivalents may not only be the perfect translations but the very general counterparts of the SL LU, too. These equivalents are capable of conveying the overall meaning of the SL headword, but, by the same expression, are appropriate as its translations only in some contexts. Therefore, cognitive equivalents must cover the prototypical senses of the headword, but not necessarily its less central or more specialised senses.

Cognitive equivalence tends to be symmetric—that is, it tends to be an invertible function: In such cases it does not matter which language is the SL and which is the TL.

Contextual (translation) equivalence As interlingual synonymy is rare, cognitive equivalents are expected to cover only the prototypical senses of the SL headwords, hence, translational gaps may remain even if cognitive equivalents are available. Contextual equivalence[3] is particularly important where no direct translations are available. As Zgusta (1971) asserts:

> when choosing a translational insertible equivalent, the main concern is given (...) to its ability to be used in a fluent, good translation of whole sentences, to be inserted into contexts of the target language...(p. 319)

Thus, it follows that—as opposed to explanatory equivalents—the TL equivalent should be a lexicalized item of the TL. Translational equivalents are particularly

[1] Cognitive equivalents are referred to as direct translations in 8.2.2

[2] Cf. "a cognitive equivalent has to be identical with the source expression on all relevant dimensions of meaning" ((Piotrowski, 1994, p. 139)

[3] In this case the term of Atkins and Rundell (2008) is relied on. The motivation behind is to avoid the confusion that the original term *translation equivalence* may casue.

useful in the case of encoding dictionaries. However, according to Adamska-Sałaciak (2010), a bilingual dictionary could never give all type of translation equivalents of a given SL item, because it is impossible to predict all the contexts in which the item can occur.

Explanatory (descriptive) equivalence This type of equivalence appears as TL gloss in Atkins and Rundell (2008). The explanatory or descriptive equivalent is chosen in order to give information about the lexical unit of the SL by means of free syntagmas. Explanatory equivalents are rather on the notional than on the purely linguistic level, therefore, they are quite general. According to Zgusta (1971) as opposed to contextual equivalence:

> ...the explanatory or descriptive equivalent is chosen in order to give more information about the lexical unit of the target language. (p. 319)

Explanatory equivalents may be especially useful in the case of decoding dictionaries, if the user is a native speaker of the TL and wants to understand SL sentences.

Functional equivalence (near-equivalence) Zgusta (1984) defines functional equivalence as follows:

> Since languages differ in all imaginable respects, the translator-lexicographer must sometimes use means quite different from those used in the original in order to obtain the same results. If the different means do produce the same effect, the texts are considered functionally equivalent. (p. 151)

Unfortunately, the terms *same results* and *same effects* used in the definition are rather vague, therefore, this description is not suited to provide us with an accurate definition of functional equivalence.

Atkins and Rundell (2008)[1] illustrate functional equivalency with the French expression *A comme André* and with its English translation *A for Able*, which are used for spelling in both languages. As they assert:

[1] They use the term near-equivalence for functional equivalence.

> 'A comme André' doesn't translate 'A for Able', but is the equivalent phrase in the TL, used in exactly the same circumstances. (p. 212)

According to Atkins and Rundell (2008) the TL expression is not a cross-lingual synonym of the SL expression in this case, but their explanation comprises obscure parts, as well, since the expression *doesn't translate* is not specific enough. Moreover, this definition is contradictory, as it is hard to see the difference between 'being completely interchangeable'—a defining criterion for cognitive equivalence—and 'being used exactly under the same circumstances'—used in the above definition.

In our view, the purported implication of this example is that although the two expressions are used exactly under the same circumstances in both languages, the TL expression is not a compositional translation of the SL expression. The French word *comme* is basically translated as *like, as, just as, just like* into English and *André* is definitively not translated as *Abel* into English. In this case, functional equivalency is to be conceived of as a translation that fits some of the TL contexts, but the parts of the expression does not contribute with their lexical meaning to the whole meaning of the TL expression. This interpretation of the example is compatible with the observation of Atkins and Rundell (2008, p. 213), according to which the SL and TL items are often culturally equivalent in functional equivalency.

Note that if the notion of compositionality is involved in the definition of functional equivalence, the essential feature of functional equivalency is that the SL LUs are not translated with their lexical meaning (cf. *André* is translated as *Abel*), but instead they are translated in a way that best suits the TL context(s). This in turn is the same as contextual equivalence, thus, we do not see any reason to distinguish between the two categories, therefore, the category of functional equivalence is disregarded in the rest of the dissertation.

Our conclusion seems to be confirmed by Adamska-Sałaciak (2010), who claims that:

> It could be treated as a subtype of translational equivalence rather than a type in its own right. To my mind, however, the presence vs absence of word-level correspondence seems an important enough criterion to distinguish between the two. (p. 399)

3.2.2.1 Gradual nature of translation relation

If the translation relation may be better interpreted as closeness or similarity than interchangeability or identity, the gradual nature of translation relation has to be accounted for somehow. In the present section two scales will be introduced.

Explanatory equivalence and translation equivalence According to Adamska-Sałaciak (2010), explanatory and translation equivalence form a continuum. If a translation equivalent has no explanatory value at all, its use is limited to rather restricted contexts. On the other end of the scale, if the translation is wholly explanatory, i.e. is given as a definition, it impedes the production of idiomatic texts in the TL. Thus, as Adamska-Sałaciak (2010) notes:

> The boundary between the two kinds of equivalence is not sharp. Rather, Zgusta treats translational insertability and explanatory power as properties which one and the same equivalent may possess to different degrees: an ideal equivalent should be both insertable and explanatory. Irrespective of whether such an equivalent is available or not, the complementary properties of insertability and explanatory power should both be present in a bilingual dictionary entry. (p. 394)

However, we need to clarify what it means that a translation exhibits some degree of 'explanatory power'. Recall what we said about explanatory equivalents in Section 3.2.2: Explanatory equivalents may be especially useful in the case of decoding dictionaries, if the user is a native speaker of the TL and wants to understand SL sentences. That is, explanatory equivalents play an important role in a decoding setup.

Since our focus is on encoding dictionaries, we do not discuss explanatory equivalents and explanatory power in more detail here.

Cognitive equivalence and translation equivalence However, cognitive equivalence and translation equivalence might be conceived of as two ends of a continuum, too. While the SL and TL expressions may appear exactly in the same contexts in the case of perfect cognitive equivalency, the translation equivalent may show up only in a restricted set of contexts. That is, *the number of contexts in which a translation may appear can be a good indicator of how close the translation equivalent is to cognitive equivalency.*

3.2.3 Is equivalence discovered or created? (Q3)

Although the third question—whether translational equivalence is discovered or created—seems to be of somewhat philosophical nature, it has a special significance on the topics discussed so far. The same question discussed in Chapter 2 arises here again in a multilingual context: Are the relations between SL and TL language LUs 'out there', merely awaiting discovery, or are they non-preexistent entities that need to be created during the course of dictionary compilation? The firm belief in pre-existent equivalence relation between pre-existent meanings entails the conviction that translations are objective—or at least highly intersubjective entities, readily available for expert bilingual lexicographers. From this it follows that relying on expert lexicographers' introspection should yield the same result in every case. On the other hand, if translation equivalencies are created entities, they allow for greater variation even for the same individual on different occasions.

As Adamska-Sałaciak (2010) suggests, a more subtle distinction is in order here according to the type of equivalency:

> It appears that as many as three of our four types are 'created' (constructed online) rather than 'discovered'. The only exception is cognitive equivalence, whose identification by skilled bilinguals is characterised by a high degree of intersubjective agreement, which may culminate in its objectification. A competent bilingual speaker can access a TL cognitive equivalent of a SL item offline, that is, without being provided with any context. Lexicographers have traditionally been assumed to belong in this group of skilled bilinguals, the uninformed view being that they can produce equivalents more or less effortlessly. When it comes to cognitive equivalence, this view is just about right; in all other cases, things are much more complex. (p. 400)

Thus, Adamska-Sałaciak (2010) puts a great emphasis on the role of context when finding the best translation.

3.2.4 Expectations toward the automatically attained translation relation

In what follows, we will give a brief summary of the properties of the translation relation that should be retrieved automatically.

(1) As translational synonymy/cognitive equivalency is rare, translation relation is best conceived of as *closeness*.

   (i) Closeness is *gradual*: As we saw in the previous section cognitive and contextual equivalencies constitute a scale according to insertability. In order to find the best translation, a suitable method should be found that is able to measure where the TL equivalent is situated on the scale. *As both cognitive equivalence and contextual equivalence can be defined in terms of the set of contexts in which the TL counterparts may appear, presumably, contexts of the SL and the TL should be heavily relied on.*

   (ii) Closeness should be also quantifiable so that the best translation could be selected.

   (iii) As "symmetric" translation relation is rare (i.e. when the translation is an invertible function), the automatically attained translation relation should capture this asymmetry somehow. Note that the gradual, possibly quantifiable notion of translation equivalence is incompatible with the binary view of translation pairs, therefore, with the mapping view of translation.

(2) We primarily expect the dictionary to enable the user to produce smooth translations (cf. encoding dictionaries).

3.3 Linking Monolingual Sense Inventories

3.3.1 Introduction

Translation In the previous section we investigated the translation relation. The translation relation might emerge as the result of translating the SL sense-inventory. In fact, translating a monolingual sense-inventory is the most common way of producing a bilingual dictionary.

Linking An alternative approach is *linking*. In this case we start out of two independently characterized monolingual sense-inventories. In the next step the entries of the two databases are aligned. The essential benefit of linking over translation is that the monolingual databases are designed in a way that enables reusability via *reversibility* or via *the hub-and-spoke model*. Reversibility will be discussed in section 3.3.2, while the hub-and-spoke model will be covered in Section 3.3.3.

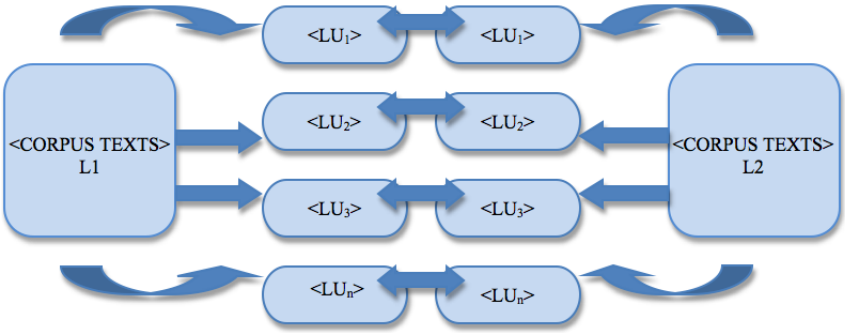


Figure 3.5: Linking: SL and TL LUs are independently characterised. In the next phase the corresponding LUs are linked.

As Figure 3.5 indicates, during the course of linking lexicographers are ideally provided with two neatly characterized monolingual sense-inventories, one for the SL and one for the TL. This means in fact that dictionaries compiled through linking are expected to be symmetrical, that is, SL and TL are expected to be interchangeable (see Martin, 2007). This entails that the translation relation should be also symmetrical in this case. Consequently, although the translation relation considered to be prevalently asymmetric in translated dictionaries, as opposed to them, translation relation is expected to be symmetric in the case of linking dictionaries. As we will see, this expectation completely corresponds to the "invertible function" interpretation of symmetric translation relation. In Section 3.3.2 the main preconditions of reversibility of a bilingual dictionary will be considered.

3.3.2 Reversibility

Reversibility and symmetry A bilingual dictionary is said to be reversible if the source language and the target language are interchangeable. Symmetry of translation and reversibility of the dictionaries are tightly related notions. The dictionary is reversible, iff "ρ translation relation is symmetric", i.e. iff ρ is an invertible function. In

this case for every $a \in A$ if $\rho(a) = b$ then $\rho^{-1}(b) = a$, that is, $\forall a \in A\ \rho^{-1}(\rho(a)) = a$. In other words, ρ should be a one-to-one mapping between the SL vocabulary A and the TL vocabulary B.

Reversibility and the linked LUs Now, we have seen that the reversibility of a dictionary means that the translation relation ρ is an invertible function. From a dictionary user perspective it is also required that the linked entities should be of the same type (cf. Martin, 2007).

With regard to the types and properties of the linked entities, Veldi (2010) indicates three main reasons that impede the successful reversion:

(1) The TL side of the dictionaries tend to comprise explanations rather than cross-lingual synonyms or near-synonyms. In this case the SL side of the reversed dictionary would comprise definitions instead of headwords.

(2) The provided equivalents are inaccurate or vague.

(3) Lexical poverty with regard to the range of possible equivalents, resulting in a low coverage of the TL vocabulary.

Reversibility and cognitive equivalency Recall that cognitive equivalents are the best candidates to be included in a reversible dictionary. This is basically due to the fact that the translation relation ρ of cognitive equivalents is a one-to-one mapping. Moreover, cognitive equivalence relation links entities strictly of the same type.

Nevertheless, as cognitive equivalency is only rarely available, reversibility is a property that should be anticipated in the design of a bilingual dictionary, according to Veldi (2010). In what follows, we will consider a project which aimed at compiling a wealth of reversible dictionaries by means of linking.

3.3.2.1 The CLVV project and linking

In the CLVV project twenty dictionaries were constructed during the period 1993-2007. In all cases Dutch was either the SL or the TL. The average dictionary comprises 45.000 entries with rich microstructures. From a meta-lexicographic point of view the dictionary projects had two main objectives:

(1) To put the linking method into practice and construct reversible dictionaries

(2) To prove that on the basis of two reversible dictionaries further dictionaries can be generated in a semi-automatic way (hub-and-spoke model) (see 3.3.3.1 in this chapter).

Linking is enabled by the following features:

(1) Instead of associating pure word forms they associated suitably characterized LUs.

(2) The design of the monolingual Dutch database (RBN) is such that it enables the linking method (cf. 2.3.2.6).

(3) OMBI dictionary editor tool facilitates the definition of reversible meaning pairs.

OMBI dictionary editor tool Based on the detailed presentation found in Maks (2007), the main advantage of OMBI dictionary editor tool is that it helps the creation of reversible bilingual dictionaries. This tool was used in about half of the CLVV projects. As Tamm and Martin (1996) puts:

> While the editing function is busy creating a bilingual database X to Y, and as such taking in translations from X to Y, OMBI simultaneously stores the reversed counterparts, thereby building a reverse database Y to X. The end result is a non-directional bilingual database, from which databases and/or dictionaries in both directions can be automatically derived at a subsequent stage.

Accordingly, the editor is composed of two main components:

(1) Two language components where each language is described as a fully autonomous monolingual resource, without being tailored as a source or a target language. The resources consist of form units, lexical units and example units.

(2) An interlingual component which is a collection of links between lexical units and example units of the two languages. Additional constraints on reversibility might be also defined, therefore unreversible links can be also used, if it is necessary.

Moreover, OMBI also classifies the translations into different classes according to their reversibility. Although the editor helps in the development of reversible dictionaries, the need of post-editing still remains.

Linking and less resourced languages Unfortunately, linking is not directly applicable in the case of less resourced language pairs to produce bilingual dictionaries. The main reason for this is that suitable databases usually are not available for such languages. Note that even if neatly characterized sense-inventories were available for that languages, linking should be performed to compile bilingual dictionaries. The objective of the hub-and-spoke model is to decrease the number of linking procedures needed to create bilingual dictionaries. Accordingly, in what follows, the hub-and-spoke model will be described.

3.3.3 Linking monolingual databases via a hub

3.3.3.1 Hub-and-spoke model

The aim of the hub-and-spoke model (Martin, 2007) is to generate multilingual dictionaries from *reversible* bilingual dictionaries. On an abstract level, the hub-and-spoke model follows the steps bellow:

(1) The first step is the generation of a reversible bilingual dictionary with languages A and B. ($A \leftrightarrow B$)

(2) The next step is adding a third language in a reversible fashion to the language A, thus generating the links $A \leftrightarrow C$.

(3) Finally, the links between C and B should be inferred by means of derivation rules, creating the links $B \leftrightarrow C$.

In this case, language A was the *hub language* and languages B and C were the *spoke languages*. Therefore, spoke-languages are not linked directly to each other but via the hub-language.

Note that the hub-and-spoke model works properly only on reversible dictionaries. The underlying reasons are the following:

(1) Let ρ_{AB} the translation relation translating from A to B, whereas ρ_{AC} translates from A to C. It is quite easy to see that both ρ_{AB} and ρ_{AC} should be invertible functions, that is, one-to-one mappings mapping every element $a \in A$. This means that $\rho_{AB}(a) = b$ and $\rho_{AB}^{-1}(b) = a$. Similarly, $\rho_{AC}(a) = c$ and $\rho_{AC}^{-1}(c) = a$.

Now we are looking for the translation relations ρ_{BC} mapping from B to C and ρ_{CB} mapping from C to B so that $\rho_{BC}(b) = c$ and $\rho_{CB}(c) = b$ if ρ_{AB} and ρ_{AC} are given. Then, $\rho_{CB}(c) = \rho_{AB}(\rho_{AC}^{-1}(c)) = b$ and $\rho_{BC}(b) = \rho_{AC}(\rho_{AB}^{-1}(b)) = c$.

Were ρ_{AB} and ρ_{AC} non-invertible functions, c and b could not be uniquely linked[1].

(2) The entities linked should be cognitive equivalents—recall, the best candidates for reversible dictionaries—or neatly characterized LUs of the same level of semantic specification in all three languages.

3.3.3.2 The CLVV project and the hub-and-spoke model

In the framework of the CLVV project (Martin, 2007) the hub-and-spoke model was put into practice by the derivation of a Danish-Finnish dictionary from the Dutch-Finnish and Dutch-Danish dictionaries. Consequently, Dutch played the role of the hub and Danish and Finish were the spoke-languages. As for Dutch, they relied on the Reference Database of Dutch, while the corresponding Danish and Finnish monolingual databases were produced in parallel with the bilingual dictionaries. The most important requirements for the successful merging were met:

(1) The entries of spoke language databases are of the same structure.

(2) They show the same level of semantic specification.

(3) The core of the example units is largely alike.

The linking process between the spoke languages is semi-automatic because post-editing is needed, but the amount of labour is reduced drastically.

3.3.3.3 Linking wordnets and framenets via a hub

When considering the hub-and-spoke model the question naturally arises if it is possible to exploit sense-inventories produced for multiple languages for that purpose.

[1]However, there is one case, when this statement does not hold, namely, if there is a homonymy in B (for instance, *crane* in a spoke-language English) which similarly appears as homonymy in C (for instance, *Krahne* in a spoke-language German), as well, but which is designated by different word forms in the hub-language A.

For its wide coverage regarding various languages, wordnets seem to be a reasonable choice. National wordnets are usually the results of translating the Princeton WordNet and thus, these wordnets are aligned to Princeton WordNet. Therefore—although not one-to-one mapping—the linking is given beforehand. However, as national wordnets tend to be the result of translation, we may face the problem of lexical poverty in the TL side (cf. p. 75). This entails the unfortunate consequence that linking wordnets of lesser-used languages is likely to result in poor dictionaries. This adverse finding impels us to not exploit wordnets for our purposes, even if wordnets cover a great number of languages and for that reason applying the hub-and-spoke model to wordnets might seem to be a promising thread of research.

As opposed to wordnets, due to their corpus-based nature, FrameNet-type databases seem to be the ideal candidates for the hub-and-spoke model. Unfortunately, the restricted number of currently available framenets[1] makes it pointless to apply the model.

3.4 Conclusion

After considering the possible types of sense-inventories in Chapter 2 and elaborating the types and properties of translation relation in Chapter 3, we are now in the position of giving an overview of the possible expectations toward a suitable methodology. But before doing so, we first summarise the projects discussed so far in Figure 3.6 below.

3.4.1 Types of dictionaries

Sense-inventories and corpus data The upper half of the figure indicates the SL sense-inventories according to their relation to corpus data, as discussed in the previous chapter. Accordingly, *introspective*, *corpus-based* and *corpus-driven* approaches are distinguished. Similarly, the bottom part represents the TL sense-inventories, which also can be *corpus-based* or *corpus-driven* with regard to their relation to corpus data. Nevertheless, instead of 'introspective' the category label 'NO' is relied on, indicating that in several cases the sense-inventory is merely the result of translation and such, cannot be considered as an independent sense-inventory.

[1] Currently Chinese, Brasilian, German, Spanish and Japanese FrameNets are listed on the project's website at `https://framenet.icsi.berkeley.edu/fndrupal/framenets_in_other_languages`

Translation or linking The TL sense-inventory might be the result of translating the SL data-base, or might be characterized independently of the SL sense-inventory. In the latter case the relation between the two is linking. Various possibilities are depicted in the figure below, where the thick line arrows represent the process of translation, while the dashed double arrows stand for linking. Some examples are also listed under each nodes. Note that as Figure 3.6 indicates, the two properties, i.e relation to corpus data and the type of correspondence between the SL and TL LUs are slightly interrelated characteristics.

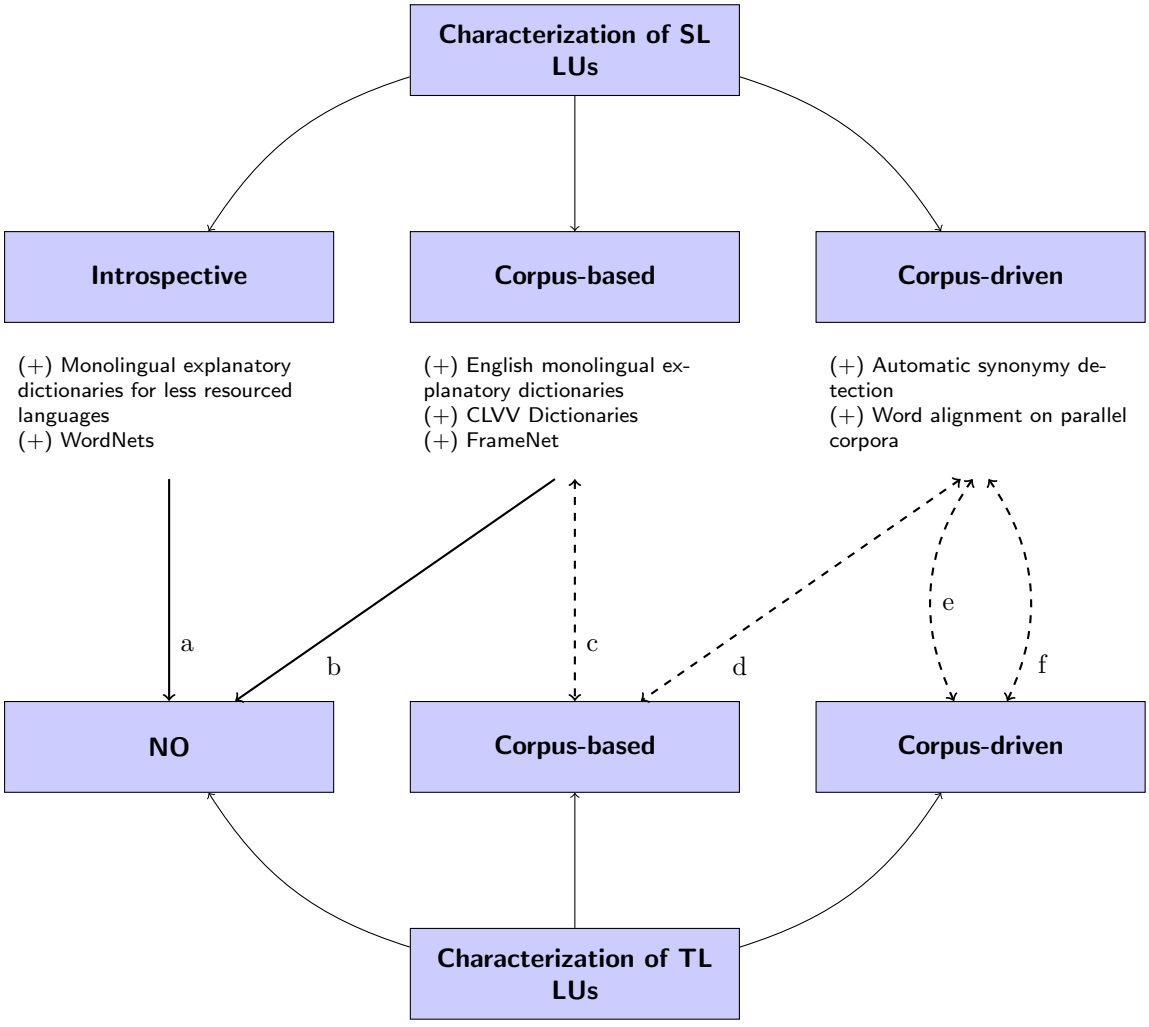


Figure 3.6: Possible approaches to building a bilingual dictionary

Dictionaries of type (a) and (b) Accordingly, edge (a) represents dictionaries

typically for lesser used language pairs (e.g. Bojtár, 2007) or WordNets being translated from the PWN. Edge (b) typically denotes dictionaries with a well-resourced SL and a lesser used TL. Dictionaries of type (a) and (b) are not really apt for being reversed, as the TL language vocabulary does not cover the TL language. This kind of dictionaries tend to serve decoding purposes.

Dictionaries of type (c) and (d) On the other hand, edges (c) and (d) are dictionaries where both the SL and TL sense-inventories are characterized independently and due to the corpus-based method the TL vocabulary of the dictionary should comprise the important lexemes of the TL. This makes linking possible, which in turn is the prerequisite of reversibility. We are not aware of type (d) dictionaries but compiling dictionaries in this way seems to be theoretically possible. CLVV dictionaries represent type (c) dictionaries. Linking framenets may also result in type (c) dictionaries, but owing to the limited scope of languages for which framenets are available, only a few language pairs—typically well-resourced ones—can be covered. Both (c) and (d) type dictionaries may be useful for encoding.

Dictionaries of type (e) and (f) Although theoretically it is possible to link two automatically characterized databases manually, linking took place automatically in both experiments discussed in Sections 2.3.3.4 and 2.3.3.5. Here, albeit edges (e) and (f) are of the same type and map from the same node to the same node, they represent different techniques. In the first case the sense-inventories are characterized independently by means of vectors and then are mapped through a combination of rotating and scaling. Note that this technique does not produce sense-inventories, that is, it retrieves translation pairs without treating polysemy.

As opposed to (e), (f) is a quite widespread technology in the field of NLP to produce bilingual lexicons for computational purposes. Similarly to dictionaries of type (e), word alignment does not create sense-inventories, rather it operates on mere word forms. Nevertheless, as opposed to (e) dictionaries, (f) dictionaries as bilingual dictionaries are able to handle multiple meanings.

As we will shortly see, (f)-type dictionaries are the focus of the remaining part of thesis. In fact, we aim to prove that *dictionaries of type (f) are suitable for not only computational purposes but for human end-users, too.* Moreover, such automatically generated dictionaries have certain benefits over traditional or corpus-based dictionaries. For a more detailed presentation see Chapter 4.

Dictionary types and the hub-and-spoke model Due to the requirements put

forward by Martin (2007) discussed in Section 3.3.3.2 in more detail, dictionaries which are the result of translation do not easily lend themselves to be spoke-hub dictionaries (e.g. the TL vocabulary for the lesser-used language is not reliable). Thus, neither dictionaries of type (a) nor of type (b) are suitable for generating hub-hub dictionaries. This entails the unfortunate consequence that linking wordnets of lesser-used languages is likely to result in poor dictionaries. This adverse finding impels us to not exploit wordnets for our purposes, even if wordnets cover a great number of languages and for that reason might seem to be a promising thread of research.

As opposed to this, due to their corpus-based nature, FrameNet-type databases seem to be the ideal candidates for the hub-and-spoke model. Unfortunately, the restricted number of currently available framenets makes it pointless to apply the model.

The CLVV dictionaries were designed in a way that enables the semi-automatic generation of a hub-hub dictionary. However, the creation of the suitable databases is quite a tedious task. Consequently, the idea to use the hub-and-spoke model in the generation of dictionaries for lesser used languages was rejected.

3.4.2 Expectations toward a suitable methodology

As it was discussed in Chapters 2 and 3, the suitable methodology has to meet expectations concerning the relation to language data, the characteristics of translation relation and has to be economical.

Translation relation In Chapter 3 four types of translation relation were distinguished: Cognitive equivalence, contextual equivalence, explanatory equivalence and functional equivalence. We found that the translation relation exhibits the following properties and should meet the following expectations:

(1) The translation relation is supposed to be primarily *asymmetric* except for cognitive translation relation, which turned out to be restricted to special semantic domains, such us natural kinds or artifacts that the two cultures have in common. We saw that the relation conception of dictionaries implies that the "symmetry" of ρ translation relation is best to think of as an invertible function. Therefore, that in this framework ρ is asymmetric if either it is a relation and not a function or if it is a non-invertible function. The automatically attained translation relation should be able to reflect *asymmetry* of the translation relation.

(2) The translation relation should be thought of as *closeness* instead of identity or interchangeability. In fact, strict interchangeability holds only in the case of cognitive equivalence.

(3) If the translation relation is best conceived of as closeness, i.e. it is a gradual notion, it should be *quantifiable*, so that the best translation could be selected among the possible translation candidates. However, due to its gradual nature, the translation relation cannot be thought of as a a relation in the mathematical sense, as, in fact, mathematical relation is a binary decision indicating whether a given LU-pair is part of the dictionary or not. Thus, a better mathematical notion has to be sought in order to quantify translation equivalence, which, at the same time, is able to reflect the asymmetric nature of translation relation.

(4) Closeness has two aspects to quantify over:

   (a) *Explanatory power*: It was mentioned that mere explanatory equivalents and mere contextual equivalents form a scale, as most contextual translation exhibit some degree of explanatory power, as well. Nevertheless, we did not discuss explanatory power separately, as explanatory equivalence plays a central role primarily in a decoding setup.

   (b) *Translational insertibility*: Translational insertibility is a good measure of the perfect translation, i.e. translation equivalency: Perfect translations are cognitive equivalents, which are interchangeable in every possible contexts. On the other end of the scale we find mere contextual translations, that are insertable only into a single context. Most translations are somewhere between these two extremities, and the automatically attained translation relation should be able to estimate where the translation is situated on the scale.

Relation to language data

(1) The reliance on intuition should be decreased.

   As it was discussed in Chapter 2, traditional lexicography is prevalently based on intuition or citation collecting. As opposed to it, the corpus-based approach heavily relies on language data, but it works best in the presence of an underlying linguistic theory that helps lexicographers to interpret the linguistic phenomena in question in the same way. Meaning-Text Theory, Levin's theory of verbs, FrameSemantics and Theory of Norms and Exploitations were covered in Section 2.3.2. However,

such theories may have their own unarticulated presuppositions, which again calls into play lexicographers' intuition when making decisions on vague issues.

Some of the assumptions can be eliminated through relying on unlabeled data, therefore, applying an unsupervised learning technique seems to be a reasonable choice. Consequently, *a data-driven approach should be exploited to construct bilingual dictionaries by possibly relying on an unsupervised learning technique.*

(2) Contexts should play a great role when characterizing meanings both in the TL and in the SL.

(3) Since we are planning to build encoding dictionaries[1], ample example sentences from real language use are also required to give hints on the proper use of the target expressions.

Economical considerations On top of the above considerations the proposed technique should satisfy certain economical requirements, too.

(1) The proposed method has to be economical, consequently hand-crafted databases are not suitable for our purposes. This excludes corpus-based methods.

(2) The method is expected to be language independent and easily re-applicable, thus, a consistent methodology is searched for, which is more or less independent of the peculiarities of SL and TL.

(3) The resulting dictionaries should be easily reversible.

The main finding of this thesis is that *the automatic estimation of conditional probabilities on the basis of parallel corpora, which is a widely used method in the natural language processing community but primarily for machine translation purposes, is able to meet the above requirements.*

The rest of the thesis is devoted to this topic. Chapter 4 explores to what extent conditional probability may serve as translation relation, especially in the case of encoding dictionaries.

[1]Recall that according to Melčuk (2006) they are more complex than decoding ones, thus easily convertable into decoding dictionaries by adding some extra features regarding look-up mechanisms.

4

Encoding Dictionaries and Conditional Pobability

> Zgusta (1984, p. 147) stipulated that, ideally, a bilingual dictionary should provide "real lexical units of the target language which, when inserted into the context, produce a smooth translation" [...] As we know – and as Zgusta knew better than anyone else – these are higly unrealistic expectations.
>
> ——— Adamska-Sałaciak (2010, p. 392-393)

4.1 Introduction

The previous chapters were centered around the dictionary building process: In Chapter 2 various sense-inventories were investigated primarily with regard to their relation to corpus data, while in Chapter 3 the main properties of translational relation were elaborated. We have also made a distinction between decoding and encoding dictionaries, claiming that the latter ones are more intricate to produce, thus the creation of encoding dictionaries may impose additional constraints on the sense-inventories A and B and on the translation relation ρ. Therefore, we are now in the position of considering our mottos:

4. ENCODING DICTIONARIES AND CONDITIONAL PROBABILITY

> In order to say what a meaning *is*, we may first ask what a meaning *does*, and then find something that does that. (Lewis, 1970)

> In order to say what a translation relation *is*, we may first ask what the translation relation *does*, and then find something that does that. (Based on Lewis, 1970)

Hence, the present chapter is made up of two main parts: First, we examine what kind of properties an SL sense-inventory should necessarily exhibit to serve as a basis for an ideal encoding dictionary. As a by-product we will come up with an interesting interpretation of meaning. Secondly, we aim to show that conceiving translation relation as conditional probability conforms to these criteria.

SL sense-inventories and word sense disambiguation In an encoding environment we are aware of the meaning of the source expression and want to find the contextually best translation for it, i.e. the translation that produce the "smoothest translation" when put into TL contexts. Recall that contrary to decoding dictionaries, in the case of encoding ones linguistic competence plays a much greater role, as in this case we cannot rely on common sense knowledge or situational information readily available in decoding environments.

Thus, high-quality encoding dictionaries—more than decoding ones—require an SL sense-inventory, where the meanings are or can be contextually anchored. But how do we know that we are "aware of the meaning of the SL expression"? From now on, the meaning of an expression will be considered known, if most native speakers assign the same meaning to it, when put into sufficiently specified contexts. This task is called word-sense disambiguation (WSD). Accordingly, an appropriate SL sense-inventory should enable us to obtain good results on word sense disambiguation tasks. There are a couple of ways to measure the quality of word sense disambiguation.

Inter-annotator agreement In the present discussion we confine ourselves to quantifying the results in terms of *inter-annotator agreement*: Accordingly the sense inventories should be characterized in a way that high inter-annotator agreement may be achieved on sense disambiguation tasks. In Section 4.2 four word sense disambiguation experiments will be presented, yielding the result that none of the investigated sense inventories corresponds to our expectations.

Main characteristics of SL sense-inventories Given the results of the word sense disambiguation experiments some conclusions can be drawn regarding the main char-

acteristics of SL sense-inventories for encoding dictionaries. First, a sense-inventory should consist of ample *contextual information*, which can be relied on when selecting the best sense for the given context. Moreover, ideally, senses should be divided in a way that facilitates the unique assignment of word occurrences to a specific meaning. Such a division is called *partition* and it will be discussed in Section 4.2.2.5 in more detail. Since corpus-based monolingual dictionaries are compiled on the basis of great amount of language data they might be suitable to serve as the starting point for high-quality encoding dictionaries. However, the production of hand-crafted SL sense-inventories requires so much effort that is usually not available in the case of lesser used languages. Therefore, instead of exploiting neatly characterized SL sense-inventories another alternative should be considered.

Extracting bilingual dictionaries on the basis of word alignment In the second part of this chapter our basic objective is to prove that contrary to what Adamska-Sałaciak (2010) believes Zgusta's view on bilingual lexicography is not unrealistic, at all. On top of that, retrieving "real lexical units of the target language" can be achieved completely automatically. For that purpose we introduce word alignment on parallel corpus—a technique, which is widely known in the natural language processing community, but as far as we know, has not been used in real lexicographic project until now.

Section 4.3 introduces word alignment on parallel corpora through giving an intuitive picture of it. The following issues are addressed in this section:

(1) To what extent does the automatically attained translation relation satisfy the expectations put forward in Chapter 3? (Section 4.3.2)

(2) How are translation equivalents retrieved during the process of word alignment? (Section 4.3.3)

(3) How can be translation equivalents conceived of? (Section 4.3.4)

(4) To what extent does the proposed technique meet the expectations put forward in Chapter 2? (Sections 4.3.6 and 4.3.7)

4.2 Word Sense Disambiguation Tasks

In Subsection 4.2 four studies (Véronis, 2003; Kuti, Héja, and Sass, 2010 and Héja, 2008) will be presented. All the experiments aim at determining to what extent it is possible to select the most appropriate meaning of a word in context on the basis of sense distinctions in traditional sense inventories[1]. We presume that contexts are never underspecified, that is, theoretically it is possible to determine the right meaning based on the provided contexts.

The most appropriate meaning can be thought of as the one that is selected with high agreement. Accordingly, the experiments presented below measure the agreement among annotators i.e. *inter-annotator agreement*. In what follows, a concise description of the inter-annotator measures will be presented based on Artstein and Poesio (2008).

4.2.1 Measures of inter-annotator agreement (ITA)

In what follows, we shortly introduce some possible measures of ITA based on Artstein and Poesio (2008).

Observed agreement One simple approach to determine ITA is the percentage agreement or observed agreement (A_o) for two annotators, where I denotes the set of items i of cardinality $\mathbf{i}$. Furthermore, $agr_i = 1$ if the two coders assign i to the same category, and 0 if the two coders assign i to different categories.

$$A_o = \frac{1}{\mathbf{i}} \sum_{i \in I} agr_i \tag{4.2.1}$$

However, this measure of ITA does not account for cases where agreement is due to chance. One of the two factors that influences chance agreement is the number of categories used in the annotation task: The fewer categories are used to classify a certain phenomenon the higher agreement by chance might be expected. Since various sense-inventories contain diverse divisions of senses, our measure has to handle such cases to be able to compare the usefulness of sense-inventories above chance.

Chance-corrected agreement The calculation of the chance-corrected inter-annotator

[1]That is, how difficult it is to assign the right meaning to words in context without explicitly given distributional information.

agreement coefficients takes chance agreement into consideration (A_e):

$$A_o = \frac{A_o - A_e}{1 - A_e} \tag{4.2.2}$$

Here 1–A_e measures how much agreement over chance agreement is attainable, while A_o–A_e tells us how much agreement over chance agreement was actually found. In cases where agreement is lower than expected, this measurement unit can take a negative value. The closest the obtained value is to 1, the higher the possibility that the agreement between the annotators is not by chance.

Chance-corrected agreement for two coders Two widely used chance-corrected coefficients of ITA between two annotators are Scott's π (Scott, 1955) and Cohen's κ (Cohen, 1960).

Scott's π assumes that if coders were operating by chance alone, their assignment would yield the same distribution for each coder, thus A_e does not reflect individual annotator bias. Cohen's κ only differs from Scott's π in that it presupposes separate distributions for each of the coders.

Thus, in the case of Scott's π the prior distribution is estimated on the basis of the observed assignments, i.e. the total number of assignments to the categories by both coders divided by the overall number of assignments where n_k stands for the total number of assignments to category k and i for the number of items to be assigned. The estimation of the prior distribution:

$$\frac{n_k}{2i} \tag{4.2.3}$$

Then, given the assumption that coders act independently, expected agreement is determined as follows, where K designates the set of categories:

$$A_e^{\pi} = \frac{1}{(2i)^2} \sum_{k \in K} n_k^2 \tag{4.2.4}$$

Chance-corrected agreement for multiple coders However, being invented for two annotators, Scott's π is not apt to measure agreement among multiple coders. Therefore, we relied on Fleiss's multi-π (Fleiss, 1976) throughout our analysis, which is a generalization of Scott's π for multiple coders. The basic idea behind this coefficient is that A_o cannot be thought of as the percentage agreement defined above. This is due to the fact that in the case of multiple annotators necessarily there will be items on which some coders agree and others disagree. The proposed solution is to compute

pairwise agreement as the proportion of agreeing judgment pairs and the total number of judgement pairs for that item. The overall A_o will be the mean of the pairwise agreement for all items. Here i stands for the number of items, c for the number of coders, and n_{ik} for the number of times an item is classified in category k. I denotes the set of items while K denotes the set of categories, for instance the senses in the sense inventory.

$$A_o = \frac{1}{(ic)(c-1)} \sum_{i \in I} \sum_{k \in K} n_{ik}(n_{ik} - 1) \tag{4.2.5}$$

In the case of multiple coders A_e i.e. the agreement by chance might be conceived of as the probability that two arbitrary coders would make the same judgement for a particular item by chance. Holding the same presuppositions about the distribution of the judgements as Scott, A_e is calculated in the same way as in the two coder case, except for the fact that instead of 2 coders c coders make the assignments, that is c assignments need to be considered, when calculating the mean.

An additional advantage of Fleiss's multi-π and Cohen's κ is that they are both insensitive to categories that were never selected by any of the annotators, therefore the results do not reflect how many categories the annotators could originally choose from.

4.2.2 Studies

4.2.2.1 Véronis' first experiment

Experimental setup Véronis (2003) first experiment was concerned with agreement on polysemy – that is, the extent to which coders agreed that a word was polysemous. Six fourth-year linguistic students were asked to decide whether a word in the context of one paragraph has multiple meanings or only one single meaning in the case of 600 occurrences of 600 French words (200 nouns, 200 verbs, 200 adjectives). Besides, the answer *'I don't know'* was also available. Note, that if a context was underspecified regarding polysemy, the relevant answer would be '*I don't know*'. The low proportion of such answers (4.05%) implies that the majority of contexts were specific enough to make a decision on polysemy.

Results Interestingly, in spite of the low rate of *'I don't know'* answers there was a considerably low agreement regarding the polysemous nature of the words in contexts:

0.67 for adjectives, 0.36 for nouns and 0.37 for verbs in terms of the extended version of Cohen's κ to multiple coders.

According to Véronis (2003) these results show that:

> individual informants had no trouble making spontaneous judgements, but different informants tended to make different judgements.

4.2.2.2 Véronis' second experiment

Experimental setup The 20 words from each category perceived by the coders in this first experiment to be most polysemous were then used in a second study. 3724 occurrences of the words in question had to be sense-tagged in the context of a one-paragraph text on the basis of the Petit Larousse explanatory dictionary by 6 different forth-year students in linguistics. Annotators were instructed to chose either one sense, or several senses if they felt that more than one was appropriate in the given context. They could also choose no sense at all, if they felt that none of the senses listed in the dictionary fit the context. Therefore, in this experiment multiple choices or zero choice might have been an indicator if a context was not specific enough for the selection of a single sense. The inter-annotator agreement was computed using the generalized version of Cohen's κ to multiple coders.

Results The inter-annotator agreement in terms of Cohen's κ for multiple coders was relatively low for all the three investigated POS-categories: 0.41 in the case of verbs and adjectives and 0.46 for nouns. Considering the fact that usually 0.8 is accepted as a threshold for reliable agreement (see Artstein and Poesio, 2008), the obtained values imply that Petit Larousse senses are not suitable for the sense-tagging of tokens in their contexts.

4.2.2.3 Experiments of Kuti et al.

We carried out a similar study for Hungarian verbs (Kuti, Héja, and Sass, 2010) based on two different sense inventories yielding approximately the same conclusion.

Experimental setup One of the sense inventories used was the Hungarian Explanatory Dictionary, which is the official reference work as a Hungarian monolingual dictionary. Another sense inventory tested was the Hungarian WordNet (HuWN) (Miháltz et al., 2008). HuWN is a lexical database, modeled partly upon the Princeton WordNet

2.0 (PWN) (Fellbaum, 2005) for English. The basic unit of HuWN, as of all wordnets, is a synset and not that of traditional dictionaries, i.e. a lexeme. It is important to note that when deciding on what verb senses should be incorporated into the Hungarian verbal WordNet, automatically extracted information about argument structures was taken into account, as well. Therefore, the sense distinctions in HuWN are partially based on distributional information. Five different annotators sense-tagged 30 occurrences of 15 verbs in one-sentence context on the basis of both sense-inventories. The possible answers were the following: any of the sense-labels in the inventories, *'neither sense fits*', '*I don't know*'. Just as in the case of Véronis' first experiment the rate of *'I don't know*' answers was considerably low, implying that in the majority of cases the one-sentence contexts were specific enough. The inter-annotator agreement was determined using Fleiss's multi-π (Artstein and Poesio, 2008).

Results The average Fleiss's multi-π was 0.3 in the case of the Hungarian Explanatory Dictionary and 0.483 when HuWN was used as sense inventory. Thus, the order of the inter-annotator agreement value was comparable to Véronis' results for both databases. These results clearly show that none of these sense inventories can be exploited to find reliably the relevant meanings of headwords in contexts. That is, such databases cannot be trustworthily used for finding the best translations in contexts.

4.2.2.4 Automatic WSD of verbs in context based on PWN

In this experiment[1] the verbal part of Princeton WordNet 3.0 (PWN 3.0) (cf. 2.3.1.2) was used as a monolingual dictionary.

Princeton WordNet In WordNet each synset consists of the headword and its synonyms, the description of the sense and a unique identifier of the given synset. Example sentences are also listed. This is how a node of PWN 3.0 looks like:

introduce, present, acquaint (ID) (Definition: cause to come to know personally)
`"permit me to acquaint you with my son"; "introduce the new neighbors to the community"`

Figure 4.1: A synset of Princeton WordNet 3.0

PWN 3.0 comprises 11529 verbal lemmata with 25047 meanings and 10759 example sentences.

[1]This experiment was performed by the author during an internship at XEROX Research Centre Europe under the supervision of Caroline Brun in 2008.

Syntactic analysis In the first step the deep syntactic parsing of the example sentences was performed with Xerox Incremental Parser (Aït-Mokhtar and Chanod, 1997). Based on the deep syntactic analysis word sense disambiguation rules were generated automatically in the next step. These rules assigned the ID of the relavant synset to verbs occurring in similar sentences.

For instance, the synset of *check* was converted into the disambiguation rule presented in Figure 4.2.

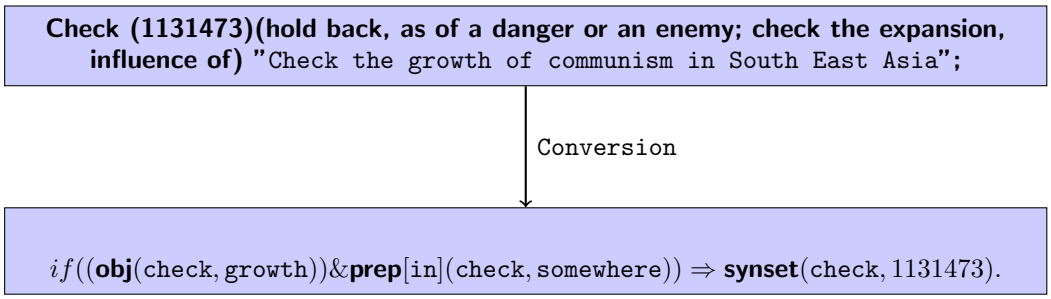


Figure 4.2: A disambiguation rule

Disambiguation rule 4.2 states that if the verb *check* has:

(i) an object whose lemma is *growth* and

(ii) an adjunct with the semantic role PLACE and

- the adjunct is attached to the verb through the preposition *in* then

the verb *check* has the meaning marked with 1131473 in PWN 3.0.

Annotating the test corpus with senses In the third step a test corpus was analyzed with the automatically generated disambiguation rules. The Semcor corpus (Miller et al., 1993) was used for that purpose. Semcor 1.6 was the first version, which was the result of tagging a part of the Brown Corpus (Francis and Kučera, 1979) with WordNet 1.6 senses. Later versions are automatic updates by mapping synsets in subsequent versions of WordNet to synset tags in Semcor[1]. Semcor is made up of 352 texts. In 186 texts all of the open class words (nouns, verbs, adjectives and

[1]Note, that in our experiment in the annotation phase PWN 3.0 was relied on, while PWN 2.1 senses were used for testing. This is due to the fact that Semcor 3.0 was not available at that time

adverbs) are annotated with part-of-speech tags[1], lemma and WordNet synset, while in the remaining 166 texts only verbs (41,497 occurrences) are annotated with lemma and synset. These 166 texts (17,308 sentences) were used for testing purposes.

In the test phase, the verbs in the test corpus were automatically tagged with PWN 3.0 senses on the basis of example sentences in PWN 3.0. It is important to emphasize that the disambiguation rules were rather specific, since beside syntactic patterns they also referred to specific lemmata. Hence, our previous expectation was to obtain only a few but correctly disambiguated verbs.

Example The disambiguation rule in Figure 4.2 assigns the synset 1131473 to *check*.

Example:	**The White House is taking extraordinary steps to check the rapid growth of juvenile delinquency in the United States.**

Figure 4.3: An example sentence of SEMCOR 2.1

Results Interestingly, the results were much worse than expected. They were also compared to a baseline method which assigns the most frequent meaning—the first meaning in PWN 3.0—to the target word. With the baseline technique 35% of all the verb occurrences were tagged with the right meaning. As opposed to this, the generated disambiguation rules assigned the right meaning only to the 30,7% (296) of all verbal occurrences. Thus, according to our results, applying the disambiguation rules result in a considerably lower precision and recall than using the baseline rules, where recall equals to 100%.

To validate our surprisingly unfavourable results the precision of the syntactic analyzer was checked, too. The generated disambiguation rules were reran on the original PWN 3.0 example sentences. A result of 98,7% was obtained in terms of precision, which confirms that precision should have been relatively high.

Discussion Manual check of 100 hits showed that low precision is mainly due to wrong PWN annotations in the manually annotated SEMCOR 2.1. For instance, *check* in the example sentence above was annotated with an other meaning in spite of the specific contextual hints. This observation is also supported by the relatively high precision of the baseline technique, which reflects that human coders tend to annotate with the first sense in a sense-inventory. This in turn raises the question of how to rank

[1]The POS tags were assigned by the Brill tagger (Brill, 1992).

meanings in a dictionary.

4.2.2.5 Discussion

The above studies have yielded two important conclusions:

(1) Native speakers tend to strongly believe that they are aware of the precise meanings of their mother tongue's lexemes. Nevertheless, their intuitions only rarely coincide in practice.

(2) An SL sense-inventory that easily lends itself to serve as a basis of a bilingual dictionary should guarantee high inter-annotator agreement while annotating senses in a corpus. Hence, contextual information should play a much greater role in characterizing meanings to provide explicit anchors during the disambiguation.

These observations are in accordance with the expectations put forward in Section 3.4.2. Certainly, these experiments are unable to prove that hand-crafted sense inventories are not suited for obtaining high inter-annotator agreement. However, the results underpin that distributional data have to be carefully explored and taken into consideration when constructing such databases. As it was mentioned earlier, since building sense inventories that exploit linguistic information as much as possible is rather expensive, this approach is typically not affordable in the case of lesser-used languages. That is why it would be particularly important to find a method that facilitates the creation of such monolingual databases or to use a different dictionary building method.

Dividing the "translation space" Based on the experiments discussed above we may raise the following expectations toward an ideal encoding dictionary:

(1) Each SL headword should be characterized in a way that *each occurrence* of that headword could be *clearly assigned to a unique meaning*. That is, there is no such occurrence that may be assigned to two different meanings.

(2) It is also presupposed that meanings are non-overlapping entities.

The necessity of requirement (2) is illustrated with some definitions of the English verb *have* in PWN 3.0. In accordance with the overall design of PWN, the various meanings of the verb are characterized by various synsets.

Synset 1: have or possess, either in a concrete or an abstract sense `"She has $1,000 in the bank"; "He has got two beautiful daughters";`

Synset 2: have ownership or possession of `"How many cars does she have?"`

Synset 3: have a personal or business relationship with someone `"have an assistant"`

Figure 4.4: Occurrences of *have* in various synsets of Princeton WordNet 3.0

As the example sentences show, both Synset 2 and Synset 3 can be subsumed under Synset 1. Since these synsets are strongly overlapping, a similar disambiguation task for *have* may yield high inter-annotator agreement only by chance.

Partition over a set Expectations (1) and (2) closely resemble to the set theoretic notion of partition described in 4.3.4 in more detail. The partition of a set A is made up of mutually exclusive (non-overlapping) subsets A_i so that the union of the subsets A_i equals to A. In our case A is made up of the occurrences of the SL word. The partition over A comprises A_is so that every A_i consists of some occurrences of the SL word. Every occurrence belongs to a subset A_i and no occurrence belongs to two subsets. These subsets may be labelled somehow through words occurring in the same contexts or through salient contexts. Trivial examples of such partitions are homonyms.

Example—homonymy Let us take a word form, such as the English *nail*. This word have (at least) two completely unrelated meanings: One sense refers to the 'pin-pointed piece of metal' and the other sense to 'a horny covering on the upper surface of the tip of the fingers and the toes'. Except for underspecified contexts[1] every occurrence of *nail* can be uniquely assigned to any of the two senses. Since the meaning of each occurrence is distinct and clear-cut, it is not difficult to find the right translation that suits the relevant contexts.

Example—related senses However, in the majority of cases the various senses of a word are related to each other in a way or another. Therefore, the question arises, how a partition may be created in the case of more intricate polysemies. Recall the result of the experiment discussed in Section 2.3.3.3. Here interchangeable adjectives were detected based on a corpus along with their typical contexts. Some examples were:

NAGY SZÉP (big nice) eredmény (result), siker (success), teljesítmény (achievement)

NAGY MÉLYSÉGES (big profound) bánat (sorrow), fájdalom (pain)

[1] Underspecified contexts are not considered here.

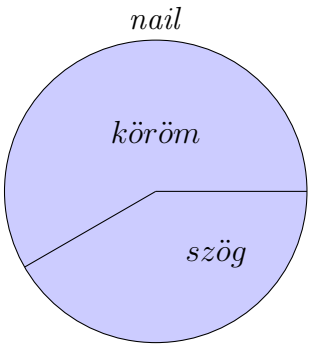


Figure 4.5: Partition over the occurrences of *nail*

Here, the results are creating a partition over the occurrences of the Hungarian word *nagy*. The subsets of occurrences are labelled both by means of possible near-synonyms[1] and by means of salient contexts. For instance, if *nagy* is occurring before *bánat* or *fájdalom* it means the same as *mélységes*. On the other hand if *nagy* is occurring before the nouns *eredmény*, *siker*, *teljesítmény* it appears with a slightly different meaning and has *szép* as a synonym. Since the contexts are non-overlapping, we are about to create a partition over the occurrences of *nagy*[2]. These results are correlating with the English translations: *profound* is a possible translation of the Hungarian word *nagy* but only in the context of *pain* and *sorrow*. Hence, such neatly characterized SL sense-inventories might serve as a basis for high-quality encoding dictionaries.

Producing high-quality encoding dictionaries Therefore, there are two alternatives to produce high-quality encoding dictionaries:

(1) In accordance with the dictionary building process described in Section 2.2.2, at first the senses and subsenses of the SL word should be characterized in a way that they create a partition over the occurrences of the SL word, for example by means of contextual anchors or near-synonyms. *In this case meaning may be conceived of as labels on partitions over SL word occurrences. Labels are composed of near-synonyms and contexts.* In the next step these senses should be translated so that they fit into the possible target sentences. Nevertheless, in the absence of such a neatly characterized database it is difficult to see how partition could be created.

(2) Another alternative is that we disregard word senses and try to retrieve translations

[1]Near-synonyms refer here to words that are interchangeable in certain contexts.

[2]Of course, at the present stage of research this partition is far from covering all occurrences of *nagy*.

by directly creating a partition of TL word forms over the SL word form. This is what word alignment on parallel corpora does.

Partition over meanings Therefore, organizing SL senses into *partitions* could be indispensable for high-quality encoding dictionaries. The experiment described in 2.3.3.3 can be a first step into that direction. Partition could be also essential in verifying the systemacity of natural languages. As Gendler Szabó (2013) puts it:

> And do all who understand 'halfway closed' and 'firmly believed' also understand 'halfway believed' and 'firmly closed'? As Johnson (2004) argues, the claim that natural languages are systematic presupposes a natural non-overlapping linguistic categorization of all the expressions. The existence of such a categorization is a bold empirical hypothesis.

The proposed method might provide answer to this question, therefore, we consider this thread of investigation a promising future research direction.

In the next section we turn to monolingual dictionaries and to how they handle the issue of ranking various senses of the headword. From our perspective, this question plays an important role, as in the traditional approaches, which follow the methodology in Section 2.2.2, the ranked senses determine the order of translations.

4.2.3 How to order meanings in a monolingual dictionary?

The senses in the SL sense-inventory should be ranked to present the senses in a consistent way within the dictionary. The criterion of ranking obviously determines the order of translations, too. As Atkins and Rundell (2008) describes SL senses might be ranked according to three different aspects.

Historical order This method presents the senses of a headword in the order in which they entered the language. However, probably this aspect is rather indifferent to the ordinary user of a bilingual dictionary, thus arranging the senses along this criterion does not seem to be a user-friendly approach.

Frequency order Frequency order reflects which meaning occurs in the language the most frequently. In spite of the apparent benefits this approach has serious shortcomings. As Atkins and Rundell (2008) puts it:

> The attraction of this method is its apparent objectivity. Further, it can plausibly be argued that the meanings which are encountered most frequently are the ones that users are most likely to look up – so it makes sense to show them first. In practice, this frequency-based approach is a good deal less straightforward than it sounds. First, it requires a well-balanced corpus. Second, determining the relative frequencies of the meanings of a polysemous word can never be an exact science because word senses are not objectively stable entities.

This observation is confirmed by the word sense disambiguation experiments described in Subsection 4.2.2. To determine which are the most frequent senses in a database, the database should comprise ample distributional information, anchoring each sense to the contexts where the given sense may appear. Such a database is a prerequisite for a properly sense-tagged corpus. In the absence of such a database it is not possible to select the most frequent meaning. Moreover, the higher a sense is ranked under a headword by the human annotator[1], the higher the probability that it will be assigned to the occurrences of the headword. Thus, without satisfactory contextual information the original structure of the sense-inventory influences the occurrence of the senses in question in the corpus.

Semantic order In this case the core meaning of a headword comes first. The core meaning is the one that is the most central meaning of a word. The judgement is based on intuition, that is why, this is the least scientific ordering according to Atkins and Rundell (2008). Interestingly, this ranking method is favoured by most dictionaries partly because it is relatively easy to apply, and partly because it is felt to give the user the most satisfying account of meaning.

How to order meanings? Because the order of meanings in the SL side determines the order of translations in the TL side of a bilingual dictionary, the ranking of senses is an important issue from the perspective of bilingual dictionaries, too. However, the problem of SL sense ordering would be eliminated completely, if we could leave out SL senses from the dictionary compilation process. Again, as indicated on page 97 dictionaries generated on the basis of word alignment fully meets this expectation.

Translation relation However, the proposed technique should not only treat the SL meanings somehow, but should be able to retrieve the best translation for each SL

[1]The relatively high precision of default rules on Semcor 2.1 in the fourth experiment is in accordance with this statement.

word. This is an additional requirement toward the method to be applied. In the next section word alignment on parallel corpora will be introduced and we will also discuss how it is able to handle the problems that have been raised so far.

4.3 Estimation of Conditional Probabilities on the Basis of Parallel Corpora

The main finding of this thesis is that *automatic estimation of conditional probabilities on the basis of parallel corpora* is a technique which may significantly contribute to bilingual lexicography. Although this technique is widely known and exploited in the machine translation community, as far as we know, no lexicographic projects have made use of it to compile bilingual dictionaries.

In the present section we consider to what extent the proposed method is able to fulfill the requirements formulated in the previous and the present chapters. Accordingly,

(i) How are parallel corpora able to ensure high inter-annotator agreement?

(ii) To what extent could conditional probability be interpreted as translation relation?

(iii) Is the proposed method able to create a partition over the SL word? (So that the translations could be ranked automatically?)

(iv) Is the best translation found?

(v) Is the proposed method data-driven?

(vi) To what extent does the proposed technique pay off from an economic point of view?

The following sections (4.3.1, 4.3.2, 4.3.4, 4.3.6 and 4.3.7) are devoted to the discussion of these issues. The algorithm itself is described in Chapter 5 in more details.

4.3.1 Conditional probability

Parallel corpus and inter-annotator agreement In the framework of the proposed technique the inter-annotator agreement may be interpreted as measuring the

agreement of translators of corpus texts. That is, automatically attained translation-candidates show how frequently a TL expression is assigned to the SL expression, thus capable of indicating the commonly used, recurrent translations. Therefore, automatically attained conditional probabilities measure the inter-annotator agreement among human translators. In what follows, we will consider conditional probability in more detail.

Basic notions of probability theory Before focusing on conditional probability let us introduce some basic notions of probability theory. Let Ω be the sample space. The sample space is the set of all possible outcomes. Thus, in the case of coin tossing $\Omega = \{haid, tail\}$. For tossing two coins, $\Omega = \{(head, head), (head, tail), (tail, head), (tail, tail)\}$. An event is defined as any subset A of the sample space Ω.

Kolmogorov axioms

(i) $0 \leq P(A) \leq 1$ for all $A \in \Omega$

(ii) $P(\emptyset) = 0$ and $P(\Omega) = 1$. Hence, the probability of the entire sample space is 1, and the probability of the null event is 0.

(iii) If $\{A_i : i \in I\}$ is a countable collection of pairwise disjunct sets, then $P(\cup_{i \in I} A_i) = \sum_{i \in I} P(A_i)$

Conditional probability Conditional probability is defined as follows: Let Ω be the sample space and A and B events, so that $A, B \subseteq \Omega$. Then,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{4.3.1}$$

In general, conditional probability measures the probability of event A given that (e.g. by evidence) event B has occurred. If we wish to measure the probability of event A knowing that event B has occurred we need to examine event A as it is restricted to event B. That is, in the case of conditional probability, Ω is restricted to B. Since both A and B are events in Ω, A restricted to B is $A \cap B$. Whenever $P(B) > 0$ with the original probability measure on the original sample space (Ω, P), B must be the sure event in the restricted space (B, P_B) and thus $P_B(B) = 1$.

To derive $P_B(A) = P(A|B)$ so that $P(B|B) = 1$, $P(A \cap B)$ should be re-scaled by dividing by $P(B)$.

This results in Formula 4.3.1 whenever $P(B) > 0$. If $P(B) = 0$, conditional probability is not defined.

Visualization An easy way to visualize conditional probabilities is as relative areas in Venn diagrams: The area of each circle represent the probability of the event in question. $P(A|B)$ represents the percentage of the area of B that is occupied by A. As indicated in Figure 4.6, if the sample space is restricted to B, $P(B) = 1$, the area of circle B amounts to 1. In this case, $A \cap B$ is the probability of A is occurring provided B.

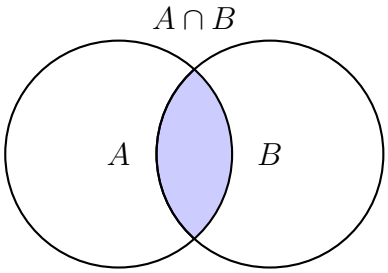


Figure 4.6: $P(A|B)$ if $P(B) = 1$

Kolmogorov axioms and conditional probability This approach results in a probability measure that is consistent with the original probability measure and satisfies all the Kolmogorov Axioms.

(i) $0 \leq P(A|B) \leq 1$

- $P(A|B) = 0$, if $A \cap B = \emptyset$;
- $P(A|B) = 1$, if $A \supseteq B$;

(ii) Any countable sequence of mutually exclusive events $A_1, A_2, ...$ and a conditioning event B satisfies:

$P(A_1 \cup A_2 \cup \cdots \cup A_n|B) = \sum_{i=1}^{n} P(A_i|B)$.

4.3.2 Conditional probability as translation relation

Recall that according to our expectations translation relation should be *gradual*, *quantifiable*, *asymmetric* and should be *automatically attainable*. From another perspective,

translation relation is expected to measure *translational insertibility* and *explanatory power*. In our point of view, conditional probability easily lends itself to measure translation relation.

Gradual, quantifiable As $0 \leq P(A|B) \leq 1$, these properties obviously hold.

Asymmetry Conditional probability is asymmetric, that is, $P(A|B) \neq P(B|A)$. $P(A|B) = \frac{P(A \cap B)}{P(B)}$, while $P(B|A) = \frac{P(A \cap B)}{P(A)}$.

Automatically attainable In Chapter 5 we will investigate to what extent conditional probability can be attained automatically for the present task.

Translational insertibility Conditional probability is estimated through actual translational insertions present in the parallel corpus (cf 4.3.3).

However, accidental gaps surely occur in every parallel corpus, since the corpus cannot contain all word form pairings, therefore, many possible translational insertions does not occur. Consequently, accidental and necessary gaps in actual translational insertions cannot be told apart in this framework. Thus, the measure of conditional probability should be complemented somehow to give a better estimation of translational insertibility. This is part of our future task.

Translational and cognitive equivalency Actual translational insertions may be also used to find out where each translation is situated on the translational equivalent—cognitive equivalent continuum. Recall the definition: Two words were said to be cognitive equivalents if they were interchangeable in every possible contexts. That is, in a perfect cognitive equivalency the frequency of the source word equals to that of the target word and $P(A|B) = \frac{P(A \cap B)}{P(B)} = 1$, i.e. the corresponding circles are completely overlapping. Figure 4.7 depicts a translation pair that is made up of "almost" cognitive equivalents.

Explanatory power Although it has not been investigated in the present research, explanatory power may be measured through giving an estimation to the semantic relatedness of the translation candidates. Explanatory power may be defined in terms of the semantically closely related translation alternatives. Semantic relatedness may be measured on the basis of the set of contexts the translation candidates share (cf. Harris distributional hypothesis). This research direction could help to give a better estimation to translational insertibility, too.

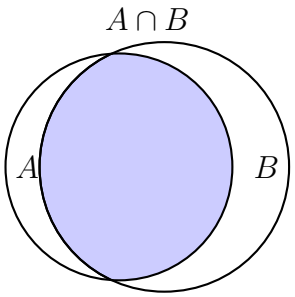


Figure 4.7: A and B are "almost" cognitive equivalents

Conditional probability as translation relation If so, how should conditional probability be interpreted to serve as a measure of translation relation? Let Ω be the sample space. That is, $\Omega = X \times Y = \{(x, y) \mid x \in X \wedge y \in Y\}$, where $X = \{SL\ word\ forms\}$ and $Y = \{TL\ word\ forms\}$. As there is no corpus that could contain all the word forms of a language, the parallel corpus yields only a more or less appropriate approximation to Ω.

Let the possible translations of b be $a_1, a_2, ..., a_k$. Therefore, $B, A_k \subseteq \Omega$, so that $B = \{(b, y) \mid b \in X \wedge y \in Y\}$ and $A_k = \{(x, a_k) \mid x \in X \wedge a_k \in Y\}$.

It is important to emphasize that as opposed to what we have asserted so far, B and A_k are mappings between pure word forms without meaning, thus in this case, we seek to map occurrences of word forms onto each other, instead of linking LUs. Note that the notation is changed in the following discussion: A and B will denote mappings between the SL and TL vocabularies, instead of denoting the SL and TL vocabularies.

Example As Figure 4.8 indicates, the Hungarian word form *ámbár* is translated as *though* and *albeit* while the English word form *though* is translated as *bár* and *ámbár*. The area of each of the circles represents the frequency of the corresponding word form. Thus, for instance *though* occurs slightly more often in the parallel corpus as *bár* and *though* occurs much more frequently than *albeit*.

Relying on conditional probabilities makes it possible to disregard the actual frequencies of the word forms: When searching for the possible translations of *ámbár*, we presume that the occurrences of this word make up the complete sample space, thus $P(ámbár) = 1$. Therefore, the possible translations of *ámbár* divide this sample space: The likelihood of the translations is depicted by the intersection of the corresponding sets (*ámbár* $\cap$ *though* is greater than *albeit* $\cap$ *ámbár*), regardless of the area of the

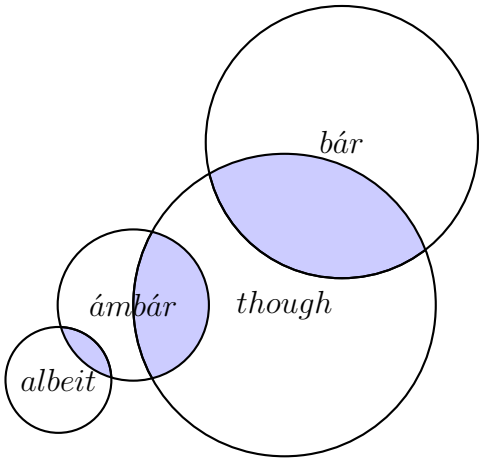


Figure 4.8: Venn diagram of the translations of *ámbár*

corresponding circles. Using conditional probability as the translation relation yields an additional advantage: *Ámbár* and *though* appear—at least partially—in the same set of sentences they are each other's translations. However as Figure 4.8 indicates this translation relation is not symmetric. *Ámbár* is more often translated as *though* as vice versa. *Though* has many others translations one of them is *bár*.

Let us suppose that we are seeking the possible translations of *though*. In that case the sample space is re-defined: It is constrained to the occurrences of *though*.

Parallel corpus, word alignment and dictionary extraction Now we are seeking the translations of an SL word b. That is, we want to determine the probability of each ordered pair (b, a_k). To put it differently, the value of $P(A_k|B)$ should be assessed. For that purpose, the actual frequencies of word alignment links in the parallel corpus have to be relied on.

Let us introduce some additional notational conventions: Let b_i denote the ith occurrence of b in the SL part of the parallel corpus, where b is a type in the SL vocabulary. Similarly, let a_{kj} stand for the jth occurrence of a_k in the TL part of the parallel corpus, where a_k is a type in the TL vocabulary.

A parallel corpus is a set of translated texts made up of aligned or parallel sentences. During the process of word alignment alignment links between the relevant SL and TL words should be found within the scope of the aligned sentences.

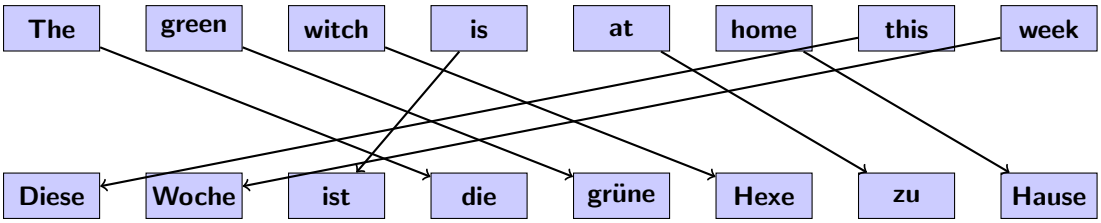


Figure 4.9: Alignment between an English sentence and its German translation. (From Jurafsky and Martin, 2008, Fig. 25.4)

4.3.3 Calculating $P(A_k|B)$ based on the parallel corpus

In the present section our primary aim is to give a basic intuition on how conditional probability may serve as translation relation and how it may be calculated on the basis of parallel data. Though unrealistic, for the sake of simplicity we will introduce some simplifying assumptions that help us to form a picture on how conditional probability may be calculated in the presence of a parallel corpus.

(1) Word alignment is present in the parallel corpus Let us suppose that the word alignment links are given beforehand, that is, the exact mapping between b_is and a_{kj}s is readily available.

(2) Equally likely outcomes Let us suppose that all the possible outcomes – $(x, y) \in \Omega$ – are equally likely. If it was true, (1) each target word is equally likely to be the translation of a given source word and (2) all the alignments were equally likely.

(3) Finite number of possible outcomes As the number of word types both in the SL and TL corpus is finite, the set of possible outcomes is also finite: $|X| \times |Y|$ From (2) and (3) it follows that the probability of a given translation pair (x, y) can be modeled by a discrete uniform distribution.

(4) Every SL word must be associated to exactly one TL word Let b_i be a specific occurrence of b and a_{kj} a specific occurrence of a_k. Every b_i is assigned a specific translation a_{kj}. So that

(i) At least one a_{kj} is assigned to every b_i.

(ii) No b_i is assigned two or more different a_ks.

In the presence of assumptions (1), (2) and (3) we can use the following equation to calculate $P(E) = P(A_k|B)$:

$$P(E) = \frac{number\ of\ desirable\ outcomes}{number\ of\ all\ possible\ outcomes} \tag{4.3.2}$$

Example Figure 4.10 represents a parallel corpus and the corresponding alignment links[1]. In the present example we calculate the probability that b (the Hungarian SL word *nagy*) is translated as a_k (into any of the English words: *large*, *profound* and *serious*) based on the given word alignment *wa*. For the sake of simplicity let us suppose that there is a one-to-one mapping between the SL and TL words. This equals to adding an extra condition:

(1) No a_{kj} is assigned two or more different b_is[2]

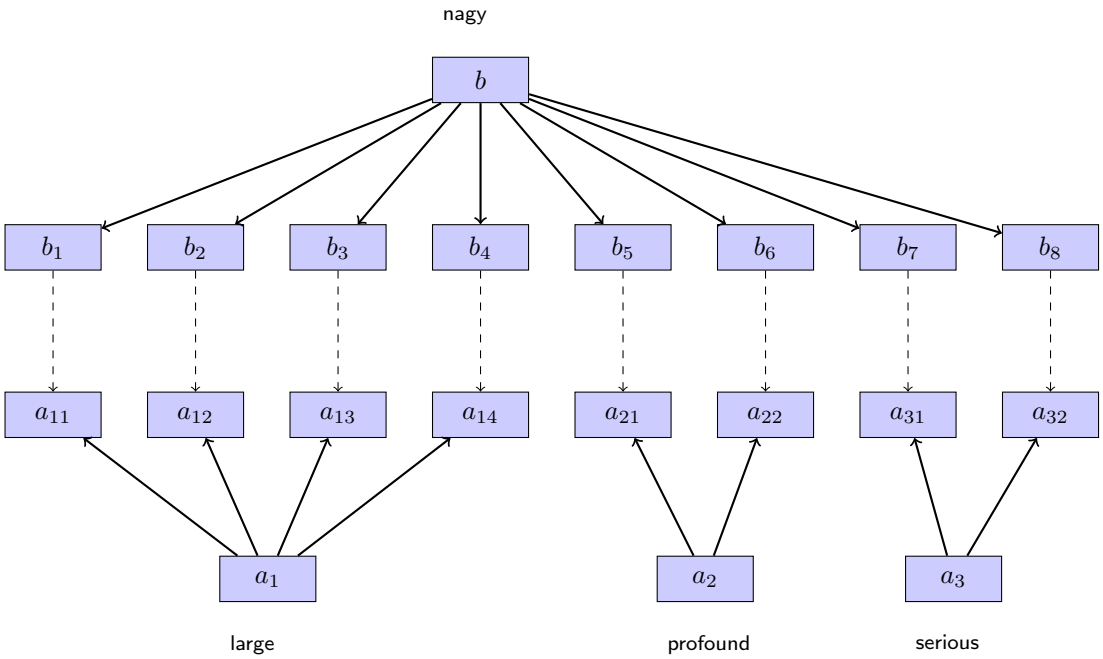


Figure 4.10: Partition over B

As we are interested only in the translations of b, the sample space is constrained to B. Therefore $P(B) = 1$.

[1]Note that the actual sentence boundaries are of no importance here.
[2]This condition is not necessary for creating a partition over B.

Recall, that $B = \{(b, y) \mid b \in X \wedge y \in Y\}$ and $A_k = \{(x, a_k) \mid x \in X \wedge a_k \in Y\}$. Thus,

$$A_1 \cap B = (b, a_1)\ and\ A_2 \cap B = (b, a_2)\ and\ A_3 \cap B = (b, a_3) \tag{4.3.3}$$

The probabilities of these events are estimated on the basis of the word alignment wa in the parallel corpus. As word alignment wa is given and the sample space is constrained to B, $(X, Y)_{wa,\,B}$ random variable takes the following values:

$$(X, Y)_{wa,\,B} = \{(b_1, a_{11}); (b_2, a_{12}); (b_3, a_{13}); (b_4, a_{14}); (b_5, a_{21}); (b_6, a_{22}); (b_7, a_{31}); (b_8, a_{32})\} \tag{4.3.4}$$

$$(X, Y)_{wa,\,A_1 \cap B} = \{(b_1, a_{11}); (b_2, a_{12}); (b_3, a_{13}); (b_4, a_{14})\} \tag{4.3.5}$$

$$(X, Y)_{wa,\,A_2 \cap B} = \{(b_5, a_{21}); (b_6, a_{22})\} \tag{4.3.6}$$

$$(X, Y)_{wa,\,A_3 \cap B} = \{(b_7, a_{31}); (b_8, a_{32})\} \tag{4.3.7}$$

As $P(B) = 1$, in accordance with 4.3.2:

$$P(A_1|B) = P(A_1 \cap B) = \frac{4}{8} = 0.5 \tag{4.3.8}$$

$$P(A_2|B) = P(A_2 \cap B) = \frac{2}{8} = 0.25 \tag{4.3.9}$$

$$P(A_3|B) = P(A_3 \cap B) = \frac{2}{8} = 0.25 \tag{4.3.10}$$

4.3.4 Partition over the SL word form

Recall that in Section 4.2.2 we have drawn the conclusion that either the SL sense-inventory should exhibit certain properties to be suitable to serve as a basis for an ideal encoding dictionary. Accordingly, each headword should be semantically characterized in a way that each occurrence of the given headword[1] can be uniquely assigned to any of the meanings belonging to the headword. Another alternative was also discussed: Namely, disregarding word senses and trying to retrieve translations by creating a

[1] If the context is not underspecified

partition of TL word forms over the SL word form directly. In this section we discuss the second option in more detail.

Partition over set B To put it more precisely, a partition of a set B is a collection of mutually exclusive subsets $A_k \subseteq B$ so that every element $b \in B$ occurs in exactly one of the A_ks. This equals to the following conditions:

(1) The intersection of any two distinct sets in the collection of sets As is empty.

(2) $\cup_{k=1}^{l} A_k = B$: The union of sets in the collection of sets A are said to cover B.

(i) That is, there is no $b \in B$ so that $b \notin A_k$, where $1 \leq k \leq l$.

(ii) And conversely, there is no $a \in A_k$, where $1 \leq k \leq l$, so that $a \notin B$.

(3) If $b \in A_k$ and $b \in A_j$, then $k = j$, that is, b is assigned to no more than one A_k.

Partition over the occurrences of source word b Just as in Example 4.10, the occurrences of the source word b are represented by $b_1, b_2, ..., b_n$. The given word alignment is denoted by wa. a_{kj} refers to the jth occurrence of a_k. Accordingly, given the specific wa alignment in Example 4.10, $O_a = \{a_{11}, a_{12}, ..., a_{1j}; a_{21}, a_{22}, ..., a_{2m}; ...; a_{k1}, a_{k2}, ..., a_{km}\}$ occurrences of the TL word forms creating a partition over the $O_b = \{b_1, b_2, ..., b_n\}$ set made up of the occurrences of the SL word b.

B and A_k should be re-defined in this case:

$B = \{(b_i, a_{kj}) \mid b \in X : SL\ vocabulary \wedge a_k \in Y : TL\ vocabulary\}$, where $0 \leq i \leq n$; $0 \leq k \leq l$ and $0 \leq j \leq m$.

$A_k = \{(b_i, a_{kj} \mid b \in X : SL\ vocabulary \wedge a_k \in Y : TL\ vocabulary\}$, where $0 \leq i \leq n$ and $0 \leq j \leq m$, thus, $A_k \subseteq B$.

(1) A_1, A_2 ... A_k are pairwise disjunct events, as $a_1, a_2, ..., a_k$ are different word forms.

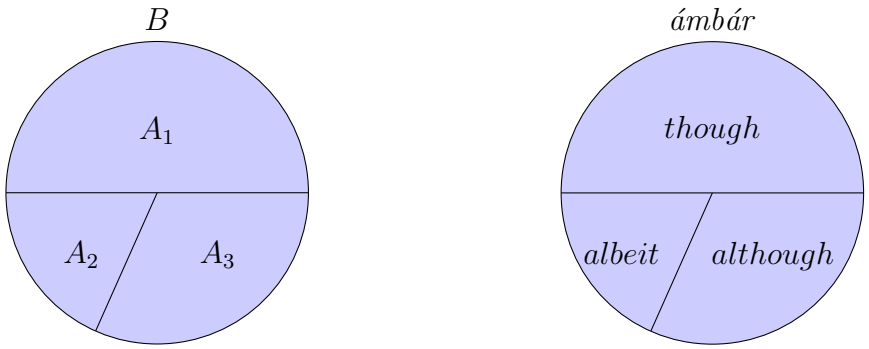
(2) $\cup_{k=1}^{l} A_k = B$: The union of sets in the collection of sets A are said to cover B.

(i) That is, there is no $(b_i, a_{kj}) \in B$ so that $(b_i, a_{kj}) \notin A_k$, where $1 \leq k \leq l$. $(A_1 \cap B) \cup (A_2 \cap B) \cup ... \cup (A_k \cap B) = B$. And conversely,

(ii) There is no $(b_i, a_{kj}) \in A_k$, where $1 \leq k \leq l$, so that $(b_i, a_{kj}) \notin B$. This follows from (4/i), namely at least one a_{kj} is assigned to every b_i.

(3) If $(b_i, a_k) \in A_k$ and $(b_i, a_j) \in A_j$, then $k = j$, that is, b_i is assigned to no more than one A_k. The specific word alignment wa guarantees that no b_is is assigned two different a_ks. In fact, the criterion of one-to-one mapping is too strong: relying only on assumption (4/ii) would suffice.

Example Figure 4.11a represents the partition created on the basis of the parallel corpus in Figure 4.10. Let us investigate the occurrences of the Hungarian word *ámbár*. As Figure 4.11b indicates *ámbár* has three different English translations: *though*, *although* and *albeit*. The partition clearly indicates the most frequently used alternative.



(a) Partition over B **(b)** Partition over *ámbár*

Translational insertions and semantic relatedness Although these translations are near-equivalents and thus might show up in a variety of the same contexts, in the texts they are never aligned to the very same SL word at the same time. Thus, this approach gives a model of translational insertibility based on actual translation insertions. However, it is not able to predict whether the TL words are interchangeable in certain contexts. Thus, the proposed technique gives certain information on translational insertibility, but more research is needed to be able to tell apart necessary and accidental gaps.

In Sections 4.3.3 and 4.3.4 we presented how conditional probability of each translation can be calculated in the presence of certain assumptions. We have also claimed that since the occurrence of different TL words as translations of a give SL word are mutually exclusive events, conditional probabilities create a partition over the occurrences of the source word form if certain assumptions hold, thus word alignment on parallel corpus may serve as an ideal basis for encoding dictionaries. In the next section the validity of these assumptions will be examined.

4.3.5 Presuppositions revisited—complicating the picture

This section focuses on the presumptions stipulated in the previous section. Unfortunately, most of them are excessively oversimplifying, thus not reflecting the basic properties of translated texts. Although many of the suppositions do not hold, the picture that has been formulated so far gives us a clear impression of what we are expecting from an ideal algorithm. In what follows, we will focus on to what extent the stipulated presuppositions hold with regard to the investigated phenomena—translated texts—and with regard to the algorithm which automatically attains conditional probabilities on the basis of parallel corpus. Unfortunately, word alignment on parallel corpora is not perfectly suitable to create a partition over a given SL word, but gives an estimation of it.

Word alignment links are more intricate The mapping properties stipulated in (4), namely that every SL word must be associated to exactly one TL word, do not necessarily hold in translated texts for multiple reasons resulting in one-to-zero, zero-to-one, many-to-one or one-to-many alignments, any of which may "distort the partition".

Translational divergencies The fact that the TL text is translation that has to be grammatically well-formed, using the lexical and phrasal stock of the TL, which is necessarily different than that of the SL may lead to more complex mappings during word alignment. Here we confine ourselves to mentioning some of these phenomena, such as a pro drop language in either side of the parallel corpus or referring by anaphors.

Morphological divergencies Various languages widely differ with regard to their morphological properties, compounds, multi- word syntactic structures expressed with a single word resulting in a many-to-one alignment.

Possible outcomes are not equally likely Since the appearance of the TL words is not independent of the words in the SL corpus, the possible outcomes are not equally likely. Therefore, translation probabilities cannot be calculated as given in Formula 4.3.2.

Word alignment links are not readily available in the parallel corpus Recall that we presupposed that the alignment links were present in the parallel corpus (cf. Figure 4.10). Unfortunately, a parallel corpus comprises only aligned sentences but

lacks word alignment. Moreover, to be able to align the words in sentences we need to know "how often" an SL word is translated as a given TL word, that is, how the translation space of each B of the SL is divided. On top of that, to calculate the frequency of each translation pair, we need the word alignment. This is exactly what Tiedemann[1] calls a chicken and egg problem, which will be addressed in Chapter 5. The corresponding algorithm is described in Section 5.3.2 in more detail.

4.3.6 Relation to corpus data

Reliance on intuition As word alignment on parallel corpus is an unsupervised technique, it seems to avoid the intuition related difficulties discussed in Sections 2.3.1, 2.3.2 and 4.2. Nevertheless, as we will see in Chapter 6, intuition comes into play again when fine-tuning the parameters and evaluating the results.

Contexts Instead of intuition context should play a great role when characterizing meanings both in the TL and in the SL. In the proposed approach contexts might be exploited in multiple ways. First, contexts are inherently taken into account when the automatically generated dictionary is produced throughout word alignment. Secondly, the contexts in which a translation pair may appear help to characterize the sub-meanings (cf. Table 4.1). Third, based on the supplied contexts automatic extraction of multiword items is also possible (for further details see Chapter 7). Moreover, relevant example sentences, in which the translation pair in question appears are also provided, thus giving additional hints on the use of the target expression. Which, in turn, is particularly useful for encoding purposes.

4.3.7 Economical considerations

Corpus-driven methodology Since the basic objective is to find a method that is able to facilitate the production of bilingual dictionaries for lesser-used languages, the available financial assistance is rather poor. Consequently, producing hand-crafted databases and thus corpus-based lexicography is not a feasible approach. Hence, a methodology is searched for that is capable of eliminating human contribution as much as possible: Corpus-driven techniques, especially unsupervised learning algorithms are particularly apt for our purposes. As word alignment works on parallel corpora tagged

[1] `http://stp.lingfil.uu.se/~joerg/mt09/f5_SMTintro-2x2.pdf`

only with a minimal linguistic mark-up, it does not require profound lexicographic work.

Language-independent The proposed method should be language independent, thus easily re-applicable. Word alignment partially meets this requirement. Obviously, the relevant parallel corpora have to be compiled. As word alignment works best on a stemmed parallel corpus—especially in the case of highly inflective languages—a minimal language dependent annotation is needed (cf. Chapter 6). But once the minimal linguistic annotation is provided the whole workflow could be considered language-independent.

Easily reversible In order to be economical the resulting dictionary should be easily reversible. This presumes that the vocabulary of the TL-side of the dictionary is representative, too. Being generated from parallel texts, the SL and TL vocabularies are largely alike. On the other hand, conditional probability is asymmetric, thus it is able to reflect the differences arising from switching the TL and the SL.

4.4 Difficulties

Beside the essential improvements the proposed method contributes to bilingual lexicography, there are certain difficulties that should be overcome to produce full-fledged proto-dictionaries of a suitable size.

Corpus size As will be described in Section 6.2.1 in more detail, the main bottleneck of the method is the scarcity of parallel texts available for medium-density languages, due to which the production of an appropriate-size parallel corpus proved to be rather tedious. Hopefully, with the increasing number of texts accessible in electronic format this task will become much more straightforward in the future.

Fortunately, this shortcoming might be at least partially compensated by a flexible selection method of the automatically generated translation candidates. The method selects subsets of translation pairs of different sizes through setting certain parameters. Therefore, a dictionary query system was designed and implemented which makes possible to determine the scope of translation candidates to be included in the proto-dictionary. The dictionary query system will be described in Chapter 8 in more detail.

Multiword expressions The proposed method is not capable of handling any kind of multiword expressions (idioms, names, collocations and verbal constructions) in itself.

Based on the provided parallel sentences manual lexicographic work is able to compensate for this shortcoming. Another possibility is to include the automatic detection of multiword expressions in the workflow. This thread of research will be elaborated in Chapter 7.

Different sub-senses of a headword As it was noted in the previous section (4.3.4), the proposed method creates a partition over the SL word form based on the alignment links of different TL word forms. Thus, the SL word form is partitioned by various TL word forms. That is, the technique is not able to differentiate either between the senses and subsenses of the SL word or between those of the TL word. The proposed technique only allows for the automatic identification of word senses of SL expressions if different translations are accessible in the TL corpus.

However, as ample empirical data is also supplied, various senses and sub-senses can be characterized manually based on the retrieved contexts. Accordingly, dictionaries relying on such information can provide positive evidence for the user that all of these sub-senses are translated with the same lemma into the target language. Different sub-senses of the Hungarian and Lithuanian counterparts of *to be born* are told apart manually in Example 4.1.

In certain situations the various meanings of words cannot be attained automatically, even in the case of completely unrelated senses (e.g. the German counterpart of the English word *nail* is *der Nagel*, regardless if it denotes the body part or the thin pointed peace of metal). Nevertheless, ignoring such cases does not pose a problem for bilingual dictionaries, since several such dictionaries follow the same practice. (e.g. Collins-Robert French Dictionary; Atkins, 1996).

4.5 Conclusions

The present chapter is made up of two main parts. In the first part we have investigated the properties that an SL-sense inventory has to exhibit to be a suitable basis for a high-quality encoding dictionary. In accordance with the view of Mel'čuk, namely, that encoding dictionaries are more intricate to produce and may be easily used for decoding purposes, our basic objective is to find a method that is able to facilitate the automatic generation of encoding dictionaries.

Hypothesis We have set out of the hypothesis that it is impossible to find the trans-

$Lemma_s$	$Lemma_t$	$p(w_t\|w_s)$	$Freq_s$	$Freq_t$
Születik	Gimti(-sta,-ė)	0.579	169	174
HU: Ő **1870-ben született**. LT: Jis **gimė 1870 metais**. EN: He **was born in 1870**.				
HU: De Fache mintha **erre született** volna. LT: Bet Fasas, regis, tiesiog **tam gimęs**. EN: As if Fas **were born to** do this.				
HU: Úgy látszik, **szerencsétlen csillagzat alatt születtél**. LT: Turbūt **gimei po nelaiminga žvaigžde**. EN: It seems that you **were born under an unlucky star**.				
HU: ..., mert **ikrei születtek**. LT: ..., nes **jai gimė dvynukai**. EN: ..., because twins **were born to her**.				
HU: Maga **úriembernek született**. LT: Tu **gimei džentlemanu**. EN: You were **born a gentleman**.				
HU: ..., hogy Buddha nem **lótuszvirágból született?** LT: ..., kad Buda **gimė** ne **iš lotoso žiedo?** EN: ..., that Budha **was born from a lotus flower**.				

Table 4.1: Sample entry of the automatically generated Hungarian-Lithuanian proto-dictionary – *to be born*

lation of an SL LU that best fits the SL contexts, if the SL LU itself cannot be uniquely assigned the right meaning in the SL context. Unique assignment means that each annotator selects the same meaning for the same target word in context. This task is called word sense disambiguation. Therefore, if unique assignment is possible, it is possible to achieve high agreement among human annotators, i.e. high inter-annotator agreement.

Experiments Four experiments were described to investigate to what extent each of the investigated sense-inventories is suitable for word-sense disambiguation tasks. The alternative meanings were listed in traditional sense-inventories. In the first, second and third experiment inter-annotator agreement was considered to measure the quality of word sense disambiguation, while in the fourth experiment the output of automatically extracted disambiguation rules were compared against manual annotation of senses in a certain test corpus. Ideally, the provided contexts should be fully specified so that

the missing information does not impede the selection of the right meaning. Although underspecified contexts cannot be completely ruled out, the '*I don't know*', '*None*' answers and the possibility of making multiple choices may cover such cases.

Results Each of the experiments has yielded rather similar results: On the basis of the provided sense-inventories the word sense disambiguation proved to be a difficult task even for human annotators. This in turn calls into question the reliability of the sense-inventories used as a basis for bilingual encoding dictionaries. How could such sense-inventories serve as basis of bilingual encoding dictionaries, if not even the meaning of the SL expression is clear enough to be assigned to a certain occurrence in context with high inter-annotator agreement?

Main characteristics of SL sense-inventories Given the results of the word sense disambiguation experiments the following conclusions can be drawn: First, a sense-inventory should consist of ample *explicit contextual information*, which can be relied on when selecting the best sense for the given context. Moreover, ideally, senses should be divided in a way that facilitates the unique assignment of word occurrences to a specific meaning. Such a division is called *partition* in set theory.

(1) Each SL headword should be characterized in a way that each occurrence of that headword could be clearly assigned to a unique meaning. That is, there is no such occurrence that may be assigned to two different meanings.

(2) It is also presupposed that meanings are non-overlapping entities.

Producing partition of meanings over word occurrences One possible approach is described in Section 2.3.3.3. The objective of the experiment described there was the unsupervised detection of synonymy classes of adjectives on the basis of a monolingual corpus. The investigated technique characterizes the senses and subsenses of the SL word in a way that they create a partition of meanings over the occurrences of the SL word, for example by means of contextual anchors or near-synonyms. *In this case meaning may be conceived of as labels on partitions over SL word occurrences. Labels are composed of near-synonyms and contexts.* To produce an encoding dictionary these senses should be translated so that they fit into the possible target sentences. Unfortunately, as this thread of research is in the first stage, a suitable database is not available at present.

Corpus-based methods and cost-efficiency Since corpus-based monolingual dictionaries are compiled on the basis of great amount of language data they might be

suitable to serve as the starting point for high-quality encoding dictionaries. However, the production of hand-crafted SL sense-inventories require so much effort that is usually not available in the case of lesser used languages. Therefore, instead of exploiting neatly characterized SL sense-inventories an other alternative should be considered.

The dictionary extraction method The second part of this chapter focuses on the automatic estimation of conditional probabilities on the basis of word alignment in parallel corpora. This unsupervised technique assigns translational probabilities to SL and TL word pairs. Translational probabilities are automatically determined on the basis of word alignment in the parallel corpus. In fact, translation probability is defined in terms of conditional probability, where the conditioning event is made up of the SL word occurrences.

Expectations toward the proposed method We have also examined whether or not word alignment on parallel corpora satisfies the requirements put forward in the previous chapters. In Chapter 2 various sense-inventories were investigated primarily with regard to their relation to corpus data, yielding the conclusion that a reversible, language independent and data-driven approach is ideal for our purposes, which is able to facilitate the creation of bilingual dictionaries in a cost-effective way. The proposed method correspond to these requirements.

In Chapter 3 the main properties of the ideal translational relation were elaborated. Accordingly, the translation relation should be:

Gradual, quantifiable and asymmetric These properties trivially hold for conditional probability.

Able to account for translational insertibility As conditional probabilities are estimated on the basis of actual translations, that is, actual translational insertions, conditional probability gives an estimation of translational insertibility.

Able to account for cognitive equivalency Cognitive equivalents are SL and TL words that are interchangeable in every possible contexts. The automatically determined conditional probability along with the corresponding frequencies serve as estimates where the given translation pair is situated on the cognitive equivalence-translational equivalence continuum.

Able to create a partition over the SL form It was also shown that the occurrences of the translation candidates create a partition over the SL word form provided that certain presupposition hold. Although some of the assumptions is

not true, it is a good intuitive way to conceive of the algorithms as if it created a partition over the occurrences of the SL word form. This entails the fact that the method ranks the translation candidates according to how likely they are, based on automatically determined translational probabilities. This, in turn, makes it possible to determine which translation of a given lemma is the most frequently used.

Automatically attainable An additional strength of the proposed method that the translation candidates are retrieved completely automatically.

In Chapter 5 the most widely used word alignment algorithms will be introduced, which can be classified into two main categories: association approaches (5.3.1) and estimation approaches (5.3.2). Section 5.3.4 describes the main advantages and drawbacks of these techniques.

5

Selecting the Alignment Techniques

5.1 Introduction

Word alignment methods enable the unsupervised learning of word pairs from sentence-aligned corpora. As stated above, one of the main advantages of using word alignment for the purpose of dictionary creation is that it helps to eliminate human intuition during dictionary building. Moreover, it exploits parallel corpora, that is, as opposed to other techniques it does not presume the existence of refined resources (e.g. monolingual explanatory dictionaries, sense-inventories characterized on the basis of monolingual corpora, wordnets).

Word alignment aims at finding alignment links between words in a parallel corpus. Bilingual lexicon extraction goes further: its goal is to identify the lexical word type links based on alignment between word tokens. Thus, dictionary extraction might be decomposed into three basic steps:

(1) The sentence alignment of the parallel corpus.

(2) The sentence alignment is extended to word alignment.

(3) Some criterion is used (e. g. frequency) to select the aligned pairs for which there is enough evidence to include them in a bilingual dictionary.

Thus, in what follows we shortly describe the most important sentence alignment methods (5.2), then the most widely used dictionary extraction techniques will be presented (5.3). There are two basic approaches to dictionary extraction on the basis of parallel corpora: Estimation approaches, which aim at estimating the conditional probabilities between SL and TL word pairs, and association approaches that in general measure that how independent the SL word and the TL word are in the parallel sentences. Note that based on the conclusions we have drawn so far we will focus on estimation approaches (5.3.2), but association approaches will be also shortly described (5.3.1)

5.2 Sentence Alignment Techniques

As Manning and Schütze (1999) puts it:

> In the sentence alignment problem one seeks to say that some group of sentences in one language corresponds in content to some group of sentences in the other language, where either group can be empty, so as to allow insertions or deletions. (p. 467)

That is, the problem of sentence alignment cannot be considered a trivial task, as the structure of the target text might be rather different from that of the source text: Sometimes even whole paragraphs are omitted or inserted yielding 1:0 or 0:1 alignments. Sentences might also appear in reversed order on the target side, these are crossing dependencies, making 2:2, 2:3, 3:2 alignments necessary. Sentences might be also merged or split, which necessitates 1:2, 2:1, 1:3, 3:1 alignments. Manning and Schütze (1999) suggests that around 90% of alignments are usually of type 1:1.

The task The objective of statistical approaches of sentence alignment is to find the alignment A with the highest probability given the two parallel texts S and T:

$$argmax_A P(A|S,T) = argmax_A P(A,S,T) \tag{5.2.1}$$

Let $B_1, ..., B_k$ a sequence of aligned beads. If we suppose that the probability of a bead—a set of sentences that is aligned with an other set of sentences—is independent

of the probabilities of other beads depending only on the sentences in the bead, then:

$$P(A, S, T) = \prod_{k=1}^{k} P(B_k) \tag{5.2.2}$$

The question now is how to estimate B_k. For the rest of Section 5.2 let us suppose that we have two parallel texts $S = (s_1, ..., s_I)$ and $T = (t_1, ..., t_J)$

Expectations Considering the fact that our primary objective is to compile dictionaries for any medium density language pairs, the suitable alignment system should meet the following criteria:

(1) It must be able to handle as many alignment types as possible, as parallel resources turned out to be rather scarce for less resourced languages.

(2) The alignment algorithm should be robust and fast.

(3) The method has to be language independent.

(4) It should be able to handle Unicode input.

Sentence aligner Manning and Schütze (1999) tell apart length-based methods (eg. Gale and Church, 1993) and lexical methods (eg. Chen, 1993; Kay and Röscheisen, 1993). We have decided to use a hybrid sentence aligner *hunalign* (Varga et al., 2005), which makes use of both length and lexical information. The main advantage of this technique that it is able to detect one-to-many alignments (eg. 1:3 or 3:1) with a high confidence value, the lack of such alignments is an obvious shortcoming of the methods described in Gale and Church (1993) and Chen (1993). From our perspective, the main problem of the selected method might be that seed dictionaries are not always available in the case of medium density languages. Nevertheless, in the lack of an initial dictionary, the hybrid algorithm falls back to surface identity of words. The seed dictionary will be generated on the basis of this initial sentence alignment. As they put it:

> To summarize our results so far, the pure sentence length-based method does as well in the absence of a dictionary as the pure matching-based method does with a large dictionary. Combining the two is ideal, but this route is not available for the many medium density languages for which bilingual dictionaries are not freely available. However, a core dictionary

> can automatically be created based on the dictionary-free alignment, and using this bootstrapped dictionary in combination with length-based alignment in the second pass is just as good as using a human-built dictionary for this purpose. In other words, the lack of a high-quality bilingual dictionary is no impediment to aligning the parallel corpus at the sentence level.

In addition, hunalign is able to deal with UTF-8 encoded input, which makes it particularly suitable for our purposes.

5.3 Dictionary Extraction Techniques

The dictionary extraction techniques can be classified into two broad categories: *Association approaches* and *estimation approaches*. In what follows, an overview of the basic properties of both kinds of extraction methods will be given.

5.3.1 Association approaches

Association approaches are also called *heuristic approaches* or *hypothesis testing approaches* in the literature. As noted by Och and Ney (2003) a common idea behind statistical association measures is to test if two words co-occur significantly more often than it would be expected if they would co-occur purely by chance. To test the independence hypothesis they count co-occurrence frequencies in the aligned regions and use some association measure to determine how independent two words are. According to Melamed (2000, p. 227.) association approaches can be decomposed into four basic steps:

(1) The selection of a similarity function S between word types of the SL and between word types of the TL.

(2) Computing the corresponding association scores for a set of word type pairs occurring in the parallel corpus.

(3) Sorting the word pairs in descending order according to the association score.

(4) Choosing a certain threshold: The word pairs with an association score above the threshold will be included in the lexicon.

The Dice coefficient or some variants of it are used most frequently as similarity function.

$$dice(i,j) = \frac{2C(e_i, f_j)}{C(e_i) + C(f_j)} \tag{5.3.1}$$

Here, $C(e, f)$ denotes the co-occurrence count of e and f in the parallel training corpus, while $C(e)$ and $C(f)$ refer to the counts of e and f in the source and target sentences. However, according to Och and Ney (2003):

> ...the use of a specific similarity function seems to be completely arbitrary. The literature contains a large variety of different scoring functions... (p. 24)

Varma (2002) gives a detailed comparison of various measures and tests of association such as Pointwise Mutual Information, Dice coefficient, Log-likelihood Ratio, Pearson's Chi-square test, Odds ratio, T-score and Fischer's Exact Test. Ribeiro et al. (2000) consider the performance of 23 similarity measures in a dictionary extraction context.

According to Melamed (2000, p. 227.) such techniques have to face and additional problem: Since association scores in step (2) are computed independently of each other, wrong translation candidates may be assigned relatively high association scores, too. This is due to the fact that there are commonly co-occurring SL and TL lemmata that are not each others' translations. Such alignments are called indirect associations and are typical in the case of collocations. The aim of the competitive linking algorithm described in Melamed (2000) is to increase the precision by getting rid of indirect associations. As we will see in Section 6.2.3, since our objective is to generate resources for human use (i.e. for lexicographers), the automatic treatment of indirect associations is not our main concern.

5.3.2 Estimation approaches

Word alignment and noisy channel model Estimation approaches to word alignment are inspired by statistical machine translation. Statistical machine translation is an application of the noisy channel model from information theory (Shannon, 1948) to the task of machine translation. In what follows, I will give a brief outline of how the

noisy channel model can be used for the purpose of word alignment based on Tiedemann (2003), Hiemstra (1996), Manning and Schütze (1999) and Jurafsky and Martin (2008). For the sake of clarity let us suppose that we are translating from French into English. That is, in lexicographical terms the source language is French and the target language is English. However, as Jurafsky and Martin (2008) asserts:

> ...applying the noisy channel model to machine translation requires that we think of things backwards [...]. We pretend that the foreign (source language) input F we must translate is a corrupted version of some English (target language) sentence E, and our task is to *discover the hidden (target language) sentence E that generated our observation sentence F.*

Hence, the usual terminology of statistical machine translation is just the opposite of what is usual in lexicography. As for subsection 5.3.2 we stick to the terminology of statistical machine translation, but will get back to the more traditional version later on.

When the noisy channel model is applied to the task of machine translation, the source language S and the target language T are considered to be random variables that produce sentences. Translation is modeled as a transmission of a source language sentence s through a noisy channel that transforms it into a t target language sentence. In a noisy channel model the target sentence t is considered to be the observable part of the system and the task of the model is to find the original input string s that has been transmitted through the channel in order to produce t target sentence. Therefore, our goal is to determine the most probable source language ŝ, given t target language sentence:

$$\hat{s} = argmax_s P(s|t) = argmax_s \frac{P(t|s)P(s)}{P(t)} \tag{5.3.2}$$

Because $P(t)$ is independent of s and, therefore, is constant for all possible source language sentences, $P(t)$ might be omitted from the equation. So, the basic equation is as follows:

$$\hat{s} = argmax_s P(t|s)P(s) \tag{5.3.3}$$

Language model In equation 5.3.4 the probability $P(s)$ is the prior probability or the language model. It expresses the probability that the translator will translate the source language sentence s. As Jurafsky and Martin (2008) points out, language model might be conceived as a measure of fluency, that is, it measures how often a given sentence shows up in a monolingual text. Both distributions $P(s)$ and $P(t|s)$ are enormously

complex. The next step is to define models for these probability distributions.

One important property of estimation approaches is how $P(s)$ is modeled. A possible solution is to presume that words are independent in the sentences and estimate $P(s)$ as the product of the probability of each word in the sentence. That is, assuming that s sentence consists of l words, s_1, s_2, ..., s_l,

$$P(s) = P(s_1) \cdot P(s_2) \cdot \ldots \cdot P(s_l) \tag{5.3.4}$$

might be a good estimation. The parameters of these models are then estimated based on a monolingual corpus. In this case the model ignores any sequence and position information, thus, this is a zero-order model.

Translation model In equation 5.3.4 the distribution $P(t|s)$-—the translation model-—can be looked upon as a source language-target language dictionary (one that has information about all possible sentences in the source and the target language), thus it gives a relatively high probability to sentences that are each others translations. As the distribution $P(t|s)$ is enormously complex in this framework translation model is learnt via *statistical alignment models* on the basis of equation 5.3.5[1].

$$P(t|s) = \sum_{A} P(t, A|s) \tag{5.3.5}$$

The word alignments that are learnt from parallel corpora can be used as the starting point to build bilingual dictionaries. Since the objective of this thesis is to investigate whether existing statistical methods can be of great help in the creation of bilingual dictionaries for human use, I do not want to dive into the details of statistical machine translation. Thus, in what follows I give only a short overview of the main properties of statistical alignment models based on Och and Ney (2003).

Statistical alignment models Statistical alignment models make some simplifying assumptions on the nature of possible alignments in order to be able to compute the best alignment between s and t. For instance Model 1, Model 3 and HMM model (Brown et al. (1993) equally presumes that each target word comes exactly from one source word. Moreover, Model 1 also presupposes a very unlikely hypothesis, namely, that each alignment is equally possible.

[1]Here A refers to a specific alignment.

The following table lists some of the important properties of the models implemented in GIZA++, the word alignment tool that was applied in our research (Och and Ney, 2003, p. 29). The importance of these properties, especially fertility model and model deficiency is given by the fact that these properties may "spoil" the partition.

Model	Alignment model	Fertility model	Deficient
Model 1	uniform	no	no
Model 2	zero-order	no	no
HMM	first-order	no	no
Model 3	zero-order	yes	yes
Model 4	first-order	yes	yes
Model 5	first-order	yes	no
Model 6	first-order	yes	yes

Table 5.1: Overview of the alignment models

Fertility The fertility-based alignment models contain a probability $p(\phi|e)$ that the **target/generating** word e is aligned to ϕ words. As Och and Ney (2003) put it:

> By including this probability, it is possible to explicitly describe the fact that for instance the german word "*übermorgen*" produces four English word (*the day after tomorrow*). In particular, the fertility $\phi = 0$ is used for prepositions or articles that have no direct counterpart in the other language. (p. 26.)

Deficiency In the case of deficient models the probabilities of all valid alignments do not sum to unity.

5.3.3 Estimating the model parameters (The EM algorithm)

Estimating the model parameters—The expectation maximization (EM) algorithm) We will now briefly describe the EM-algorithm following Manning and Schütze (1999). Let us recall what we called a 'chicken and egg problem' in the previous section: Translation probabilities were estimated on the basis of word alignment and word alignment was calculated on the basis of translation probabilities. The EM-algorithm eliminates this problem by starting with a random initialization of translation

probabilities $P(w_f|w_e)$, These random translation probabilities are then iteratively updated by maximizing the likelihood function until the process converges at a maximum.

$$Z_{w_f,w_e,} = \sum_{(e,f)s.t.w_e \in e, w_f \in f} P(w_f|w_e) \tag{5.3.6}$$

In the E-step we compute the expected number of times we will find w_f in the French sentence, given we have w_e in the English sentence. The M-step re-estimates the translation probabilities from these expectations:

$$P(w_f|w_e) = \frac{z_{w_f,w_e}}{\sum_\nu z_{w_f,\nu}} \tag{5.3.7}$$

where the summation ranges over all pairs of aligned sentences such that the English sentence contains w_e and the French sentence contains w_f. The correctness of the algorithm is proved in Dempster et al. (1977). The EM algorithm was first introduced to analyze parallel corpora by Brown et al. (1990).

Most estimation approaches are the extensions of the IBM models described in Brown et al. (1993) such as the algorithm described in Hiemstra (1996) or Model 6 described in Och and Ney (2003). Och and Ney (2003) gives also a detailed comparison of the different IBM models and concludes that alignment models with a first-order dependence and a fertility model yields significantly better results than simple heuristic models and estimation approaches with zero-order dependence. There is also a free software GIZA++ (designed and written by Och) for training purposes.

5.3.4 Pros and cons

According to Och and Ney (2003) the main advantage of heuristic models is their simplicity. They are easy to implement and understand, but the specific similarity function seems to be completely arbitrary. Moreover, their results support that certain kinds of statistical models yield significantly better results than simple heuristic methods. This is probably due to the fact that statistical alignment models are more coherent: The general principle for coming up with an association score between words results from a statistical estimation theory and the parameters of the models are adjusted such that the likelihood of the models on the training corpus is maximized. In the previous chapters it was also argued that conceiving translation relation as conditional probability is

a reasonable choice for it meets the expectations put forward by bilingual lexicography toward the ideal translation relation. On top of that, such a conception of translation relation renders it into a quantifiable notion, thus enabling the automatic selection of the best translation candidate.

Tiedemann (2003) notes that different alignment strategies might be chosen to suit particular needs. According to him estimation methods should be preferred when coverage also plays a great role, such as machine translation. Since our objective is to find a method that is suitable for the automation of the generation of dictionaries from scratch, coverage plays a great role. Mainly that is why we have decided to work with GIZA++ .

In Chapter 6 the construction and evaluation of the Hungarian-Slovenian and Hungarian-Lithuanian core dictionaries will be presented.

6

Proof-of-Concept Experiments: One-Token Units

6.1 Introduction

In Chapter 6 the compilation process of proto-dictionaries for two lesser used language pairs will be introduced: Hungarian-Lithuanian and Hungarian-Slovenian. While the former language pair was the main focus of our research, the latter one was used to set certain parameters to filter the results (cf. 6.2.2). This parameter setting was then applied in the case of the Hungarian-Lithuanian proto-dictionary.

Since the creation of reversed dictionaries is rather straightforward using word alignment, four proto-dictionaries were available by the end of the workflow. Additional dictionaries for well-resourced languages were also created so as to be able to compare the amount of effort needed to compile proto-dictionaries for lesser used and well-resourced language pairs. French-Dutch and English-Hungarian dictionaries were also produced from readily available parallel corpora. Section 8.4.1 presents the evaluation results of the French-Dutch proto-dictionary, this language pair was used in the experiments described in Chapter 7, as well. The English-Hungarian (and vv.) proto-dictionaries were generated for testing purposes.

Accordingly, this chapter discusses two proof-of-concept experiments, yielding the conclusion that although word alignment on parallel corpora seems to be a rather promising approach to facilitate the work of bilingual lexicographers from several points of view,

the collection and normalization of parallel texts is rather tedious. Once the texts are collected and normalized and the tool-chain is set up, the creation of the dictionaries can be completely automatized.

6.2 Workflow

The workflow comprised three main stages. First, resources and language-specific tools were collected to create the parallel corpora (6.2.1).

Secondly, the proto-dictionaries were generated. The generation phase comprised the following steps: (i) word-alignment was performed both for the Hungarian-Slovenian and for the Hungarian-Lithuanian language pairs, (ii) the aligned sentences were provided where the translation pair candidates show up, (iii) a bilingual speaker determined the basic parameters of the Hungarian-Slovenian proto-dictionary through a preliminary evaluation of it (cf. 6.2.2).

Thirdly, in the evaluation phase the parameters determined on the basis of the Hungarian-Slovenian proto-dictionary were re-applied to the Hungarian-Lithuanian proto-dictionary. Next, two bilingual speakers performed a more detailed evaluation of the Hungarian-Lithuanian proto-dictionary. The exact criteria for evaluation were also defined during this evaluation phase (cf. 6.2.3).

Figure 6.1 depicts the detailed workflow. Nevertheless, only the phases listed above will be discussed in more detail.

6.2.1 Creation of parallel corpora

Collection of texts Since the objective of the project was to create dictionaries for everyday language vocabulary, we decided to focus on the genre fiction and news while collecting texts for our corpora. One of the main difficulties the project had to face was the scarce availability of general-domain parallel texts. As collecting direct translations both between Hungarian and Slovenian and between Hungarian and Lithuanian yielded only a moderate success, texts translated from a third language, mainly English, French and German made also part of our parallel texts. Although national digital archives

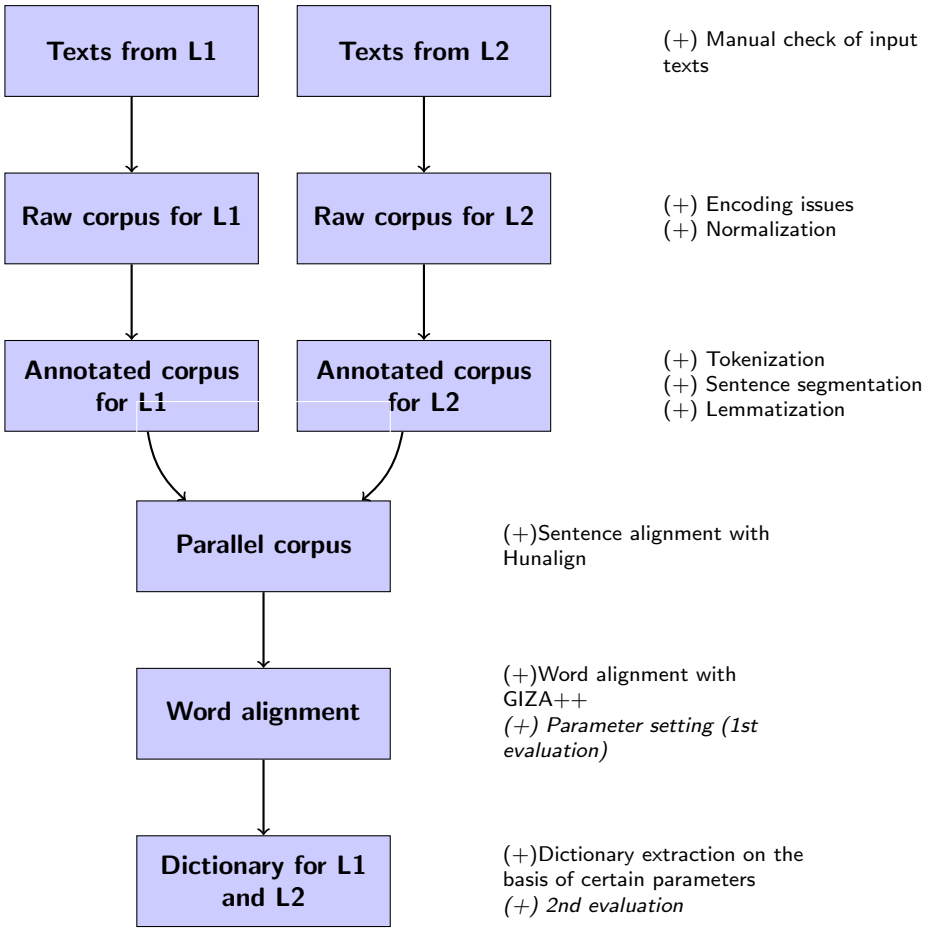


Figure 6.1: The basic process of proto-dictionary generation

such as the Digital Academy of Literature[1] and the Hungarian Electronic Library[2] do exist in Hungary providing us with a wealth of electronically available texts, similar resources have not been found, either for Slovenian or for Lithuanian.

Hungarian-Slovenian Several translators, authors and publishers were contacted to gather Slovenian counterparts of the available Hungarian texts[3]. Bilingual web pages were also used. The European Comission's news website[4] proved to be a particularly

[1] `http://www.pim.hu`
[2] `http://mek.oszk.hu`
[3] This work was accomplished by Bence Sárossy.
[4] `http://ec.europa.eu/news/`

valuable resource[1].

Hungarian-Lithuanian As for Hungarian and Lithuanian, the Lithuanian Centre of Computational Linguistics, Vytautas Magnus University[2] provided us with sentence segmented and morphologically disambiguated texts. We selected the Lithuanian texts from the Lithuanian National Corpus (Rimkutė et al., 2007) and from the Lithuanian-English parallel corpus (Rimkutė et al., 2008) for which Hungarian counterparts were available. The annotated texts were manually checked to detect missing parts and insertions[3]. Just as in the case of Slovenian, we obtained a great amount of parallel texts from the European Comission's news website.

Text processing tools Basic text-processing tasks (i.e. tokenization, sentence segmentation and lemmatization–with disambiguation) were accomplished by the means of language-specific tools accessible for all these three languages. As for Lithuanian, the majority of texts have already been annotated. The European Comission's news texts formed the only exception, which were analyzed with the same tool as all the other Lithuanian texts at the Lithuanian Centre of Computational Linguistics (Vytautas Magnus University)[4].

All Slovenian texts were processed with the online tool-chain available at the website of the Jožef Stefan Institute[5] (Erjavec et al., 2005).

The Hungarian annotation was provided by the part-of-speech tagger of the Research Institute for Linguistics, HAS (Oravecz and Dienes, 2002).

Creation of parallel corpora We used hunalign (Varga et al., 2005) to align sentences for both language pairs. The texts were converted into a lemmatized format—i.e. instead of inflected wordforms each corpus comprised only the lemmata derived from the morhologically disambiguated texts. These lemmatized versions served then as the input texts for sentence alignment so that we could eliminate as much as possible the problem of data sparseness resulting from rich morphology.

As a result of sentence alignment, we have produced two parallel corpora of different sizes. Table 6.1 shows the corpus size for each of the language pairs. The second column uses translational units (TUs) as a measure of corpus size instead of sentences.

[1]The texts from this resource were gathered by the author.
[2]The texts were provided by Rūta Marcinkevičienė and by Andrius Utka.
[3]This work was done by Iván Mittelholcz.
[4]The analysis was performed by Andrius Utka.
[5]http://nl.ijs.si/jos/analyse/

This is due to the fact that translations in parallel texts might merge or split up source language sentences, thus recognizing only one-to-one sentence mappings often entails loss of corpus data. Hunalign is able to overcome this difficulty by creating one-to-many or many-to-one alignments (i.e. 1:2, 1:3, 2:1, 3:1) between sentences.

Size of Hungarian-Lithuanian parallel corpus		
Hungarian	2,121,000 tokens	147,158 TUs
Lithuanian	1,765,000 tokens	147,158 TUs
Size of Hungarian-Slovenian parallel corpus		
Hungarian	666,000 tokens	38,574 TUs
Slovenian	733,000 tokens	38,574 TUs

Table 6.1: Size of the parallel corpora

6.2.2 Creation of proto-dictionaries

The present section describes how the list of translation candidates was generated, and how the most likely translation candidates were selected to produce the proto-dictionaries.

List of translation pair candidates The creation of proto-dictionaries follows two main steps. The first step is word alignment for which the freely available tool GIZA++ (Och and Ney, 2003) was used. To perform word alignment GIZA++ assigns translational probabilities to SL and TL lemma pairs. As it was described in Chapter 5 in more detail, the translational probability is an estimation of the conditional probability of the target word given the source word, $P(W_{target}|W_{source})$ by means of the EM algorithm.

Parallel contexts and lemma frequencies The retrieved lemma pairs with their translational probabilities and the provided contexts served as the starting point for the proto-dictionaries. Note that the provided aligned contexts were made up of the original sentences—and not of the lemmatized versions. These contexts played an important role in the evaluation, as the evaluation was based on the presumption that *if there is at least one TU where the translation is correct, than the translation is right.* And conversely, *a translation is wrong if it never occurs as a right translation in the corresponding TUs.*

Furthermore, both the SL lemma and TL lemma frequencies were included into the proto-dictionaries.

Parameters based on the Hungarian-Slovenian proto-dictionary After the generation of the proto-dictionaries we investigated how to filter the results to get rid of the wrong translations while keeping the right ones. For doing so, the Hungarian-Slovenian proto-dictionary was evaluated. A bilingual speaker[1] distinguished between the right and the wrong translations relying on the provided contexts.

Out of approximately 80,000 translation pair candidates the pair candidates with a translation probability between 0.5 and 1 were selected, yielding 13,790 translation candidate pairs. A sample of these 13,790 translation candidate pairs, 5749 lemma pairs were classified into two categories. Table 6.2 shows the number of evaluated translation candidates according to their translation probability: Every translation pair in the translation probability ranges [5; 0.7) and [0.7; 1) was evaluated.

Translational probability	**Number of lemma pairs**	**Evaluated lemma pairs**
1	10233	2192
[0.7; 1)	2110	2110
[0.5; 0.7)	1447	1447

Table 6.2: The number of evaluated lemma pairs in each range of $P(tr)$

It was investigated how the lemma frequency influences the proportion of the right translations in each of the translation probability ranges. More precisely, the translation candidate pairs were told apart into two categories: A pair candidate was assigned to the first category if both the SL *and* the TL lemma frequencies were greater than 5, and it was assigned to the second category if *both* the SL *and* TL lemma frequencies were below 5. These two categories were evaluated in all three translation probability ranges ($p(tr) = 1$; $0.7 \leq p(tr) < 1$ and $0.5 \leq p(tr) < 0.7$ — Table 6.3, Table 6.4 and Table 6.5, respectively).

Table 6.3 indicates the results if $P(tr) = 1$. In this case only 1,564 translation pair candidates were taken into consideration since all the other translation candidate pairs were made up of either punctuation marks or words from a third language, different from both the SL and the TL.

[1]This work was carried out by Bence Sárossy.

	Number of lemmata	Number of right translations	Proportion of right translations
SL L-frequencies < 5 TL L-frequencies < 5	1086	204	18 %
SL L-frequencies > 5 TL L-frequencies > 5	15	10	66 %

Table 6.3: Proportion of the right translation pairs depending on the SL and TL lemma frequencies if $p(tr) = 1$

Similarly, Table 6.4 presents how the SL and TL lemma frequencies affect the proportion of right translations if $0.7 \leq p(tr) < 1$

	Number of lemmata	Number of right translations	Proportion of right translations
SL L-frequencies < 5 TL L-frequencies < 5	336	84	25 %
SL L-frequencies > 5 TL L-frequencies > 5	662	504	76 %

Table 6.4: Proportion of the right translation pairs depending on the SL and TL lemma frequencies if $0.7 \leq p(tr) < 1$

As Table 6.4 shows, the lemma frequencies have a great effect on the proportion of the right translations. This observation is confirmed by the next evaluation domain, where $0.5 \leq p(tr) < 0.7$, as well.

	Number of lemmata	Number of right translations	Proportion of right translations
SL L-frequencies < 5 TL L-frequencies < 5	508	74	14 %
SL L-frequencies > 5 TL L-frequencies > 5	429	245	57 %

Table 6.5: Proportion of the right translation pairs depending on the SL and TL lemma frequencies, if $0.5 \leq p(tr) < 0.7$

Additional parameters Consequently, beside translation probabilities we took the lemma frequencies into consideration, as well. Thus, we relied in the following three parameters when filtering the results:

(1) Translational probability

(2) Source language lemma frequency

(3) Target language lemma frequency

Lemma frequencies had to be taken into account for at least two reasons:

(1) A minimal amount of data was necessary for the word alignment algorithm to be able to estimate the translational probability.

(2) In the case of rarely used TL lemmas the alignment algorithm might assign high translational probabilities to incorrect lemma pairs if the source lemma occurs frequently in the corpus and both members of the lemma pair recurrently show up in aligned units. This phenomenon is illustrated with two examples in Table 6.6 below:

Hungarian Lemma (SL)	**L-Frequency**	**Lithuanian Lemma**	**L-Frequency**	$\mathbf{P(w_t|w_s)}$
arcizom ('muscle in the cheeks')	5	jis (he, him, it)	60667	0.8523
ádáz ('grim')	23	su (with)	8562	0.7971

Table 6.6: Incorrect candidates with high translational probabilities

To filter out such cases an additional constraint was introduced for the Hungarian-Lithuanian language pair: translation candidates where one of the members occurs at least 100 times more than the other were ignored.

Parameter setting The evaluation of a sample Hungarian-Slovenian proto-dictionary (5749 lemma pairs) has yielded the following findings:

(1) Source language and target language members of lemma pairs should occur at least 5 times in order to have reliable amount of data when estimating probabilities.

(2) If the translational probability is at least 0.5, slightly more than 65% of the translation candidates with the corresponding parameters were right translations based on the evaluation results.

(3) As is described above, in the case of Hungarian-Lithuanian a further constraint was added: We also excluded translation candidates where either the Lithuanian or the Hungarian lemma occurred more than 100 times than the other in the whole parallel corpus.

Number of translation candidates Table 6.7 indicates the number of translation candidates that correspond to the parameters determined through the preliminary evaluation. The second column of the table shows the number of expected correct translations, assuming that 65% of the translation candidates with the corresponding parameters are correct.

	Number of Translation Candidates above the Threshold	Expected Number of Right Translations
Hungarian-Slovenian	4969	3230
Hungarian-Lithuanian	4025	2616

Table 6.7: Expected number of right translations

Considering the fact that we do not intend to create perfect dictionaries, but proto-dictionaries facilitating lexicographers' work, it seems reasonable to target this value (65%), since it is much easier to throw out wrong translations than make up new ones. Based on these parameters a detailed manual evaluation of the core Hungarian-Lithuanian dictionary was performed.

Increasing the coverage of the proto-dictionaries Unfortunately, the obtained numbers of expected translation candidates stay far below the targeted size of a medium-sized dictionary ($15,000 - 25,000$ entries). At present we see three possibilities to increase coverage:

(1) The first possibility is to augment the size of the parallel corpora. Hopefully, the amount of electronically available texts is continuously increasing, at least for medium density languages.

(2) Another alternative is to refine the parameters used for filtering (i.e. SL and TL lemma frequency and translation probability). As it will be described in 8.4.1 in more detail, in the case of higher SL lemma frequencies even lower translation probabilities yield also a high proportion of right translations. The refined evaluation of the French-Dutch and Hungarian-Lithuanian proto-dictionaries (p 201) confirmed this hypothesis, thus this is a viable approach to increase the coverage of proto-dictionaries.

(3) Finally, monolingual corpora could be also exploited to increase the coverage of proto-dictionaries. One interesting research question is whether the proto-dictionaries could be complemented with the automatically detected near-synonymy classes (described in 2.3.3.3) on the basis of the provided contexts.

To summarize, the augmentation of parallel corpora and the refinement of parameters will be definitely part of our future work.

6.2.3 Evaluation of the Hungarian-Lithuanian proto-dictionary

Right and wrong translations Recall that throughout the evaluation of the Hungarian-Slovenian proto-dictionary right and wrong translations were distinguished by a bilingual speaker, who had to decide on the basis of the provided contexts. *We considered a translation right, if there was at least one parallel sentence where the translation was correct. If there was no such context at all, the translation was considered wrong.* In the case of the Hungarian-Lithuanian language pair a more detailed evaluation was carried out.

Acceptable and unacceptable translation units Throughout the evaluation of the Hungarian-Lithuanian proto-dictionary we distinguished between *lexicographically acceptable* and *lexicographically unacceptable* translation units instead of right and wrong translations.

Lexicographically acceptable translation units We say that *a translation unit is lexicographically acceptable if (a) at least once the TL member of the translation unit occurred as the translation of the SL member of the translation unit in the parallel texts (cf. right translations) or (b) on the basis of the provided contexts any or both member of the translation unit can be extended into expression(s) that form right translations in at least one context pair.*

Although post-editing is needed in the latter case, these translation pairs are lexicographically acceptable, since the linked members of the translation unit can be extended into a right translation pair in at least one context. This means that the post-edited translation unit was used as translation in at least one context, thus it may be included in encoding dictionaries.

Lexicographically unacceptable translation units We say that *a translation unit is lexicographically unacceptable if (a) there is no context where the TL member of the translation unit occurred as the translation of the SL member in the parallel texts (cf. wrong translations) or (b) the translation pair is out of scope of our dictionaries.* Such translation pairs are not acceptable, since our objective was to compile general purpose dictionaries.

The eligibility of this classification is clearly verified by the fact that there are completely right but too specific translation pairs that are absolutely of no use for our purposes (e.g. *tünde*, *elfas*, 'elf'). Another example for the eligibility of the classification can be a partial match between an SL compound and a TL lemma, which forms part of the corresponding MWE (e.g. *kétszáz – šimtai*, where *šimtai* is part of the multi word expression *du šimtai*, 'two hundred'). Although post-editing is needed in this case, to consider the translation unit lexicographically acceptable is in accordance with our original purpose, e.g. providing lexicographers with material that facilitate their work, since in this case the generated resources are manually checked by bilingual experts.

Two bilingual speakers[1] evaluated the proto-dictionaries. The annotators followed an annotation guide presented in Figure 6.2. The guide describes the following evaluation categories:

Categories The evaluation was based on the two main categories: *Lexicographically acceptable* and *lexicographically unacceptable* translation units. Acceptable translation units were made up of perfect translations (category 1) and of those categories where post-editing was needed, and the provided contexts furnished enough information for post-editing (categories 2, 3a, 4). Semantic relatedness is an additional case, which was considered to be lexicographically acceptable, as well. The reason for this is that if semantically related words show up as translations in parallel texts, then they can be used as translations in certain contexts (category 6). This category needs post-editing, as well. On the other hand, lexicographically unacceptable translation units were those candidates where the TL word was never the translation of the SL word, and could not be extended into a corresponding MWE (categories 5, 3b). Finally, the translation pairs that were out of the scope of the proto-dictionaries were also excluded (category 7).

[1]This task was accomplished by Beatrix Tölgyesi and Justina Lukaseviciute.

(1) The translation unit is a perfect translation (e.g. gyümölcs–vaisius 'fruit').

(2) The morphological analysis (i.e. the lemma) or the POS-tag is wrongly assigned to any member of the translation unit or the POS-categories of the members are different. Otherwise the translation is good and comprehensible on the basis of the assigned example sentences

(e.g. emberi (ADJ)–žmogus (N) + GEN. That is, the Hungarian adjective 'human' is translated into Lithuanian as a noun ('man') with the genitive case marker).

(3) Any member of the translation unit is a compound with a multi word expression equivalent. Only one word of the multi word expression was retrieved. (e.g. adatbázis–duomenų bazė, 'data base') Two cases were distinguished:

- **(a) If any of the supplied example sentence comprises the relevant multi word expression, based on which the relevant translation equivalents can be detected manually.**
- **(b) The relevant translation equivalents cannot be detected manually in the example sentences[a].**

(4) Each member of the translation equivalence is a collocation. Though the retrieved words are not the corresponding pairs, the corresponding collocations can be manually obtained on the basis of the provided example sentences.

(e.g. in the case of the collocate pair bíborosi testület–kardinolų kolegija 'cardinal college', the Hungarian lemma bíborosi 'cardinal' was linked to the Lithuanian lemma kolegija 'college').

(5) Completely wrong translation candidates due to (a) mismatched sentences or (b) to loose translations

(e.g. festetlen 'unpainted' is translated as megzti(-zga,-zgė) 'knit' on the basis of a wrongly aligned sentence pair).

(6) The translation is not perfect but there is still some kind of a semantic relation between the source language word and the target word, for instance hyponymy or hyperonymy.

(e.g. lúdtoll (literally: goose-feather) plunksna (literally: 'feather', 'pen'): intended meaning in both cases: quill pen).

(7) The vocabulary is not relevant for the purpose of the particular dictionary or any dictionary in general

(e.g. unimportant proper names: Abdul-Abdulas).

[a]Note that although we did not encounter such cases throughout the evaluation, translation units of this type cannot be completely excluded.

Figure 6.2: The categories used for the evaluation of the Hungarian-Lithuanian and the French-Dutch proto-dictionaries

Evaluation methodology and the results Out of the 4025 translation units with the parameters determined above 863 pairs were manually evaluated. Throughout the evaluation three intervals were distinguished based on the value of the translation units' translational probability. The translational probability of 520 candidates was within the range $[0.5; 0.7)$ and 280 candidates' translational probability lay within $[0.7; 1)$. The proportion of the number of translation candidates within these intervals reflects their actual proportion in our proto-dictionary. All the translation candidates with translational probability 1 (63 pairs) were included in the evaluation. Table 6.8 indicates the result of the evaluation.

	Lexicographically Acceptable Translation Units		**Lexicographically Unacceptable Translation Units**	
$P(tr)$	OK	Post-editing	Irrelevant	Wrong
$[0.5; 0.7)$	52.1%	32.9%	2.3%	12.7%
Sum	**$\sum$ 85%**		**$\sum$15%**	
$[0.7; 1)$	65.3%	31.9%	0.6%	2.2%
Sum	**$\sum$97.2%**		**$\sum$2.8%**	
1	38%	13%	49%	0%
Sum	**$\sum$51%**		**$\sum$49%**	

Table 6.8: Results of the Hungarian-Lithuanian proto-dictionary

If we consider the sum of right and lexicographically acceptable translation units, we can state that 85% of the translation pairs is acceptable in the probability range between 0.5 and 0.7. This value goes up to 97,2% in the range between 0.7 and 1. Interestingly, only 51% of translation pairs with the highest probability (1) are acceptable, and only 38% of them are right. This is due to the high proportion of not relevant proper names in this probability range. Based on this evaluation of the sample, we might expect that 3549 translation units out of 4025 should be lexicographically acceptable, which yields a better coverage than our original hypothesis (Table 6.7).

Coverage Despite the improved results, the coverage of our proto-dictionary has to be further augmented as it was discussed on page 137. This issue will be given a closer look in Chapter 8.

6.3 Treatment of Multiple Meanings

As it was pointed out earlier in Chapter 4, one of the main benefits of the proposed method is that it enables the extraction of all the relevant translations available in the corpora, thus diminishing the role of human intuition during lexicographic process. Furthermore, it ranks the extracted translation candidates on the basis of their translational probabilities. These features imply that the proposed technique copes with related meanings more efficiently than traditional lexicography or lexicography based on monolingual corpora.

In this chapter we present two examples to illustrate the above statements. Atkins and Rundell (2008) claim that

> there is a strong correlation between a word's ferquency and its [semantic] complexity (p. 61)

Taking this citation as our starting point, we concentrated on cases where Lithuanian lemmas occur at least 100 times in the corpus. In parallel with the augmentation of frequency, we decreased the threshold of translational probability: we set it to 0.02 instead of 0.5. With these parameters we obtained 6500 translation candidates for 1759 Lithuanian lemmas.

6.3.1 Example 1: *Puikus*

Table 6.9 illustrates that the proposed method is able to extract various translations ranked according their likelihood. The translation candidates below support our hypothesis: in the case of more frequent words, translation candidates even with lower probabilities might yield correct results.

The order of the translation candidates might be stunning at first sight for someone who speaks Hungarian, for *remek* which turned out to be the second most probable translation of the Lithuanian *puikus*, is stylistically marked when it modifies a noun. However, the provided examples account for this oddity. In one third of the examples *remek* occurs as a one-word response, which form is quite extensively used in Hungarian. (e.g. *–Puiku, - atsakė balsas. –Remek – válaszolta a hang.* '–All right – the voice answered.')

$Lemma_s$	$Lemma_t$	$p(w_t\|w_s)$	*English translation*
puikus	jó	0.128	'good'
puikus	remek	0.071	'great', 'all right'
puikus	tökéletes	0.052	'perfect'
puikus	szép	0.048	'nice'
puikus	pompás	0.035	'splendid'
puikus	jól	0.035	'well'
puikus	nagyszerű	0.035	'great'
puikus	finom	0.028	'fine'
puikus	gyönyörű	0.02	'marvelous'

Table 6.9: Example 1: Hungarian equivalents of the Lithuanian word *puikus* sorted by the translational probability.

6.3.2 Example 2: *Aiškiai*

As it was discussed earlier, the proposed technique seems to be particularly apt to support the creation of encoding dictionaries. If multiple translations are present, it is essential that the choice among them be guided by explicit linguistic criteria. The provided parallel data could be of great help for lexicographers in describing the relevant conditions under which a target language expression could occur. Table 6.10 illustrates the role of the context in finding the right translational equivalent:

	$Expression_s$	$Expression_t$	*English gloss*	*Literal English*
	aiškiai	tisztán	pure+ly	'clearly'
Contexts	PERCEPTION VERBS: *lát, látszik, hall*			'see', 'seem', 'hear'
	aiškiai	világosan	clear+ly	'clearly'
Contexts	PERCEPTION VERBS: *lát, látszik, hall* COGNITION VERBS: *megért, gondolkodik* COMMUNICATION VERBS: *beszél, válaszol*			'see', 'seem', 'hear' 'understand' 'think' 'speak' 'answer'
	aiškiai	láthatóan	visible+ly	'visibly'
Contexts	COMMUNICATION VERBS: *beszél, válaszol*			'speak' 'answer'
	aiškiai	jól		well
Contexts	PERCEPTION VERBS: *lát, látszik, hall*			'see', 'seem', 'hear'

Table 6.10: Example 2: Characterization of the Lithuanian adverb *aiškiai* on the basis of the provided contexts

Although due to its size our corpus is not well suited for providing sufficient data for the complete description of these terms, on the basis of the contexts several conclusions can be drawn. First, *tisztán*, *világosan* and *jól* can modify verbs of perception. *Láthatóan* is clearly distinguishable, as it usually refers to the fact that the emotional change a person underwent was overt. *Világosan* is also commonly used with verbs of cognition and verbs of communication with the same meaning, i.e. the content of the communication is clearly comprehensible. As opposed to this, with verbs of communication *tisztán* would mean that the speech conveying the message was clearly pronounced. This kind of information can be of great help for a Lithuanian speaker who wants to make utterances in Hungarian.

6.4 A Uniform Corpus Representation

Motivation As it was described earlier in the present chapter, during the first experiment the parallel corpora was stored in a rather simple format: *Sentence-alignment* was carried out on the *lemmatized versions* of the texts. These versions were relatively easy to create and to deal with for the purpose of dictionary extraction of one-token units. However, it soon turned out that a more structured input format would be more desirable both for technical reasons (1, 2) and to enhance the quality of the resulting proto-dictionaries (3, 4):

(1) Regular expressions are quite difficult to maintain, inconsistencies due to different punctuation conventions over various texts and languages are hard to cope with.

(2) Regarding that are basic objective is to create encoding dictionaries, natural language sentences, in which the translation pairs occur, play an important role in our proto-dictionaries. In the basic version relevant sentences were retrieved from the original texts, which were stored separately. Relying on data stored in XML-format renders this task more straightforward: We need only search in the very same files for the stemmed and inflected forms of a word. The original sentences are generated from the latter information: From the *msd* attribute values.

(3) The automatically attained translation pairs should be assigned to typical text types in which they occur.

(4) As the method itself deals only with one-token expressions an additional module is required to retrieve multiword expressions (such as collocations or verbal struc-

tures) and their translations for them (cf. Chapter 7). Language independent methods do exist to extract the desirable multiword expressions, thus, a cross-linguistically "quasi-uniform" morphological annotation, which can be processed alike, should be established to preserve the language independent characteristics of the tool-chain.

Thus, we have converted the parallel corpus into XML-format, which contains all the relevant information in a structured way, which in turn can be extracted in different ways when needed.

6.4.1 Workflow — uniform corpus representation

(1) **Creating the XML version for each of the texts** The XML parallel corpus was created on the basis of the morphologically disambiguated texts. Basically, a simplified version of the TEI-compliant Hungarian National Corpus served as standard and the following structural tags were used: *w* for words, *s* for sentences and *p* for paragraphs. The s-tag has an obligatory attribute: *sid* (sentence identifier), which assign a unique identifier to each of the sentences in the corpus. w-tags have two attributes: *lemma* and *msd*. The value of the attribute *lemma* is the stem of the given word form, while the morphosyntactic description of the word form is assigned to the attribute *msd*. The content node comprises the word form itself.

```
<s sid="1.732">
<w lemma="Mohón" msd="Adv">Mohón</w>
<w lemma="vár" msd="V">várták</w>
<w lemma="az" msd="Det">az</w>
<w lemma="első" msd="Num.NOM">első</w>
<w lemma="oktatás" msd="N.ACC">oktatást</w>
<w lemma="." msd="SPUNCT">.</w>
</s>
```

Figure 6.3: A Hungarian sentence in XML format: "*They eagerly awaited the first education*"

(2) **Harmonizing the morphological annotation** Since morphological analyzers vary from language to language, using different annotations, the morphological information has to be harmonized so that it can be processed in a uniform way later on, regardless of the previous processing steps. Adding this step is necessary for multiple reasons: First, the resulting dictionaries should be improved through

including gender and part-of-speech information. Secondly, an additional module has to be set up which is able to handle multiword expressions, such as collocations or verbal structures. The extraction of such expressions rests upon part-of-speech information and upon various markers of the syntactic structure, such as case or word order. Thus, for multiple reasons, it is essential to make part-of-speech and case information directly accessible. We also took gender information into consideration wherever it was present (e.g. in the case of Lithuanian and Slovenian).

(3) **Aligning texts** In the next step lemmatized versions were created and aligned by means of Hunalign (Varga et al, 1995).

(4) **Creating parallel XML corpus** The lemmatized and aligned sentences were looked up in the XML-corpora to create the aligned XML-corpora. A *tuid* (translation unit identifier) attribute has been also introduced to map the relevant sentences of the parallel corpus. Figure 6.4 examplifies a many-to-one alignment: The Hungarian sentences with *id* attributes 1.731 and 1.732 are aligned with the Lithuanian sentence with *sid* attribute 1.694.

```
<s sid="1.732" tuid="731-732,694-694">
<w lemma="Mohón" msd="Adv">Mohón</w>
<w lemma="vár" msd="V">várták</w>
<w lemma="az" msd="Det">az</w>
<w lemma="első" msd="Num.NOM">első</w>
<w lemma="oktatás" msd="N.ACC">oktatást</w>
<w lemma="." msd="SPUNCT">.</w>
</s>
```

Figure 6.4: - An aligned Hungarian sentence in XML format: "*They eagerly awaited the first education*"

Unfortunately, texts of XML-format could not have been directly aligned, since sentence alignment works best on the lemmatized versions of texts. Even the addition of a single ID number to every sentence distorts the quality of the alignment significantly, mostly in the case of short sentences. Therefore, (3) and (4) are necessary steps in the alignment process.

6.4.2 Results

Table 6.11 indicates the sizes of the resulting XML parallel corpora in terms of tokens, sentences and translation units.

Hungarian-Lithuanian That is, the Hungarian-Lithuanian parallel corpus comprises 4,813,956 Hungarian and 4,141,521 Lithuanian lemmata, 319,489 Hungarian and 320,678 Lithuanian sentences and 304,419 aligned translation units[1].

Hungarian-Slovenian As for Hungarian and Slovenian, the parallel corpus comprised 723,857 Hungarian and 809,448 Slovenian tokens, 40,926 Hungarian and 42,659 Slovenian sentences. It consisted of 38,791 aligned translation units.

Hungarian-English The Hungarian-English XML parallel corpus was created mainly for evaluation purposes. Since it was created on the basis of Hunglish 1.0 (Varga et al., 2005), the workflow differed somewhat from that of discussed in the previous section. As Hunglish 1.0 consisted of aligned sentences, only annotation was performed. The Hungarian sub-corpus was annotated with the HNC-tagger, while the English part of the parallel corpus was analyzed with the TreeTagger (Schmid, 1994), using the tag-set of Penn Treebank. The Hungarian-English XML parallel corpus contains only one-to-one alignments of Hunglish 1.0 and the subtitle sub-corpus has been completely omitted.

As a result, a parallel corpus of XML-format has been created comprising 6,921,127 Hungarian and 8,312,795 English tokens and 494,044 sentences both in the English and the Hungarian side. Because the resulting parallel corpus consisted of only 1-to-1 mappings, the number of translation units was equal to that of the sentences.

	Hungarian	*Lithuanian*	*Hungarian*	*Slovenian*	*Hungarian*	*English*
Tokens	4,813,956	4,141,521	723,857	809,448	6,921,127	8,312,795
Sentences	319,489	320,678	40,926	42,659	494,044	494,044
TUs	304,419		38,791		494,044	

Table 6.11: The sizes of the resulting XML parallel corpora in terms of tokens, sentences and translation units

[1]Note that the original parallel corpus was extended: 27 novels were added to the parallel corpus. The Hungarian versions were gathered by Iván Mittelholcz by contacting publishers, the Lithuanian counterparts form part either of the Lithuanian National Corpus or the Lithuanian-English parallel corpus.

6.5 Conclusion

In this chapter the generation process of two proto-dictionaries for lesser used languages is described. We have verified that word alignment on parallel corpora is able to facilitate the cost-effective creation of bilingual encoding dictionaries.

This methodology meets the expectations put forward in the previous chapters. It is *corpus-driven*, thus it diminishes the role of intuition in the dictionary building process.

Economical considerations Once the parallel corpus is available, word alignment on parallel corpus decreases the amount of human labour needed to produce a bilingual dictionary. It turned out that the most time-consuming part of the workflow is the collection and normalization of parallel texts.

Reversibility Once a suitable parallel corpus is available, the creation of the reversed proto-dictionary is rather straightforward. This is due to the asymmetric nature of the automatically attained translation relation, which is conditional probability. The reversed proto-dictionaries—Lithuanian-Hungarian and Slovenian-Hungarian were generated, too.

Language independency Since hunalign and GIZA++ are language independent tools, sentence alignment and word alignment are readily re-applicable for any language pair. We have generated four proto-dictionaries on the basis of two readily available parallel corpora: for French and Dutch and for Hungarian and English.

Multiple meanings The method is able to rank polysemious meanings, that is, the automatically retrieved translation probability indicates how the translation space is divided among the various translation candidates, i.e. which is the most frequent translation of the source word. This feature of the applied method was illustrated with the example of the Lithuanian source word *puikus* which was assigned 9 Hungarian translations.

Encoding dictionaries An additional requirement toward the method was that it should enable the creation of encoding dictionaries. In our view, the automatically retrieved natural example sentences are of great help when trying to find the translation that "produces the smoothest translation" among the possible translation candidates. Obviously, this step is not wholly automatic, but the retrieval of competing translation candidates and the relevant contexts for each of these candidates may help lexicographers (and end-users) to focus their attention on the relevant linguistic facts. This

standpoint was underpinned by the example of *aiskiai*, the Hungarian translations of which were clearly characterized based on the retrieved contexts.

Evaluation Instead of correct and incorrect translation pairs, the evaluation is based on lexicographically acceptable and lexicographically unacceptable translation units. The detailed evaluation of the resulting proto-dictionaries showed that the proportion of correct translation pairs depends upon the frequency of the source and target lemmata and on the automatically attained translation probabilities: The higher the translation probability, the greater the proportion of lexicographically acceptable translation units. The coverage of the proto-dictionaries is far below of what was expected (cf. Chapter 1). However, refinement of parameters might result in an increased number of lexicographically acceptable translation units. This will be discussed in Chapter 8 in more detail.

Uniform XML format The parallel corpora were converted into XML-format with a simple and uniform morphological annotation. This conversion made the following workflow more simple, through rendering possible the uniform processing of the various parallel corpora. Moreover, it is easier to generate parallel corpora with various levels of annotation (containing only lemmata or part-of-speech information is also included, etc.). This flexibility is especially important when extracting multiword expression translation candidates, since in this case finding the relevant morphological/syntactical information is not straightforward (cf. 7).

Difficultites Unfortunately, some difficulties have to be addressed, too. First, at the present state of research the size of the parallel corpus is not great enough to ensure an appropriate coverage of dictionaries. This problem might have several solutions, either introducing related methods, such as retrieving translation pairs based on comparable corpora or fine-tuning the parameters used for filtering (cf. Chapter 8).

Secondly, the method in its present form enables only the retrieval of one-token translation pairs. Thus, it does not handle collocates or verbal expressions. Since such structures are inherently part of natural languages and are essential for the production of idiomatically correct translations, they have to be included in the proto-dictionaries. In the next chapter we will investigate to what extent the automatic extraction of translations of verbal structures is possible.

6.6 Appendix: The Morphosyntactic Annotation

Hungarian The morphosyntactic annotation present in the Hungarian National Corpus served as the starting point of the conversion. The morphosyntactic annotation comprises no gender information, as in Hungarian distinctions based on grammatical gender do not exist. Hungarian cases and part-of-speeches are listed in Tables 6.12 and 6.13, respectively.

Abbr.	Case	Morpheme	Example
NOM	nominative	—	The dog barked.
ACC	accusative	-t, -at, -et, -ot, -öt	I saw a dog.
DAT	dative	-nak, -nek	I gave a bone to the dog.
ILL	illative	-ba, -be	I went into the theater.
INE	inessive	-ban, -ben	I am in the theater.
ELA	elative	-ból, -ből	I am coming from school.
ALL	allative	-hoz, -hez, -höz	I am going to John.
ADE	adessive	-nál, -nél	I am at the house.
ABL	ablative	-tól, -től	I am coming from Mary.
SUB	sublative	-ra, -re	I sat onto a chair.
SUP	superessive	-n, -on, -en, -ön	I am sitting on a chair.
DEL	delative	-ról, -ről	We are talking about him.
INS	instrumental	-val, -vel	I am eating with a fork.
FAC	factive	-vá, -vé	It became sweet.
FOR	formative	-ként, -képp(en)	He works as a teacher.
TEM	temporal	-kor	He arrived at five.
CAU	causalis	-ért	We are fighting for her.
TER	terminative	-ig	You have to pay until May.
SOC	sociative	-stul, -stül	Including interest
ESS	essivus formalis	-ul, -ül	The cave served as a house for him.

Table 6.12: Hungarian case suffixes

Abbr.	Case
Det	Determiner
NU	Postposition
Pro	Pronoun
Num	Numerical
Con	Conjunction
A	Adjective
N	Noun
V	Verb
Dig	Digit
Adv	Adverb
Clit	Clitic
MIA	Future participle
MIB	Past participle
MIF	Present participle
Int	Interjection
S	Sentence-word
Abb	Abbreviation
V.INF	Infinitive
V.HIN	Adverbial participle
ELO	Prefix

Table 6.13: Hungarian part-of-speeches

Lithuanian As for Lithuanian we have considered three types of information: part-of-speech category, gender and case, based on Rimkutė et al. (2007).

Abbr.	**Case**
NOM	Nominative
GEN	Genitive
DAT	Dative
ACC	Accusative
INS	Instrumental
LOC	Locative
VOC	Vocative

Abbr.	**Gender**
m	Masculine
f	Feminine
c	Common gender
n	Neuter gender

Table 6.14: Lithuanian cases and genders

Abbr.	POS
N	Noun
N	Proper noun
N	Uninflective proper nouns (*'Don'*, van, Sanct)
A	Adjective
V	Verb
Pro	Pronoun
V.PART	Participle *'walking'*
N	Gerund *'on the walk home'*
V.HPART	Half participle *'when speaking'*
Num	Numeral
Con	Conjunction
Adv	Adverb
Particle	Particle *'also'*
Prep	Preposition
Int	Onomatopoeic interjection
Int	Interjection
Abb	Acronym
Abb	Abbreviation
V.INF	Infinitive
V.INF2	Second infinitive *'at a run'*
Num rom	Roman numbers
UNKNOWNTAG	Unrecognized
Idiom	Idiom *'rest eternal'*
Con	Connective idiom *'et cetera'*
idPS	P.S.
Prep	Prepositional idiom *'inter alia'*
Pro	Pronominal idiom *'nevertheless'*

Table 6.15: Lithuanian part-of-speech categories

English The tagset used for tagging the English sub-corpus of Hunalign is the Penn Treebank Tagset, described for example in Marcus et al. (1993).

Con	CC	Coordinating conjunction
Num	CD	Cardinal number
Det	DT	Determiner
Expl	EX	Existential there
FW	FW	Foreign word
Prep	IN	Preposition or subordinating conjunction
A	JJ	Adjective
A	JJR	Adjective, comparative
A	JJS	Adjective, superlative
LITEM	LS	List item marker
MOD	MD	Modal
N	NN	Noun, singular or mass
N	NNS	Noun, plural
N	NP	Proper noun, singular
N	NPS	Proper noun, plural
Adv	PDT	Predeterminer (all, half, nary, quite, such)
POS	POS	Possessive ending ('s)
Pro	PP	Personal pronoun
Pro	PP$	Possessive pronoun
Adv	RB	Adverb
Adv	RBR	Adverb, comparative
Adv	RBS	Adverb, superlative
Particle	RP	Particle
SYM	SYM	Symbol
TO	TO	to
Int	UH	Interjection
V	VB	Verb, base form
V	VBD	Verb, past tense
V.PART	VBG	Verb, gerund or present participle
V.PART	VBN	Verb, past participle
V	VBP	Verb, non-3rd person singular present
V	VBZ	Verb, 3rd person singular present
Det	WDT	Wh-determiner (that, what, whatever, which, whichever)
Pro	WP	Wh-pronoun (what, who, whoever, whom, whomever)
Pro	WP$	Possessive wh-pronoun (whose)
Adv	WRB	Wh-adverb (how, however, whenever, when, where, whereby, wherever, why)
WPUNCT	WPUNCT	wpunct
SPUNCT	SPUNCT	spunct

Table 6.16: Tag-set in the Penn Treebank

Slovenian The tagset used for tagging the Slovenian sub-corpus of the Hungarian-Slovenian parallel corpus is based on Erjavec et al. (2005).

Abbr.	Case
NOM	Nominative
GEN	Genitive
DAT	Dative
ACC	Accusative
INS	Instrumental
LOC	Locative

Table 6.17: Slovenian cases

Abbr.	POS
Pro	Pronoun
Num	Numeral
Con	Conjunction
A	Adjective
N	Noun
V	Verb
Adv	Adverb
Int	Interjection
Abb	Abbreviation
Prep	Preposition
Residual	Residual
Particle	Particle

Table 6.18: Slovenian part-of-speech categories

7

Extracting Parallel Verbal Structures

7.1 Introduction

Chapter 7 describes a solution to the treatment of multiword expressions within the given framework. The problem is addressed through the extraction of parallel verbal structures, that is, verbal structures and their translation candidates.

Verbal structure In accordance with Sass (2011), the term verbal structure is conceived of in a rather broad sense here. It refers to the verbal head along with its salient complementation pattern, where the complementation pattern comprises frequently occurring constituents regardless of their syntactic status, thus, complements and adjuncts are both considered. A verbal structure might also consist of lexically bound constituents.

Example For instance, the French expression *donner lieu á* 'give rise to' is a verbal structure comprising a lexically bound (*lieu*) and a lexically unbound constituent. In the latter case only the preposition (*à*) – the syntactic function marker – is inherently part of the verbal structure.

The algorithm The research described here heavily relies on the work presented in Sass (2011) most elaborately. Throughout his PhD work his goal was to invent an algorithm which is able to extract verbal structures on the basis of corpus data. A suitable

algorithm is expected to find out which constituents belong to the verbal structure. It also should be able to determine whether the lexical head of each constituent is inherently part of the given verbal structure or not. This entails that the algorithm cannot be based on predefined syntactic patterns:

(1) The algorithm is expected to be language independent also operating for languages with various (or free) word orders.

(2) Verbal structures are made up of different number of constituents.

(3) It cannot be determined in advance whether a head is lexically bound or not.

Main features of the extraction algorithm The algorithm of Sass (2011) exhibits two features which are indispensable for our purposes:

(1) The algorithm is language independent that is, it does not make use of any language specific feature.

(2) Verbal structures are extracted in an unsupervised way.

Language independency is essential so that the same algorithm could be applied to any language. Since the production of training data usually is not affordable in the case of less resourced languages an unsupervised learning algorithm should be applied.

The exact operation of the algorithm will be described in Subsection 7.2.

Example Table 7.1 gives some examples of French and Dutch verbal structures. In fact, this entry is part of the automatically generated verbal proto-dictionary. Four different Dutch translations were assigned to the French verbal structure *prendre médicament* 'take medicine'. The most likely Dutch translation is the literary translation *geneesmiddel innemen*. The next translation *geneesmiddel gebruiken* 'use medicine' has a slightly different meaning but might be used in similar contexts. Although the third translation *start met gebruik van* 'start the use of' is not a whole verbal structure, it can also be considered as a relevant translation. Finally, although the least likely translation *staan onder invloed van drug*— literally: 'stand under the influence of drug'—has intuitively a different meaning, there are contexts where the TL expression can be used as the translation of the SL expression.

Relation to phrase alignment Nowadays phrase-alignment is a hot topic in the field of statistical machine translation. Phrase alignment techniques start out of word

$Expression_{source}$	$Expression_{target}$	p_{tr}	F_s	F_t
prendre médicament=obj	neem-in genees-middel=obj	0.377	53	32
	gebruik genees-middel=obj	0.102	53	21
	start gebruik=met:cmp met:cmp-van	0.097	53	28
	sta invloed=onder:particle drug=van:cmp	0.05	53	11

Table 7.1: A sample entry from the French-Dutch verbal proto-dictionary.

alignment and they try to find out how complete phrases are aligned by applying some suitable heuristics. As verbal structures may be constituted by long-distance dependencies, especially in the case of free constituent order languages—where verbal complements and adjunct are free to mingle—phrase-based methods used in machine translation does not work here. Thus, verbal structures should be detected independently of word alignment in a preprocessing phase. The workflow is made up of the following steps:

Workflow

(1) In the first step a list of verbal structure candidates are extracted from the corpora.

(2) Then, the tokens of each occurrence of the verbal structures are merged.

(3) In the rest of the workflow they could be handled as one-token units.

Structure of Chapter 7 In Section 7.2 the applied extraction method for verbal structures is described. Then two experiments were carried out. The first one is a proof-of-concept investigation with the aim to verify the viability of the approach. In this case a shallow parsed parallel corpus was exploited. The corresponding workflow and the results are described in Section 7.3. In Section 7.4 we have enhanced our method and utilized a deep parsed parallel corpus instead of the shallow parsed parallel corpus. Section 7.5 describes the conclusions and points out possible further research directions.

7.2 Description of the Extraction Method

7.2.1 Conversion of input corpora

Clause boundary and noun phrase detection The extraction method is based on the following presuppositions:

(1) The verb and its whole complementation frame (adjuncts, complements and their

syntactic markers) lie within one clause.

(2) One clause contains the complement structure of only one verb.

Hence, the input corpus has to have *clause boundary annotation*. Moreover, for a number of reasons the extraction of relevant complementation frames presupposes that *noun phrase information* is also present in the corpus.

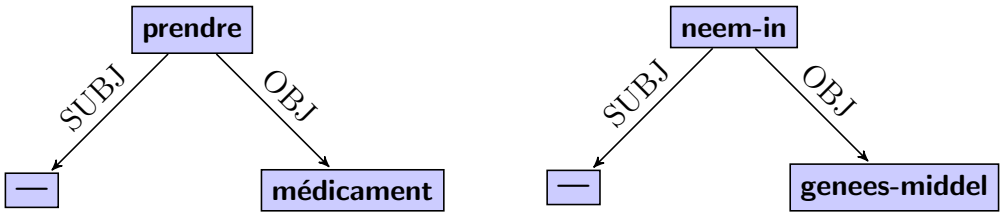
(1) Since it is usually one of the head nouns that is a bound element in the verbal structure (if there is any at all) we need to know which constituent is the head noun.

(2) Being part of the complementation frame, syntactic markers (prepositions, case suffixes, special syntactic positions) between head nouns and the verb should also be detected.

Uniform representation of corpus texts One serious expectation from the verbal structure extraction algorithm is that it should be language independent, that is, it should be able to detect verbal structures regardless of the specificities of the particular language. Various languages greatly differ with regard to how they mark syntactically the relation between the constituents and the verb. Roughly speaking, the complementation pattern of a verb comprises subjects, objects, complements and adjuncts, each of which can be marked by certain positions in the sentence (e.g. subjects and objects in bounded constituent order languages) or by prepositions, postpositions and case suffixes. The algorithm should operate alike, regardless of how the complementation pattern of the verb is expressed. Thus, corpus sentences are converted into partial dependency trees, thus providing a uniform representation of diverse language data.

Dependency grammar In the proposed framework the central element of a sentence is the verb. Complements and adjuncts are dependent on the verb. The specific relations are marked by case suffixes or postpositions in Hungarian. In other languages word order, prepositions may also mark the syntactic function. The algorithm considers only those function markers that directly appear in the corpus. It does not take semantic information (e.g. thematic roles) into account. In dependency trees constituent heads (i.e. the verbal head and the heads of the corresponding NPs) are represented by nodes and syntactic relations by edges regardless of the exact nature of the syntactic function marker (case suffix, postposition, preposition, subject or object).

Example Figure 7.1a and 7.1b are two verbal structures represented as dependency trees. In both cases there are two dependents of the verb, a subject and an object. The object is lexically bound. In fact, the representations of the two verbal structures are the same, except for the lexical elements.



(a) *Prendre médicament* **(b)** *Genesmiddel innemen*

Figure 7.1: Dependency trees

Multi-level dependencies However, in many verbal structures the dependent of a dependent element—an adjunct of the lexical head, etc—might be also salient, thus has to be included into the verbal structure. These structures are multi-level dependencies. Although Sass (2011) converts multi-level dependencies into one-level dependencies, in the present analysis we used multi-level dependencies to be able to represent that an element belongs to a non-verbal node (cf. Figure 7.2).

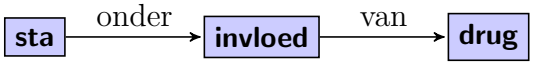


Figure 7.2: Multi-level dependency

Length of verbal structures The length of a verbal structure is the sum of the number of bound lexical elements and the number of dependency labels in the given verbal structure. The subject node (if any) and the subject relation is not considered. Thus the length of "*prendre médicament*" and "*geneesmidel inneemen*" is 3, while the length of "*staan onder invloed van drug*" is 5.

7.2.2 The algorithm

Once the morphosyntactically annotated corpus is converted into the desired representation, the extraction of verbal structures takes place.

The algorithm The technique determines the salient complementation frames of a verb by counting the frequent subframes. It follows three main steps:

(1) The preliminary verbal structures are generated on the basis of all clauses in the corpus.

(2) Every nominal head of these verbal structures is deleted in all possible variations.

(3) The resulting verbal structures are ranked according to their length. Verbal structures occurring less than five times in the corpus are omitted and their frequency is added to the frequency of the first matching verbal structure in the ranked list.

The repeated execution of the last step results in a list of the relevant verbal structures.

As a result, we obtain verbal structures, such as the French *mettre accent sur* or its Dutch equivalent *leggen nadruk op* 'put emphasis on'.

In what follows, we will examine whether this approach can be extended to retrieve parallel verbal structures that could be included in the proto-dictionaries.

7.3 Extracting Parallel Verbal Structures from a Shallow Parsed Parallel Corpus

In this section a proof-of-concept investigation is described to prove that the proposed technique is suitable to detect translations for verbal structures. A similar experiment is described in Sass (2010). Throughout the experiment a predefined class of verbs were focused on and the relevant verbal structures were also manually selected. Thus, the workflow described in this section is not fully automatic.

Workflow The workflow is made up of the following steps:

(1) Manual selection of 20 frequent verbs and their translations

(2) The conversion of the input corpus[1]

  (i) Clause boundary detection

  (ii) Noun phrase annotation

  (iii) Conversion into partial dependency trees

(3) Extraction of the most frequent verbal structures for both Dutch and French

[1]Steps (2) and (3) were performed by Bálint Sass.

(4) Creating the proto-dictionary

  (i) Merging each occurrence of the verbal structures into one-toke expressions in the corpus, so that they could be treated as one-token lemmata

  (ii) Performing word alignment

  (iii) Filtering

(5) Evaluation.

7.3.1 Semi-automatic extraction of verbal structures

7.3.1.1 The scope of investigated verbs

In the first step 20 frequent polysemous French verbs were selected manually (e.g. *mettre* 'put'). Then one Dutch translation was assigned manually to each of these French verbs based on a French-Dutch dictionary (e.g. in the case of the French verb *mettre* the Dutch verb *leggen* was selected as translational equivalent). Table 7.2 lists the selected French verbs and their Dutch equivalents. Table 7.2 also indicates the number of the different retrieved verbal structures. As it is presented in 7.2, for some verbs (*enlever*, *rester*, *voir*, *vergaan*, *zien*) not any verbal structures could have been retrieved meeting the criteria defined below in Section 7.3.2. In the next step, French and Dutch verbal structures were extracted automatically from the relevant monolingual part of the parallel corpus applying the algorithm of Sass (2011).

7.3.1.2 Conversion of the input corpus

Input corpus The input corpus was the French-Dutch sub-corpus of the Dutch Parallel Corpus (Clercq and Perez, 2010). The subcorpus consists of 3,605,791 French tokens, 3,214,756 Dutch tokens and 186,945 aligned units. In our experiment both noun phrase detection and clause boundary segmentation were carried out through rather simple approximative rules for both languages. Some of the main principles behind the rules are listed below (see Sass, 2010, p. 102).

French verb	Structure types	Dutch translation	Structure types	English
donner	12	geven	31	give
effectuer	3	teweegbrengen	0	carry out
enlever	0	verwijderen	1	remove
faire	31	doen	12	do
mener	2	leiden	4	lead
mettre	26	leggen	5	put
montrer	4	wijzen	1	show
obtenir	5	behalen	2	obtain
offrir	1	aanbieden	2	offer
ouvrir	1	openen	1	open
passer	3	vergaan	0	pass
porter	3	brengen	14	bring
prendre	23	nemen	23	take
recevoire	2	krijgen	12	receive
rendre	3	maken	19	make (sb ADJ)
rester	0	blijven	1	stay
tenir	4	houden	11	hold
traiter	1	behandelen	2	treat
trouver	3	vinden	6	find
voire	0	zien	0	see

Table 7.2: French verbs and their Dutch translations. The number of verbal structure types.

Clause boundary detection clause boundaries are indicated by any of the following:

(1) Every sentence boundary

(2) Conjunctions

(3) The Dutch *te* and the French *pour* introducing a subordinate clause

(4) Relative pronouns

(5) Certain punctuation marks (comma, colon and semicolon) if they occur between two verbs in the same sentence.

Noun phrase detection We have relied on the following rules while detecting noun phrases:

(1) Nouns are considered as heads of noun phrases

(2) Reflexive pronouns (Dutch *zich* and French *se*) are also considered to be heads of noun phrases

(3) The syntactic relation is indicated by the preposition at the beginning of the noun phrase

If no preposition is present in that position,

- the noun phrase directly before the verb are taken to be subjects,
- the noun phrase directly after the verb are considered to be objects.

In the next step, the sentences of the morphologically annotated DPC was converted into partial dependency trees, and the automatic extraction method retrieved verbal structures comprising any of the listed verbs.

7.3.2 Manual selection of relevant verbal structures

Lexicographically interesting translations Since our purpose is to find translations to be included in a dictionary, we have manually selected verbal structures that presumably have interesting translations. However, recall that the described methodology is expected to be language independent in order to be easily re-applicable to different language pairs. This requirement entails that we cannot decide in advance—on the basis of monolingual data—whether a verbal structure will be interesting from a translation point of view and should be kept in a dictionary, or conversely, its translation is fully compositional and, thus, should be neglected. For example, although the French verbal structure *mettre l'appareil hors tension* 'cut off power to the device' can be considered compositional, its Dutch translation-—*uitschakelen*-—does not preserve the original structure, hence this translation should be included in the proto-dictionary.

The applied criteria Consequently, we had to make use of certain hints when determining whether an expression is worth keeping or not. The following cues were relied on:

(1) The meaning of the source expression is not transparent (e.g. *faire mouche* 'hit the bull's eye'.

(2) The translation of the verbal structure will presumably not be translated into the target language in a compositional way:

(a) The verbal structure is institutionalized (e.g. *mettre l'appareil hors tension* 'power off the device').

(b) The syntactic function of at least one of the nouns in the verbal structure is a verb modifier in the Hungarian translation of the original expression (cf. É. Kiss et al., 2003) (e.g. *donner conseil*, 'tanácsot ad' – advise)

On the other hand, some distributional criteria were also set:

(4) Verbal structures were kept only if they contained at least one noun.

(5) In this experiment we confined ourselves to the examination of nouns occurring in any relation to the verb except for the subject relation.

(6) We did not strive to keep the complete verbal frame, that is, empty suffixes without a typical bound noun were omitted in some cases. The reasons behind this are twofold:

(a) Omitting empty case suffixes increases the number of the occurrence of the given frame type, thus reducing the problem of data sparseness.

(b) As the input corpus of the word alignment component contained neither partial syntactic annotation nor clause boundary information, the right preposition could have not been easily recognized in the parallel corpus at that stage of research.

Types of verbal structures Verbal structures matching the following syntactic patterns serve as input for the word alignment algorithm. (V: verb; N_ACC: the syntactic function of the noun is object; ACC: an object with an unbound nominal head; N_PREP: the noun appears with a preposition; PREP: a preposition with an unbound nominal head):

(1) V + N_ACC

(2) V + N_PREP

(3) V + ACC + N_PREP

(4) V + N_ACC + PREP

(5) V + N_PREP + PREP

(6) V + N_PREP + N_PREP + PREP

By the end of this stage 126 French verbal structures and 146 Dutch verbal structures were available. In the next phase the selected verbal structures are detected in both side of the parallel corpus and are merged into a one-token expression so that they can be aligned by the alignment component.

7.3.3 Creating the proto-dictionary

Merging In the rest of the workflow the selected multi-word expressions were treated as one-token units so that they could serve as direct input to the alignment component.

The input corpus for word alignment has been lemmatized but does not contain clause boundary or noun phrase information. The first step of this stage is the detection of the listed verbal structures in the corpus. It is important to note that always the longest matching verbal structure was selected. Then the longest matching verbal structures were merged into one-token expressions.

The 126 selected French verbal structures occurred 7805 times in the French part of the parallel corpus, while the 146 Dutch verbal patterns occurred 8029 times in the Dutch part of the parallel corpus.

Aligning Just as in the case of the preceding research phases, word alignment was carried out with GIZA++ alignment tool (Och and Ney, 2003)[1].

Filtering However, as the assigned translational probability strongly varies, at this stage we have many wrong translation candidates. Therefore, some constraints had to be introduced to find the best translation candidates without the loss of too many correct pairs. Our previous experiments (eg. Héja, 2010) have shown that exploiting corpus frequency data and translational probability facilitates filtering. Thus, the data illustrated in Table 7.3 have to be included in the proto-dictionary:

Based on previous evaluation of Hungarian-Lithuanian and Hungarian-Slovenian proto-dictionaries entries in the proto-dictionary need to meet the following general criteria:

(1) Source language and target language members of lemma pairs should occur at least 5 times in order to have reliable amount of data when estimating probabilities.

[1]Recall that this algorithm assigns translational probabilities to source language and target language lemma pairs. The translational probability is an estimation of the conditional probability of the target word given the source word, $P(W_{target}|W_{source})$ by means of the EM (expectation maximization) algorithm (Dempster et al., 1977). The retrieved lemma pairs together with their translational probabilities served as the starting point for the proto-dictionaries.

$Expression_s$	$Expression_t$	$p(w_t\|w_s)$	$Freq_s$	$Freq_t$
mettre_à_jour 'update'	actualiseren 'update'	0.047	105	39

Table 7.3: Data in the verbal proto-dictionary

(2) The translational probability is equal to or greater than 0.5.

(3) The ratio of the frequency of the source and target lemmata (or expressions) may not be higher than a certain threshold. The reason for this is that in the case of rarely used source lemmata the alignment algorithm might assign high translational probabilities to incorrect lemma pairs if the target lemma occurs frequently in the corpus and both members of the lemma pair recurrently show up in aligned units. Table 7.4 illustrates this phenomenon.

$Expression_s$	$Expression_t$	$p(w_t\|w_s)$	$Freq_s$	$Freq_t$
mettre_vie_en_danger 'jeopardize sy's life'	rekening_houden 'take into consideration'	0.877	24	577

Table 7.4: Wrong translation candidates

As before, since the objective of this work is to provide lexicographers with empirical data, instead of simply telling apart wrong translation candidates from right ones, we have decided to distinguish between lexicographically acceptable and lexicographically unacceptable translation units throughout the evaluation process.

You may recall that the evaluation of one-word units has yielded the result that with the above parameters (frequency of source and target lemmata is at least 5 and the translation probability is at least 0.5) about 90% of the retrieved translation units is lexicographically acceptable. However, the aim of this method is not to increase precision but to optimize both precision and coverage, and to find the most useful trade-off between the two. Considering the fact that coverage correlates inversely with precision and manual post-editing has to be carried out anyway, the 90% precision is too high.

Since our basic objective at this stage is not to find the best parameter setting but to examine whether our original methodology can facilitate lexicographic work by extending it to verbal structures, we set the parameter values so that we could keep a

relatively high number of lexicographically acceptable translation units, even if it leads to low precision.

Consequently, we included every translation candidate in our proto-dictionary where both the source and the target lemma occurred more than 5 times. The translational probability dropped from 0.5 to 0.2.

7.3.4 Results

Partial evaluation of our results has shown that the proposed method is suitable for the extraction of translations even in the case of verbal structures. With the parameters determined above, we obtained 906 translation candidate pairs, where at least one member of the pair is a verbal structure. We had 113 different French verbal structures in our proto-dictionary and 127 different Dutch verbal structures. We focused on the French verbal expressions throughout the evaluation.

294 translation candidates were manually checked. The evaluation was based on three categories: Besides right and wrong translation candidates we also distinguished partially correct translation candidates. The latter refers to translation candidates where any member of the pair is an unrecognized multiword expression, thus resulting only in partial alignment.

Out of the 294 translation candidates 57 were right translations (19%) and 28 translation candidates turned out to be partially correct (9,5%).

Possible improvements Based on the manual evaluation of the candidate pairs we intend to introduce an additional parameter to filter out wrong translation candidates: The number of sentences in which a given candidate pair shows up. An additional finding was that the number of wrong translation candidates could be significantly reduced if clause boundary and noun phrase information were exploited not only during the extraction of verbal structures but in the alignment phase, as well. The experiment described in Section 7.4 is focusing on this issue.

However, it is important to note that the main objective of our investigation was not to achieve the best results, but to determine whether our method is apt to retrieve verbal structure translation pairs or not. Precision can be considerably increased by changing the values of the parameters. Table 7.5 and 7.6 present two examples from the proto-dictionary.

$Expression_s$	$Expression_t$	$p(w_t\|w_s)$	$Freq_s$	$Freq_t$
mettre_à_jour	bijwerken	0.65	105	60
FR: Comment les met-on à jour ? NL: Hoe worden ze bijgewerkt ? EN: How can these be updated?				
mettre_à_jour	actualieseren	0.047	105	39
FR: De plus, un PGR mis à jour doit être soumis : NL: Bovendien dient een geactualiseerd RMP ingediend te worden : EN: In addition, an updated RMP has to be submitted:				
mettre_à_jour	aanpassen	0.037	105	442
FR: Mise à jour de la liste des produits admis au remboursement NL: Aanpassing van de lijst van de voor vergoeding aangenomen producten EN: Adjustment of the list of products admitted for reimbursement				
mettre_à_jour	update	0.03	105	34
FR: Toutes les informations au sujet du changement y ont été publiés avec de fréquentes mises à jour . NL: Alle informatie met betrekking tot de omslag is erop gepubliceerd , met regelmatige updates . EN: All information concerning changes is published there with regular updates				

Table 7.5: Example 1: Dutch translations for the French expression *mettre à jour* with one-sentence contexts

7.3.5 Discussion

The aim of this section was to describe an experiment which confirms that the methodology developed for the extraction of one-token translation candidates from parallel corpora can be extended to retrieve of multi-word verbal structures.

In the first monolingual phase we retrieved verbal structures independently from both the source and the target language in a semi-automatic way. Verbal structures for predefined source language verbs and for their translations were listed automatically on the basis of corpus data and the relevant verbal structures were then manually selected. In the next step every occurrence of the selected verbal structures was merged in the parallel corpus into a one-token unit, so that they could serve as input for the alignment algorithm.

Although due to the parameter setting the precision of our results was rather low, the retrieved translation candidates confirm that the proposed methodology is suitable to detect translations of verbal structures.

Manual evaluation has shown that considerable improvement can be achieved by developing the noun phrase and clause boundary detection for both languages. In addition, this information has to be exploited throughout the alignment process, too.

The coverage of our proto-dictionary can be significantly increased if we considered all

$Expression_s$	$Expression_t$	$p(w_t\|w_s)$	$Freq_s$	$Freq_t$
prendre_en_considération	nemen_in_aanmerking	0.186	93	73
FR: Les offres qui dérogent à cette date ne sont pas prises en considération . NL: Offertes die hiervan afwijken worden niet in aanmerking genomen . EN: Offers deviating from the indicated date will not be considered.				
prendre_en_considération	houden_rekening	0.167	93	438
FR: De plus, un PGR mis à jour doit être soumis : NL: Bovendien dient een geactualiseerd RMP ingediend te worden : EN: In addition, an updated RMP has to be submitted:				
prendre_en_considération	nemen_in_overweging	0.022	93	35
FR: La date de conclusion à prendre en considération pour le choix ... NL: De datum van sluiting die in overweging moet worden genomen voor de keuze ... EN: Termination date that has to be considered for the choice ...				

Table 7.6: Example 2: Dutch translations for the French expression *prendre en considération* with one-sentence contexts

verbs occurring in the parallel corpus instead of a predefined verb list. We expect that such an improvement would decrease the number of partial matches and by doing so, further augment the number of perfect translations. These improvements are included in the next experiment described in the following section.

7.4 Extracting Parallel Verbal Structures from Deep Parsed Parallel Corpus

Focus of the research The focus of this experiment is to improve the results of the previous experiment so that the retrieved parallel verbal structures could be included in the proto-dictionaries. For doing so, both precision and recall should be augmented. In the present section we investigate whether the use of a deep parsed parallel corpus is able to improve the results. Accordingly, the main threads of the research were determined as follows:

(1) Instead of a predefined list of verbs all sufficiently frequent verbs should be considered

(2) Every verbal structure should be considered, not only those with a bound nominal head.

(3) Instead of exploiting parallel corpora with shallow syntactic annotation we have used a deep parsed corpus.

(4) Noun phrase and clause boundary information was considered throughout the alignment process, too.

According to our expectations the proposed method can contribute to the design of the microstructures of the verbal entries by supplying information on the relevant complement structures and the typically co-occurring nominal heads.

7.4.1 Workflow

The workflow is the same as in the case of shallow syntactic parsing except for the fact that instead of the approximative clause boundary and noun phrase detection rules deep syntactic analysis was performed for both French and Dutch.

Therefore, the preprocessing step is made up of four phases:

(1) In the first phase the deep-syntactic analysis of each side of the Dutch-French parallel corpus was performed[1].

(2) Then the resulting phrase structures were converted into partial dependencies which is the required input format of the next step[2].

(3) Thirdly, monolingual verbal structures were extracted from each side of the parallel corpus[3].

(4) Finally, the selected verbal structures were merged into one-token expressions in both side of the parallel corpus comprising noun phrase and clause boundary information so that they could serve as input to the alignment algorithm[4].

We have utilized the same parallel corpus as previously, the Dutch-French subcorpus of the DPC containing 186,945 aligned units. Figure 7.3 depicts the complete workflow.

[1]The deep-parsed French subcorpus was provided by Eric Villemonte de la Clergerie. The deep syntactic analysis of the Dutch subcorpus was carried out by the author.

[2]The Dutch subcorpus was converted into partial dependencies by the author. The French subcorpus was converted into partial dependencies by Dávid Takács.

[3]Bálint Sass extracted the verbal structures.

[4]These tasks were accomplished by the author.

Syntactic parsers The Dutch corpus was parsed with the HPSG parser Alpino (Bouma et al., 2001) while the French corpus was parsed with the hybrid TIG/TAG parser (Villemonte de la Clergerie, 2010).

Alpino is a rule-based syntactic analyzer which supplies detailed syntactic annotation: Annotates the phrase boundaries and assigns syntactic function labels to the phrases.

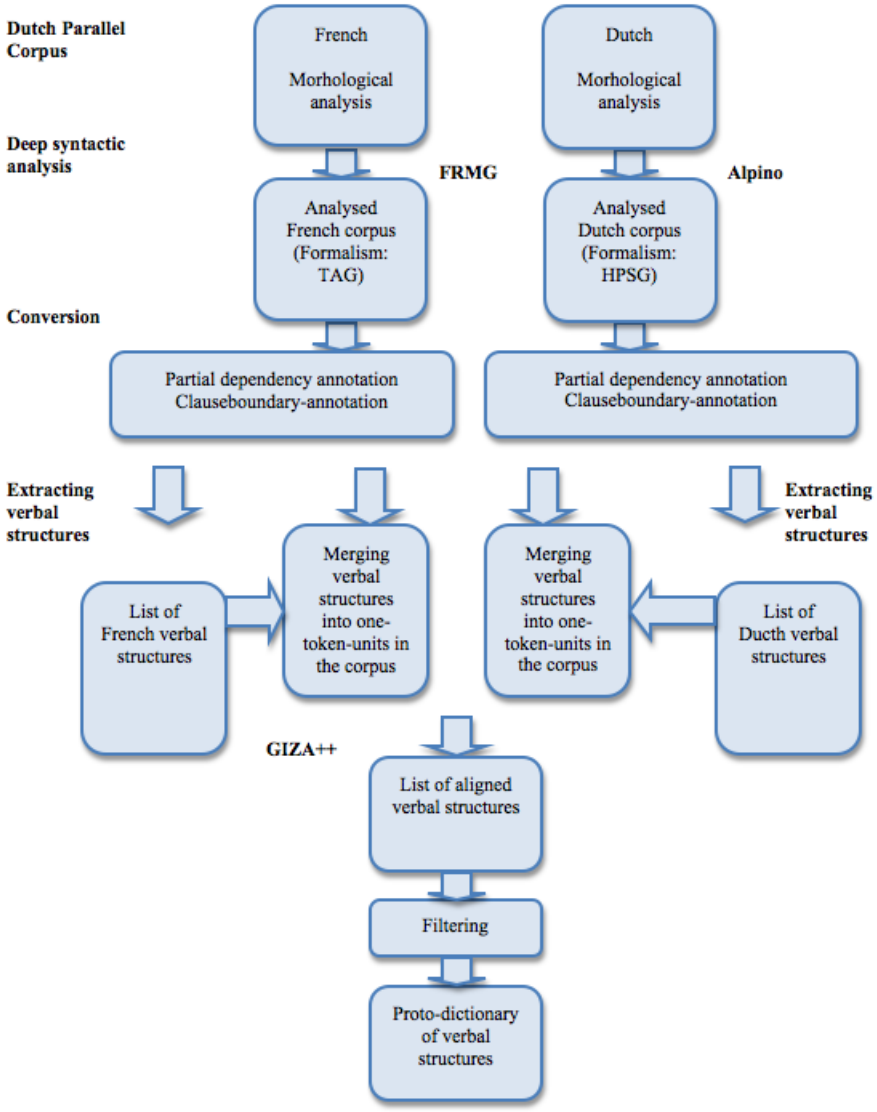


Figure 7.3: Extracting translation candidates for verbal structures on the basis of a deep parsed parallel corpus

Alpino recognizes also the verbal complement structures and particles. It also annotates the inner structure of phrases: It labels the head of the phrase and its dependents. From our perspective, an important feature of Alpino is that it recognizes the clause boundaries and assigns the relevant labels (main clauses, subordinate clauses, coordinations) to them. Unfortunately, albeit the French FRMG parser also performs a deep syntactic analysis, the annotation does not contain clause boundary segmentation.

Clause boundary detection for French Since the algorithm extracts verbal structures on the basis of clause boundary information, an extra module had to be added to detect clause boundaries for French. The module consists of the following rules:

(1) Participial structures constitute a clause in themselves.

(2) Relative pronouns always indicate a clause boundary.

(3) Infinitives always indicate a clause boundary if the infinitive is directly preceded by a preposition (*de*, *pour*, *sans*, *en vue de*, *à*, etc.)

(4) Coordinate conjunctions between two finite verbs always indicate a clause boundary (*et* 'and', *puis* 'then', *ou* 'or', etc.)

(5) Subordinate conjunctions between two finite verbs always indicate a clause boundary (*que* 'that', *quand*, *pendant que* 'when', etc.)

(6) If there is no clasueboundary between two finite verbs, the first coma, semicolon or colon should designate a clause boundary.

The rules are to be applied one after the other. The last rule applies in all the cases where there are two finite verbs without a clause boundary between them.

Conversion into partial dependency trees In the next step the HPSG and TIG/TAG annotations were converted into partial dependency trees. To increase coverage not only finite verbs and their dependencies were taken into account but passive and participial structures, as well. These structures were converted into clauses containing one finite verb. These conversions rest upon the detailed syntactic annotation, which marks the derivative forms of verbs, too (e.g. passive, various participial structures).

Passive–active Passive structures were converted into active ones for both French and Dutch. One such conversion is exemplified in Figure 7.4.

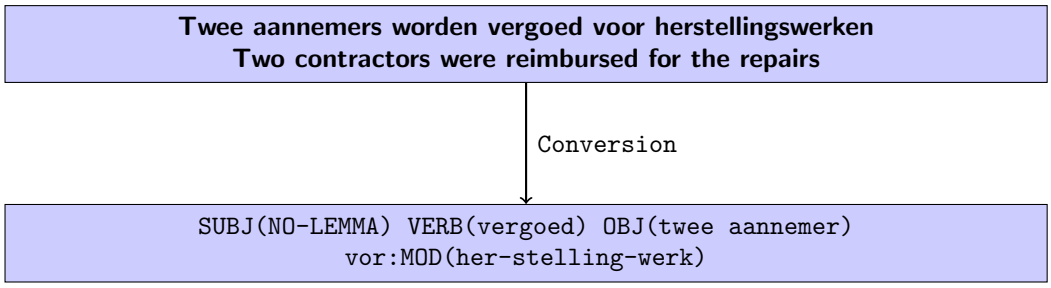


Figure 7.4: A passive-active conversion

Participial structures in a modifier position As is indicated in Figure 7.5, if participial structures appeared in a modifier position, they were converted into separate clauses. The syntactic functions of the constituents depend on the type of the participial structure.

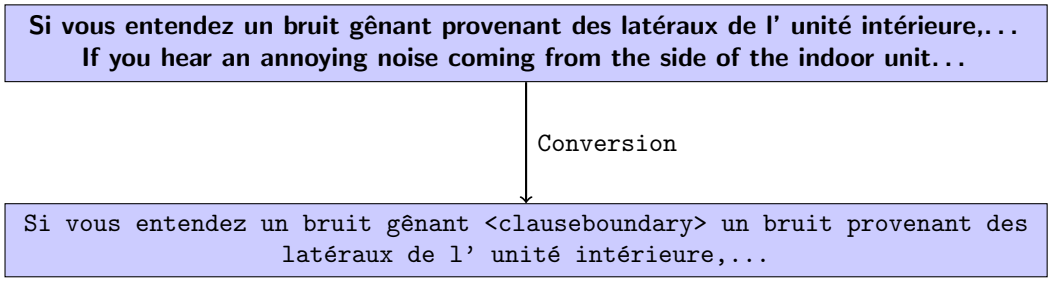


Figure 7.5: Participial structure as an additional clause

Present perfect Auxiliaries (*est*, *avoire*, *zijn*, *hebben*) were deleted in the present perfect (passé composé) verbal structures

Feature selection Finally, we had to determine the set of features yielded by the deep syntactic annotation to be considered while extracting verbal structures. Here, two contradictory requirements had to be met: On the one hand, exploiting more features characterizes the verbal structures more precisely. On the other hand, keeping too many features might considerably impair the results, since irrelevant syntactic labels increase the diversity of data. Thus, we have experimented with two different feature set.

(1) The verb and the heads of the direct dependents were kept. Adjectives and complements dependent on the dependent heads were also preserved, while determiners were omitted. Only the first constituent of a coordinated construction was kept (except for coordinated clauses).

(2) Only the verbal lemma and the head of the direct object were taken into account. Verbal structures only with different verbal lemma or different direct object were kept separately, otherwise they were merged and the corpus frequencies were recounted.

7.4.2 The automatic extraction of verbal structures

Similarly to the previous experiment the verbal structures were extracted fully automatically by applying the method of Sass (2011) described in Section 7.2 in more detail. However, there are two main differences between the previous experiment and the current one:

(1) Instead of a predefined list of verbs all verbs were considered in the corpus.

(2) The restricted scope of complementation patterns (those consisting of at least one nominal head) was extended to all types of complementation frames.

Examples In Tables 7.7, 7.8, 7.9 and 7.10 some examples of the automatically attained verbal structures are presented.

Expression	*Freq*	*English*
gebruik obj1	470	use sg
gebruik niet=mod:ADV obj1	159	do not use sg
gebruik obj1 obj1_ADJ	104	use a sort of sg
gebruik obj1 als=predc:CP	95	use sg as ...

Table 7.7: The four most frequent structure of the Dutch verb *gebruiken*

Table 7.7 reflects an unfortunate consequence of the too detailed feature set. Accordingly, some complementation patterns are completely irrelevant: Based on the results,

gebruik niet=mod:ADV obj1 and *gebruik obj1 obj1_ADJ* are both frequently occurring frame types. However, as they are fully compositional, it is not really feasible to store them in a dictionary as separate entries. Yet, both frame types could be converted into the first one (*gebruik obj1*) by omitting the irrelevant constituents, thus increasing the frequency of that frame type and the likelihood of extracting the right translation equivalents. Table 7.8 also presents an irrelevant verbal frame as a result of the deep syntactic analysis.

Expression	*Freq*	*English*
geef obj1	170	gives sg
geef obj1 obj1_ADJ	80	give a sort of sg
geef aan:obj2 obj1	78	give sg to sy
geef obj2 obj1	72	give sy sg

Table 7.8: The four most frequent structure of the Dutch verb *geven*

As Table 7.8 shows, if the object modifying adjective was not taken into account, then the most frequent complementation patterns of *geven* were exactly the expected ones.

Verbal structures in 7.9 comprise a lexical element, too. This example illustrates an other undesirable effect of deep parsing: The analyzer might annotate the same surface structure in two different ways, which in turn results in a reduced amount of data available. In the first case the Dutch *op* 'on' is dependent of the direct object of the verb, while in the second case it depends directly on the verb as a free modifier. A further issue is that our feature set is not able to grasp that *een* (Dutch determiner) appears obligatorily in this structure. Thus, although our feature set is too large from one perspective, i.e. it retrieves irrelevant verbal structures, from another viewpoint it is too small to cover all the relevant verbal structures. The exact characterization of the relevant features is a further research issue.

Expression	*Freq*	*English*
doe beroep=obj1 obj1_op	72	make a proposal for
doe beroep=obj1 op:mod	39	make a proposal for

Table 7.9: Ambiguous analysis of the Dutch expression *een beroep doen op*

Table 7.10 illustrates how partial dependency trees characterize the verbal structures relying on the selected features. For instance, the structures of the expressions *prendre médicament*, *gebruik genees-middel* and *neem-in genees-middel* are identical: the nouns are dependent on the corresponding verbs, and the dependency relation is an object relation. In the case of the expression *start gebruik met van*, the noun is dependent on the verb, too, but the dependency relation labeled by the preposition *met* (*with*). Besides, we also learn that the *met* dependency-relation is a complement-type dependency relation. Moreover, the noun *gebruik* has also a lexically free dependent, this dependency relation is labeled with the preposition *van* (of)[1] Finally, *staa onder invloed van drug* is analyzed as if "onder invloed" was a compound expression, being in *particle* relation to the verb, whereas, the noun *drug* is dependent of the verb, as well. The corresponding relation is labeled with the preposition *van* of type complement. Nevertheless, it is important to emphasize that the number of dependency levels is of no importance here, as far as the partial dependency tree determines the surface form of the given expression.

Expression	*Freq*	*English*
neem-in genees-middel=obj	32	Take medicine
gebruik genees-middel=obj	21	Use medicine
start gebruik=met:cmp met:cmp-van	28	Start with the use of
sta onder-invloed=particle drug=van:cmp	11	Stand under influence of drug

Table 7.10: A sample of Dutch verbal structures

Lexicographically interesting translation candidates In the next step those verbal structures were selected which we want to be included in the dictionary. Note, that it is only in the case of MWE—MWE translations, where we have to decide, whether a translation is worth keeping or not. Both SL MWE—TL one-token and SL one-token—TL MWE translations form definitively part of the proto-dictionaries.

Recall, that in the previous experiment some hints (e.g. institutionalization) were applied to help the manual selection of lexicographically interesting translation candidates. Unfortunately, the scope of these hints is rather restricted and the rules can

[1]In the latter case our notation could be interpreted as if the dependency relation *van* did depend on the dependency relation *met:cmp*. Note that this convention was introduced only for convenience.

not be easily formalized. Moreover, the selection process should be performed fully automatically. Note, that since the dictionaries are bilingual, compositionality should not be conceived as a function of only one language, but interlingually as a function of two languages. On top of that, because the proposed technique claimed to be language independent, it might easily occur that an expression of the source language A is translated compositionally into target language B, while it cannot be translated compositionally into target language C. Therefore, alternative approaches should be considered.

(1) Frequency-based approach One obvious solution is to filter verbal structures on a *frequency basis*. In this case we start out from the hypothesis that the most frequent phenomena of a language should be recorded in a dictionary, regardless of their semantic transparency. As frequently occurring expressions are included in the dictionary, this choice is in accordance with the editing principles of encoding dictionaries.

(2) Default translations Another possibility is to come up with a heuristic to automatically filter *lexicographically uninteresting* complementation patterns relying on default translations[1]. According to our hypothesis an MWE is lexicographically interesting if it has at least one non-default translation, regardless of the existence of a valid default translation. Thus, the notion of default translation should be grasped by means of data present in the parallel corpus: The default translation is the one with the highest translational probability. Regarding the language-dependent nature of the task, it is only after the generation of the corresponding one-token dictionary that we can decide whether a construction is worth keeping or not: If the TL MWE is the result of the default translations of the parts of the SL MWE without altering the corresponding syntactic relations, then the TL translation is not interesting lexicographically.

Example The French expression *poser une question* (raise a question) is translated to Dutch as *vraag stellen* (state a question) in a lexicographically uninteresting way, since the most frequent translation of *poser* is the Dutch verb *stellen* in the corresponding one-token dictionary. As opposed to this, *poser une problème* is translated as *problem hebben*, that is, in a lexicographically interesting way.

During this experiment the first alternative was applied, i.e. verbal structures were kept on a frequency basis. However, the second approach offers a more insightful solution

[1]The basic idea came from Dávid Takács.

to the problem of interlingual compositionality, therefore we intend to use that on the long run. Accordingly, only those automatically attained verbal structures were kept which occur at least 5 times of the relevant side of the parallel corpus. This criterion was met by 289 Dutch verb in 5804 different frame types and 391 French verbs with 5987 various frame types.

Expression	*Dutch*	*French*
Number of verb types	289	391
Number of frame types	5804	5987

Table 7.11: Number of verb and frame types for Dutch and French

7.4.3 The creation of proto-dictionaries

Merging and aligning In the third step verbal structures were recognized in the parallel corpus, merged into one-token units and aligned with GIZA++. In the previous experiment only verbal structures comprising a bound lexical head were considered. The underlying reason was that the input corpus for merging contained neither clause boundary information nor syntactic annotation, thus, the identification of the relevant prepositions turned out to be impossible. This in turn results in mismatch between verbal frames and their occurrences in the corpus.

The objective of the current experiment is to find translation equivalents to *all* the frequent complementation patterns. Whereas in the previous experiment 126 French verbal structure occurred altogether 7805 times and 146 Dutch verbal structure appeared 8029 times in the parallel corpus, in the current experiment we had 170,229 matching French verbal structures and 207,610 matching Dutch verbal frames. After merging the occurrences of verbal frames in the corpus, the expressions were treated as one-token expressions and were aligned with GIZA++ (Och and Ney, 2003).

Proto-dictionaries The proto-dictionaries were based on the frequencies both of the source and target words and the assigned translation probabilities. As the translation probability may take on any value between 0 and 1, at this stage ample unacceptable translation units were available. Thus, we have relied on the same filters as before to get rid of the wrong translation candidates. Table 7.1 is repeated here to demonstrate how an entry of the verbal proto-dictionary looks like.

$Expression_{source}$	$Expression_{target}$	p_{tr}	F_s	F_t
prendre médicament=obj	neem-in genees-middel=obj	0.377	53	32
	gebruik genees-middel=obj	0.102	53	21
	start gebruik=met:cmp met:cmp-van	0.097	53	28
	sta onder-invloed=particle drug=van:cmp	0.05	53	11

Table 7.12: A sample entry from the French-Dutch verbal proto-dictionary.

Parameter setting Based on previous results (Hungarian-Lithuanian, Hungarian-Slovenian) the following general requirements were formulated:

(1) Both source language and target language units has to occur at least 5 times in the parallel corpus. This condition is necessary to have enough data to be able to reliably estimate the translation probabilities.

(2) In the case of more frequent source lemmata a lower translation probability might yield approximately the same proportion of correct or lexicographically acceptable translation units as a higher translation probability in the case of less frequent lemmata.

7.4.4 Evaluation

At first, we have chosen a parameter setting that presumably results in a high proportion of right or lexicographically acceptable translation units. If there exists a parameter setting that yields high precision results, the recall can be increased by the refinement of the parameters.

100 translation candidate pairs were selected among the candidate pairs which occur at least 100 times by setting the minimum translation probability to 0.44.

Figure 7.6 represents the distribution of the French-Dutch verbal structure candidates as a function of the logarithmic frequencies of the source words and the corresponding translation probability values. The translation candidates falling within the area of the black rectangular[1] were evaluated.

Evaluation criteria During the evaluation two different aspects were considered:

[1]Frequency is greater or equals to 100 and the translation probability is at least 0.44

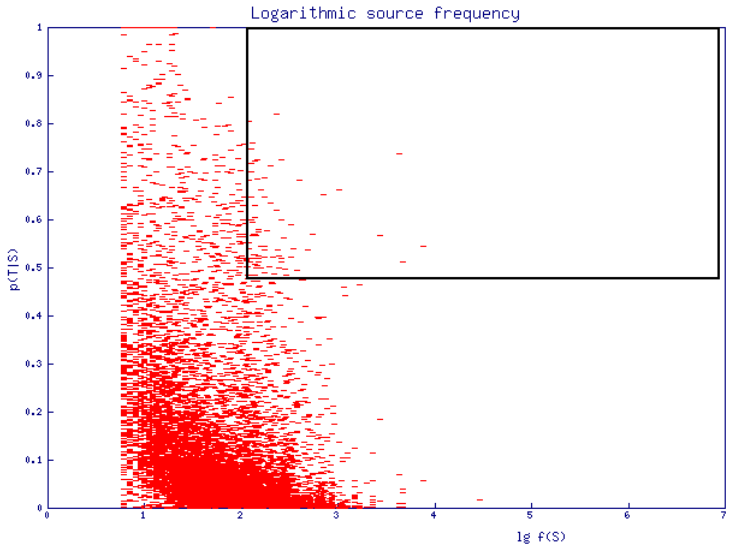


Figure 7.6: The evaluated French-Dutch verbal structures

(1) Whether the translation candidate expression consisted of the right verb.

(2) Whether complete verbal frames were matched.

Results Out of the 100 evaluated examples 46 structures were perfect translations: That is the translation was a complete verbal structure with the right verb. In the rest of the cases the right verbs were extracted as translations, but any or both of the frames was incomplete. In 24 cases both frames were incomplete, while 21 source verbs and 9 target verbs were retrieved with incomplete frames. The evaluated frames contained mostly only one dependent prevalently an object without a lexical head, but verbs with more than one dependent were also aligned, for example:

$Expression_{source}$	$Expression_{target}$	$Englishtranslation$
avoir besoin=obj1 de:cpl	hebben obj1 nodig=predc:ADJ	sy needs sg

Table 7.13: Translation equivalents with one dependent and one lexical head

Accordingly, the French expression *avoir besoin de qc* is assigned the Dutch translation *hebben nodig ob*.

Increasing the proportion of complete frames The results raised the question of how the proportion of complete frames could be increased among translation pairs.

A possible solution might be to omit the "incorrect" frames out of the automatically generated frame list by means of some suitable heuristics. If so, we have to decide automatically what counts as an "incorrect" frame. Since our objective is to build general purpose dictionaries, too long frames[1] should be excluded. The intuition behind this is that long frames occur typically in highly specialized languages (e.g. medical or IT texts), which contradicts to our original objective, i.e. to compile general purpose dictionaries .

The longest French frame is of length 24 and occurs 14 times in the corpus. Because verbal structures were ranked according to their length during the merging process, matching and aligning the too long frames results in sparser data considering the shorter and more general frames.

Figure 7.7 shows the number of French frame types of the same length. It also indicates the number of occurrences for each frame type: the x-axis indicates the frame length while y-axis indicates the corpus frequency of each of the frames.

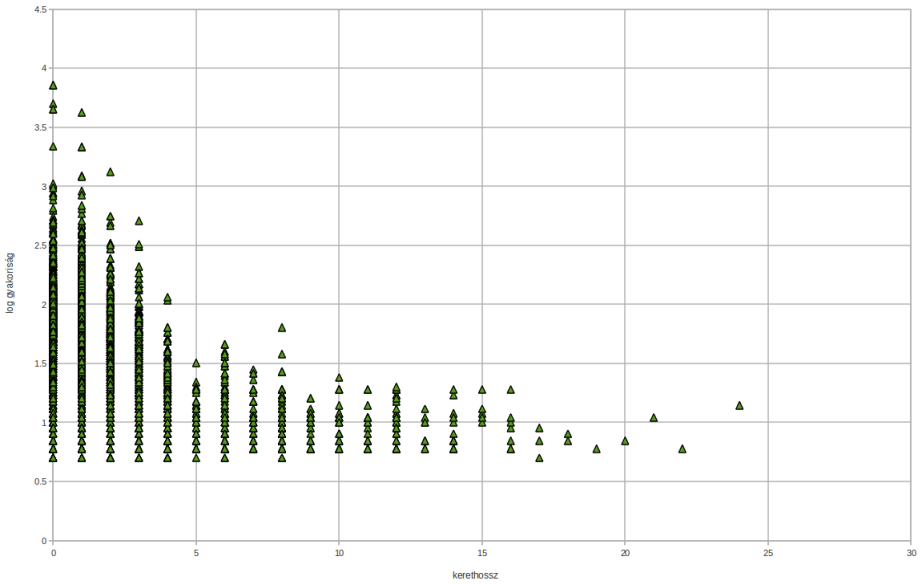


Figure 7.7: The distribution of French frames according to their length and corpus frequency

According to Figure 7.7 among the frame of length 8 some are quite frequent, thus, these frames could be worth including into the workflow, whereas the longer ones are

[1]Recall, that the length of a frame equals to the number of the bound lexical heads plus the number of the dependency labels.

rare enough to be rather specific. Nevertheless, manual evaluation of eight-long frames clearly indicated that these are fairly specific, too. The manual evaluation yielded the conclusion that frames of length 5 or less should be included in the workflow.

Evaluation of verb+object structures To verify the hypothesis that omitting longer frame decreases the proportion of incomplete frames in the proto-dictionary, a second evaluation was performed on a different data-set[1]. The list of verbal structures is made up of frames consisting of a verbal lemma and a direct object, whether it is lexically bound or not. Verbal structures only with different verbal lemmata or different direct objects were kept separately. Otherwise, the originally different verbal structures were subsumed under the same frame type and the corpus frequencies were recounted.

The evaluation was carried out with the same parameter setting as before. In this case 60% (as opposed to 46%) of the evaluated source lemmata had at least one right and complete translation. 31% of the source lemmata was assigned only incomplete equivalents, while 9% of the source lemmata had only wrong translations. Our hypothesis proved to be correct since the proportion of the right but incomplete translations dropped significantly in the second case.

The characterization of a frequent verb The evaluation was performed from another perspective, too. We have also investigated to what extent a frequently occurring verb is characterized in terms of the number of frames in which it shows up and the number of translations which are assigned to it.

For doing so, the French verb *mettre* was given a closer look. 65 different frames were extracted for the 5706 corpus occurrences of *mettre*. We have selected those translation candidates where both the source and target language expressions occur at least five times and the translation probability is higher or equal to 0.02. In Figure 7.8 the x-axis indicates the logarithmic frequency of the source expression while y-axis indicates the translation probability. The distribution of translation candidate pairs headed by *mettre* in the source language fall within the black rectangle.

The 65 French verbal structures is assigned 132 Dutch verbal frames so that they form 151 translation pairs. The 132 Dutch frames occur 5611 times in the corpus. The French-Dutch translation pairs were manually evaluated. A yes-or-no decision was made on the correctness of the translation. The translation was considered to be right, if there was at least one context in the corpus where the given Dutch frame was a correct translation. Incomplete frames were accepted if they could be completed based

[1]Dávid Takács evaluated the verb+object structures.

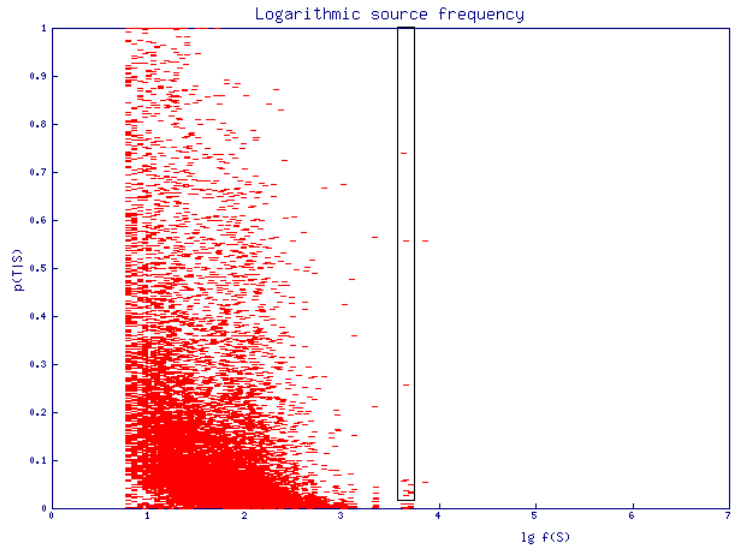


Figure 7.8: The distribution of the French-Dutch verbal structure candidates comprising the French verb *mettre*

on the concordance. 62% of the 151 translation candidates proved to be correct.

Incomplete French and Dutch frames were both marked. Only 10 out of the 65 *mettre*-frames were not assigned any correct translations, at all. That is, 55 frames—85% of the *mettre*-frames—was assigned one or more correct translation.

Discussion The results clearly indicate that in the presence of ample empirical data the proposed method is able to supply information regarding the verbal structures and their translations to be included in the proto-dictionaries. However, several translation pairs are made up of incomplete frames. In such cases concordance have to be exploited to include complete verbal phrases in the proto-dictionaries. However, it also turned out that a drastic decrease in the length of verbal frames increases the proportion of complete frames in the proto-dictionary.

The problem of data sparsity Generally speaking, the difficulty that should be addressed in the long-run is the data sparsity problem. Sparse data is a consequence of diverse data, that is the too detailed characterization of verbal frames results in too many frames. In this case, less corpus data is available for each of the frames during the alignment process.

Decreasing the number of frames types Therefore, it is necessary to decrease the number of different frames. This can be done in more then one way: First, we

may impose an upper limit to the length of verbal structures. Manual evaluation of the frames showed that including maximum 5-long frames in the verb frame list is a reasonable choice. Secondly, too detailed syntactic characterization of the verbs yields too many frames. Thus, the feature space may be also reduced. This thread of research is part of our future task.

7.5 Conclusion

General framework Our basic motivation was to investigate to what extent it is possible to handle multi-word expressions in the proposed framework, that is, to retrieve parallel verbal expressions from parallel corpora completely automatically. A preprocessing module was introduced (1) to detect monolingual verbal expressions in each side of the parallel corpus (2) to merge each occurrence of the verbal expressions so that they could be treated as one-token units in the following stages of the retrieval process. Instead of extracting aligned phrases directly from word alignment, as it is usual in the statistical machine translation literature, a preprocessing module was introduced, so that we could handle non-adjacent and adjacent dependencies alike.

Workflow

(1) Each input corpus was converted into partial dependencies, as this is the required input format for the verb structure extraction algorithm.

(2) The list of verbal structures was generated for each language.

(3) Detection of the longest matching verbal expressions in each side of the parallel corpus.

(4) Merging each occurrence of the verbal expressions so that they could be treated as one-token units in the following stages.

(5) Performing word alignment as in the case of one-token expressions.

Shallow parsed parallel corpus The objective of the first experiment was to prove that such an approach is able to extract parallel verbal structures. Noun phrases and clause boundaries were detected by means of rather simple and approximative rules. Only a limited set of verbs was considered. The evaluation yielded the conclusion that although the results are promising both precision and recall should be increased.

Deep parsed parallel corpus To improve both precision and recall the second experiment was performed on the basis of a deep parsed parallel corpus. This time all verbs occurring in the parallel corpus were considered. The evaluation showed that there were "too many" different frame types. This resulted in sparse data during the alignment, thus the number of frame types had to be decreased. First, too long frames were omitted, as they are typical of highly specialized languages. Manual evaluation showed that maximally frames of length 5 should be included into the proto-dictionary. Secondly, the deep syntactic analysis present in the corpus makes possible a more intricate description of verbal frames. This sometimes yields non-existent verbals structures that could be subsumed under a more general frame. Thirdly, the syntactic analysis may be ambiguous: The same phenomena might be analyzed in different ways (cf. 7.9), thus uselessly producing two different frame types.

8

The Dictionary Query System

8.1 Introduction

In the previous chapters it was argued that word alignment on parallel corpora is particularly useful for lexicographic purposes. Beside its cost-efficiency the data-oriented nature of the proposed method is worth mentioning here. Nevertheless, some difficulties also have arisen, for example, it is quite hard to produce completely clean proto-dictionaries of appropriate size, at least for medium-density languages. As a consequence, at the first stage of the research we confined ourselves to providing lexicographers with the most suitable data to facilitate their work instead of creating full-fledged dictionaries for end-users.

Extending the scope of users One natural improvement of our work would be to extend the usability of the generated data, so that it could furnish information even for end-users. Thus, a dictionary query system has been designed and implemented which is able to compensate for the disadvantages and exploit the advantages of the proposed method.

Macrostructure and microstructure Regarding the organization of a dictionary macro- and microstructures can be distinguished. The macrostructure is the headword list the dictionary is made up of, while the microstructure is the structure of the entries. The macrostructures and microstructures of dictionaries widely differ as a function of the target group the dictionary is designed for.

Macro- and microstructures of proto-dictionaries The macrostructure of the

dictionary is determined by the size and domain of the parallel corpus. The microstructure of the dictionary is partially determined by the annotation present in the parallel corpus and the recognized multi-word expressions. The dictionary query system (DQS) also affects both macro- and microstructure. In the present chapter it will be discussed how the dictionary query system makes accessible the macro- and microstructures of proto-dictionaries.

8.2 What type of information should be ideally included in the dictionary?

Before presenting the DQS we will shortly discuss what type of information should be ideally included in a dictionary according to Atkins and Rundell (2008). Besides, we will also consider whether the corresponding piece of information can be treated in the proposed framework.

8.2.1 Lemma headword

Pronunciation This type of information definitively does not form part of the proto-dictionaries.

Variant form Variant form shows a slight variation of the headword lemma, such as an alternative spelling (e.g. *emphasize* or *emphasise*). Alternatives form might show up in the proto-dictionaries if the texts of the parallel corpus comprises these alternative word forms, but they will not appear as alternatives within a headword.

Frequency In recent corpus-based dictionaries frequency information is supplied to give hint on the importance of each headword to the user. Lemma frequencies are estimated on the basis of large corpora. In the case of proto-dictionaries frequency information plays a crucial role, not only in the case of headwords (source lemma frequencies), but also in the case of the translation candidates (translation lemma frequencies). Moreover, proto-dictionaries also indicate the frequencies of the translation candidates in terms of translation probabilities.

Inflected form Dictionaries consist of information on the inflected forms of headwords, especially in the case of irregular inflection. As most taggers and syntactic analyzers are aware of irregular word forms and link them to their headword, this type

of information is accessible in the proto-dictionaries. However, at the present stage of implementation inflectional paradigms are not queriable in the DQS.

8.2.2 Meaning and translation in bilingual dictionaries

Although under different names, Atkins and Rundell (2008) list the same translational categories as we discussed in Chapter 3. There we also accepted the view that three out of the four categories form continuous scales rather than being distinct types of translations. Namely, translations are situated somewhere on the cognitive equivalent—translational equivalent scale. Translations also exhibit different amount of explanatory power: The one end of the scale is being merely a contextual translation, which perfectly fits into the given context, but does not describe the source word expression, whereas the other end of the scale is constituted by definitions that provide an exhaustive description on the meaning of source expressions but may not be used to produce idiomatically correct translations in the target language. In Chapter 4 we saw how each of these types are treated within the given framework. In what follows, we briefly consider the type of translation equivalents as described in Atkins and Rundell (2008).

Direct translation Direct translations correspond to cognitive equivalents: in Chapter 4 we saw that the corpus frequency data along with translation probabilities provide a strong hint where a translation candidate is situated on the direct translation—translational equivalent continuum. For instance, if the translational probability of a translation pair candidate equals to one and corresponding frequencies are the same the SL and TL words are likely to be cognitive equivalents (cf. Section 4.3.2).

Contextual translation As cognitive equivalency is rare and is constrained to certain semantic classes, the majority of SL words could be assigned only a more or less satisfying direct translation. Recall that direct translations may be conceived of in terms of contextual translations: Direct translations are those contextual translations that fit every context. In this case direct and contextual translations form a scale, and a measure may be assigned to each of the translation candidates where they are situated on this scale.

Proto-dictionaries indicate this type of information in terms of SL and TL lemma frequencies and translational probabilities.

Near-equivalent In the case of near-equivalents, the SL and TL items are often culturally equivalent. Near-equivalents are used when there is no real TL equivalent of the

SL headword or phrase. Thus, near-equivalency is the same as functional equivalency in Section 3.2.2: In the case of functional equivalency the word-level lexical meaning of an expression does not correspond directly to the word-level meaning, but since the expressions play similar roles in a conversation they are said to be functionally equivalent. For instance, the English expression *"A for Abel"* is a functional equivalent of the French expression *"A comme André"*. Actually, because it does not rely on lexical meanings, the proposed framework does not distinguish functional equivalents from other type of translations, rather it distills meaning from parallel distributions and hence near-equivalence (or functional equivalence) can be conceived of as a multi-word contextual equivalence.

However, although the proposed method rejects the notion of lexical meaning, namely, that words' meanings are intersubjective entities, readily available for native speakers (cf. Section 2.3.3), it still undertakes to predict if a translation of an MWE is lexicographically interesting or no (cf. Sections 7.4.2 and 8.3.2).

TL gloss As we saw in Section 3.2.2, TL glosses or definitions are at one end of the explanatory power—translational equivalency scale. Definitions comes into play when there is no direct translations or near-equivalents. A TL gloss explains the meaning of the SL expression to the TL user, but is of no use for encoding purposes, or at least it cannot produce idiomatically correct translation. As word alignment is based on translated texts, giving translation by means of TL glosses is no option for us. However, though it does not form part of the present research, explanatory power may be measured in terms of retrieved synonyms (cf. 4.3.2)

8.2.3 Sense indicators

As Atkins and Rundell (2008) puts it:

> A 'sense indicator' is a component designed to lead people as quickly as possible to the right part of the entry. (They are therefore a special kind of navigation aid.) Sense indicators are rare in monolingual dictionaries for native speakers, who can see from the definitions and examples the various senses of the headword. This is not the case, however, for learners of the language, and the sense indicator is an essential part of entries for learners. There are two main types of sense indicator: specifiers (in monolingual and bilingual dictionaries) and collocators (mainly in bilinguals). (p. 214)

Domain labels will be also discussed in this section, as domain labels—if chosen appropriately—may indicate the relevant sense of the headword.

Domain label Domain label specifies the typical domains in which a translation occurs. Since the genre of texts is preserved, this information is easily accessible. However, as the basic objective of the present research is the automatic compilation of *general domain* bilingual dictionaries, we strove to gather general domain texts, mainly of literature. Thus, domain labels are not that informative in our case as it could be if more specific domain texts were included in the parallel corpus.

Specifiers and collocators Specifiers mark sense distinctions in the SL side. They may contain many different types of information, such as superordinates, synonyms, typical modifiers, paraphrases, and so on. Collocators are used to help users to find the best translation. A collocator is a word representing a group of words belonging to the same word class and similar in meaning. Both specifiers and collocators are words from the language of the encoding user, i.e. the source language. For instance, in Figure 8.1 the descriptions of subsenses between brackets are specifiers, whereas indicators of typical subjects are collocators appearing between square braquets.

Clear (verb) I. (become transparent, unclouded)[liquid, sky] s'éclaircir;
II. (disappear)[smoke, fog, cloud] se dissiper;

Figure 8.1: Specifiers and collocators of the headword *clear* in the Oxford-Hachette French Dictionary (Ormal-Grenon and Pomier, 2001)

Since the various SL lemma senses of proto-dictionaries are not specified before the assignment of translations, but through translations, specifiers are not provided.

Considering collocators, typical collocations may serve as collocators. However, as at the present stage of research the detection of parallel collocates has not been completed, this feature is not accessible yet. The only exceptions are French and Dutch verbal structures, where typical objects may serve as collocators.

8.2.4 Grammar

Grammar information should be included in every bilingual dictionary, for it indicates how the TL word or expression should be combined with other words to form grammatical or rather idiomatically correct TL sentences.

Word-class marker As part-of-speech categories provide general hints about how words are used and combined with other words they appear in every dictionary.

In the case of proto-dictionaries part-of-speech categories are indicated as subscript of the given lemma. However, it should be kept in mind that part-of-speech categories were tagged completely automatically by means of morphological taggers or syntactic analyzers, thus, wrong part-of-speech categories may occur.

Construction Providing the set of constructions in which a word may appear makes an idiomatically more correct language use possible than mere part-of-speech information. Thus, a dictionary must contain all the constructions the learner must know in order to use the word flexibly and fluently for four open word classes (verbs, nouns, adjectives and adverbs). The constructions often reflect corpus evidence.

Because the detection of collocations with pre-defined syntactic patterns is a common practice in language technology, the inclusion of such constructions into the proto-dictionaries probably is rather straightforward. At the present stage of research only constructions with syntactic patterns were considered: *Verb + object* structures are queriable in the Dutch-French (and vv.) dictionaries. *Adjective + noun*, *noun + noun* and *adverb + verb* structures were included in the Hungarian-English (and vv.), Slovenian-Hungarian (and vv.) and Lithuanian-Hungarian (and vv.) dictionaries without any specialized querying facility. These constructions were extracted following some standard techniques[1], but no detailed evaluation has been performed.

Grammar label Grammar labels are a set of complementing information that is needed beyond word-class markers and constructions to produce grammatical utterances. For English nouns countability and proprer noun are commonly used grammar labels. As for verbs, their semantic type might be indicated (activity, accomplishment, achievement, or stative), and in the case of adjectives their predicative or attributive use is usually indicated.

Grammar labels are not available in the proto-dictionaries. Such information in the parallel corpus may be only the result of deep syntactic analysis. Unfortunately, language analysis tools providing such detailed information are usually not available in the case of less resourced languages.

[1]For collocation extraction the UCS tool-kit (Evert, 2004) was used, the candidates were filtered on the basis of both MI-score and Z-score.

8.2.5 Contexts

Multiword expressions Although Atkins and Rundell (2008) discusses multiword expressions (eg. idioms, collocations, compounds, phrasal verbs) and constructions separately we do not see the reason for differentiating between them. Thus, we do not discuss them separately.

Examples Examples form important part of learners' bilingual dictionaries. However, primarily for the reason of limited content the printed versions of dictionaries may contain only a constrained set of examples.

As the problem of limited space does not apply in the case of electronic dictionaries, ample example sentences could be included in such dictionaries. Moreover, in the proposed framework all the example sentences come from corpus texts. Nevertheless, the great amount of examples might render it difficult to the end-users to select the relevant facts about translation.

8.2.6 Vocabulary types

Vocabulary type might be determined by domain labels. As it was previously mentioned, text domains are encoded in the corpus annotation. The proportion of occurrences of translations in different domains are indicated in DQS. However, other types of vocabulary indicators common in usual dictionaries are not present in the database, such as register (formal, informal, very informal), style (literary, newspaper), time (obsolete) , attitude (derogative, pejorative, appreciative) and meaning types (literal, metaphorical). Thus, vocabulary types, except for domain labels, are not indicated in the proto-dictionaries.

8.2.7 Usage

Subject-oriented usage note The basic idea behind subject-oriented usage notes is that repeated information should be avoided in dictionaries. So the specific piece of information is described in a subject-oriented usage note and is cross-referenced from all the relevant headwords. For instance, in the Oxford-Hachette French Dictionary Ormal-Grenon and Pomier (2001) concerns how to translate into French various constructions containing names of countries and continents.

Local usage note Local usage notes specify uncommon or even wrong usages of the headword.

Neither subject-oriented nor local usage notes do form part of the proto-dictionaries. On the one hand, being electronic dictionaries repeated information does not pose a problem for them. On the other hand, proto-dictionaries are made up of only existing language phenomena, thus cannot contain wrong usages that never occur. Nevertheless, uncommon usages could be included at the cost of a lower precision (cf. 8.4.2).

8.2.8 Other lemmas

Secondary headword Secondary headwords are subsumed under the headword: Lemmata given secondary headword status are mainly multi-word expressions, such us compounds or phrasal verbs.

In proto-dictionaries secondary headwords appear in the French-Dutch and Dutch-French proto-dictionaries, when in the case of verb+object structures all the relevant constructions are listed both under the verbal and nominal headword that make up the construction (cf. 8.3.2). This look-up method has to be extended to the other multi-word expressions, too.

Cross-reference As Atkins and Rundell (2008) claim:

> The cross-reference component tells the user that more information relating to the current headword will be found at the other entry, this can be done in a number of different ways, both directly and indirectly. [...] Every dictionary has its own palette of admissible ways of cross-referring from one entry to another. (p. 238.)

One important novelty of proto-dictionaries is that the cross-referencing system is extended to the reverse dictionaries, too. Therefore, if we are not sure about the exact meaning of a translation candidate, we can easily get to know more about it by clicking on the translation candidate and thus getting the translations of the translation.

8.2.9 Discussion

We claim that as opposed to usual dictionaries, in the proposed framework new information may be presented in new ways. In the present section we gave an overview of what type of data should be ideally included in a dictionary based on Atkins and Rundell (2008). Some of these are obviously not accessible for the proposed approach and others—may be less straightforward for traditional and corpus-based lexicography—are readily available (frequency ranks among translations, ample example sentences, the proportion of translations over various text genres, reversibility, etc.). In other cases the required information is not readily available but certain heuristics may be developed to give hints on it (e.g. explanatory power, cross-lingual semantic relation).

Furthermore, some features were also implemented that are not familiar to the ordinary dictionary users, such as hints on cross-lingual usability or the feature of immediate reversibility. Being parts of the Dictionary Query System, these characteristics will be described in the next section, which is made up of two subsections. The first presents the Dictionary Browser (8.3) while the second discusses the Cut Board (8.4).

8.3 DQS: Dictionary Browser

As earlier has been mentioned, the proposed method has several benefits compared to more traditional approaches:

(1) A parallel corpus of appropriate size guarantees that the most relevant translations be included in the dictionary.

(2) Based on the translational probabilities it is possible to rank translation candidates ensuring that the most likely used translation variants go first within an entry.

(3) All the relevant example sentences from the parallel corpora are easily accessible facilitating the selection of the most appropriate translations from possible translation candidates.

(4) The reversible dictionary can be generated easily, thus, it is readily available.

Accordingly, the Dictionary Query System presents some novel features: The innovative representation of the generated bilingual information helps to find the best translation

for a specific user in the Dictionary Browser Window[1]. The Section 8.3.1 discusses the basic features of the Dictionary Browser, which is extended with some additional facilities to query two-token expressions described in Section 8.3.2 in more detail.

8.3.1 One-token units

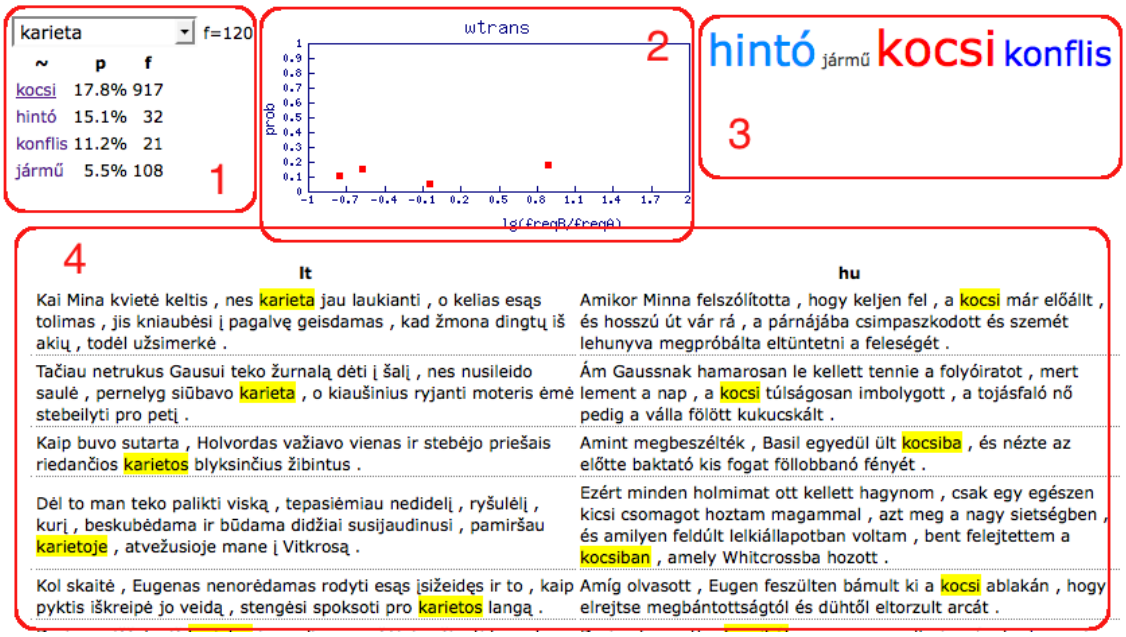


Figure 8.2: The Dictionary Browser

As Figure 8.2 illustrates, the Dictionary Browser displays four different types of information.

(1) List of the translation candidates ranked by their translation probabilities. This guarantees that most often used translations come first in the list (from top to bottom). Absolute corpus frequencies are also displayed.

[1]The layout of the Dictionary Browser and the underlying MySQL database scheme was implemented by Dávid Takács. The author produced the required input data.

(2) A plot displaying the distribution of the possible translations of the source word according to translation probability and the ratio of corpus frequency between the source word and the corresponding translation candidate.

(3) Word cloud reflecting the scope of use of the TL lemmata compared to that of the SL lemmata. Words in the word cloud vary in two ways.

First, their *size* depends on their translation probabilities: the higher the probability of the target word, the bigger the font size is.

Secondly, *colours* are assigned to target words according to their frequency ratios relative to the source word: less frequent target words are cool-coloured (dark blue and light blue) while more frequent target words are warm-coloured (red, orange). Target words with a frequency close to that of the source word get gray colour.

According to our hypothesis the frequency ratios provide the user with hints about the scope of use of the TL lemmata compared to that of the SL lemmata, which might be particularly important when creating texts in a foreign language. For instance, the Lithuanian lemma *karieta* has four Hungarian eqivalents: "kocsi" (word with general meaning, e.g. 'car', 'railway wagon', 'horse-drown vehicle'), "hintó" ('carriage'), "konflis" ('a horse-drawn vehicle for public hire'), "járműű" ('vehicle'). The various colours of the candidates indicate whether the scope of use of the TL lemma is more specific or more general than that of the SL word: The red colour of "kocsi" marks that the target word may show up in a greater set of contexts than the source word. Conversely, the dark blue colour of "konflis" shows that the target word appear in a more restricted set of contexts. Probably, the proportion of the number of contexts in which the SL and TL lemmata may occur provide us with hints about the cross-lingual semantic relation of the two lemmata. This hypothesis should be tested in the future which makes part of our future work.

(4) Provided example sentences with the source and target words highlighted, displayed by clicking one of the translation candidates.

Moreover, in the bottom of the window appears the number of translation pairs in each text genre present in the parallel corpora (literature, law and software documentation.)

8.3.2 Two-token expressions

Verbal structures of the form *verb + object* were also uploaded into the query system. The extraction process and the evaluation of parallel *verb + object* structures was described on page 182. As a result, both French and Dutch verbal structures can be queried. Please note, that in this case the verbal structures are constrained to verb + object structures, thus may not contain other essential parts of the verbal structure, such as determiners or prepositions. Yet, they are of great use when searching for multi-word expressions and the correct form of the verbal structures can be inferred from the example sentences provided.

Figure 8.3: Querying verb + object structures based on objects

As Figure 8.3 shows, the French word *fiche* "sheet" and "plug" has two different Dutch translations: *stekker* "pin contact, male connector" and *fiche* "sheet". Moreover, the dictionary browser also displays the verbs of which the French noun *fiche* is a frequent object complement. The relevant translations and the concordance can be displayed by clicking any of them.

All the retrieved bigrams appear in the Dictionary Browser, even if there is no translation for them with the actual parameter setting. *Verb + object* structures without translations appear struck through. Bigrams are queriable both according to verbs and objects

Lexicographically interesting translations In the case of the French-Dutch *verb + object* structures (and vv.) the proto-dictionaries also indicate whether the translation is lexicographically interesting or not (cf. page 177.) Lexicographically interesting translations are marked with an exclamation mark. For instance: *poser une question* (raise a question) is translated as *vraag stellen* (state a question) in a lexicographically uninteresting way, since the most frequent translation of *poser* is the Dutch verb *stellen*. As opposed to this, *poser une problème* is translated as *problem hebben*, that is, in a lexicographically interesting way.

Figure 8.4: Querying verb + object structures based on verbs

8.4 DQS: Cut Board

Recall that according to Atkins and Rundell (2008):

> A new dictionary designed for electronic as well as print publication-—a rare bird in the reference publishing world, because of the cost involved – opens exciting possibilities of totally new information presented in new ways. Key features of such a dictionary will be 'customizability' and 'personalizability': in this model, the 'dictionary' is essentially a collection of lexical resources (possibly multilingual), which users can select from and configure according to their needs. (p. 239.)

The automatically constructed bilingual databases extended with the Dictionary Query System at least partially corresponds to what the dictionary of the future is, as one of the key features is *customizability*. In fact, the proto-dictionaries may be tailored to suit one's needs, and this feature makes proto-dictionaries useful for end-users, too.

8.4.1 Fine-tuning the parameters

In Chapters 6 and 7 we saw that the coverage of the results proved to be a serious bottleneck for the selected approach. As it was pointed out several times a possible solution to this problem could be the refinement of the parameters. Two more evaluation rounds were carried out to prove this hypothesis.

Secondly, translation synonymy is rare in general language (cf. 3.2), thus other semantic relations, such as hyponymy or hyperonymy were also considered.

Parameters We considered three parameters when searching for the best translations: The translational probability, the SL lemma frequency and the TL lemma frequency (p_{tr}, F_s and F_t, respectively). The lemma frequency had to be taken into account for at least two reasons.

(1) A minimal amount of data was necessary for the word alignment algorithm to be able to estimate the translational probability.

(2) In the case of rarely used TL lemmas the alignment algorithm might assign high translational probabilities to wrong translation pairs if the source lemma occurs frequently in the corpus and both members of the lemma pair recurrently show up in aligned units.

Results of the first evaluation Results of the first evaluation showed that translation pairs with relatively low frequency and with a relatively high translational probability yielded cc. 85% lexicographically acceptable translation pairs in the case of the Hungarian-Lithuanian proto-dictionary. Although the precision was rather convincing, it has also turned out that the size of the resulting proto-dictionaries might be a serious bottleneck of the method (Héja, 2010). Whereas the targeted size of the dictionaries is between 15,000 and 25,000 entries, the proto-dictionaries comprised only 4,025 Hungarian-Lithuanian and 7,007 French-Dutch translation candidates above the predefined threshold values. Accordingly, the coverage of the proto-dictionaries had to be augmented. Note that a repeated experiment on an increased size of Hungarian-Lithuanian parallel corpus, which comprised 262,423 TUs yielded only 5,521 translation pairs with the suitable parameters[1].

[1]The original Hungarian-Lithuanian parallel corpus consisted of 147,158 TUs. This means that a nearly double-sized corpus provided us only 37% more translation candidates above the predefined threshold values.

Fine-tuning the parameters According to our hypothesis in the case of more frequent source lemmata even lower values of translation probability might yield the same result in terms of precision as in the case of lower frequency source lemmata. Hence, the different evaluation domains need to be determined as a function of source lemma frequency. That is:

(1) The refinement of the parameters yields approximately the same proportion of lexicographically acceptable translation units as the basic parameter setting.

(2) The refinement of the parameters ensures a greater coverage.

Results of the refined evaluation[1] Detailed evaluation of the French-Dutch translation candidates confirmed the first part of our hypothesis. The evaluation was based on the same evaluation guide as the first evaluation of the Hungarian-Lithuanian proto-dictionary (cf. Figure 6.2) and it used the same classification. Accordingly, lexicographically acceptable and lexicographically unacceptable translation pairs were distinguished.

We have chosen a parameter setting in accordance with (1) (see Table 8.1). 6934 French-Dutch translation candidates met the given conditions. 10 % of the relevant pairs was manually evaluated. The results are presented in Table 8.1. For instance, the first evaluation range (1^{st} row of Table 8.1) comprised translation candidates where the source lemma occurs at least 10 times and at most 20 times in the parallel corpus. With these parameters only those pairs were considered where the translation probability was at least 0.4. As the 1^{st} and 2^{nd} rows of Table 8.1 show, using different p_{tr} values as cut-off parameters give similar results (87%), if the two source lemma frequencies also differ.

F_s	p_{tr}	Acceptable
$10 \leq LF \leq 20$	p ≥ 0.4	83%
$100 \leq LF \leq 200$	p ≥ 0.06	87%
$500 \leq LF$	p ≥ 0.02	87.5%

Table 8.1: Evaluation results of the refined French-Dutch proto-dictionary.

The manual evaluation of the Hungarian-Lithuanian translation candidates yielded the

[1]The French-Dutch translation pairs were evaluated by Annemieke Hoorntje and Piroska Lendvai, and the Hungarian-Lithuanian translation pairs were evaluated by Beatrix Tölgyesi.

same result. We have used this proto-dictionary to confirm the 2nd part of our hypothesis, i.e. that the refinement of these parameters may increase the size of the proto-dictionary. Table 8.2 presents the results. *Expected* refers to the expected number of lexicographically acceptable translation units, estimated on the basis of the evaluation sample. 800 translation units were evaluated altogether, 200 from each evaluation domain. As Table 8.2 shows, it is possible to increase the size of the dictionary through refining the parameters: With fine-tuned parameters the estimated number of acceptable translation units was 13,605 instead of 5,521, which is quite close to the targeted size of the proto-dictionaries.

F_s	p_{tr}	Translation units	Evaluated	Acceptable	Expected
$5 \leq$ LF < 30	p > 0.3	6713	200	64%	4,296
$30 \leq$ LF < 90	p > 0.1	5181	200	80%	4,144
$90 \leq$ LF < 300	p > 0.07	3401	200	89%	3,026
$300 \leq$ LF	p > 0.04	2725	200	79%	2,139
					13,605

Table 8.2: Evaluation results of the refined Hungarian-Lithuanian proto-dictionary.

8.4.2 Trade-off between precision and recall

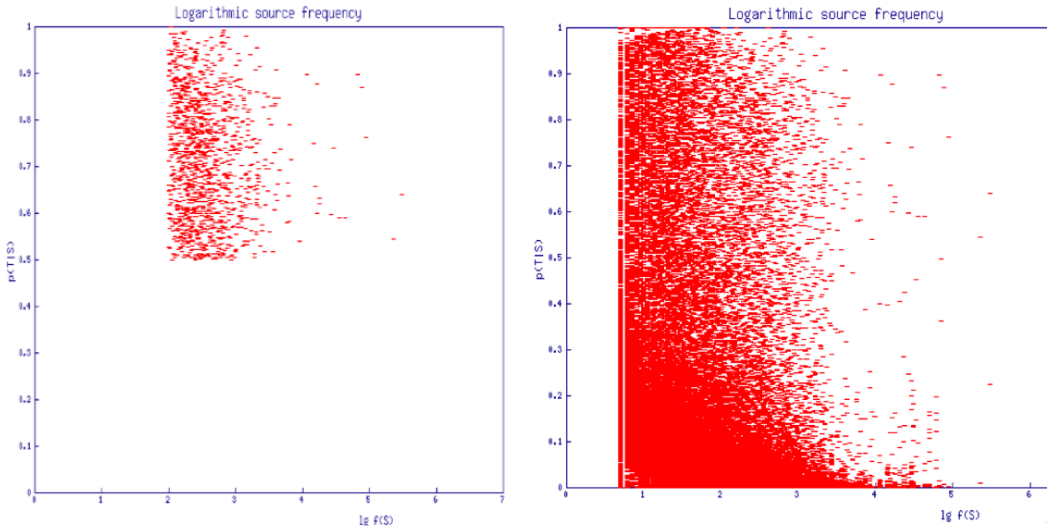
Another alternative to increase the size of the dictionaries is to include more translation pairs even at the cost of lower precision.

Usage scenarios This leads us to the notion of usage scenarios: We think that various parameter settings match well different user needs. For instance, when the settings are strict, that is, the minimal frequencies and probabilities are set high, the dictionary will contain less translation pairs, resulting in high precision and relatively low coverage, with only the most frequently used words and their most frequent translations. Such a dictionary is especially useful for a *novice language learner*. On the other end of the scale *professional translators* are able to judge whether a translation is correct or not. They might be rather interested in special uses of words, lexicographically acceptable but not perfect translation units, and more subtle cross-language semantic relations. At the same time, they can easily catch wrong translations which are the side-effect of the method looking at the concordance provided along with the translation pairs. This

kind of work may be supported by a proto-dictionary with increased recall even at the cost of a lower precision. Thus, the Dictionary Query System should be customizable to support various user needs.

8.4.3 Customization: Cut Board

Proto-dictionaries may be customized on the Cut Board[1]. The Cut Board is an interface where the number of the translation candidates can be determined as a function of the parameters used for filtering (SL and TL lemma frequency, translation probability and quotient of the TL and SL lemma frequencies).



(a) A few translation candidates **(b)** Ample translation candidates

Figure 8.5: Different parameter settings

Strict parameter setting In Figure 8.5a the frequency of the source lemma and that of the target lemma has to be at least 100. The threshold for translation probability is set to 0.5. The resulting French-Dutch proto-dictionary included 1352 translation pairs.

Relaxed parameter setting Figure 8.5b shows a relaxed parameter setting resulting in a much greater proto-dictionary. The frequency of the source lemma and that of the target lemma has to be greater or equal to 5. Translation probability has to exceed

[1]The Cut Board was implemented by Dávid Takács.

or be equal to 0.001. With these parameters there are 104,039 translation pairs in the French-Dutch proto-dictionary.

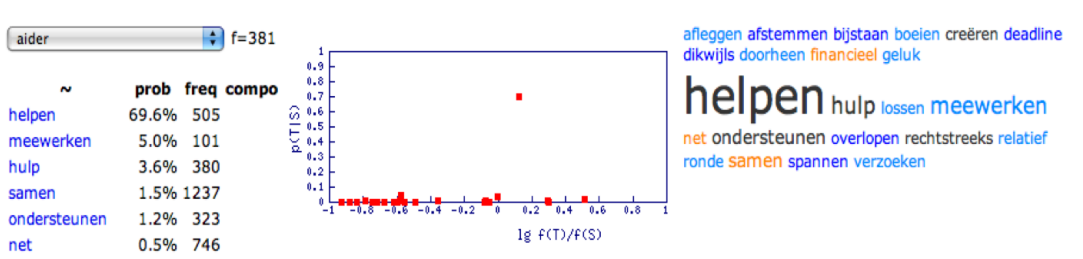


Figure 8.6: Translation candidates: Relaxed parameter setting

Consequently, the Cut Board determines which translation candidates appear in the Dictionary Browser. Then, if the French verb *aider* is queried in the Dictionary Browser, the different proto-dictionaries (i.e. those depicted in Figure 8.5b and 8.5a) include different set of translation pairs. In the first case, the Dictionary Browser comprising 23 translation candidates, some of which are wrong ones (cf. Figure 8.6), while in the second case, the Dictionary Browser comprising a sole and right Dutch translation equivalent (cf. Figure 8.7).

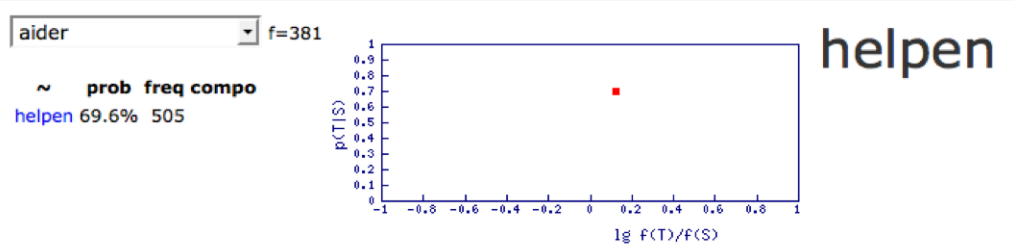


Figure 8.7: One translation candidate: Strict parameter setting

Translation probability as a function of frequency In the previous examples the translation probability was cut only at one certain value. That is, one important finding of the evaluation was not considered, namely that higher frequency source lemmata yield more lexicographically acceptable translations even with lower translation probabilities. To exploit the potential in that observation and to be able to maximize the coverage of the proto-dictionaries, the user should be able to set various translation probabilities as a function of the source lemma frequency. These parameters can be set in the following interface:

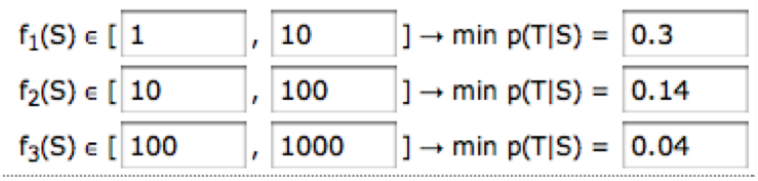


Figure 8.8: Parameter setting as a function of the source lemma frequency

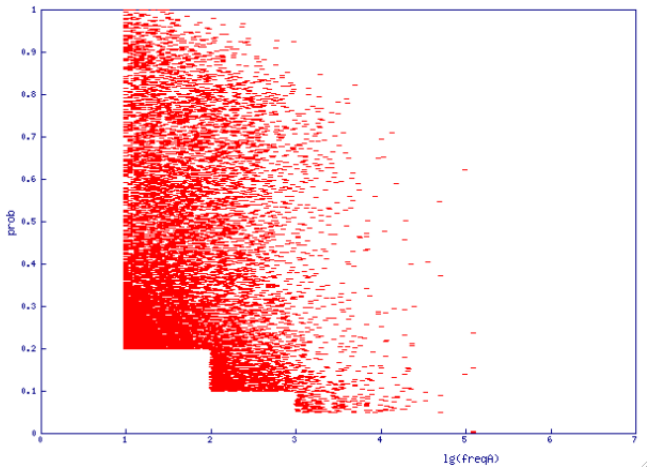


Figure 8.9: The customized dictionary: The distribution of the Lithuanian-Hungarian translation candidates.

The distribution of the translation pairs corresponding to the parameters in 8.8 is presented in Figure 8.9. Figure 8.9 visualizes the distribution of the logarithmic frequency of the source words (x-axis) and the relevant translation probability (y-axis) for each word pair, selected by the given custom criteria. With these parameters there are 22,545 translation pairs in the French-Dutch dictionary.

Filtering translation pairs with frequencies of different orders Recall that measuring translation relation in terms of conditional probabilities results in high probability values if the frequency of the SL word is high while the frequency of the target word is low and both words recurrently show up in the same sentences. This phenomena was illustrated in Figure 6.7. The Cut Board makes it possible to get rid of such translation pairs. Accordingly, Figure 8.10 visualizes the distribution of the logarithmic frequency ratio of the target and source words (x-axis) and the corresponding translation probability for each word pair (y-axis), selected by the given custom criteria.

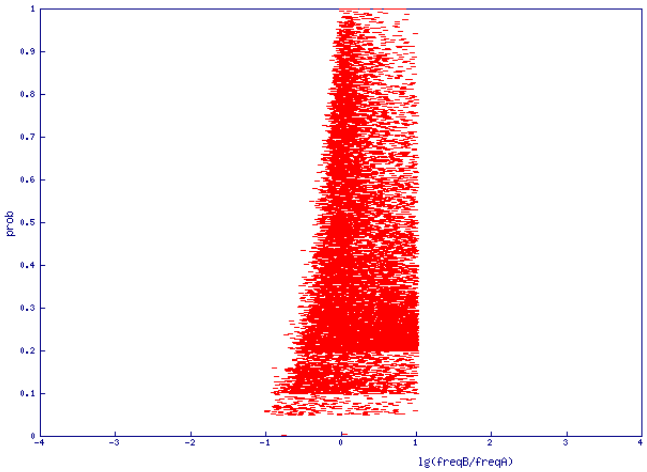


Figure 8.10: The customized dictionary: The distribution of the Lithuanian-Hungarian translation candidates.

Customization of proto-dictionaries Proto-dictionaries are customizable by the following criteria:

(1) Maximum and minimum ratio of the relative frequencies of the source and target words (left and right boundary on Plot 8.9).

(2) Overall minimum frequency of either the source and the target words (left boundary on Plot 8.10).

(3) Overall minimum translation probability (bottom boundary on both plots).

(4) Several more cut off intervals can be defined in the space represented by Plot 8.10: Word pairs falling in rectangles given by their left, right and top boundaries are cut off.

After submitting the given parameters the charts are refreshed giving feedback to the user and the parameters are stored for the session, i. e. the dictionary page shows only word pairs fitting the selected criteria.

8.4.4 Implementation

The DQS is available at `http://efnilex.efnil.org`. It is based on the LAMP web architecture. We use a relational database to store all the data: The multilingual corpus

text, sentences and their translations, the word forms and lemmata and all the relations between them. The implementation of such a data structure and the formulation of the queries is straightforward and efficient. The data displayed in the dictionary browser as well as the distributional dataset presented on the charts is selected on-the-fly. The size of the database is linear with the size of the corpus.

8.5 Conclusions and Future Work

In the present chapter the Dictionary Query System has been introduced that has been implemented to make the automatically generated dictionaries valuable resources not only for lexicographers but for end-users, too. Accordingly, the design of the DQS aims at compensating for the shortcomings of the proto-dictionaries and preserve the advances of the technique at the same time. This entails that some information usually present in dictionaries also appear in the proto-dictionaries, while others could be included only in the manual post-editing phase.

Information ideally present in a dictionary In Section 8.2 we gave an overview of the information ideally present in a dictionary based on Atkins and Rundell (2008). We have found that as opposed to usual dictionaries, new information may be presented in new ways in the proposed framework. Some of the relevant data are obviously not accessible for the proposed approach (pronunciation, specifiers, grammar labels, vocabulary types except for domain labels). Others—may be less straightforward for traditional and corpus-based lexicography—are readily available (frequency ranks among translations, ample example sentences, the proportion of translations over various text genres, etc.). Furthermore, some features were also implemented that are not familiar to the ordinary dictionary users, such as hints on the scope of cross-lingual usability or the feature of immediate reversibility.

Dictionary Query System (DQS) In order to demonstrate the generated proto-dictionaries, we have designed and implemented an online dictionary query system, which exploits the advantages of the data-driven nature of the applied technique. The DQS is made up of two components: In Dictionary Browser the proto-dictionaries can be queried, while the proto-dictionaries can be customized in the Cut Board.

DQS: Dictionary Browser The DQS provides different visualizations of the possible translations based on their translation probabilities and frequencies, along with their relevant contexts in the corpus. Some new assumptions can be formulated which

connect the statistical properties of the translation pairs, e.g. their frequency ratios and the cross-language semantic relations between them. Based on the generated dictionaries such hypotheses may be further examined in the future. At the present stage of research one-token units are queriable for all the four language pair we have dealt with. Besides, the French-Dutch (and vv.) verb + object structures can be queried both on the basis of verbs and objects.

DQS: Cut Board One bottleneck of the proposed method is that with the initial parameter setting the coverage of the proto-dictionaries is not sufficient. However, based on the evaluation of the French-Dutch and the Hungarian-Lithuanian proto-dictionaries we found that higher frequency source lemmata yield the same result in terms of precision with lower translation probabilities as less frequent source lemmata with higher translation probabilities. That is, fine-tuning the parameters increases coverage.

Nevertheless, the exact parameter-setting depends on the users' needs: The proto-dictionary for novice language learners should be made up of the most frequent SL and TL lemmata without wrong translations, whereas professional translators may be interested in intricate translations of less common words. Thus, the different user needs may be formulated in terms of precision and coverage: Accordingly, a proto-dictionary for a novice language learner should be a high precision and low coverage dictionary, while a proto-dictionary for a professional translator has to exhibit low precision with a great coverage. The Cut Board makes the proto-dictionaries customizable by enabling a cascaded filtering technique.

9

Conclusions and Future Work

9.1 Conclusions

Objective The basic objective of the present PhD thesis was to investigate to what degree language technology methods are able to facilitate the creation of bilingual dictionaries. This question is particularly important in the case of less resourced languages, for which due to low demand it does not pay off for publishers to invest into the creation of dictionaries.

Main finding Thesis I concerns the main finding of the present work:

(I) Although word alignment techniques on parallel corpora are widely used for the purpose of machine translation and until recently they have been hardly—if at all—used in lexicographic projects, THE AUTOMATIC LEARNING OF TRANSLATION PAIRS ON THE BASIS OF PARALLEL CORPORA USING CONDITIONAL PROBABILITIES is particularly apt for lexicographic purposes for *theoretical*, *practical* and *economical* reasons.

9.1.1 Summary of the dissertation

The essential theoretical achievement of the present thesis is concisely described in Thesis II:

(II) THE AUTOMATIC LEARNING OF TRANSLATION PAIRS ON THE BASIS OF PAR-

ALLEL CORPORA USING CONDITIONAL PROBABILITIES has certain benefits over both traditional and corpus-based lexicography, inasmuch the proposed method is able to define some of the fundamental notions of lexicography *in terms of quantifiable corpus data*.

In the rest of the present section each chapter will be briefly considered along with the corresponding theses, where roman numerals indicate the theses.

9.1.1.1 Chapter 2: Compiling the headword list

In Chapter 2 *traditional*, *corpus-based* and *corpus-driven* monolingual dictionaries were investigated with regard to their relation to corpus data.

Traditional lexicography We found that the fundamental assumption of traditional lexicographic approaches is that the basic building blocks of dictionaries are word form–meaning pairs. Traditional lexicography also presumes that (1) the meanings of the words are fairly stable across different contexts. Moreover, (2) word–meaning pairs are stored in the mental lexicon and can be assessed by means of introspection. From (1) it follows that contextual information does not play a great role in traditional lexicography, while (2) has two questionable implications. First, everyone has a strong belief that they know exactly the meaning(s) of a word. Secondly, this knowledge is largely alike across the members of a language community. That is, meanings are objective or at least highly intersubjective entities.

Corpus-based lexicography As for corpus based approaches, the great number of relevant projects indicates that this has been the dominant methodology in the field of lexicography recently. The basic assumptions shared by the different theories are the following: The meaning of words' is highly dependent on the contexts in which they occur. This view on meanings is compatible with (at least) two interpretations. First, word forms do not have meanings in themselves, but they have meaning potentials. Alternatively, words tend to be highly polysemious and show up with different meanings in different contexts. In either case, the various meanings can change significantly across different contexts. Because of the increased role of contexts in the description of meanings, introspection in itself is not enough to be able to list the relevant headwords of a dictionary and to provide a sufficient characterization of them. Therefore, the lexicographic intuition should be underpinned by corpus evidence. Moreover, a sound linguistic theory is needed to draw lexicographers' attention to the lexicographically

relevant facts. In addition, for the same purpose, the exploitation of lexical profiling tools turned out to be indispensable, too, as corpus size increased.

Corpus-driven lexicography With regard to the basic assumptions, corpus-driven lexicography was found to be rather similar to corpus-based lexicography. However, data plays a greater role than in corpus-based lexicography, at least as far as corpus-driven techniques are suitable to handle greater amount of data than corpus-based approaches. This is primarily due to the fact that the methodology has changed and unsupervised learning techniques became widely used for natural language processing tasks. The aim of unsupervised learning techniques is to learn hidden structures from unlabeled data. Thus, unsupervised techniques can eliminate unarticulated theoretical presumptions present in the labeling itself. However, human intuition cannot be completely excluded: It comes into play again when selecting the investigated phenomenon, coming up with a representation set up, fine-tuning the parameters and throughout the evaluation, as well.

By the end of Chapter 2 we found that a suitable method should be corpus-oriented, as in this case the description of LUs is underpinned by corpus evidence. However, a corpus-based methodology requires much human effort, which is usually not affordable in the case of lesser used languages. Consequently, corpus-driven approaches should be preferred.

9.1.1.2 Chapter 3: The translation phase

Various types of ρ Chapter 3 focused on translation relation ρ. Based on the literature it was found that several sub-types of translation relation may be distinguished and, at the same time, translation relation tends to be asymmetric and gradual.

The scale spans from the perfect translation equivalency, where the corresponding SL and TL expressions are interchangeable in every possible context, to translation equivalents that are fully dependent on the context. Explanatory equivalents (i.e. definitions) were also distinguished.

(IV) We investigated how the symmetry of translation relation can be interpreted. We found that if $\rho \subseteq A \times B$, i.e. if ρ is a relation in the mathematical sense, then the symmetry of ρ translation relation is best interpreted as ρ being an invertible function mapping from the SL vocabulary A to the TL vocabulary B. It was also found that this definition is consistent with the cases when "sym-

metric translation relation" is exploited in practical lexicography, such as when designing reversible dictionaries or when applying the hub-and-spoke model.

Main characteristics of ρ As translational synonymy/cognitive equivalency is rare, translation relation is best conceived of as *closeness*. This in turn entails that ρ translation relation should be gradual and quantifiable. We found that the gradual, possibly quantifiable notion of translation equivalence is incompatible with the binary view of translation pairs, therefore with the mapping view of translation.

(V) We accept that the translation relation is best to think of as a gradual notion, and we propose that we should be able to compare two possible translations of a given SL expression and to select the better one. That is, *translation relation should be quantifiable*. The strength of the translation equivalence could be measured by the number of contexts in which a TL expression appears as translation. This claim corresponds to the fact that perfect translation equivalents are defined as translation pairs that are interchangeable in every context, while on the other end of the scale, contextual translations appear only in a rather constrained set of contexts. Accordingly, we do not consider perfect translational equivalence (cognitive equivalence) and contextual equivalence separate types of translational equivalence, instead, we propose that they are the two ends of the very same scale.

(VI) In our view, instead of mathematical relation, ρ translation relation should be conceived of as *conditional probability*, $P(b|a)$, which gives an estimation of how many times the occurrences of $a \in A$ are translated as an occurrence of $b \in B$ on the basis of sentence aligned parallel corpora. We claim that conditional probability is a suitable mathematical construction to represent and to quantify over translation relation for multiple reasons. First, as opposed to the binary notion of mathematical relation, conditional probability is able to reflect the gradual nature of translation relation. Secondly, conditional probability captures the fact that translation relation tends to be asymmetric, as well. Thirdly, this mathematical construction is also able to reflect that translation relation is symmetric in the case of perfect translational equivalence.

9.1.1.3 Chapter 4: Encoding dictionaries

Encoding dictionaries and inter-annotator agreement By definition, in an encoding environment we are aware of the meaning of the SL expression and want to

find the contextually best translation for it. We consider the meaning of an expression known, if most native speakers assign the same meaning to it, when put into sufficiently specified contexts. That is, we expect the native speakers to achieve high inter-annotator agreement when finding the relevant meaning of a word in context. In Chapter 4 four experiments were discussed yielding the conclusion that neatly characterized sense-inventories are indispensible to achieve high inter-annotator agreement in the sense-annotation task.

(VII) We found that the neatly characterized sense-inventory should exhibit certain properties to enable the annotators to achieve high agreement on the annotation task. Namely,

(i) The sense-inventory has to comprise abundant contextual information that enable annotators to select the appropriate meaning on the basis of explicit distributional information.

(ii) Each SL headword in the sense-inventory should be characterized in a way that *each occurrence* of the given headword could be clearly assigned to a unique meaning. That is, there is no such occurrence that may be assigned to two different meanings.

(iii) It is also presupposed that meanings are non-overlapping entities.

That is, if our presupposition holds, a suitable SL sense-inventory for a high-quality encoding dictionary should be characterized in a way that *the various meanings of a word form create a partition in the mathematical sense over the occurrences of that word form.*

Partition of meanings over SL word occurrences Although still in an early stage of research, a possible methodology of creating a partition of meanings or sub-meanings over the occurrences of the SL word was introduced. The algorithm learns adjectival near-synonymy classes along with their contexts from corpus data yielding the conclusion that in this case *meaning may be conceived of as labels on partitions over SL word occurrences. Labels are composed of near-synonyms and contexts.*

(VIII) Unfortunately, such neatly characterized data-base usually is not available. Therefore, in the absence of such a sense-inventory an alternative way of creating high-quality encoding dictionaries should be found. Another possibility is that

we disregard word senses and try to retrieve translations by creating a partition of TL word forms directly over a given SL word form. The conditional probabilities $P(b_i|a)$ create a partition of occurrences of the possible translations b_i on the set of occurrences of the SL word form a. Moreover, translation pairs of the form a-b_i are linked on the basis of their natural contexts. Thus, conceiving of translation relation as conditional probability turned out to be suitable to create high quality encoding dictionaries.

(III) Albeit the general view in lexicography takes form-meaning pairs as the atomic building blocks of dictionaries, it is argued that, if the proposed method is used, word forms (in the sense of lemmata) may serve as the basic units for bilingual encoding dictionaries. That is, in this case we do not have to address the rather difficult problem of how to characterize meanings of word forms, as it falls back to the problem of how to characterize mere word forms in a bilingual dictionary.

9.1.1.4 Chapter 5: Selecting the alignment techniques

This chapter gave a brief summarization both of the sentence alignment and of the word alignment techniques. We found that hunalign is a suitable tool for our purposes, as it is able to handle both many-to-one and one-to-many alignments, it is language independent, and it is able to process UTF-8 input texts. As for dictionary extraction techniques, association and estimation approaches were distinguished. Association approaches test if two words co-occur significanty more often than it would be expected merely by chance. Although several association measures were introduced in the literature to test the independence hypothesis of SL and TL word occurrences, the use of a specific association measure seems to be hard to interpret as translation relation. As opposed to this—as it was shown in the previous chapters—using conditional probability to extract translation pairs is motivated both by arguments from the fields of linguistics and translatology. Thus, translation pairs were extracted using one of the estimation approaches.

9.1.1.5 Chapter 6: Proof-of-concept experiments

In this chapter we have verified that the proposed method is able to facilitate the cost-effective creation of bilingual dictionaries.

This methodology meets the expectations put forward in the previous chapters. It is *corpus-driven*, thus it diminishes the role of intuition in the dictionary building process.

Economical considerations Once the parallel corpus is available, word alignment on parallel corpus significantly decreases the amount of human labour needed to produce a bilingual dictionary. It turned out that the most time-consuming part of the workflow is the collection and normalization of parallel texts.

Language independency Since hunalign and GIZA++ are language independent tools, sentence alignment and word alignment are readily re-applicable for any language pair.

Reversibility In the presence of the parallel corpus, the creation of the reversed proto-dictionary turned out to be rather straightforward. This is due to the asymmetric nature of conditional probability. The reversed proto-dictionaries—Lithuanian-Hungarian and Slovenian-Hungarian were generated, too.

These observations correspond to Thesis XI:

(XI) Owing to the data-driven nature of the proposed technique, the amount of human effort needed to compile bilingual dictionaries is significantly decreased. The extraction method and the Dictionary Query System are language-independent, thus, only the language dependent resources and tools need to be collected again when preparing dictionaries for new language-pairs. Once the required resources and tools are collected, the generation of the reversed dictionary is a straightforward process.

Multiple meanings The method is able to rank polysemious meanings, that is, the automatically retrieved translation probability indicates how the translation space is divided among the various translation candidates, i.e. which is the most frequent translation of the source word. This corresponds to Thesis Xb:

(Xb) Representing translation relation as conditional probability makes it possible to rank translations according to how likely they are. Presenting translation candidates in such a way is an obvious advance compared to the usual ordering techniques applied in bilingual dictionaries.

Encoding dictionaries An additional requirement toward the method was that it should enable the creation of enconding dictionaries. In our view, the automatically

retrieved natural example sentences are of great help when trying to find the translation that produces the idiomatically best translation among the possible translation candidates. Obviously, this step is not wholly automatic, but the retrieval of competing translation candidates and the relevant contexts for each of these candidates helps lexicographers (and end-users) to focus their attention on the relevant linguistic facts.

Evaluation Instead of right and wrong translation pairs, the evaluation was based on lexicographically acceptable and lexicographically unacceptable translation units. The detailed evaluation of the resulting proto-dictionaries showed that the proportion of lexicographically acceptable translation pairs depends upon the frequency of the source and target lemmata and on the automatically attained translation probabilities: The higher the translation probability, the greater the proportion of lexicographically acceptable translation units. The coverage of the proto-dictionaries was far below the targeted size. However, refinement of parameters resulted in an increased number of lexicographically acceptable translation pairs (cf. Chapter 8).

Uniform XML format The parallel corpora were converted into XML-format with a simple and uniform morphological annotation. This conversion made the following workflow more simple, through enabling the uniform processing of the various parallel corpora. Moreover, it is easier to generate parallel corpora with various levels of annotation.

Limitations Unfortunately, some limitations had to be overcome, too. First, at the present state of research the sizes of the parallel corpora are not big enough to ensure an appropriate coverage of dictionaries. This problem might have several solutions, either introducing related methods, such as retrieving translation pairs based on comparable corpora or fine-tuning the parameters used for filtering. In Chapter 8 we found that a refined parameter setting can partly compensate for the limited coverage as lower translation probabilities may yield high quality candidates in the case of frequent SL lemmata. Secondly, the method in its present form enables only the retrieval of one-token translation pairs. Thus, it does not handle collocates or verbal expressions. Since such structures are inherently part of natural languages and are essential for the production of idiomatically correct translations, they have to be included in the proto-dictionaries.

9.1.1.6 Chapter 7: Extracting parallel verbal structures

We investigated to what extent it was possible to retrieve parallel verbal expressions from parallel corpora. For doing so, a preprocessing module was introduced (1) to detect monolingual verbal expressions in each side of the parallel corpus, (2) to merge each occurrence of the verbal expressions so that they could be treated as one-token units in the following stages of the retrieval process. In the third phase the merged verbal structures were aligned.

Experiments Three experiments were performed relying on parallel corpora with different annotation levels. In the first experiment only a limited set of verbs was considered and the parallel corpus was parsed with rather simple and approximative rules yielding a shallow syntactic annotation. The evaluation showed that although the results are promising both precision and recall should be increased.

The second experiment was performed with a deep parsed parallel corpus to improve precision and recall and this time all verbs occurring in the parallel corpus were taken into account. This experiment yielded the conclusion that there are too many different frame types resulting in sparse data during the alignment. Thus, the number of frame types should be decreased. There are more ways to do that: First, too long frames should be omitted, as they turned out to be typical of highly specialized languages. Manual evaluation has shown that maximally frames of length 5 should be included in the proto-dictionary. Secondly, the deep syntactic analysis present in the corpus gives a more intricate description of the verbal frames. This sometimes yields non-existent verbals structures that should have been subsumed under a more general frame. Moreover, the syntactic analysis may be ambiguous, which also increases the number of different frame types.

To decrease the data sparsity the frame types were redefined in a third experiment: Verbal structures only with different verbal lemma or different direct object were kept separately, otherwise they were merged and the corpus frequencies were recounted. The resulting parallel verbal structures were made accessible online.

9.1.1.7 Chapter 8: The Dictionary Query System

The main practical finding of this dissertation is that a suitable dictionary query system (DQS) is capable of rendering proto-dictionaries a useful resource not only for lexicographers but for end-users, too. Therefore, a DQS was designed and implemented that

displays some novel features compared to traditional dictionaries. Beside the generated proto-dictionaries the practical results of the present thesis concern the novelties of the dictionary query system.

Information ideally present in a dictionary In Section 8.2 we gave an overview on the information ideally present in a dictionary. It was found that as opposed to usual dictionaries, new information may be presented in new ways in the proposed framework. Some of the relevant data are obviously not accessible for the proposed approach. Others—may be less straightforward for traditional and corpus-based lexicography—are readily available. Furthermore, some features were also implemented that are not familiar to the ordinary dictionary users, such as hints on the scope of usability of the TL lemma or the feature of immediate reversibility.

Dictionary Query System (DQS) In order to demonstrate the generated proto-dictionaries, we have designed and implemented an online dictionary query system, which exploits the advantages of the data-driven nature of the applied technique. The DQS is made up of two components: The Dictionary Browser makes the proto-dictionaries queryable, while the Cut Board makes the proto-dictionaries customizable.

DQS: Dictionary Browser The DQS provides different visualizations of the possible translations based on their translation probabilities and frequencies, along with their relevant contexts in the corpus. Some new assumptions can be formulated which connect the statistical properties of the translation pairs. Based on the generated dictionaries such hypotheses may be further examined in the future.

(Xc) As opposed to traditional dictionaries, the DQS gives a hint on the scope of usability of the translation based on some very simple heuristics. To know whether the TL word may show up in a more restricted or a more general set of contexts than the SL word is essential in the case of encoding dictionaries.

At the present stage of research one-token units are queryable for all the four language pair we have dealt with. Besides, the French-Dutch (and vv.) verb + object structures can be queried both on the basis of verbs and objects.

DQS: Cut Board One bottleneck the proposed method has to face is that the coverage of the proto-dictionaries was insufficient with the initial parameter setting. However, the refined evaluation of the French-Dutch and the Hungarian-Lithuanian proto-dictionaries led us to the conclusion that higher frequency source lemmata even with lower translation probabilities also yield good results in terms of precision. That is, fine-tuning the parameters increases the coverage. This observation corresponds to

Thesis IX:

(IX) We found that a cascaded filtering technique significantly increases the coverage of the resulting proto-dictionaries: In the case of more frequent lemmata even lower values of conditional probabilities may yield lexicographically acceptable translations. Hence, fine-tuning the parameters results in a bigger proto-dictionary.

Nevertheless, the exact parameter-setting depends on the users' needs: The proto-dictionary for novice language learners should be made up of the most frequent SL and TL lemmata without wrong translations, whereas professional translators may be interested in intricate translations of less common words.

Becase of the different user scenarios we did not strive ourselves to find the "best" parameter setting. Instead, by presetting different selection criteria on the Cut Board the contents of the dictionaries are customizable to suit various usage scenarios.

(Xa) The most important novelty of DQS is that it is *customizable*. That is, the user can select the sub-part of the proto-dictionary that suits most their needs. We think that various parameter settings match well different user needs. The scope of various users may span from novice language learners to professional translators. Keeping the most frequently occurring translation pairs results in a low-coverage but high-precision proto-dictionary, which are appropriate for novice language learners. On the other end of the scale, selecting more relaxed parameters generates a proto-dictionary with a greater coverage but with a lower precision. Such a proto-dictionary may suit the needs of professional translators who may be interested in special uses of words, in lexicographically acceptable but not perfect translation units. At the same time, they can easily catch wrong translations, therefore, low precision does not pose a problem for them. Thus, the customizability feature of the Dictionary Query System supports various user scenarios.

9.2 Future Work

Our most important future tasks consist in increasing the coverage of the dictionaries and customization.

Increasing coverage First, one obvious way to increase the size of the protodictionaries is to augment the size of the parallel corpora. For this purpose, web crawlers could be used to build parallel corpora automatically.

An alternative way to include more translation pairs into the proto-dictionaries is to complement the automatically compiled proto-dictionaries with translation pairs extracted from monolingual corpora, which are made up of several orders greater amount of texts than parallel corpora. For doing so, the clique-based automatic synonymy detection method (cf. 2.3.3.3) should be improved.

Moreover, instead of the cascaded evaluation technique a curve should be applied to filter out improper results. The parameters of the curve could be learnt by means of some suitable classification method, such as logistic regression.

Customization A possible future work is to further evaluate the dictionaries in real world use cases. In our view customizability is a key feature that may increase the coverage of dictionaries. Predefined cut-off curves should be applied so that users could select the proto-dictionary that suits best their needs, i.e. the one with the best precision-recall trade-off for their needs.

9.3 Related Publications

9.3.1 Related Publications in English

- Héja, Enikő and Takács, Dávid. Automatically Generated Customizable Online Dictionaries. In: Daelemans W. et al. (eds.) *Proceedings of the Demonstrations at the 13th Conference of EACL*, The Association for Computer Linguistics, Avignon, France, April 23-27, 2012, p. 51-57.

- Héja, Enikő and Takács, Dávid. An Online Dictionary Browser for Automatically Generated Bilingual Dictionaries. In: Fjeld R. V. and Torjusen, J. M. (eds.) *Proceedings of the 15th EURALEX International Congress.* Oslo, Norway, 7-11 August 2012, p. 468-477.

- Héja, Enikő and Takács, Dávid. Automatically Generated Online Dictionaries. In: Calzolari N. et al. (eds.) *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC2012)*, European Language Resources Association (ELRA), Istanbul, Turkey, May 23-25, 2012, p. 2487-2493.

- Héja, Enikő. Dictionary Building based on Parallel Corpora and Word Alignment. In: Dykstra, A. and Schoonheim, T., (eds):*Proceedings of the XIV. EURALEX International Congress*, Leeuwarden/Ljouwert, The Netherlands, July 6-10, 2010, p. 341-352.

- Héja, Enikő. The Role of Parallel Corpora in Bilingual Lexicography. In: *Proceedings of the LREC2010 Conference*, La Valletta, Malta, May 17-23, 2010, p. 2798-2805.

- Kuti, Judit, Héja, Enikő and Sass, Bálint. Sense disambiguation - "Ambiguous sensation"? Evaluating sense inventories for verbal WSD in Hungarian. In: *Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, La Valletta, Malta, May 23, 2010, p.

9.3.2 Related Publications in Hungarian

- Héja, Enikő, Takács, Dávid and Sass, Bálint. Igei bővítménykeretek fordítási ekvivalenseinek kinyerése mélyen elemzett párhuzamos korpuszból. (Extracting Verb Centered Constructions and their Translation Equivalents from a Deep Parsed Parallel Corpus) In: Tanács Attila, Vincze Veronika (szerk.) *MSZNY2011 VIII. Magyar Számítógépes Nyelvészeti Konferencia.* SZTE, Szeged, 2011., p. 47-59.

- Héja, Enikő and Sass, Bálint. Többszavas kifejezések kezelése a párhuzamos korpuszokra épülő szótárkészítési módszertanban. (The Treatment of Multiword Expressions in a Parallel Corpus based Lexicographic Framework) In: Tanács Attila, Vincze Veronika (szerk.) *MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia*, SZTE, Szeged, 2010., p. 80-90.

- Héja, Enikő and Dávid, Takács. Melléknevek szűk szemantikai osztályainak detekciója a Magyar Nemzeti Szövegtárban jelentésegyértelműsítés céljából (Detection of Adjectival Near-Synonymy Classes for Word Sense Disambiguation) In: Tanács Attila, Vincze Veronika (szerk.) *MSZNY2010, VII. Magyar Számítógépes Nyelvészeti Konferencia*, SZTE, Szeged, 2010., p. 360-362.

References

A. Adamska-Sałaciak. Examining Equivalence. *International Journal of Lexicography*, 23(4):387–409, Dec. 2010. 27, 61, 63, 64, 66, 67, 68, 69, 70, 71, 72, 85, 87

J. Ah-Pine and G. Jacquet. Clique-based clustering for improving named entity recognition systems. In *EACL*, pages 51–59, 2009. 51, 54

S. Aït-Mokhtar and J.-P. Chanod. Incremental finite-state parsing. In *Proceedings of Applied Natural Language Processing*, pages 73–79, Washington, DC., April 1997. 93

R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4): 555–596, 2008. 88, 91, 92

S. Atkins. Analysing the verbs of seeing: a frame semantics approach to corpus lexicography. In S. Gahl, C. Johnson, and A. Dolbey, editors, *Proceedings of the Twentieth Annual Meeting of the Berkeley Linguistics Society*. Berkeley Linguistics Society, Berkeley Linguistics Society, 1994. 36

S. Atkins. *Collins-Robert French-English, English-French Dictionary*. Bilingual Dictionaries Series. Dictionnaires Le Robert, 1996. ISBN 9780062755216. 114

S. Atkins and M. Rundell. *The Oxford Guide to Practical Lexicography*. Oxford University Press, 2008. ISBN 0199277710. 15, 16, 17, 18, 21, 33, 36, 40, 63, 65, 68, 69, 70, 98, 99, 142, 188, 189, 190, 193, 194, 195, 199, 207

S. Atkins, C. Fillmore, and C. Johnson. Lexicographic relevance: selecting information from corpus evidence. *International Journal of Lexicography*, 16.3:251–280, 2003a. 36

S. Atkins, M. Rundell, and H. Sato. The contribution of framenet to practical lexicography. *International Journal of Lexicography*, 16.3:333–357, 2003b. 36

C. F. Baker and J. Ruppenhofer. FrameNet's frames vs. levin's verb classes. In J. Larson and M. Paster, editors, *Proceedings of 28th Annual Meeting of the Berkeley Linguistics Society*, pages 27–38, 2002. 33, 36

E. Bojtár. *Litván—magyar szótár Lietuviu—vengru kalbu zodynas (Lithuanian–Hungarian Dictionary)*. Lietuvos Kalbos Institutas, Vilnius, 2007. ISBN 978-9955-704-35-5. 19, 81

G. Bouma, G. van Noord, and R. Malouf. Alpino: wide-coverage computational analysis of dutch. *Computational Linguistics in the Netherlands 2000. Selected Papers from the 11th CLIN Meeting.*, pages 45–59, 2001. 171

E. Brill. A simple rule-based part of speech tagger. In *ANLP*, pages 152–155, 1992. 94

P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *COMPUTATIONAL LINGUISTICS*, 16(2):79–85, 1990. 127

P. F. Brown, V. J. Pietra, S. A. D. Pietra, and R. L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311, 1993. 125, 127

R. Carter. *Vocabulary: Applied Linguistic Perspectives*. Routledge linguistics classics. Routledge, 1998. ISBN 9780415168649. 28

R. Chapman. *Roget's International Thesaurus*. Harper Colophon Books. Crowell, 1977. ISBN 9780690000108. 23

S. F. Chen. Aligning Sentences in Bilingual Corpora using Lexical Information. In *Proceedings of the 31st Annual Conference of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, 1993. Association for Computational Linguistics, Association for Computational Linguistics. URL `http://portal.acm.org/citation.cfm?id=981576&dl=`. 121

N. Chomsky. *Syntactic Structures*. Mouton & Co., The Hague, 1957. ISBN 3-11-017279-8. Reprinted 1985 by Springer, Berlin and New York. 25

S. Cinkova and P. Hanks. Validation of corpus pattern analysis - assigning pattern numbers to random verb samples. Validation Manual, 2010. 37

O. D. Clercq and M. M. Perez. Data collection and ipr in multilingual parallel corpora. dutch parallel corpus. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. 161

J. Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. 89

D. Cruse. *Lexical Semantics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1986. ISBN 9780521276436. URL `http://books.google.hu/books?id=xDSBaet2uSsC`. 23

A. P. Dempster, N. M. Laird, and R. D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–22, 1977. 127, 165

K. É. Kiss, F. Kiefer, and P. Siptár. *Új magyar nyelvtan*. Osiris tankönyvek. Osiris, 2003. URL `http://books.google.hu/books?id=QZ1YuAAACAAJ`. 164

REFERENCES

T. Erjavec, C. Ignat, B. Pouliquen, and R. Steinberger. Massive multi-lingual corpus compilation: Acquis communautaire and totale. In *Proceedings of the 2nd Language Technology Conference*, pages 32–36, April 2005. 132, 153

S. Evert. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD thesis, IMS, University of Stuttgart, 2004. 192

C. Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998. 22

C. Fellbaum. Wordnet and wordnets. In K. Brown, editor, *Encyclopedia of Language and Linguistics*, pages 665–670, Oxford, 2005. Elsevier. 92

C. Fillmore. The hard road from verbs to nouns. In M. Chen and O. Tzeng, editors, *In Honor of William S-Y. Wang*, pages 105–129. Pyramid Press, Taiwan, 1994. 33

C. Fillmore. Frame semantics. In K. Brown, editor, *Encyclopedia of Language and Linguistics*. Elsevier, Oxford, 2005. 33

C. J. Fillmore and B. T. S. Atkins. Framenet and lexicographic relevance. In *Proceedings of the First International Conference on Language Resources and Evaluation, Granada, Spain*, 1998. 36

C. J. Fillmore and C. F. Baker. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop*, Pittsburgh, June 2001. NAACL, NAACL. 36

J. R. Firth. *Papers in linguistics 1934-1951*. Oxford University Press, London, 1957. 42, 45

J. L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1976. 89

N. Francis and H. Kučera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979. 23, 93

W. A. Gale and K. W. Church. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 1993. 121

D. Geeraerts. The lexicographical treatment of prototypical polysemy. Cognitive linguistics research, pages 327–344. Bod Third Party Titles, 2006. ISBN 9783110190427. 21

Z. Gendler Szabó. Compositionality. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2013 edition, 2013. 25, 98

P. Hanks. Compiling a monolingual dictionary for native speakers. *Lexikos*, 20:580–598, 2010. ISSN 1684-4904. 22, 27, 42, 49, 59

P. Hanks and J. Pustejovsky. A pattern dictionary for natural language processing. *Revue française de linguistique appliquée*, 10(2):580–598, 2005. 24, 32, 33, 36, 37, 38

Z. Harris. Distributional structure. *Word*, 10(23):146–162, 1954. 25, 42, 48, 50, 57

E. Héja and D. Takács. Melléknevek szűk szemantikai osztályainak detekciója a magyar nemzeti szövegtárban jelentés-egyértelműsítés céljából. In *VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010)*, pages 360–362, 2010. 50

D. Hiemstra. Using statistical methods to create a bilingual dictionary. Master's thesis, University of Twente, 1996. 124, 127

E. Héja. The outlines of a hybrid approach to word sense disambiguation, June 2008. Presentation at Intern's day, Xerox Research Centre Europe, Grenoble, France. 88

E. Héja. The role of parallel corpora in bilingual lexicography. In N. C. C. Chair), K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, and D. Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7. 165, 200

T. Janssen. *Compositionality*. ILLC research report and technical notes series. Inst. for Logic, Language and Computation, 1996. URL `http://books.google.hu/books?id=t3KgHAAACAAJ`. 25

E. Jezek and P. Hanks. What lexical sets tell us about conceptual categories. *Lexis*, 4:7–22, 2010. ISSN 1951-6215. 20

D. Jurafsky and J. H. Martin. *Speech and Language Processing (2nd Edition) (Prentice Hall Series in Artificial Intelligence)*. Prentice Hall, 2 edition, 2008. ISBN 0131873210. xiv, 106, 124

M. Kay and M. Röscheisen. Text-translation alignment. *Comput. Linguist.*, 19(1):121–142, Mar. 1993. ISSN 0891-2017. URL `http://dl.acm.org/citation.cfm?id=972450.972457`. 121

S. Keinys, K. J., P. J., J. Pikčilingis, N. Sližienė, U. K., and V. V. *Dabartinės lietuvių kalbos žodynas (Lithuanian Explanatory Dictionary)*. Vilnius: Mokslo ir enciklopedijų leidykla, 1993. 19

A. Kilgarriff and I. Kosem. Corpus tools for lexicographers. In S. Granger and M. Paquot, editors, *Electronic Lexicography*, pages 1–6. Oxford University Press, Oxford, 2012. 40, 41

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. The sketch engine. In *Proc. EURALEX 2004*, pages 105–116, Lorient, France, 2004. 41

A. Komlósy. Régensek és vonzatok. In F. Kiefer, editor, *Struktrális magyar nyelvtan 1. Mondattan (Structural Grammar of Hungarian 1. Syntax)*, pages 299–527. Akadémiai Kiadó, Budapest, 1992. 31

J. Kuti, E. Héja, and B. Sass. Sense disambiguation – "ambiguous sensation"? evaluating sense inventories for verbal wsd in hungarian. In *Proceedings of LREC 2010 Workshop on Exploitation of Multilingual Resources and Tools for Central and (South-) Eastern European Languages*, pages 32–36, La Valletta, Malta, May 2010. 88, 91

N. LaRoche and L. Urdang. *The Synonym Finder*. Rodale, 1981. 23

REFERENCES

B. Levin. *English verb classes and alternations : a preliminary investigation*. The University of Chicago Press, 1993. 31, 32

B. Levin and M. Hovav. *Argument Realization*. Research Surveys in Linguistics. Cambridge University Press, 2005. ISBN 9780521663762. 31

D. Lewis. General semantics. *Synthese*, 22(1-2):18–67, 1970. v, 86

D. Lindemann. Bilingual lexicography and corpus methods. the example of german-basque as language pair. *Procedia-Social and Behavioral Sciences*, 95:249–257, 2013. 2

J. Locke. *An Essay Concerning Human Understanding (1689)*. 1841. 45

I. Maks. Ombi: The practice of reversing dictionaries. *International Journal of Lexicography*, 20(3):259–274, 2007. 76

D. C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999. ISBN 0262133601. 120, 121, 124, 126

M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330, 1993. 152

W. Martin. Government policy and the planning and production of bilingual dictionaries : The 'dutch' approach as a case in point. *International Journal of Lexicography*, 20(3): 221–237, 2007. 74, 75, 77, 78, 82

D. Melamed. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249, 2000. 122, 123

I. Mel'čuk, N. Arbatchewsky-Jumarie, and A. Clas. *Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques*. Number v. 1 in Dictionnaire explicatif et combinatoire du français contemporain: recherches lexico-sémantiques. Presses de l'Université de Montréal, 1984. ISBN 9782760606593. 29

I. A. Mel'cuk and A. K. Zolkovskij. *Explanatory combinatorial dictionary of modern Russian: opyty semantiko - sintaksiceskogo opisanija russkoj leksiki*, volume 14 of *Wiener Slawistischer Almanach : Sonderband*. Ges. zur Foerderung slawist. Studien, Wien, 1984. 29

I. Melčuk. *Explanatory Combinatorial Dictionaries*. Polimetrica, 2006. 18, 29, 30, 44, 84

I. Mel'čuk, N. A. Jumarie, L. Iordanskaja, and S. Mantha. *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques III*. Presses de l'université de Montréal, Montréal(Quebec), Canada, 1992. 29

I. Mel'čuk, N. A. Jumarie, L. Iordanskaja, S. Mantha, and A. Polguère. *DEC : Dictionnaire explicatif et combinatoire du français contemporain, recherches lexico-sémantiques IV*. Presses de l'université de Montréal, Montréal(Quebec), Canada, 1996. 29

M. Miháltz, C. Hatvani, J. Kuti, G. Szarvas, J. Csirik, G. Prószéky, and T. Váradi. Methods and results of the hungarian wordnet project. In A. Tanács, D. Csendes, V. Vincze, C. Fellbaum, and P. Vossen, editors, *Proceedings of the IVth Global WordNet Conference*, pages 311–321, Las Palmas, 2008. 91

T. Mikolov, Q. V. Le, and I. Sutskever. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168, 2013. 56

G. A. Miller, C. Leacock, R. Tengi, and R. T. Bunker. A semantic concordance. In *Proceedings of the workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7. 93

F. J. Och and H. Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003. 13, 122, 123, 125, 126, 127, 133, 165, 178

C. Oravecz and P. Dienes. Efficient stochastic part-of-speech tagging for hungarian. In *Proceedings of the Third International Conference on Language Resources and Evaluation*, pages 710–717, Las Palmas, 2002. 132

J.-B. Ormal-Grenon and N. Pomier. *The Oxford-Hachette French dictionary. Le grand dictionnaire Hachette-Oxford; 3rd ed.* Oxford Univ., Oxford, 2001. xiv, 191, 193

T. Piotrowski. *Problems in Bilingual Lexicography*. Wydawnictwo Uniwersytetu Wrocławskiego, Wrocław, 1994. 68

S. Ploux and B. Victorri. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *TAL*, 1(39):146–162, 1998. 50

A. Polguère. Towards a Theoretically-motivated General Public Dictionary of Semantic Derivations and Collocations for French. In *Proceedings of EURALEX'2000*, Stuttgart, Germany, 2000. URL `http://www.olst.umontreal.ca/FrEng/APolEURALEX.pdf`. 29

J. Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, MA, 1995. 32

A. Ramos. Semantic description of collocations in a lexical database. In F. Kiefer, G. Kiss, and J. Pajzs, editors, *Papers in Computational Lexicography*, COMPLEX, pages 17–27. Linguistics Institute and Hungarian Academy of Sciences, 2005. 29

A. Ribeiro, G. Pereira Lopes, J., and J. Mexia. Extracting equivalents from aligned parallel texts: Comparison of measures of similarity. In M. Monard and J. Sichman, editors, *Advances in Artificial Intelligence*, volume 1952 of *Lecture Notes in Computer Science*, pages 339–349. Springer Berlin Heidelberg, 2000. ISBN 978-3-540-41276-2. doi: 10.1007/3-540-44399-1_35. URL `http://dx.doi.org/10.1007/3-540-44399-1_35`. 123

E. Rimkutė, V. Daudaravičius, and A. Utka. Morphological annotation of the lithuanian corpus. In *45th Annual Meeting of the Association for Computational Linguistics; Workshop Balto-Slavonic Natural Language Processing 2007 Conference Proceedings*, pages 94–99, Praga, 2007. 132, 151

REFERENCES

E. Rimkutė, V. Daudaravičius, A. Utka, and J. Kovalevskaitė. Bilingual parallel corpora for english, czech and lithuanian. In *The Third Baltic Conference on Human Language Technologies 2007 Conference Proceedings*, pages 319–326, Kaunas, 2008. 132

J. Ruppenhofer, M. Ellsworth, M. R. L. Petruck, C. R. Johnson, and J. Scheffczyk. FrameNet II: Extended theory and practice. Technical report, ICSI, 2010. URL `http://framenet.icsi.berkeley.edu/book/book.pdf`. 35

P. Rychlý and A. Kilgarriff. An efficient algorithm for building a distributional thesaurus (and other sketch engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 41–44, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL `http://dl.acm.org/citation.cfm?id=1557769.1557783`. 48

G. Saldanha. Principles of corpus linguistics and their application to translation studies research. *Tradumática*, 7, 2009. 47

B. Sass. A unified method for extracting simple and multiword verbs with valence information and application for hungarian. In *Recent Advances in Natural Language Processing, RANLP 2009, 14-16 September, 2009, Borovets, Bulgaria*, pages 399–403, 2009. URL `http://aclweb.org/anthology/R/R09/R09-1072.pdf`. 47

B. Sass. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból [extracting parallel multiword verbs from parallel corpora]. In *VII. Magyar Számítógépes Nyelvészeti Konferencia*, pages 102–110, 2010. 160, 161

B. Sass. *Igei szerkezetek gyakorisági szótára – egy automatikus lexikai kinyerő eljárás és alkalmazása (The Frequency Dictionary of Verbal Structures – an authomatic lexical extraction method and its application)*. PhD thesis, PPKE ITK, 2011. 14, 47, 155, 156, 159, 161, 174

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, pages 44–49, Manchester, UK, 1994. 147

W. A. Scott. Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19(3):321–325, 1955. 89

C. E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27(1):379–423, 1948. 123

A. Tamm and W. Martin. Ombi: An editor for constructing reversible lexical databases. In M. Gellerstam, editor, *Proc. Euralex '96*, pages 678–688, Göteborg, 1996. Göteborg University. 76

J. Tiedemann. *Recycling Translations - Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing*. PhD thesis, Uppsala University, 2003. 124, 128

E. Tognini-Bonelli. *Corpus Linguistics at Work*. Studies in corpus linguistics. J. Benjamins, 2001. ISBN 9789027222763. 46

L. Urdang. *A Basic Dictionary of Synonyms and Antonyms*. Orient Paperbacks, 1983. ISBN 9788122200362. 23

H. van der Vliet. The referentiebestand nederlands as a multipurpose lexical database. *International Journal of Lexicography*, 20(3):239–257, 2007. 39

D. Varga, L. Németh, P. Halácsy, A. Kornai, V. Trón, and V. Nagy. Parallel corpora for medium density languages. In *Proceedings of the RANLP 2005*, pages 590–596, Borovets, Bulgaria, 2005. 13, 121, 132, 147

N. Varma. Identifying word translations in parallel corpora using measures of association. diploma thesis, University of Minnesota, 2002. 123

E. Veldi. Reversing a bilingual dictionary: a mixed blessing. In *Proc. EURALEX 2010*, pages 861–865, Leeuwarden, The Netherlands, 2010. 75

E. Villemonte de la Clergerie. Convertir des dérivations TAG en dépendances. In *17e Conférence sur le Traitement Automatique des Langues Naturelles - TALN*, 2010. 171

P. Vossen. Eurowordnet: A multilingual database of autonomous and language-specific wordnets connected via an inter-lingual index. *International Journal of Lexicography Vol.17*, 2:161–173, 2004. 22, 23

T. Váradi. The hungarian national corpus. In *In Proceedings of the Second International Conference on Language Resources and Evaluation, Las Palmas*, pages 385–389, 2002. 52

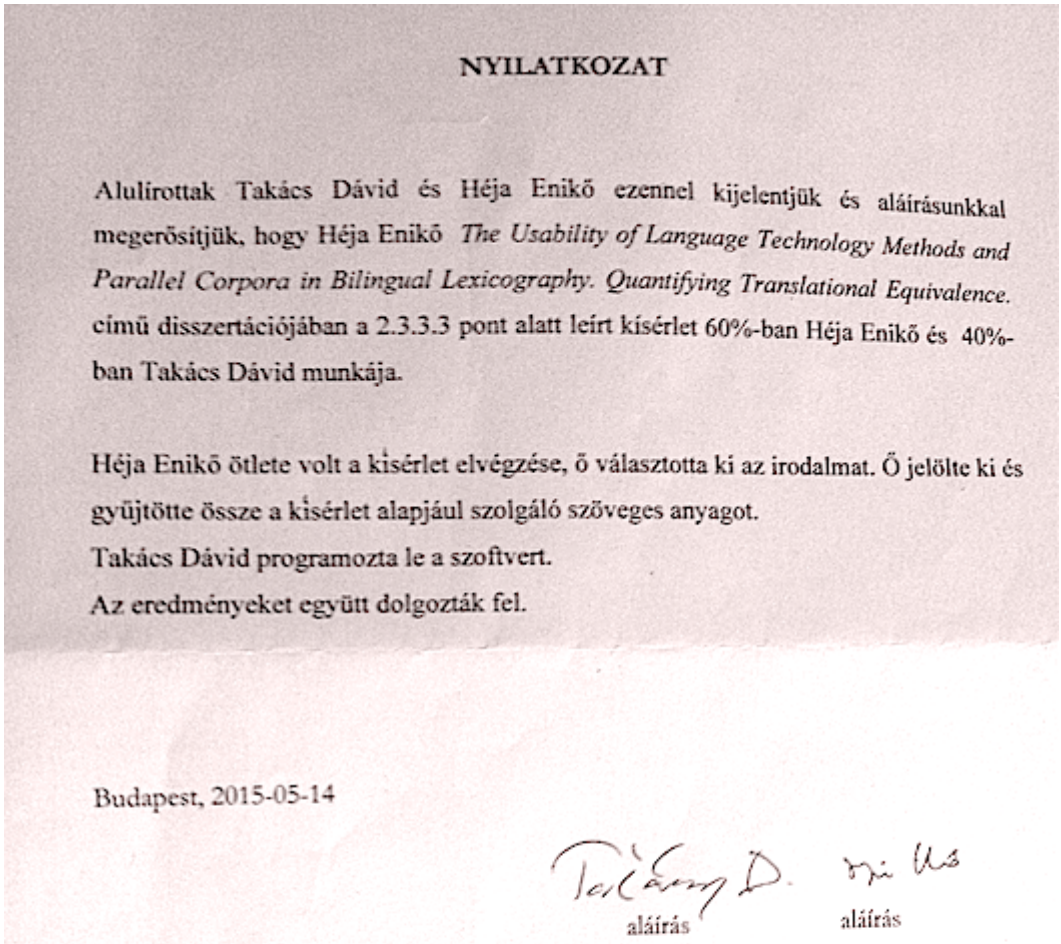
J. Véronis. Sense tagging: does it make sense? In A. Wilson, P. Rayson, and T. McEnery, editors, *Corpus Linguistics by the Lune: a festschrift for Geoffrey Leech*. Peter Lang, Frankfurt, 2003. 88, 91

L. Wittgenstein. *Lectures on Philosophy, 1932-35*. Blackwell, 1979. 45

L. Zgusta. *Manual of Lexicography*. Academia, Prague, 1971. 68, 69

L. Zgusta. Translational equivalence and the bilingual dictionary. In R. R. K. Hartmann, editor, *LEXeter '83 Proceedings*, pages 147–154, Tübingen, 1984. Max Niemeyer. 69, 85

Declaration—Nyilatkozatok



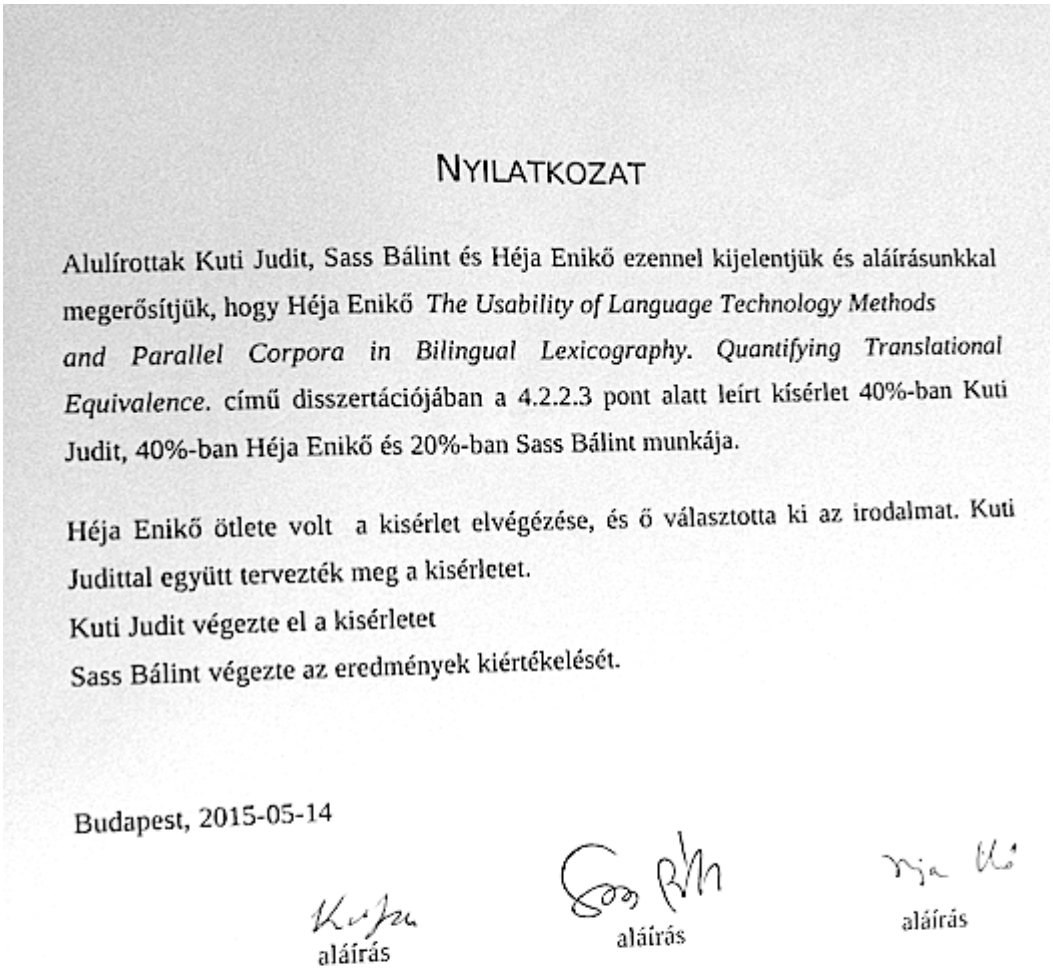
NYILATKOZAT

Alulírottak Takács Dávid és Héja Enikő ezennel kijelentjük és aláírásunkkal megerősítjük, hogy Héja Enikő *The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence.* című disszertációjában a 2.3.3.3 pont alatt leírt kísérlet 60%-ban Héja Enikő és 40%-ban Takács Dávid munkája.

Héja Enikő ötlete volt a kísérlet elvégzése, ő választotta ki az irodalmat. Ő jelölte ki és gyűjtötte össze a kísérlet alapjául szolgáló szöveges anyagot.
Takács Dávid programozta le a szoftvert.
Az eredményeket együtt dolgozták fel.

Budapest, 2015-05-14

aláírás aláírás



NYILATKOZAT

Alulírottak Kuti Judit, Sass Bálint és Héja Enikő ezennel kijelentjük és aláírásunkkal megerősítjük, hogy Héja Enikő *The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence.* című disszertációjában a 4.2.2.3 pont alatt leírt kísérlet 40%-ban Kuti Judit, 40%-ban Héja Enikő és 20%-ban Sass Bálint munkája.

Héja Enikő ötlete volt a kisérlet elvégézése, és ő választotta ki az irodalmat. Kuti Judittal együtt tervezték meg a kisérletet.
Kuti Judit végezte el a kisérletet
Sass Bálint végezte az eredmények kiértékelését.

Budapest, 2015-05-14

aláírás aláírás aláírás

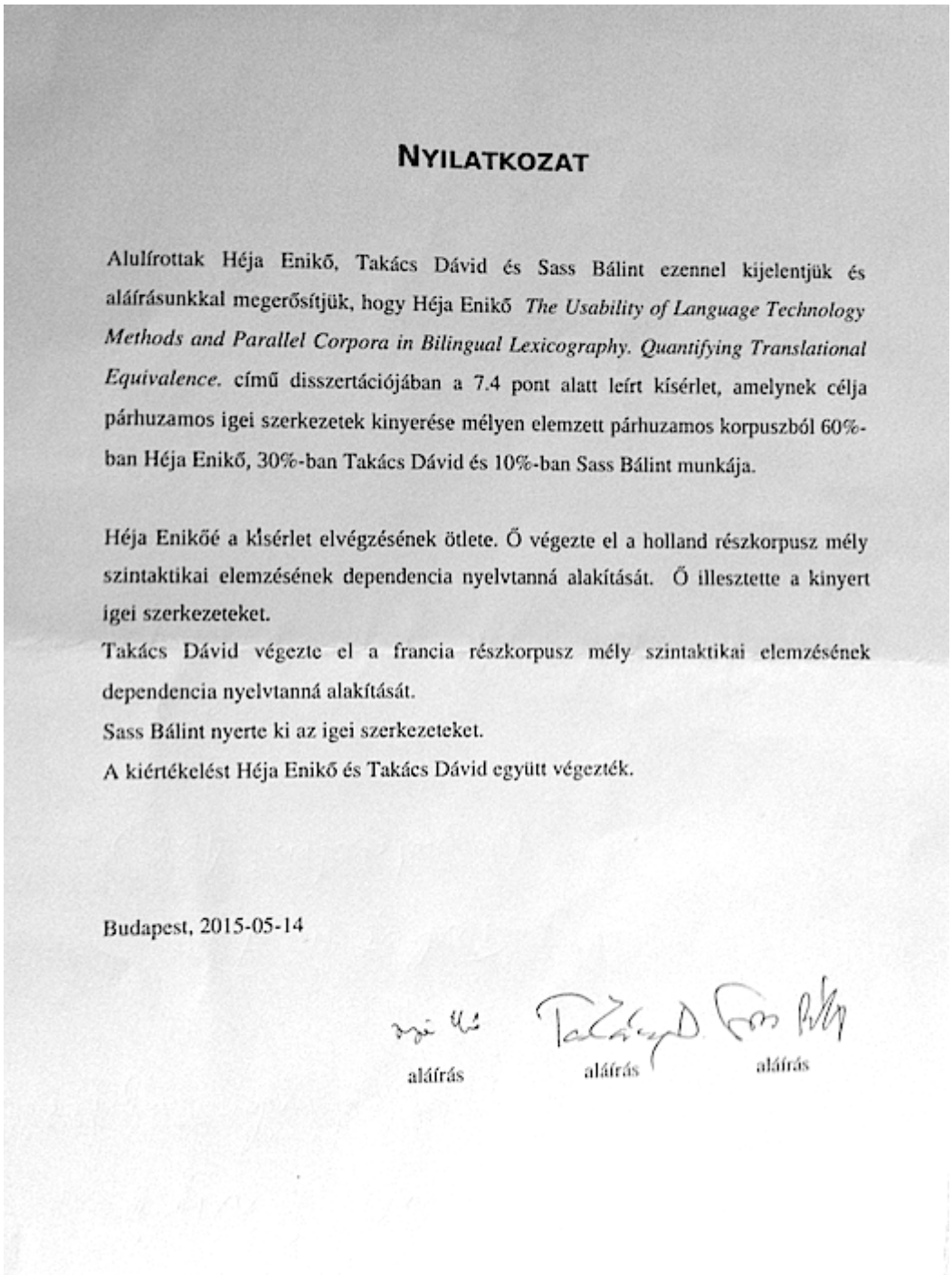
Nyilatkozat

Alulírottak Sass Bálint és Héja Enikő ezennel kijelentjük és aláírásunkkal megerősítjük, hogy Héja *Enikő The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence.* című disszertációjában a 7.3.1 pont alatt leírt párhuzamos igei szerkezetek kinyerését célzó kísérlet sekélyen elemzett párhuzamos korpuszból 70%-ban Héja Enikő és 30%-ban Sass Bálint munkája.

Sass Bálint végezte az igei szerkezetek kinyerését és az ehhez szükséges sekély szintaktikai elemzést. Minden más Héja Enikő munkája.

Budapest, 2015-05-14

aláírás aláírás



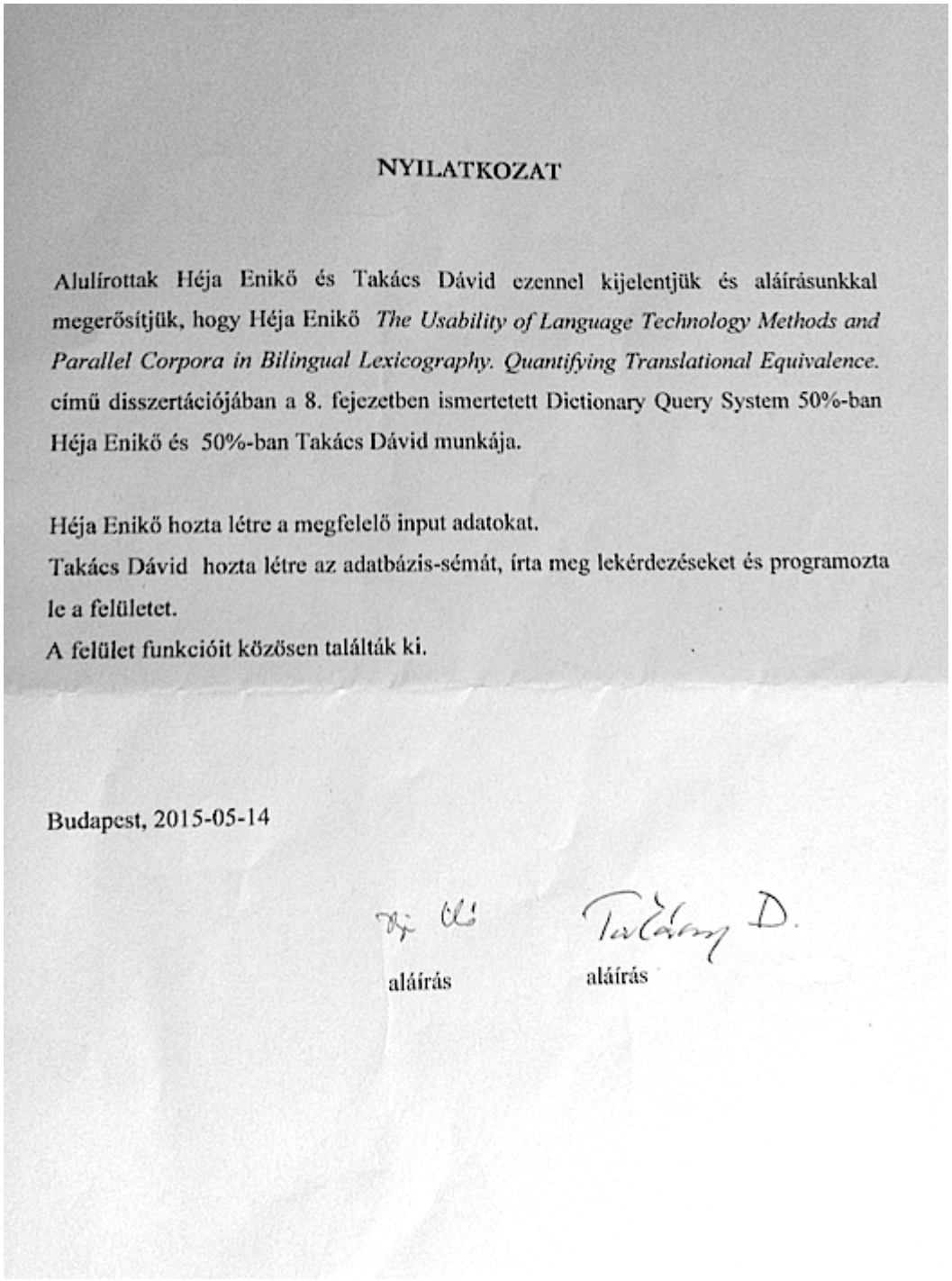
NYILATKOZAT

Alulírottak Héja Enikő, Takács Dávid és Sass Bálint ezennel kijelentjük és aláírásunkkal megerősítjük, hogy Héja Enikő *The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence.* című disszertációjában a 7.4 pont alatt leírt kísérlet, amelynek célja párhuzamos igei szerkezetek kinyerése mélyen elemzett párhuzamos korpuszból 60%-ban Héja Enikő, 30%-ban Takács Dávid és 10%-ban Sass Bálint munkája.

Héja Enikőé a kísérlet elvégzésének ötlete. Ő végezte el a holland részkorpusz mély szintaktikai elemzésének dependencia nyelvtanná alakítását. Ő illesztette a kinyert igei szerkezeteket.
Takács Dávid végezte el a francia részkorpusz mély szintaktikai elemzésének dependencia nyelvtanná alakítását.
Sass Bálint nyerte ki az igei szerkezeteket.
A kiértékelést Héja Enikő és Takács Dávid együtt végezték.

Budapest, 2015-05-14

aláírás aláírás aláírás



NYILATKOZAT

Alulírottak Héja Enikő és Takács Dávid ezennel kijelentjük és aláírásunkkal megerősítjük, hogy Héja Enikő *The Usability of Language Technology Methods and Parallel Corpora in Bilingual Lexicography. Quantifying Translational Equivalence.* című disszertációjában a 8. fejezetben ismertetett Dictionary Query System 50%-ban Héja Enikő és 50%-ban Takács Dávid munkája.

Héja Enikő hozta létre a megfelelő input adatokat.
Takács Dávid hozta létre az adatbázis-sémát, írta meg lekérdezéseket és programozta le a felületet.
A felület funkcióit közösen találták ki.

Budapest, 2015-05-14

aláírás aláírás