Improving Statistical Alignement And Translation Using Highly Multilingual Corpora

Camelia Ignat

École Doctorale Mathématiques, Sciences de l'Information et de l'Ingénieur

UdS – INSA

THÈSE

présentée pour obtenir le grade de

Docteur de l’Université de Strasbourg
Discipline : Informatique
Spécialité : Traitement Automatique des Langues

par

Camelia IGNAT

Improving Statistical Alignment and Translation Using Highly Multilingual Corpora

Amélioration de l'alignement et de la traduction statistique par utilisation de corpus parallèles multilingues

Soutenue publiquement le 16 Juin 2009

Membres du jury

Directeur de thèse :	M. François Rousselot, MC, Université de Strasbourg
Rapporteur externe :	M. Dan Tufiş, PR, Université de Bucarest
Rapporteur externe :	M. François Yvon, Professeur, Université de Paris-Sud
Examinateur :	M. Pierre Collet, Professeur, Université de Strasbourg

LGECO - LICIA **Unité EA 3938**

Improving Statistical Alignment and Translation Using Highly Multilingual Corpora

Ph.D. Thesis Dissertation

Camelia Ignat

Strasbourg, June 2009

Acknowledgements

Firstly, I would like to thank the advisor of this thesis, M. François Rousselot, for his confidence in me and for encouraging and supporting me all these years. Besides my advisor, I would like to thank the rest of my thesis committee, M. Dan Tufiş, M. François Yvon and M. Pierre Collet, for accepting to be part of the jury and for their willingness and patience.

I am grateful to the NLP group, from the Research Institute for Informatics and the Romanian Academy, and especially to Dan Tufiş, who initiated me into the field of Computational Linguistics and has provided assistance in numerous ways during my research activity with valuable advice and interesting ideas.

I warmly thank the Language Technology group of the JRC, where I worked for five years, for the stimulating and fruitful experience and, in particular, I am grateful to Ralf Steinberger and Bruno Pouliquen for the pleasant years of working together. I also thank all the authors of the free available tools that I used during this thesis (especially Moses toolkit).

Thanks also go to my professor, Florian Colceag, for illuminating me with a first glimpse of research, when I was at high school. I thank my best friend, Genţiana, and all my friends who encouraged me over all these years.

Last but not least, I would like to thank my family: my parents who trusted me, for their constant support, my daughter who "allowed" me to work on the thesis and constantly tried to "help" me and I would like particularly to thank my husband who encouraged me to finish this work, for helping me get through the difficult times, and for all the emotional support he provided. Without his understanding and support it would have been impossible to complete this thesis.

Remerciements

Je tiens tout d'abord à remercier mon directeur de thèse, M. François Rousselot pour m'avoir fait confiance et encouragée pendant toutes ces années ainsi que les autres membres de la commission d'évaluation, M. Dan Tufiş, M. François Yvon et M. Pierre Collet pour avoir accepté de faire partie du jury et d'avoir rapporté et examiné mes travaux ainsi que pour leur patience lors de la finalisation de ce mémoire.

Je voudrais également saluer les membres du groupe de recherche TAL de l'institut de recherche en informatique et de l'académie roumaine, et plus spécialement M. Dan Tufiş qui m'a fait découvrir la linguistique informatique et qui a toujours su m'apporter une aide précieuse en partageant ses idées tout au long de mes différentes activités de recherche.

Je remercie aussi toute l'équipe "Technologie du Langage" du JRC avec laquelle j'ai travaillé pendant cinq ans pour leur créativité et leur enthousiasme et plus particulièrement M. Ralf Steinberger et M. Bruno Pouliquen pour ces belles années de travail.

Je voudrais encore exprimer ma gratitude à tous les auteurs des programmes du domaine public que j'ai utilisés intensément durant cette thèse (spécialement 'Moses').

Merci à mon professeur M. Florian Colceag qui a su m'ouvrir les yeux sur la recherche lorsque j'étais encore au lycée et à ma meilleure amie Genţiana pour m'avoir encouragée pendant toute ces années.

Et un grand merci à toute ma famille : à mes parents pour m'avoir fait confiance et pour leur soutien, à ma fille pour m'avoir "authorisée" a travailler tard le soir et pendant les week-ends et particulièrement à mon mari qui m'a encouragée à terminer ce travail en m'aidant dans les moments difficiles et sans lequel je n'aurais sans doute jamais été en mesure d'achever cette thèse.

Abstract

This thesis presents the compilation of a highly multilingual parallel corpora (JRC-Acquis) and its usage to improve statistical alignment and translation by triangulation, which is the process of translating from a source to a target language via an intermediate third language. We explore heuristics to improve alignment and translation using multilingual, parallel, sentence-aligned corpora in several bridge languages. Our study offers two methods utilizing a bridge language to create a translation model, with a procedure for combining translation systems for multiple bridge languages. We present experiments showing that multilingual, parallel text in twenty-two languages can be used in this framework to improve statistical translation.

The motivation for this approach is two-fold. First, we believe that parallel corpora available in several languages provide a better training material for alignment systems relative to bilingual corpora. Word alignment systems trained on different language pairs make errors which are somewhat orthogonal. In such cases, incorrect alignment links between a sentence-pair can be corrected when a translation in a third language is available. Thus it can help to resolve errors in word alignment. We combine word alignments and translation models based on them using several bridge languages with the aim of correcting some of the alignment errors. The second advantage to this approach concerns the problem of data coverage. Current phrase-based statistical machine translation (SMT) systems perform poorly when using small training sets. When there are only small bilingual corpora between low-density language-pairs (like Romanian and Finnish), the triangulation allows the use of a much wider range of parallel corpora for training. Therefore, pivot alignment could be expected to make a positive and safe contribution in a word alignment system, i.e. increasing recall without lowering precision.

Kay[Kay, 2000] suggests that much of the ambiguity of a text that makes it hard to translate into another language may be resolved if a translation into some third language is available, and he suggests using multiple source documents as a way to inform subsequent machine translations. He calls the use of existing translations to resolve underspecification in a source text "triangulation in translation", but does not propose a method to go about performing this triangulation. The challenge is to find general techniques that will exploit the information in multiple sources to improve the quality of alignment and machine translation. [Callison-Burch et al., 2006] used pivot language for paraphrase extraction to handle the unseen phrases for phrased-based SMT. [Borin, 2000b] and [Wang et al., 2006] used pivot language to improve word alignment: [Borin, 2000b] used multilingual corpora to increase alignment coverage, and [Wang et al., 2006] induced alignment models by using two additional bilingual corpora to improve word alignment quality. Kumar, Och and Machery [Kumar et al., 2007] describe an approach to improve SMT performance where word alignment systems are combined from multiple languages by multiplying posterior probability matrices. An approach based on phrase table multiplication

is discussed in [Utiyama and Isahara, 2007, Wu and Wang, 2007]. Scores of the new phrase table are computed by combining corresponding translation probabilities in the source-pivot and pivot-target phrase-tables. [Bertoldi et al., 2008] gives a mathematical sound formulation of the various approaches and introduces two methods to train translation models through pivot languages (bridging at translation time and bridging at training time). [Cohn and Lapata, 2007] present a method that alleviates the coverage problem over source and target phrases, by exploiting multiple translation of the same source phrases.

Although related to their approach, our method is slightly different in terms of the implementation and the large coverage of languages. We propose two methods, one at the alignment level, and the other at the phrase-table level, both focusing on translation improvement. Our experiments cover a large number of language pairs and intermediate languages and constitute the basis for studying different factors that influence the alignment and translation via a pivot language: the training corpus size, the type of the intermediate language (the relatedness of the pivot language with the source and target language, poor or rich morphology). We designed a set of experiments to compare the methods proposed to demonstrate the importance of each of these features and to show how triangulated alignments or phrase-tables can be combined with the standard ones to improve the output of a statistical translation system.

The aim of this thesis is to explore how a highly multilingual parallel corpora could increase alignment and translation performances, using a bridge language. We developed methods to train and combine alignment models through pivot languages.

In pursuing the main goal, the following tasks have been distinguished:

Corpora compilation (JRC-Acquis and its sub-corpora): Documents and their multilingual translations have been collected and transformed into a format which can be used extensively and efficiently. This task involves downloading documents, format conversions, and some pre-processing, such as tokenization and sentence alignment. We selected sub-corpora that has been used in our experiments, as training data and as developement set.

Training baseline translation models: We used parallel corpora in 22 languages to create 462 translation systems for all possible language pairs. The resulting systems and their performances reveal the different challenges for the statistical machine translation.

Training alignment and translation models using a pivot language:The focus of the research presented is on the pivot methods in translation. We developed and explored two main methods (with slightly variations) for training alignment and translation models through pivot languages.

Application in SMT: experiments and evaluation: The final part contains the evaluation of our methods in statistical machine translation. We performed experiments that show the improvement brought by the usage of a pivot language and the influence of different factors on our models.

Parallel corpora are the essential data in our research and the JRC-Acquis corpus [Steinberger et al., 2006] was compiled while working on this thesis. JRC-Acquis is a unique and freely available parallel corpus containing European Union (EU) documents of mostly legal nature. It is available in 22, out of 23 official EU languages. The corpus contains 463,792 texts and a total of over one billion words. It consists of almost 21000 documents per language, with an average size of nearly 48 million words per language. Pair-wise paragraph alignment information produced by two different aligners (Vanilla and HunAlign) is available for all 231 language pair combinations. Most texts have been manually classified according to the EUROVOC subject domains so that the collection can also be used to train and test multi-label classification algorithms and keyword-assignment software. The corpus is encoded in XML, according to the Text Encoding Initiative Guidelines. Due to the large number of parallel texts in many languages, the JRC-Acquis is particularly suitable to carry out all types of cross-language research, as well as to test and benchmark text analysis software across different languages [Tufiş, 2007, Tufiş et al., 2008].

The JRC-Acquis corpus is a valuable data for our research, due to its highly multi-linguality (22 languages) and its size (to our knowledge it is the biggest parallel corpus). Furthermore, it provides resources for rare language pairs like Finnish-Maltese, or Romanian-Estonian.

We have created two different subcorpora of JRC-Acquis. The first one includes all the documents of JRC-Acquis corpus, that have been (manually) classified into “health” and “health-related” domains according to the EUROVOC thesaurus. The second one has been selected on the language availability basis: we have extracted all the documents that have translations in all the 22 languages of the JRC-Acquis corpus. They have been used in our experiments in order to study and validate the pivot approach.

Our research provides recipes to use a bridge language to construct a translation model and to combine translation models produced by multiple systems. We focus on the techniques from statistical machine translation because they form the basis of our methods, as SMT has become the dominant paradigm in machine translation in recent years and has repeatedly been shown to achieve state-of-the-art performance. Whereas the original formulation of SMT [Brown et al., 1993] was word-based, contemporary approaches have expanded to phrases. Phrased-based SMT [Koehn et al., 2003] uses larger segments of translated text, multi-word units, described as “substrings” or “blocks” since they just denote arbitrary sequences of contiguous words (and not syntactic constituents). The phrases are stored in a data structure called phrase table, as pairs of source phrase and their translations into the target language along with the value of their translation probabilities.

The phrase-table for each language pair is generated using a statistical machine translation system, Moses, that allows to train translation models automatically, based on a collection of translated texts. Moses [Koehn et al., 2007, Hoang and Koehn, 2008] is a complete out-of-the-box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models. It also contains tools for tuning these models using minimum error rate training [Och, 2003] and evaluating the resulting translations using the BLEU score [Papineni et al., 2002]. Moses uses standard external tools for some of the tasks to avoid duplication, such as GIZA++ [Och and Ney, 2003] for word alignments and SRILM[Stolcke, 2002] for language modeling.

Based on Acquis subcorpora and performing Moses tool, we trained translation models for the 22 language pairs in both directions (462 translation systems). The resulting systems and their preformances demonstrate the different challenges for statistical machine translation for different (non-traditional) language pairs.

We explore two heuristics for combining translation models using a pivot language. The first one proposes a procedure at the alignment level and the second one at the phrase table level.

As using Moses, our lexical scores are estimated on a training corpus which is automatically aligned using GIZA++ in both directions between source and target and symmetrized using the growing heuristic [Koehn et al., 2003]. Our first heuristic offers a procedure where this symmetrized alignment table between a language pair is combined with the alignment tables between the source and the pivot language, and between the pivot and the target language. Thus, we evaluate the enhancement produced by an intermediate language to alignment.

The second heuristic combines phrase tables and is evaluated in bilingual lexicon extraction and in machine translation. For a triad of languages we create the phrase table between the source and the pivot language and between the pivot and the target language. For each phrase entry we identify their translations into the intermediate language and then into the target language and we generate the triangulated phrase table. This leads to many errors and omissions in this table, but these problems can be tackled by using the triangulated phrase table in conjunction with a standard one. We suggest using the linear interpolation to combine two or more phrase tables.

We study the different factors that influence the performance of this method. Firstly, the size of the training data: on small data sets the performance gains with triangulation. The choice of the pivot language is also an important factor. The degree of relatedness of the languages in a triad seems to play a role in how well a pivot alignment will work: a high degree of similarity with the source or target language makes the intermediate language more effective. On the other hand, different pivot languages add different alignments. The more languages we add, the better the results become, i.e. different additional languages complement one another.

Our systems has been evaluated in the SMT context. Improving alignment quality is one way to improve translation models. Since the entries in the phrase table act as basis for the behaviour of the decoder - both in terms of the translation options available to it, and in terms of the probabilities associated with each entry - it is a common point of modification in SMT research. We evaluate the efficacy of using a pivot language by computing BLEU score.

We show that parallel corpora available in several languages provide a better training material for translation systems relative to bilingual corpora and it can be exploited to improve performance of an translation system. We combine translation models using several bridge languages with the aim of correcting some of the alignments errors (errors which are somewhat orthogonal) and to enhance the data coverage. We analyze the factors influencing the alignment results and translation models via a pivot language and evaluate the resulting systems in statistical machine translation.

Résumé

Notre thèse porte sur la constitution d'un corpus parallèle multilingue (JRC-Acquis) et son application à l'amélioration de l'alignement et de la traduction statistique par triangulation, processus de traduction d'une langue source vers une langue cible par le biais d'une langue tierce. Dans ce cadre, nous avons développé deux approches basées sur l'utilisation de corpus parallèles multilingues alignés au niveau des phrases dans plusieurs langues dites 'pivots'. Les deux méthodes proposées par notre étude permettent de générer un modèle de traduction par combinaison de plusieurs systèmes créés pour différentes langues pivots. Nous démontrons ainsi que des textes parallèles multilingues en vingt-deux langues peuvent améliorer sensiblement la traduction automatique.

L'intérêt de notre recherche est double. Tout d'abord, nous pensons que la mise à disposition de corpus parallèles dans un grand nombre de langues peut fournir une base d'entraînement plus performante aux systèmes d'alignement en comparaison avec les corpus bilingues classiquement utilisés. Les systèmes d'alignement au niveau des mots opérant sur des paires de langues déterminées produisent en effet des erreurs que l'on peut qualifier d'orthogonales. Or, dans de tels cas, l'alignement incorrect de deux phrases pourrait être corrigé si une traduction dans une troisième langue était disponible. Ceci permettrait alors de résoudre un certain nombre d'erreurs d'alignement. C'est dans cette perspective que nous avons combiné les alignements et les modèles de traduction issus de différentes langues pivots.

Le second avantage que nous espérons retirer de ces ressources concerne le problème de la couverture des données. Les systèmes statistiques de traduction automatique étant en général peu performants lorsqu'ils reposent sur des corpus limités, la triangulation pourrait permettre, pour des paires de langues ne disposant que de corpus à faible densité (comme la roumain et de le finnois), l'élargissement des données disponibles pour l'entraînement. L'alignement et la traduction par pivot devraient ainsi apporter une contribution significative aux systèmes de traduction en augmentant le rappel sans diminuer la précision.

[Kay, 2000] suggère que la majeure partie de l'ambiguïté d'un texte qui rend sa traduction difficile dans une autre langue peut être résolue par le recours à une troisième dont la traduction est également disponible et propose ainsi d'élargir les capacités des systèmes de traduction automatique par l'utilisation de documents 'multi-sources'. Il appelle cette utilisation de traductions tierces en vue de résoudre la sous-spécification dans un texte source 'Triangulation en traduction' mais ne propose toutefois pas de méthode concrète pour sa mise en œuvre. Notre idée consiste donc à présenter des techniques générales exploitant les informations fournies par ces 'multi-sources' de manière à accroître la qualité de l'alignement et de la traduction automatique.

[Callison-Burch et al., 2006] utilisent une langue pivot pour l'extraction de paraphrases, qui seront ensuite utilisées dans les tables de traduction, dans la perspective d'augmenter la couverture des systèmes de traduction statistique. [Borin, 2000b] et [Wang et al., 2006] se servent d'une langue pivot pour améliorer

l'alignement au niveau des mots : [Borin, 2000b] emploie des corpus multilingues pour augmenter la couverture de l'alignement et [Wang et al., 2006] induisent des modèles d'alignement en utilisant deux corpus multilingues supplémentaires dans le but d'améliorer la qualité de l'alignement des mots. Kumar, Och and Machery [Kumar et al., 2007] décrivent une méthode pour augmenter la performance des systèmes de traduction statistique, par l'intermédiaire de corpus parallèles disponibles dans différents langages pivots, dans laquelle, pour chaque langue pivot, les systèmes d'alignement sont combinés en multipliant les matrices de probabilité postérieure. Une approche basée sur la multiplication des tables de traduction est discutée dans [Utiyama and Isahara, 2007, Wu and Wang, 2007]. [Bertoldi et al., 2008] proposent une formulation mathématique des différentes approches existantes et présentent deux méthodes destinées à l'entraînement des modèles de traduction par recours à une langue pivot. Enfin, [Cohn and Lapata, 2007] réduisent le problème de la couverture au niveau des séquences de mots entre langue source et langue cible, en exploitant les tables de traduction pour différents langues pivots.

Bien que liée à leur approche, nos méthodes sont légèrement différentes quant à leur implementation et de part leur large multilinguisme. Nous proposons deux méthodes, l'une au niveau de l'alignement et l'autre au niveau des tables de traduction, les deux mettant l'accent sur l'amélioration de la traduction. Nos expériences portent plus particulièrement sur l'utilisation d'un grand nombre de paires de langues et de langues pivots et constituent une base d'étude des facteurs qui influencent l'alignement et la traduction par le biais d'une langue pivot, soit : la taille du corpus d'entraînement, les caractéristiques de la langue intermédiaire (la relation entre le pivot et la langue source / cible, morphologies pauvres ou riches). Nous avons effectué des expérimentations pour comparer nos méthodes. Nous démontrons ainsi l'importance de chacun des paramètres et analysons la façon dont ces alignements ou tables de traduction pivot peuvent être combinés avec des tables de traduction standard de manière à améliorer les résultats d'un système de traduction automatique.

L'objectif de cette thèse est d'étudier comment un corpus multilingue parallèle pourrait augmenter les performances d'alignement et de traduction, par le biais d'une langue pivot. Dans cette perspective, nous avons développé différentes méthodes pour entraîner et combiner plusieurs modèles de traduction par le biais de langues tierces.

Nous avons effectué pour cela les étapes suivantes :

La constitution des corpus parallèles (JRC-Acquis et ses sous-corpus) : Les textes de l'Acquis Communautaire et leurs traductions dans les différentes langues de l'Union Européenne ont été rassemblés et stockés dans un format facilement utilisable. Cette tâche implique le téléchargement des documents, leur conversion et des pre-traitements comme segmentation en phrases et alignement au niveau des phrases. Nous avons également sélectionné des sous-corpus qui ont été utilisés dans nos expérimentations (pour l'entraînement, le paramétrage et le test).

La construction des systèmes de traduction 'baseline' : Nous avons utilisé des textes parallèles en 22 langues pour construire 462 systèmes de traduction correspondant à toutes les combinaisons possibles de paires de langues. Les systèmes résultants et leur performance révèlent les différents défis de la traduction statistique.

La construction des modèles d'alignement et de traduction utilisant une langue pivot : Notre recherche est focalisée sur les méthodes pivots en traduction statistique. Nous avons développé et exploré deux méthodes principales (avec des légères variations) pour entraîner des modèles d'alignement et de traduction, par le biais d'une langue pivot.

Application dans la traduction automatique - Expérimentations et évaluation : La partie finale comprend l'évaluation de nos méthodes dans la traduction automatique. Nous avons effectué des expérimentations qui montrent l'amélioration apportée par l'utilisation d'une langue pivot et quelle est l'influence des différents paramètres sur nos modèles.

Les corpus parallèles constituent les données essentielles nécessaires à notre domaine de recherche. C'est dans cette perspective, que nous avons construit le corpus 'JRC-Acquis' utilisé dans le cadre de cette thèse. Celui-ci est un corpus parallèle, unique par sa taille et le nombre de langues couvertes (22 des 23 langues officielles de l'Union Européenne.). Il est disponible gratuitement et contient la plupart des documents de nature juridique de l'Union européenne (UE). Il contient au total 463.792 textes de loi, soit plus d'un milliard de mots. Par langue, il comporte en moyenne 21.000 documents, soit 48 millions de mots. L'alignement au niveau des paragraphes est issu des résultats produits par deux systèmes d'alignement (Vanilla et HunAlign) et est disponible pour l'ensemble des 231 combinaisons de langues. La plupart des textes ont été répertoriés grâce aux descripteurs EUROVOC, de manière à ce que le corpus puisse également être utilisé pour l'entraînement et l'évaluation d'algorithmes de classification automatiques et de logiciels d'attribution automatique de mots clés (keyword-based). Le corpus est encodé en XML, selon les 'Text Encoding Initiative Guidelines'. Du fait du grand nombre de textes parallèles disponibles et de si nombreuses langues, le JRC-Acquis est particulièrement adapté pour mettre en exécution des recherches multilingues, ainsi que pour tester et étalonner des logiciels d'analyse de textes multilingues.

Le corpus JRC-Acquis regroupe ainsi un ensemble de données précieuses pour notre recherche, de part son large multilinguisme (22 langues) et de part sa taille (il est, à notre connaissance, le plus grand corpus parallèle). De plus, il fournit des ressources pour des paires de langues rares, telles que finnois - maltais, ou estonien - roumain.

Nous avons crée deux sous-corpus du JRC-Acquis. Le premier comprend l'ensemble des documents du corpus JRC-Acquis qui ont été classés (manuellement) dans les domaines "santé" et "relatif à la santé" d'après le thésaurus EUROVOC. Nous avons constitué le second en nous basant sur la disponibilité par langue : nous avons extrait tous les documents possédant une traduction référencée dans les 22 langues du corpus JRC-Acquis. Nous avons utilisés ceux-ci dans nos expériences portant sur l'étude et la validation de l'approche pivot.

Notre recherche propose des méthodes basées sur l'utilisation d'une langue pivot pour produire un modèle de traduction, ainsi que sur la combinaison des modèles de traduction produits par différents systèmes. Lors de nos travaux, nous nous sommes concentrés sur les techniques de traduction automatique statistiques car celles-ci constituent la base des méthodes que nous utilisons. La traduction statistique [Brown et al., 1993] est en effet devenu le paradigme dominant en traduction automatique au cours de ces dernières années en prouvant de manière répétée sa capacité à réaliser des performances conformes à l'état de l'art actuel. Alors que les premiers algorithmes de traduction statistique se basaient sur les mots, les approches actuelles ont permis leur extension au niveau de séquences (de mots). Les systèmes de traduction à base de séquences [Koehn et al., 2003] utilisent des segments plus larges de texte traduits (des groupes de mots), définis comme 'sous-chaînes' ou 'blocs', car constitués de séquences de mots contiguës et pas sur une base syntaxique. Ces systèmes stockent alors l'ensemble des séquences dans une structure de données appelée 'table de traduction', par paires de séquences (originale/traduite) associées à une probabilité de traduction.

Nous avons généré ces tables pour chaque paire de langues en utilisant 'Moses', système de traduction statistique, qui permet de construire automatiquement des modèles de traduction sur la base d'une collection de textes parallèles. 'Moses' est un système de traduction *out-of-the-box* destiné à la recherche. Celui-ci regroupe l'ensemble des composants nécessaires à la préparation des données, à l'entraînement des modèles de langage et des modèles de traduction. Il contient également les outils nécessaires à l'optimisation de ces modèles à l'aide d'un entraînement à taux d'erreur minimum [Och, 2003] ainsi qu'un système d'évaluation des traductions obtenues reposant sur le score BLEU. 'Moses' recourt également à d'autres outils externes pour certaines tâches permettant d'éviter certaines duplications, tels que GIZA++[Och and Ney, 2003] pour l'alignement des mots ou encore SRILM [Stolcke, 2002] pour la modélisation du langage.

En nous basant sur les sous-corpus de l'Acquis et l'utilisation de 'Moses', nous avons entrainé des modèles de traduction pour les 22 paires de langues (462 systèmes de traduction). Les systèmes résultants et leur performances mettent en avant les différents défis à relever par la traduction statistique pour les paires de langues les moins étudiées.

Nous avons exploré deux heuristiques pour combiner plusieurs modèles de traduction en utilisant une langue pivot. La première propose une procédure au niveau de l'alignement et la seconde au niveau des tables de traduction.

Utilisant 'Moses' pour notre étude, les scores lexicaux ont été calculés à partir d'un corpus d'entraînement automatiquement aligné par GIZA++ dans les deux sens entre les langues source et cible, et après avoir été symétrisés selon 'l'heuristique de croissance'[Koehn et al., 2003]. Notre première heuristique propose une procédure où cette table d'alignement symétrisée entre deux langues est combinée avec les tables d'alignement entre langues source - pivot et pivot - cible. Nous évaluons ainsi l'amélioration produite par l'introduction d'un langage intermédiaire au niveau de l'alignement.

Notre seconde heuristique repose sur la combinaison des tables de traduction. Pour une triade de langues, nous construisons les tables de traduction entre les langues source - pivot, puis pivot - cible. Pour chaque phrase nous identifions leurs traductions dans la langue intermédiaire, puis dans la langue cible et générons la table de traduction triangulée. Appliquée telle quelle, cette approche basique pourrait conduire à de nombreuses erreurs et omissions mais nous parvenons à réduire ces problèmes en associant la table de traduction triangulée à une table standard par interpolation linéaire. Nous proposons ainsi d'utiliser l'interpolation linéaire de manière à combiner deux ou plusieurs tables de traduction.

Notre étude porte finalement sur les différents paramètres qui influencent les performances de nos méthodes. La taille du corpus d'entraînement est un des premiers facteurs car sur des ensembles réduits, la triangulation permet des gains de performances. Le choix de la langue pivot est également un facteur important. En effet, le degré de parenté des langues dans une triade joue un rôle sur alignement : un haut degré de similitude de la langue pivot avec la langue source ou cible augmente significativement l'efficacité de notre approche Enfin, l'ajout successif de langues pivots permet d'améliorer successivement l'alignement : plus nous utilisons de langues pivots, meilleurs sont les résultats, celles-ci se complétant les unes aux autres.

Nous avons évalué nos systèmes dans le contexte de la traduction automatique statistique. Augmenter la qualité de l'alignement, est un des moyen d'améliorer les modèles de traduction. En effet, les entrées de la table de traduction servant de base au décodeur, (tant en termes d'options de traductions offertes qu'en termes de probabilités associées) celles-ci constituent un paramètre classiquement étudié de la recherche en traduction statistique. Nous évaluons ainsi l'efficacité de l'utilisation d'une langue pivot dans ce contexte en utilisant le score BLEU.

Nous montrons dans notre thèse qu'un corpus parallèle disponible en plusieurs langues permet de fournir un meilleur matériel d'entraînement pour les modèles de traduction qu'un corpus bilingue classique et qu'il peut être exploité de cette manière pour améliorer les performances d'un système de traduction donné. Notre approche se base sur la combinaison de plusieurs modèles de traduction par recours à des langues pivots, dans le but d'une part, de corriger certaines erreurs d'alignement et d'une autre, d'améliorer la couverture des données. Nous analysons enfin les paramètres qui influencent l'alignement et les modèles de traduction lorsque nous passons par une langue pivot et évaluons nos résultats dans le domaine de la traduction automatique.

Contents

III Alignment and translation models 101

Part I

Preliminaries

Chapter 1

Introduction

Parallel corpora are a key resource as training data for statistical machine translation, and to build or extend bilingual lexicons and terminologies. Often in this context, more data is better data.

1.1 Motivation

We collected a highly multilingual parallel corpora (JRC-Acquis) and we explored how this resource can improve statistical alignment and translation. The view that is presented here is that multiple versions of a text can (and should) be seen as additional sources of information that can effectively be exploited to produce better billingual alignment.

The access to a multilingual corpora raises the following questions:

- Can anything be gained by viewing multilingual documents as more than just multiple pairs of translations?
- Can multilingual parallel translations help us to learn better about word alignment and translation models than we would with bilingual translations alone ?

Bilingual alignments have been used for a variety of purposes in a wide range of linguistics applications, and their usefulness as such is well established. However, as trilingual and multilingual aligned corpora are less widely used, their utility and distinctiveness is not as clear. Indeed one may ask whether there is any real use in mapping out translation equivalences between more than two languages. After all, in the vast majority of applications, such as machine translation, terminology and lexicography, the focus is on bilingual not multilingual correspondences.

What we intend to show is that while trilingual or multilingual text alignments may not be interesting in themselves, any additional version of a translated text should be viewed as additional information that can be leveraged to produce better bilingual alignments, and therefore a better knowledge of bilingual translational equivalences. In other words, whatever the intended application, three languages are better than two, or to put things idiomatically, the more translated languages at our disposal, the better!

Another important question is how to best combine these parallel sources of information in a principled statistical framework.

Central to our approach is triangulation, the process of translating from a source to a target language via an intermediate third language (pivot or bridge language).

The motivation for this approach is two-fold.

First, we believe that parallel corpora available in several languages provide a better training material for alignment systems relative to bilingual corpora. Word alignment systems trained on different language pairs make errors which are somewhat orthogonal. In such cases, incorrect alignment links between a sentence-pair can be corrected when a translation in a third language is available. Thus it can help to resolve errors in word alignment. We then combine word alignments using several bridge languages with the aim of correcting some of the alignment errors.

The second advantage to this approach is related to the problem of data coverage, from an application point of view. Current phrase-based Statistical Machine Translation (SMT) systems perform poorly when using small training sets. When there are only small bilingual corpora between low-density language-pairs (like Romanian and Finnish), the triangulation allows the use of a much wider range of parallel corpora for training. Therefore, pivot alignment could be expected to make a positive and safe contribution in a word alignment system, i.e. increasing recall without lowering precision.

Different pivot languages may catch different linguistic phenomena, and improve alignment and translation quality for the desired language pair in different ways.

1.2 Context

We are putting our work in the context of text alignment for statistical machine translation (SMT).

Machine translation and alignment are closely related problems [Lopez and Resnik, 2006]. State-of-the-art SMT is based on alignments between *phrases* (sequences of words in the source and target sentences). The learning step in these systems often relies on alignment between words. It is commonly assumed that the quality of the word alignment is critical for translation.

The dominant paradigm in SMT is referred to as phrase-based machine translation [Koehn et al., 2003]. In phrase-based models, the unit of translation is any contiguous sequence of words that we call a phrase. Each phrase in the target language is nonempty and translates exactly one nonempty phrase in the source language. This is done using a simple mechanism.

1. the source sentence is segmented into phrases.
2. each phrase is translated

3. the translated phrases are permuted into a final order.

The set of rules that governs this process is contained in a phrase table, which is a simple list of all source phrases and all their translations, with a number of associated statistics. The phrase table is learned from the training data.

Thus, in the phrase-based SMT framework, the translation task is split into two phases. The first phase induces word alignment over a sentence-aligned bilingual corpus and generates a translation model, and a second phase uses statistics over these predicted words to decode (translate) novel sentences. Our work deals with the first of these tasks.

The *phrase table* is at the center of the process, it is a list of phrases identified in a source sentence, together with potential translations. This can be done using word alignments by extracting all phrases that are consistent with the word alignment. The term *phrase* refers to a sequence of words characterized by its statistical, rather than grammatical, properties. Phrase in the source may overlap and also may have several translations, so that a subset of the table must, in general, be selected to make a translation of the sentence. The members of the subset must then be arranged in a specific order to give a translation. These operation are determined by statistical properties of the target language enshrined in the so-called *language model.*

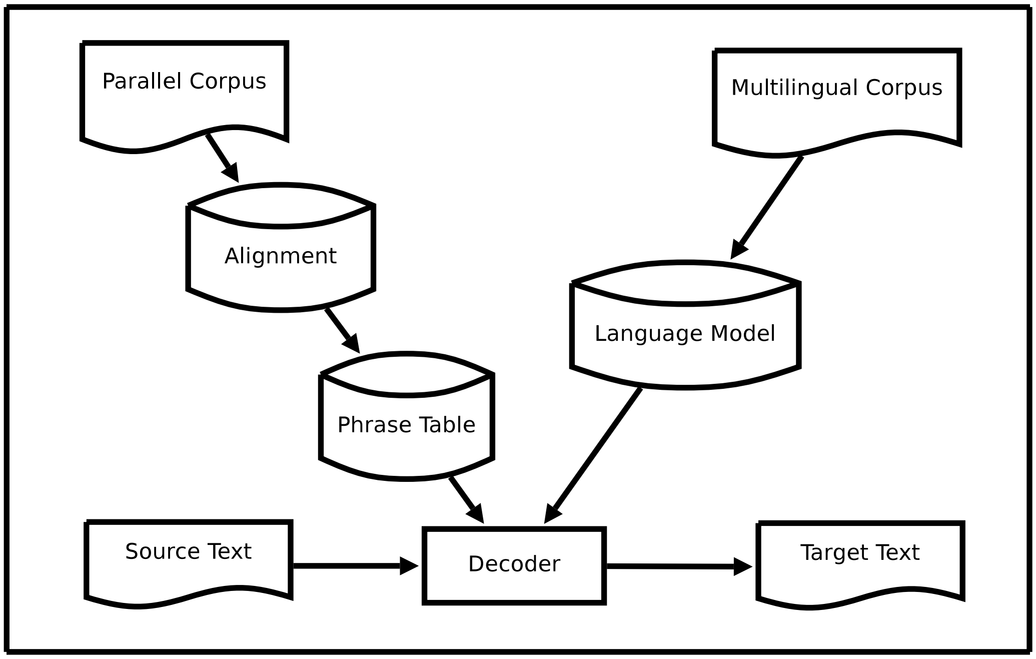


Figure 1.1: Phrase-based Statistical Machine Translation system

The constitution of the *phrase table* is determined by the translation model which captured the supposedly relevant statistical properties of a corpus consisting of paired source and target sentences. Very generally speaking, the faithfulness, or accuracy of a translation depends more on the translation model and its fluency on the language model.

Several current SMT systems work this way and our research is based on the most suited to our purposes, the freely available, open-source Moses Toolkit [Koehn et al., 2007]. We use the *phrase-based* SMT framework to develop pivot methods.

1.3 Aims and objectives

The aim of this thesis is to explore how a highly multilingual parallel corpora could increase alignment and translation performances, using a bridge language. We have developed methods for training and combining alignment models and translation models through pivot languages.

In pursuing the main goal, the following tasks have been acomplished:

1. **Corpora compilation (JRC-Acquis and its sub-corpora):** Documents and their multilingual translations have been collected and transformed into a format which can be used extensively and efficiently. This task involves downloading of documents, format conversions, and some pre-processing, such as tokenization and sentence alignment. We selected sub-corpora that has been used in our experiments, as training data and and as developement set.

2. **Training baseline translation models:** We used parallel corpora in 22 languages to create 462 translation systems for all possible language pairs. The resulting systems and their performances reveal the different challenges for the statistical machine translation.

3. **Training alignment and translation models using a pivot language**: The focus of the research presented is on the pivot methods in translation. We developed and explored two main methods (with slight variations) for training alignment and translation models through pivot languages.

4. **Application in SMT: experiments and evaluation**: The final part comprises the evaluation of our methods in statistical machine translation. We performed experiments that shows the improvement brought by the usage of a pivot language and the influence of different factors on our models.

1.4 Outline

The thesis is composed of four parts that include eight chapters presenting research that has been carried out these last few years. Some parts of the thesis elaborate on work published elsewhere [Steinberger et al., 2006, Erjavec et al., 2005]; the other parts contain recent, unpublished work that is described in detail in comparison with earlier achievements.

Part 1 : Preliminaries - Introduction and Framework

The framework, that follows this introduction, provides some background to the field of research on statistical machine translation and presents concepts that are relevant to our approach. It introduces basic terminology and includes a summary of related work. It sketches the important points and the contribution of our approach.

Part 2: JRC-Acquis and its sub-corpora - Corpora compilation and pre-processing

This part gives an overview of the parallel corpus (JRC-Acquis) and its sub-corpora, which has been collected, built and used in the thesis.

Part 3: Alignment and translation models

This part constitutes the main contribution of the thesis. Here we present the direct translation models and describe the pivot methods, followed by experiments and evaluation.

Part 5: Conclusions and further directions

This part concludes the thesis with a summary of contributions and some prospects for future work.

Chapter 2

Framework

This section describes the framework of our approach and introduces concepts from statistical machine translation that form its basis. We start by defining the term of parallel corpus as used in the thesis, in relation to other concepts of computational corpus linguistics.

2.1 Parallel corpora

In computational linguistics a corpus is a collection of spoken and written utterences of natural language usually accessible in electronic form. Often a corpus represents a particular genre of text or speech. Other corpora contain a large variety of types and genres to represent language used in a more general way.

There are several ways of classifying corpora into different types and categories according to their properties. One way is to distinguish between corpora that include only one language (*monolingual corpora*) and corpora that includes several languages (*multilingual corpora*). Multilingual corpora can be divided into parallel and comparable corpora[1]. A *parallel corpus* is a collection of texts, each of which is translated into one or more other languages. The simplest case is where only two languages are involved: one of the corpora is an exact translation of the other. Some parallel corpora, however, exist in several languages. The term *comparable corpora* refers to texts in two (or more) languages that are similar in content, but are not translations.

Parallel corpora usually contain a common source document (the original) and one or more translations of this source (target documents). Sometimes the original language is unknown (*mixed source corpora*) or the original document is not included at all (*multi-target corpora*) [Merkel, 1999].

[1] There are no multilingual corpora apart from parallel and comparable corpora; there are plenty of centres that have collected text material in several languages, and some of these collections are corpora in their own right. But unless the collections share common features of selection, at least at the level of the comparable corpus, then they are just text resources in different languages. It therefore seems unhelpful to use the term *multilingual corpus* (Sinclair).

In order to exploit a parallel text, some kind of text *alignment*, which identifies equivalent text segments, is a prerequisite for analysis. A large number of methods were proposed for aligning text at different levels. (i.e., mapping the units that translate each other). The units in question include paragraphs, sentences, words and expressions. The bilingual aligned parallel texts are sometimes called *bitext*, and the term *multitext* is used to refer to parallel text in more than one language, as mentioned in [Véronis, 2000].

Parallel corpora are a prime resource for the development of multilingual language technologies. Serving as training datasets for inductive programs, they can be used to learn models for machine translation, cross-lingual information retrieval, multilingual lexicon extraction, sense disambiguation, etc. The value of a parallel corpus grows with the following characteristics:

• **Size**: larger corpora give not only statistically more reliable counts, but also reveal phenomena that are completely lacking in smaller samples.

• **Number of languages**: the utility here grows quadratically with the number of languages, as each language can be paired with any other. While bilingual corpora usually contain at least one 'major' language, larger multilingual collections will also contain pairings of less common languages, where such a resource is of great value (Maltese-Finish for example).

• **Linguistic annotation**: can be used as a normalisation step on the raw text, hence reducing the complexity (search space) of the LT task; or to enable multiple knowledge of the text (e.g. morphosyntactic tags, collocations, predicate-argument structure) to be exploited.

• **Semantic annotation**: refers to the classification of documents (or their parts, e.g. words) into some hierarchy of concepts, which can be used to access the data (e.g. the Semantic Web paradigm).

2.1.1 Available parallel corpora

Many projects aiming at compiling parallel text corpora have sprung around the world. Parallel corpora are leveraged in the business of communication in multilingual societies, such as the United Nations, the NATO, the European Union and officially bilingual countries such as Canada.

The **Hansard corpus** (French-English) is no doubt the first and in any case the most famous of all parallel corpora. Collected during the eighties by groups such as Bell Communications Research and the IBM T.J.Watson Research Center, this corpus contains over fifty million words taken from transcriptions of debates in the Canadian Parliament between the mid-seventies and 1988. It has been used in many studies, and over the years, has become a *de facto* gold standard for developing and testing systems. However, its limitation to one type of text and to one pair of languages has made it necessary to collect other data.

The last release[2], from 2001, contains 1.3 million pairs of aligned text chunks (sen-

[2] `http://www.isi.edu/natural-language/download/hansard/index.html`

tences or smaller fragments) from the official records (Hansards) of the 36th Canadian Parliament.

Multilingual parallel corpora with translations into more than one language are available and became very popular in recent studies. Due to their high cost, aligned (and verified) texts are much less common than unaligned ones.

The two main institutions for the distribution of corpora are the Linguistic Data Consortium [3] and the European Language Resource Association [4]. Their catalogues contain some available parallel corpora.

We will present two multilingual parallel corpora, comparable with the JRC-Acquis corpus whose description is detailed in the next chapter.

The **MULTEXT-East** language resources[5] presented in [Erjavec et al., 1996] is a multilingual dataset for language engineering research and development, first developed in the scope of the EU MULTEXT-East project as mentioned in [Dimitrova et al., 1998], that has now already reached its 3rd edition [Erjavec, 2004]. This standardised (XML/TEI P4, [Sperberg-McQueen and Burnard, 2002]) and linked set of resources covers a large number of mainly central and eastern european languages and includes annotated parallel, comparable and speech corpora with morphosyntactic lexica and specifications. The most important component is the linguistically annotated corpus consisting of Orwell's novel "1984" in the english original and translations, about 100,000 words in length. The translations of "1984" have been automatically sentence aligned with the original english text, and the alignments hand-validated. The languages included are: Bulgarian, Czech, English, Estonian, Hungarian, Romanian, Slovene, Lithuanian, Serbian, and Russian. This dataset, unique in terms of languages and wealth of encoding, is extensively documented, and freely available for research purposes.

The **Europarl corpus** [6] presented in [Koehn, 2005] is a collection of the proceedings of the European Parliament, dating back to 1996. It includes versions in 11 European languages: Romanic (French, Italian, Spanish, Portuguese), Germanic (English, Dutch, German, Danish, Swedish), Greek and Finnish. Altogether, the corpus comprises of about 30 million words for each language. The corpus has been collected mainly to aid the research in statistical machine translation and it is used by the Machine Translation community for the Shared Task of Workshops in SMT (2006-2009) [7] .

[3] http://www.ldc.upenn.edu/

[4] http://www.icp.grenet.fr/ELRA/home.html

[5] http://nl.ijs.si/ME/V3/

[6] http://www.statmt.org/europarl/

[7] http://www.statmt.org/wmt06/, http://www.statmt.org/wmt07/, http://www.statmt.org/wmt08/, http://www.statmt.org/wmt09/.

2.1.2 Applications

In recent years, parallel corpora have become more widely available and serve as a source for data-driven Natural Language Processing (NLP) tasks. Their applications are extremely diverse and include compiling translation memories, deriving dictionaries and bilingual terminology lists, extracting knowledge for cross-language information retrieval, retrieving examples for computer assisting teaching or contrastive linguistics, statistical machine translation, etc.

In this subsection, we will list some applications that have been based on parallel corpora. Note that this description is not intended as a comprehensive list of tools and projects on this subject.

Sentence aligned parallel corpora are directly applicable to support translators in their daily work. **Translation Memories** have been used for a long time by human translators and sentence aligned bitexts can be used as such without any further processing. Extending the functionality of translation memories by aligning even sub-sentential parts leads to the idea of Example-Based Machine Translation [Brown, 1996].

The idea of reusing translation fragments for **Machine Translation** (MT) seems to date back to the late seventies. The research trend called *Memory-Based Machine Translation* (MBMT) or *Example-Based Machine Translation* (EBMT) began in the mid-eighties [Nagao, 1984, Sumita and Tsutsumi, 1988, Sadler, 1989a, Sadler, 1989b, Sato and Nagao, 1990, Sumita et al., 1990]. The basic idea behind this type of translation is to search a translation sample database for fragments similar to certain portions of the text to be translated, and then combine them in an appropriate way—which may require defining a set of highly complex rules. Another line of research started up at about the same time, in particular at IBM, where the goal was to get rid of this complexity by letting the machine "learn" automatically, based on statistical models. Accordingly, [Brown et al., 1988, Brown et al., 1990] who to some extent took up on [Weaver, 1949]'s initial idea, proposed a translation model for which they estimated the parameters of 40,000 sentence pairs drawn from the Hansard corpus (French-English). The results were surprisingly good for such a simple model. Various improvements discussed in [Arad, 1991, Brown et al., 1992] demonstrated the validity of the approach.

Statistical Machine Translation (SMT) systems have become even more popular due to recent improvements of translation models and the increased power of today's computer technology. SMT systems present the advantage that they can be developed very fast once there are tools and sufficient training data available for the particular language pair. SMT systems have the disadvantage that they rely on training and the statistical model. Corrections and improvements are hard to integrate in the set of estimated parameters which are usually not human readable. We will give a more detailed description of SMT framework in section 2.3.

Another obvious application of parallel corpora is the **Extraction of Bilingual Terminology**. Several systems have been developed using word alignment techniques as described above. `Termight` uses Church's character-based alignment approach char align [Dagan and Church, 1994], `TransSearch` uses IBM's model 2

[Macklovitch and Hannan, 1996], and Champollion uses Smadja's collocation aligner [Smadja et al., 1996]. Terminology extraction techniques have successfully been ported to a variety of language pairs among them less related languages such as English and Japanese [Fung and McKeown, 1997] or English and Chinese [Wu and Xia, 1994]. They have been applied in different domains, like the medical one [Deléger et al., 2009, Langlais et al., 2008].

Related to terminology extraction is the field of **Lexicography**. The use of bilingual data to build translation dictionaries has been investigated in several projects. BICORD is one example that combines information derived from a bilingual dictionary with information extracted from a parallel corpus, and shows how it can be applied to the study of verbs of movement [Klavans and Tzoukermann, 1990]. Dilemma is another lexicographic tool that re-uses existing translations [Karlgren et al., 1994]. Many more projects aim at the automatic or semi-automatic extraction of bilingual lexicons for different language pairs [Resnik and Melamed, 1997, Ribeiro et al., 2001, Ahrenberg et al., 2002, Tufiş et al., 2004a].

Many authors have worked on **Extracting Dictionaries** of single words, mostly using statistical methods [Dagan et al., 1993, Wu and Xia, 1994, Dagan and Church, 1994, Melamed, 1997b, Resnik and Melamed, 1997]. Very quickly, though, researchers began to focus on units longer than the graphic word, such as collocations, expressions, and phraseology. Complex units like these are one of the major weakness of standard dictionaries. Many authors have attempted to extract such complex units from aligned texts: [Kupiec, 1993, Smadja et al., 1996, Melamed, 1997a, Hiemstra, 1998, Gaussier, 1998].

Another field of research where parallel data can help is the field of **Word Sense Disambiguation**. Ambiguities are distributed differently in natural languages. This fact can be used for cross-lingual comparisons, which may help to disambiguate words and to identify concepts in context [Gale et al., 1992, Diab and Resnik, 2002, Tufiş et al., 2004b].

Another application of parallel corpora to be mentioned here is the **adaptation of language tools to new languages** with the help of parallel data. Robust Text Analysis tools, which exist for one language, can be ported to other languages by projecting analyses (such as part-of-speech and chunks) from one language to another in a parallel corpus [Borin, 2000a, Yarowsky et al., 2001, Borin, 2002].

Finally, the **Pivot Methods** can also be mentioned, in which a third language may be used to induce sentence or word alignments between two other languages [Simard, 2000, Borin, 2000b]. We will detail the approaches based on such a *pivot language* in the section 2.4.

2.2 Alignment

Source language documents in a translation corpus can be split into segments that correspond (monotonically) to segments in translated documents. Establishing links

Link type	English sentence(s)	French sentence(s)
1:1	A Joint Committee is hereby established which shall be responsible for the administration of the Agreement and shall ensure its proper implementation.	Il est institué un comité mixte qui est chargé de la gestion de l'accord et qui veille à sa bonne exécution.
1:2	For this purpose, it shall make recommendations and take decisions in the cases provided for in the Agreement.	A cet effet, il formule des recommandations. Il prend des décisions dans les cas prévus à l'accord.
1:1	These decisions shall be put into effect by the Contracting Parties in accordance with their own rules.	L'exécution de ces décisions est effectuée par les parties contractantes selon leurs règles propres.

Table 2.1: Sentence alignment from Acquis Communautaire corpus, between English and French version

between corresponding segments is called *alignment*.

When establishing the correspondence between two reciprocal translations, there are two levels of alignment to be considered: *sentence alignment* and *word alignment*. The sentence alignment of the parallel corpora is a prerequisite to any multilingual NLP setting. Word alignment is more complex than sentence alignment and is mainly used to build the translation models of SMT systems. The latest phrase-based or syntactically motivated translation systems use word alignment as a prerequisite step.

2.2.1 Sentence alignment

Sentence alignment is a well established task which does not exclusively refer to 1-to-1 alignments. Sentence boundaries may vary in different translations. However, it usually assumes that information at the sentence level is expressed in the same order in the original document as in its translations. With this assumption, sentence alignment can be modelled as a monotonic mapping process, i.e. an alignment without crossing links. A sample of a sentence aligned bitext is given in the table 2.1.

Several approaches to automatic sentence alignment have been proposed. The main approaches apply either *length based models* using correlations between the lengths of corresponding sentences [Gale and Church, 1991b, Gale and Church, 1993, Brown et al., 1991], or *models based on lexical anchoring*, using correspondences between words and other lexical units [Kay and Röscheisen, 1993], or combinations of both. [Langlais and El-Beze, 1997] stressed the importance of combining different sources of information (lexicon, cognates, sentence length, matching frequencies) and the necessity of having an adequate model to choose the best combination. Enhancements and combinations of sentence alignment techniques can also be found in the literature, e.g. [Simard et al., 1993].

Automatic sentence alignment is known as a task that can be accomplished with high accuracy, above 90%. The systems evaluated in ARCADE evaluation project [Véronis and Langlais, 2000] achieved a success rate of 98.5% on "clean" texts. However, improvements are still possible in the most difficult cases, when "noisy" texts, including divergent and incomplete translations, are processed.

Last, the use of more than two languages is explored in [Simard, 1999] where he shows that paired alignment is not optimal and that the simultaneous alignment of several languages can improve the overall results.

2.2.2 Word alignment

Linking corresponding words and phrases in parallel corpora is usually called *word alignment*. The type of relation between words varies in parallel texts. Texts contain many tokens that are related in complex ways (compound words, idiomatic expressions, phraseology) and no true alignment or extraction of any quality can be done at the lexical level without taking such phenomena into account.

Furthermore, the strategy of aligning words and phrases in parallel corpora depends on the task to be accomplished. Usually, word alignment aims at a complete alignment of all lexical items in the corpus, i.e. the goal is to break each bitext segment into sets of corresponding lexical items. This often leads to "fuzzy" translations relations between certain words [Merkel et al., 2002, Véronis, 1998, Och and Ney, 2000] due to lexical differences, structural and grammatical differences, paraphrased translations, spelling mistakes, and other divergent translations. The alignment between two word strings can be quite complicated. Often, an alignment includes effects such as reorderings, omissions, insertions, and word-to-phrase alignments. The degree of correspondence can be expressed in terms of alignments probabilities, which is useful for many tasks, such as Machine Translation. Bilingual lexicon extraction aims at the identification of lexical word type links in parallel corpora. These links can be inferred from word alignments.

There are generally two approaches to word alignment, the *association* [Tiedemann, 2003] or *hypothesis testing* [Hiemstra, 1998] *approach* using measures of correspondance of some kind, and the *estimation approach* using probabilistic translation models. Association approaches are also referred to as *heuristic approaches* and estimation approaches are often called *statistical alignments* [Och and Ney, 2003].

A common idea behind the **Heuristic Methods** is to test if two words cooccur significantly more often than it would be expected if they would occur purely by chance. These methods [Gale and Church, 1991a, Smadja et al., 1996, Tiedemann, 1998, Ahrenberg et al., 2000, Melamed, 2001] produce pairs of translation candidates, extracted from corresponding segments of the parallel texts, each of them being subject to an independence statistical test. The translation candidates that show an association measure higher than expected under the independence assumption are assumed to be translation pairs. The translation pairs are extracted independently and therefore the process might be characterized as a local maximization (*greedy*) one.

The **Statistical Alignment Model** or estimation approach [Brown et al., 1993, Kay and Röscheisen, 1993, Kupiec, 1993, Hiemstra, 1998] is based on building a statistical bitext model from data, the parameters of which are to be estimated according to a given set of assumptions. The bitext model allows for global maximization of the translation equivalence relation, considering not individual translation equivalents but sets of translation equivalents. Most work in this field has been inspired by the work on statistical machine translation introduced in [Brown et al., 1990]. As we chose to follow this approach for word alignment, we will describe more precisely statistical word alignment produced by `Giza++` [Och and Ney, 2003] in the section 2.3.3.1.

Combination of these two methods for word aligment systems in bitext correspondences identification were developed as well. [Tufiş et al., 2005, Tufiş et al., 2006] showed that through combining two aligners, one based on hypothesis testing approach and the other closer to the estimation approach, the results are significantly improved compared to those obtained by each individual aligner.

Pros and cons for each type of approach are discussed in [Hiemstra, 1998] and [Och and Ney, 2003]. [Och and Ney, 2003] consider that the main advantage of the heuristic models is their simplicity as they are very easy to implement and understand. Therefore, variants of the heuristic models are widely used in the word alignment literature. Nevertheless, one problem with heuristic models is that the use of a specific similarity function seems to be completely arbitrary and the literature contains a large variety of different scoring functions, some including empirically adjusted parameters. For this reason, in their view, the approach of using statistical alignment models is more coherent. The general principle is to come up with an association score between words results from statistical estimation theory, and the parameters of the models are adjusted to maximize the likelihood of the models on the training corpus.

2.2.3 Evaluation of word alignment

It is common to evaluate word alignment intrinsically, by comparison with alignments prepared by human annotators, although sometimes task-based evaluation might be preferable, depending on the purpose of the alignment experiment.

The **Automatic Evaluation** using a reference alignment (named *gold standard*) is often preferred over manual a posteriori evaluation. The main advantage of reference alignments is their re-usability once they are created, while the main difficulty is to produce representative samples of reliable reference alignments. Most of these test sets contain a few hundred sentences and are available in several languages [Melamed, 1998b, Och and Ney, 2000, Mihalcea and Pedersen, 2003]. Ideally, each sentence is aligned by multiple annotators and the results are combined in some way. In much of the reported literature, the annotations contain two sets of links. The *Sure* set S contains links about which all annotators agreed. The *Probable* set P is a superset of S that additionally contains links about which annotators disagreed or expressed uncertainty about, such as “idiomatic expressions, free translations, and missing function words” [Och and Ney, 2000].

The metrics described below have been typically used in recent literature and for the evaluation measures of the HLT-NAACL 2003 [Mihalcea and Pedersen, 2003] and ACL 2005 Workshops on “Building and Using Parallel Texts” [Martin et al., 2005]. Automatically computed alignments (alignments to be evaluated) are compared to a manually aligned reference corpus (gold standard) and scored with respect to **precision**, **recall**, **F-measure**[8] and **Alignment Error Rate** (AER).

The *precision* is defined as the proportion of computed links that are present in the reference. The *recall* is the proportion of reference links that were computed (eq. 2.1). The *F-Measure* (eq. 2.2) is a way of combining both metrics [Van Rijsbergen, 1979]. Finally, the *AER* (eq. 2.3), introduced by [Och and Ney, 2000] to take into account the ambiguity of the manual alignment task, involves unambiguous links (set S or Sure) and ambiguous links (set P or Probable). If there is a P link between two words in the reference, a computed link between these words is acceptable, but not compulsory. On the contrary, if there is an S link between these words in the reference, a computed link becomes compulsory.

The measures which are defined are the following:

$$Precision = \frac{|aligned \cap probable|}{|aligned|}, \qquad Recall = \frac{|aligned \cap sure|}{|sure|} \tag{2.1}$$

$$Fmeasure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.2}$$

$$AER = 1 - \frac{|aligned \cap sure| + |aligned \cap probable|}{|aligned| + |sure|} \tag{2.3}$$

where *aligned* is the computed alignment, *sure* is the set of unambiguous (or sure) links and *probable* is the set of ambiguous (or probable) links in the reference gold standard.

If only one type of links is considered in the alignment reference, 2.3 becomes:

$$AER_1 = 1 - \frac{2 * Precision * Recall}{Precision + Recall} = 1 - Fmeasure \tag{2.4}$$

It has been shown that the percentage of Sure and Probable links in the gold standard reference has a strong influence in the final AER result, favouring high-precision alignments when Probable links outnumber Sure links, and favouring high-recall alignments otherwise [Lambert and Castell, 2004]. A well-founded criterion is

[8] A balanced F-measure is often used to combine both precision (P) and recall (R) for a comparison of the overall performance. This is derived from the weighted F-measure, which is defined as the ratio $F_\beta = ((\beta^2 + 1) * P * R)/(\beta^2 * P + R)$. Setting $\beta = 1$ “balances” precision and recall, i.e. both rates are weighted to be equally important.

to produce Probable links only when they allow combinations which are considered equally correct, as a reference with too many Probable links suffers from a resolution loss, causing several different alignments to be equally rated. Therefore, detailed guidelines are necessary for manual annotators when creating gold standards [Lambert et al., 2005, Véronis, 1998, Melamed, 1998a].

Application-Oriented Evaluations may also be considered. For instance, in lexicon extraction, the focus is on content words, whereas function words may be neglected. The evaluation measures of the `ARCADE` [Véronis and Langlais, 2000] word alignment track were tailored towards the task of translation spotting, i.e. the search for proper translations of the given source language terms. In SMT, word alignment is measured by its contribution to parameter estimation of our translation models (see section 2.3.3.1). If one alignment method produces a better translation system than another, we might conclude that it is more accurate overall.

Nowadays, due to a lack of perfect correlation between AER and translation evaluation scores observed in many experiments, alternative word alignment evaluation metrics are being pursued [Ayan and Dorr, 2006, Fraser and Marcu, 2007]. [Fraser and Marcu, 2007] found that the use of Probable links reduced the ability of alignment metrics to predict translation accuracy and recommends an annotation style that does not contain them [Melamed, 1998b].

We will focus next on the techniques from Statistical Machine Translation, as they form the basis for our alignment method via a pivot language.

2.3 Statistical Machine Translation

> "*It is very tempting to say that a book written in Chinese is simply a book written in English which was coded into the "Chinese code." If we have useful methods for solving almost any cryptographic problem, may it not be that with proper interpretation we already have useful methods for translation?*"
> – Warren Weaver
>
> (in [Lopez, 2007] - A Survey of SMT)

Machine translation (MT) is the automatic translation from one natural language into another using computers. Interest in MT is nearly as old as the electronic computer. Popular accounts trace its modern origins to a letter written by Warren Weaver in 1949, only a few years after the Electronic Numerical Integrator And Computer (ENIAC) came online [Weaver, 1949].

Statistical Machine Translation (SMT) is an approach to MT that is characterized by the use of machine learning methods. SMT has come to dominate academic MT research, and has gained a share of the commercial MT market. Since its revival more than a decade ago when IBM researchers presented the Candide SMT system [Bro90, Bro93], the statistical approach to machine translation has seen an increasing interest among both natural language and speech processing research communities. Mainly, three factors account for this increasing interest:

- There is a **growing availability of parallel texts** (though this applies, in general, only to major languages in terms of presence in internet), coupled with increasing computational power. This enables research on statistical models which, in spite of their huge number of parameters (or probabilities) to estimate, are sufficiently represented in the data.
- The **statistical methods are more robust to speech disfluencies or grammatical faults**. As no deep analysis of the source sentence is done, these systems seek the most probable translation hypothesis for a given source sentence, assuming the input sentence is correct.
- And last but not least, shortly after their introduction, these **methods** proved at least as **good or even better as rule-based approaches** in various evaluation campaigns.

We will then firstly show the place of SMT in the general classification of MT system, before describing the main methods of this approach in the section 2.3.3.

2.3.1 Approaches to MT

Several criteria can be used to classify Machine Translation approaches, yet the most popular classification is done according to the level of linguistic analysis (and generation) required by the system to produce translations. Usually, this can be graphically expressed by the machine translation pyramid in Fig. 2.1.

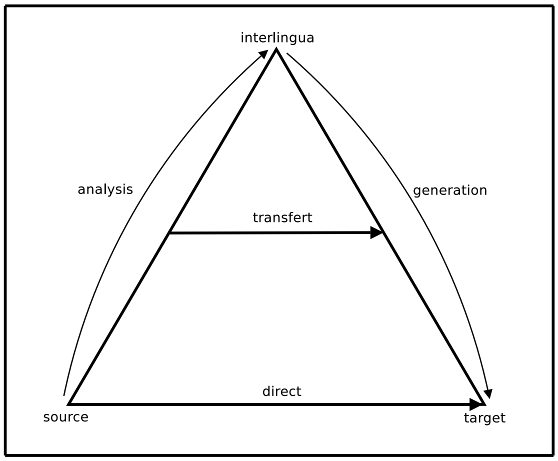


Figure 2.1: Machine Translation Pyramid

Generally speaking, the bottom of the pyramid represents those systems which do not perform any kind of linguistic analysis of the source sentence in order to produce a target sentence. Moving upwards, the systems which carry out some analysis (usually by means of morphosyntax-based rules) are to be found. Finally on top of the pyramid, a semantic analysis of the source sentence turns the translation task into generating a target sentence according to the obtained semantic representation.

Aiming at a concise survey rather than a complete review, we will next discuss each of these approaches briefly, before delving into the statistical approach to Machine Translation.

Interlingua-based translation. The *interlingua* idea is based on the mapping of the input into a language independent representation of its meaning. This approach advocates the deepest analysis of the source sentence, reaching a language of semantic representation named Interlingua. This conceptual language, which needs to be developed, has the advantage that, once the source meaning is captured by it, in theory we can express it in any number of target languages, so long as a generation engine for each of them exists. Though conceptually appealing, several drawbacks make this approach unpractical. First of all the difficulty of creating a conceptual language capable of bearing the particular semantics of all languages is an enormous task, which in fact has only been achieved in very limited domains. Apart from that, the requirement that the whole input sentence needs to be understood before proceeding onto translating it, has proved to make these engines less robust to the grammatical incorrectness of informal language, or which can be produced by an automatic speech recognition system.

Transfer-based translation. The rationale behind the *transfer-based* approach is that, once we grammatically analyse a given sentence, we can pass this grammar on to the grammatical representation of this sentence in another language. In order to do so, rules to convert source text into some structure, rules to transfer the source structure into a target structure, and rules to generate target text from it are needed. Lexical rules need to be introduced as well. Usually, rules are collected manually, thus involving a great deal of expert human labour and knowledge of comparative grammar of the language pair. Apart from that, when several competing rules can be applied, it is difficult for the systems to prioritise them, as there is no natural way of weighing them. This approach was massively followed in the eighties, and despite much research effort, high-quality MT was only achieved for limited domains [Hutchins and Somers, 1992].

Direct Translation. This approach solves translation on a word-by-word basis, and it was followed by the early MT systems, which included a very shallow morphosyntactic analysis. These approaches included initially the rule-based approach and corpus-based approaches (such as Example-Based Machine Translation and Statistical Machine Translation).

Typically, the rule-based systems are ad-hoc systems built with only one language pair in mind, that perform simple (but reliable) operations adapted to the specificities of that language pair. One of the problems of rule-based direct systems is that they hit a ceiling at which they become so complex that the addition of any rule causes as much degradation as enhancement. To reduce the complexity of the rule system, some aspects of the transfer approach can be introduced. Thus, the original versions of the `Systran` system [Toma, 1976], in operation since the seventies, used a direct approach, but the many modifications have transformed it in a rather transfer-based system.

Today, the direct translation approach has been almost abandoned, even in the framework of corpus-based approaches: although SMT initially worked on a word-to-word basis and could therefore be classified as a direct method, nowadays several engines attempt to include a certain degree of linguistic analysis into the SMT approach, slightly climbing up the aforementioned MT pyramid.

2.3.2 Corpus-based approaches

Many *corpus-based* approaches sprung at the beginning of the nineties. These systems extract the information needed to generate translations from parallel corpora that include many sentences which have already been translated by human translators. The advantage is that, once the required techniques have been developed for a given language pair, it should in theory be relatively simple to transpose them to another language pair, as long as sufficient parallel training data is available. Thus, parallel corpora form a basis for data-driven aproaches to machine translation, from which the most relevant ones are Example-Based Machine Translation [Nagao, 1984] and Statistical Machine Translation [Brown et al., 1988]. Both approaches learn subsentential units of translation from the sentence pairs in a parallel corpus and reuse these fragments in subsequent translations. Therefore one of the primary tasks for both EBMT and SMT is to identify the correspondence between sub-sentential units in their parallel corpora.

EBMT makes use of parallel corpora to extract a database of translation examples, which are compared to the input sentence in order to translate. By choosing and combining these examples in an appropriate way, a translation of the input sentence can be provided.

In **SMT**, this process is accomplished by focusing on purely statistical parameters and a set of translation and language models, among other data-driven features. The following section further introduces the statistical approach to machine translation.

2.3.3 Statistical approach to MT

SMT treats translation as a machine learning problem. This means that they apply a learning algorithm to a large body of previously translated text. The learner is then able to translate previously unseen sentences. With an SMT toolkit and enough parallel text, we can build an MT system for a new language pair within a very short period of time - perhaps as little as a day [Al-Onaizan et al., 1999, Oard and Och, 2003, Oard et al., 2003]. Workshops have shown that translation systems can be built for a wide variety of language pairs within similar time frames [Koehn and Monz, 2005, Koehn and Monz, 2006, Callison-Burch et al., 2007]. The accuracy of these systems depends crucially on the quantity, quality, and domain of the data.

In "A survey of SMT", [Lopez, 2007] consider four problems that have to be solved in order to build a functioning SMT system.

- First, one must describe the series of steps that transform a source sentence into a target sentence. This is called a **Translational Equivalence Model**. Often, they derive from concepts from automata and language theory.
- Next, in order to enable the model to make good choices when faced with a decision to resolve some ambiguity, one need to develop a **Parameterization** of the model that will assign a score to every possible source and target sentence pair that the model might consider. Taken together, translational equivalence modeling and parameterization are often combined under the rubric of *modeling*.
- The parameterization defines a set of statistics called parameters used to score the model, but we need to associate values to these parameters. This is called **Parameter Estimation**, and it is based on machine learning methods.
- Finally, when we are presented with input sentence, we must search for the highest-scoring translation according to our model. This is called **Decoding**.

The first two steps are often conflated under the term of *modeling* in the literature, following [Brown et al., 1990]. This is because early systems involved a tight coupling between the translational equivalence model and the parametrization (or mathematical model). The most popular models can be described by one of two formalisms: *Finite-State Transducers* (FST) or *Synchronous Context-Free Grammars* (SCFG); for a detailed explanation of this models see [Lopez, 2008]. For our research, we followed the phrase-based appraoch of SMT, presented by Koehn, Och and Marcu [Koehn et al., 2003, Zens et al., 2002b], which is based on FST formalism.

In the next section, we will describe word-based IBM models, which introduce many of the common problems in translation modeling. They are followed by phrase-based models.

2.3.3.1 Word-based models - IBM alignment and translation models

SMT continues to be influenced by the groundbreaking IBM approach [Brown et al., 1990, Brown et al., 1993, Berger et al., 1994]. The IBM Models are word-based models and represent the first generation of SMT models. They illustrate many common modeling concepts.

In its basic form, the result of translation is modelled as the maximum of some function which represents the importance of faithfulness and fluency. This translation approach was first described by [Brown et al., 1990, Brown et al., 1993], in terms of the *noisy channel* model. In this model, the input sentence f to be translated is considered

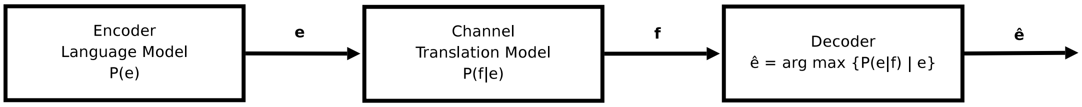


Figure 2.2: The noisy channel model in machine translation. The Language Model generates an English sentence *e*. The Translation Model transmits *e* as the Foreign sentence *f*. The decoder finds the English sentence $\hat{e}$ which is most likely to have given rise to *f*.

to be a distorted version of some target language sentence e (in this view the distortion due to noise has produced a language change). The task of the translation decoder is, given the distorted sentence f, to find the sentence $\hat{e}$ which has the best probability to have been converted into f (Fig. 2.2) [Manning and Schütze, 1999]. In this model (IBM Model 4), the process that produces e_i from f_j, takes three steps (Fig. 2.3), each step corresponding to a single transducer in a composed set [Knight and Al-Onaizan, 1998].

1. Each target word chooses the number of source words that it will generate. This number is called ϕ_i the fertility of e_i. It enables the definition of a translational equivalence between source and target sequences of different lengths.

2. Each copy of each target word produces a single source word. This represents the translation of individual words.

3. The translated words are permuted into their final order.

These steps are also applied to a special empty token ε, called the *null word* (or simply *null*). Null translation accounts for target words that are dropped in translation, as is often the case with function words.

This **Translational Equivalence Model** allows to enumerate possible structural relationships between pairs of strings, but the translation system needs a mechanism to decide between those. This mechanism comes with the parametrization (the mathematical model) that designs a function which allows us to assign a real-valued score to any pair of source and target sentences.

This is formalized by a **Generative Model** as following. [Brown et al., 1990] proposed that translation could be treated as a probabilistic process in which every sentence in one language is viewed as a potential translation of a sentence in the other language. To rank potential translations, every pair of sentences source - target[9] $(\mathbf{f}, \mathbf{e})$ is assigned a conditional probability $p\left(\mathbf{f} \mid \mathbf{e}\right)$. The best translation $\hat{\mathbf{e}}$ is the sentence that maximizes this probability. Using Bayes' theorem, [Brown et al., 1990] decomposed the probability into two components:

[9] We use the notation f (foreign or French) for the source source and e (English) for the target sentence for historical reasons, as it has been initially introduced by Brown and al. (1990) and has been used subsequently by the SMT literature.

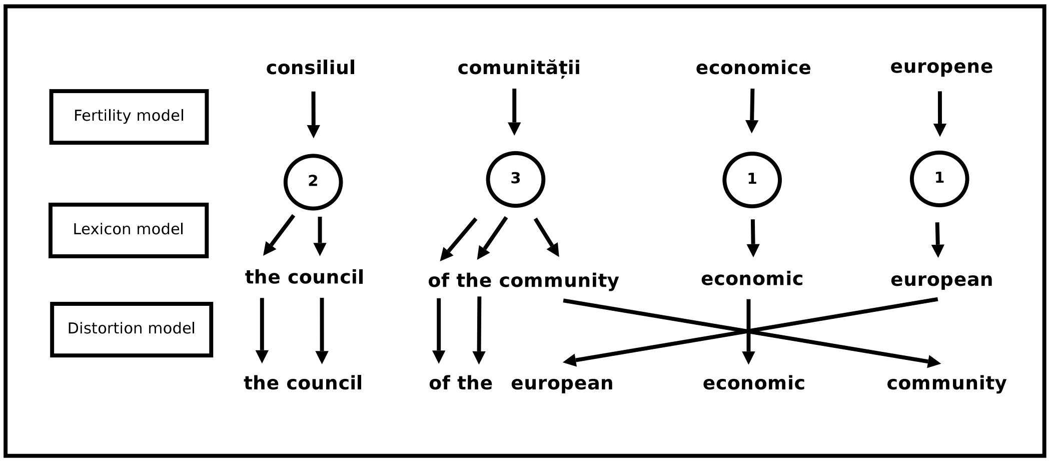


Figure 2.3: Visualization of IBM Model 4. This model of translation takes three steps. (1) Each Romanian (E) word (and the null word) selects a fertility - the number of English (F) words to which it corresponds. (2) Each Romanian (E) word produces a number of English (F) words corresponding to its fertility. Each English (F) word is generated independently. (3) The English (F) words are reordered.

$$\hat{\mathbf{e}} = arg\,\underset{e}{max}\,p\left(\mathbf{e}\,|\,\mathbf{f}\right) \tag{2.5}$$

$$\hat{\mathbf{e}} = arg\,\underset{e}{max}\,p\left(\mathbf{e}\right)p\left(\mathbf{f}\,|\,\mathbf{e}\right) \tag{2.6}$$

The two components are $p\left(\mathbf{e}\right)$ which is a language model probability, and $p\left(\mathbf{f}\,|\,\mathbf{e}\right)$ which is a translation model probability, where roughly, the first one quantifies the fluency of the language and the second quantifies the faithfulness of the translation.

Note that while the objective is to discover e given f, we actually model the reverse. The advantage of this over modeling $p\left(e, f\right)$ directly is that we can apply two independent models to the disambiguation of e. This is beneficial because the estimates for each model contain errors. By applying them together we hope to counterbalance their errors.

To implement equation 2.6, three tasks must be performed: quantify *fluency,* $p\left(e\right)$, quantify *faithfulness,* $p\left(f\ \mid e\right)$, (that means to define the parameters of the models and to estimate them) and find an algorithm which maximises the product of these two functions (the translation is defined as an optimisation problem).

The set of parameters, or probabilities of the language and translation model is to be automatically learned from parallel data (parameter estimation step). We can see the model as a stochastic process that generated the data (that is why these models are called *generative models*). In fact, we can think of the language model $p\left(e\right)$ as a stochastic model that generates target language sentences, and the translation model $p\left(f|e\right)$ as a second stochastic process that "corrupts" the target language to produce source language sentences.

The Language Model. The language model probability does not depend on the foreign language sentence f. It represents the probability that the e is a valid sentence in English. Rather than trying to model valid English sentences in terms of grammaticality, Brown et al. borrow n-gram language modeling techniques from speech recognition. These language models assign a probability to an English sentence by examining the sequence of words that comprise it. For $e = e_1 e_2 e_3 \ldots e_n$, the *language model probability* $p\left(e\right)$ can be calculated as:

$$p\left(e_1 e_2 e_3 ... e_n\right) = p\left(e_1\right) p\left(e_2 | e_1\right) p\left(e_3 | e_1 e_2\right) \ldots p\left(e_n | e_1 e_2 e_3 \ldots e_{n-1}\right) \tag{2.7}$$

This formulation disregards syntactic structure, and instead recasts the language modeling problem as the challenge of computing the probability of a single word given all of the words that precede it in a sentence. At any point in the sentence we must be able to determine the probability of a word, e_j , given a history, $e_1 e_2 \ldots e_{j-1}$. In order to simplify the task of parameter estimation for n-gram models, we reduce the length of the histories to be the preceding $n - 1$ words. Thus in a trigram model we would only need to be able to determine the probability of a word, e_j , given a shorter history, $e_{j-2} e_{j-1}$. Although n-gram models are linguistically simpleminded, they have the redeeming feature that it is possible to estimate their parameters from plain monolingual data.

The Translation Model. The design of a translation model has similar tradeoffs to the design of a language model. In order to create a translation model whose parameters can be estimated from data (which in this case is a parallel corpus) Brown et al. avoid linguistic sophistication in favor of a simpler model. They ignore syntax and semantics and instead treat translation as a word-level operation. They define the *translation model probability* $p\left(f|e\right)$ in terms of possible word-level alignments, a, between the sentences:

$$p\left(f|e\right) = \sum_a p\left(f, a|e\right) \tag{2.8}$$

Just as n-gram language models can be defined in such a way that their parameters can be estimated from data, so can $p\left(f, a|e\right)$. Introducing word alignments simplifies the problem of determining whether a sentence is a good translation of another into the problem of determining whether there is a sensible mapping between the words in the sentences (Fig. 2.4).

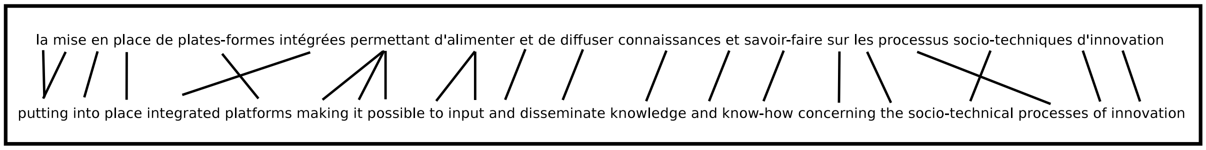

la mise en place de plates-formes intégrées permettant d'alimenter et de diffuser connaissances et savoir-faire sur les processus socio-techniques d'innovation
putting into place integrated platforms making it possible to input and disseminate knowledge and know-how concerning the socio-technical processes of innovation

Figure 2.4: Word alignments between a phrase pair in a French-English parallel corpus

Brown et al. defined a series of increasingly complex translation models, referred to as the IBM Models, which define $p\left(f, a|e\right)$. IBM Model 3 defines word-level alignments in terms of four parameters. These parameters include a word-to-word translation probability, and three less intuitive probabilities (*fertility*, *spurious word*, and *distortion*) which account for english words that are aligned to multiple foreign words, words with no counterparts in the foreign language, and word re-ordering across languages (c.f. table 2.2).

The (word) translation probabilities $t\left(f_j|e_i\right)$	The probability that a foreign word f_j is the translation of an English word e_i
Fertility probabilities $n\left(\phi_i|e_i\right)$	The probability that a word e_i will expand into ϕ_i words in the foreign language
Spurious word probability p	The probability that a spurious word will be inserted at any point in a sentence
Distortion probabilities $d\left(j|a_j, l, m\right)$	The probability that a target position j will be chosen for a word, given the index of the English word that this was translated from a_j, and the lengths l and m of the English and foreign sentences

Table 2.2: The IBM Models define translation model probabilities in terms of a number of parameters, including translation, fertility, distortion, and spurious word probabilities.

Parameter Estimation (EM algorithm). The probability of an alignment $p\left(f, a|e\right)$ is calculated under IBM Model 3 as[10]:

$$p(f,a|e) = \prod_{i=1}^{n} n(\phi_i|e_i) * \prod_{j=1}^{m} t(f_j|e_i) * \prod_{j=1}^{m} d(j|a_j, l, m) \qquad (2.9)$$

If a bilingual parallel corpus contained explicit word-level alignments between its sentence pairs, like in figure 2.4, then it would be possible to directly estimate the parameters of the IBM Models using maximum likelihood estimation. However, since word-aligned parallel corpora do not generally exist, the parameters of the IBM Models must be estimated without explicit alignment information. Consequently, alignments are treated as hidden variables. The *Expectation Maximization* (EM) framework for maximum likelihood estimation from incomplete data [Dempster et al., 1977] is used to estimate the values of these hidden variables. EM consists of two steps that are iteratively applied:

- The E-step calculates the posterior probability under the current model of every possible alignment for each sentence pair in the sentence-aligned training corpus;

[10] The true equation also includes the probabilities of spurious words arising from the "NULL" word at position zero of the English source string, but it is simplified here for clarity.

- The M-step maximizes the expected likelihood under the posterior distribution, $p(f, a|e)$, with respect to the model's parameters.

While EM is guaranteed to improve a model on each iteration, the algorithm is not guaranteed to find a globally optimal solution. Because of this, the solution that EM converges on is greatly affected by initial starting parameters. To address this problem Brown et al. first train a simpler model to find sensible estimates for the t table, and then use those values to prime the parameters for incrementally more complex models which estimate the d and n parameters described in Table 2.1.

IBM Model 1 is defined only in terms of word-for-word translation probabilities between foreign words f_j and the English words ea j which they are aligned to:

$$p(f, a|e) = \prod_{j=1}^{m} t(f_j|e_{a_j}) \tag{2.10}$$

IBM Model 1 produces estimates for the the t probabilities, which are used at the start EM for the later models.

Beyond the problems associated with EM and local optima, the IBM Models face additional problems. While equation 2.8 and the E-step call for summing over all possible alignments, this is intractable because the number of possible alignments increases exponentially with the length of the sentences. To address this problem Brown et al. did two things:

- They performed approximate EM wherein they sum over only a small number of the most probable alignments instead of summing over all possible alignments.
- They limited the space of permissible alignments by ignoring many-to-many alignments and permitting one-to-many alignments only in one direction.

[Och and Ney, 2003] undertook systematic study of the IBM Models. They trained the IBM Models on various sized German-English and French-English parallel corpora and compared the most probable alignments generated by the models against reference word alignments that were manually created. They found that increasing the amount of data improved the quality of the automatically generated alignments, and that the more complex of the IBM Models performed better than the simpler ones.

Improving alignment quality is one way of improving translation models. Thus word alignment remains an active topic in research. Some work focus on the improvement on the training procedures used by the IBM Models. [Vogel et al., 1996] used Hidden Markov Models. [Callison-Burch et al., 2004] recast the training procedure as a partially supervised learning problem by incorporating explicitly word-aligned data alongside the standard sentence-aligned training data. [Fraser and Marcu, 2006] did similarly. [Moore, 2005, Taskar et al., 2005, Ittycheriah and Roukos, 2005, Blunsom and Cohn, 2006] treated the problem as a fully supervised learning problem and applied discriminative training. Others have focused on improving alignment quality by integrating linguistically motivated constraints [Cherry and Lin, 2003].

But the most promising direction in improving translation models has been to move beyond word-level alignments to phrase-based models which are described in the next section.

2.3.3.2 Phrased-based models in SMT

Whereas the original formulation of Statistical Machine Translation was word-based, contemporary approaches have expanded to phrases. Phrase-based Statistical Machine Translation [Och and Ney, 2003, Koehn et al., 2003] uses larger segments of human translated text. By increasing the size of the basic unit of translation, phrase-based SMT does away with many of the problems associated with the original word-based formulation. In particular, [Brown et al., 1993] did not have a direct way of translating phrases; instead they specified the fertility parameter which is used to replicate words and translate them individually.

Furthermore, because words were their basic unit of translation, their models required a lot of reordering between languages with different word orders, but the distortion parameter was a poor explanation of word order. Phrase-based SMT eliminated the fertility parameter and directly handled word-to-phrase and phrase-to-phrase mappings. Phrase-based SMT's use of multi-word units also reduced the dependency on the distortion parameter. In phrase-based models less word re-ordering needs to occur since local dependencies are frequently captured. For example, common adjective-noun alternations are memorized, along with other frequently occurring sequences of words. Note that the *phrases* in phrase-based translation are not congruous with the traditional notion of syntactic constituents; they might be more aptly described as *substrings* or *blocks* since they just denote arbitrary sequences of contiguous words. [Koehn et al., 2003] showed that using these larger chunks of human translated text resulted in high quality translations, despite the fact that these sequences are not syntactic constituents.

In order to calculate a phrase translation probability it is crucial to identify phrase-level alignments between phrases that occur in sentence pairs in a parallel corpus.

Symmetrizing word alignments Many methods for identifying phrase-level alignments use word-level alignments as a starting point.

[Och and Ney, 2003] defined one of those. Their method first creates a word-level alignment for each sentence pair in the parallel corpus by outputting the alignment that is assigned the highest probability by the IBM Models. Because the IBM Models only allow one-to-many alignments in one language direction they have an inherent asymmetry. In order to overcome this, [Och and Ney, 2003] train models in both the e→f and f→e directions, and symmetrize the word alignments by combining them. At a minimum, all alignment points of the intersection of the two alignments are maintained. At a maximum, the points of the union of the two alignments are considered.

[Och and Ney, 2003] explore the space between intersection and union with expansion heuristics that start with the intersection and proceed by iteratively adding links from the union.

Their method has been reimplemented for `Moses` system, [by Koehn et al], in the following way:

- It starts with **intersection of the two word alignments**. Only new alignment points that exist in the union of two word alignments can be added. They also always require that a new alignment point connects to at least one previously unaligned word.
- Then, they **expand to only directly adjacent alignment points**, starting from the top right corner of the alignment matrix (alignment points the first English word, then for the second English word, and so on).
- This is **done iteratively** until no more alignment point can be added.
- In a final step, they **add non-adjacent alignment points**, with otherwise the same requirements.

This creates a single word-level alignment for each sentence pair, which can contain one-to-many alignments in both directions.

There are other ways to obtain symmetric alignments. [Matusov et al., 2004] present a symmetric word alignment method based on linear combination of complementary asymmetric words alignment probabilities. [Ayan and Dorr, 2006] investigate the effect of various symmetrization heuristics on the performance of phrase-based translation. However, these symmetrized alignments do not have many-to-many correspondences which are necessary for phrase-to-phrase alignments.

Phrase extraction [Och and Ney, 2004] defined a method for extracting incrementally longer phrase-to-phrase correspondences from a word alignment, such that the phrase pairs are consistent with the word alignment. Consistent phrase pairs are those in which all words within the source language phrase are aligned only with the words of the target language phrase and the words of the target language phrase are aligned only with the words of the source language phrase.

Following this approach, in `Moses`, all aligned phrase pairs that are consistent with the word alignment are collected. The words in a legal phrase pair are only aligned to each other, and not to words outside. The set of *Bilingual Phrases* (BP) can be defined formally [Zens et al., 2002a] as:

$$BP\left(f_1^J, e_1^J, A\right) = \left\{\left(f_j^{j+m}, e_i^{i+n}\right) : \forall\left(i', j'\right) \in A : j \leq j' \leq j+m \leftrightarrow i \leq i' \leq i+n\right\} \tag{2.11}$$

Probability distribution of phrase pairs Phrase-based SMT calculates a phrase translation probability $p(f|e)$ between an english phrase e and a foreign phrase f. In general the phrase translation probability is calculated using maximum likelihood estimation by counting the number of times that the english phrase was aligned with the foreign phrase in the training corpus, and dividing by the total number of times that the english phrase occurred:

$$p\left(\bar{f}|\bar{e}\right) = \frac{count\left(\bar{f},\bar{e}\right)}{count\left(\bar{e}\right)} \tag{2.12}$$

To calculate the maximum likelihood estimate for phrase translation probabilities the phrase extraction technique is used to enumerate all phrase pairs up to a certain length for all sentence pairs in the training corpus. The number of occurrences of each of these phrases are counted, as are the total number of times that pairs co-occur. These are then used to calculate phrasal translation probabilities, using equation 2.12. This process can be done with [Och and Ney, 2004]'s phrase extraction technique, or a number of variant heuristics. Other heuristics for extracting phrase alignments from word alignments were described by [Vogel et al., 2003, Tillmann, 2003, Koehn, 2004a].

As an alternative to extracting phrase-level alignments from word-level alignments, [Marcu and Wong, 2002] estimated them directly. They use EM to estimate phrase-to-phrase translation probabilities with a model defined similarly to IBM Model 1, but which does not constrain alignments to be one-to-one in the way that IBM Model 1 does. Because alignments are not restricted in [Marcu and Wong, 2002]'s model, the huge number of possible alignments makes computation intractable, and thus makes it impossible to apply to large parallel corpora. [Birch et al., 2006] made strides towards scaling [Marcu and Wong, 2002]'s model to larger data sets by putting constraints on what alignments are considered during EM, which shows that calculating phrase translation probabilities directly in a theoretically motivated way may be more promising than [Och and Ney, 2004]'s heuristic phrase extraction method.

2.3.3.3 Log-linear model and minimum error rate training

By moving from generative models to *log-linear models* (or *discriminative models*), additional context can be brought into the modeling. Log-linear models discriminate between different possible values translations e_i when presented with a particular source sentence f. They define a relationship between a set of K fixed features $h\left(e,f\right)$ of the data and the function $P\left(e|f\right)$ that we are interested in. Thus, they allow us to define an arbitrary feature that allows us to improve the translation.

Whereas the original formulation of statistical machine translation [Brown et al., 1990] used a translation model that contained two separate probabilities:

$$\hat{e} = arg\,\max_{e} p\left(e|f\right) = arg\,\max_{e} p\left(f|e\right) p\left(e\right) \tag{2.13}$$

contemporary approaches to SMT instead employ the log linear formulation [Och and Ney, 2002], which breaks the probability down into an arbitrary number of weighted feature functions:

$$\hat{e} = arg\,\underset{e}{max}\,p\,(e|f) = arg\,\underset{e}{max}\sum_{m=1}^{M}\lambda_m h_m\,(e,f) \tag{2.14}$$

The advantage of the log linear formulation is that rather than just having a translation model probability and a language model probability assign costs to translation, we can now have an arbitrary number of feature functions, $h(e, f)$ which assign a cost to a translation. In practical terms, this gives us a mechanism to break down the assignation of cost in a modular fashion based on different aspects of translation.

Most SMT systems use a log-linear model of $p\,(e|f)$ that incorporates generative models as feature functions.

In current systems the feature functions that are most commonly used include a language model probability, a phrase translation probability, a reverse phrase translation probability, lexical translation probability, a reverse lexical translation probability, a word penalty, a phrase penalty, and a distortion cost.

Estimation in log-linear models The weights, λ_m, in the log linear formulation act to set the relative contribution of each of the feature functions in determining the best translation. The Bayes' rule formulation (equation 2.13) assigns equal weights to the language model and the translation model probabilities[11]. In the log linear formulation these may play a greater or lesser role depending on their weights. The weights can be set in an empirical fashion in order to maximize the quality of the MT system's output for some development set (where human translations are given). This is done through a process known as *minimum error rate training* [Och and Ney, 2003], which uses an objective function to compare the MT output against the reference human translations and minimizes their differences. Modulo the potential of overfitting the development set, the incorporation of additional feature functions should not have a detrimental effect on the translation quality because of the way that the weights are set.

2.3.3.4 The phrase table

The decoder uses a data structure called a *phrase table* to store the source phrases paired with their translations into the target language, along with the value of feature functions that relate to translation probabilities. In our case the feature functions used are: a phrase translation probability, a reverse phrase translation probability, lexical translation probability, a reverse lexical translation probability and a word penalty. The phrase table contains an exhaustive list of all translations which have been extracted

[11] The noisy-channel approach can be obtained as a special case if we consider only two feature functions, namely the target language model $h1\,(e,f) = log\,p\,(e)$ and the translation model of the source sentence given the target $h2\,(e,f) = log\,p\,(f|e)$.

from the parallel training corpus. The source phrase is used as a key that is used to look up the translation options, These translation options are learned from the training data and stored in the phrase table. If a source phrase does not appear in the phrase table, then the decoder has no translation options for it.

Because the entries in the phrase table act as basis for the behavior of the decoder – both in terms of the translation options available to it, and in terms of the probabilities associated with each entry – it is a common point of modification in SMT research. Often people will augment the phrase table with additional entries or modify the scores associted with an existing entry, and show improvements without modifying the decoder itself. We do similarly in our pivot-based methods, which are explained in chapter 6.

2.3.3.5 Decoding

Once we have a model and estimates for all of our parameters, we can translate new input sentences. This is called *decoding*. In principle, decoding corresponds to solving the maximization problem in equation 2.15.

$$\hat{e} = arg\,\underset{e}{max}\, p\left(e|f\right) = p\left(f|e\right) \times p\left(e\right) \tag{2.15}$$

The decoder is the software which uses the statistical translation model to produce translations of novel input sentences. For a given input sentence the decoder first breaks it into subphrases and enumerates all alternative translations that the model has learned for each subphrase. The decoder then chooses among these phrasal translations to create a translation of the whole sentence. Since there are many possible ways of combining phrasal translations the decoder considers a large number of partial translations simultaneously. This creates a search space of hypotheses. These hypotheses are ranked by assigning a cost or a probability to each one. The probability is assigned by the statistical translation model and stored in the phrase table.

In word-based SMT systems, search was performed following different approaches including *optimal A* search* [Och et al., 2001], *integer programming* [Germann et al., 2001], *greedy search algorithms* [Wang and Waibel, 1998]. An important issue of these decoders is the computational complexity introduced by reordering (changes in word order) when single words are considered instead of longer units.

In phrase-based decoders, short-distance reorderings between source and target sentences are already captured within the translation units, which alleviates the reordering problem [Tillmann and Ney, 2000, Och and Ney, 2004]. `Pharaoh` [Koehn, 2004a], an efficient and freely available beam search phrase-based decoder was very successful and contributed in making SMT more accessible and more popular. Recently, `Pharaoh` has been replaced/upgraded by `Moses` [Koehn et al., 2007], which is also a phrase-based decoder implementing a beam search, allowing to input a word lattice with confusion networks and using a factored representation of the raw words (surface forms, lemma, part-of-speech, morphology, word classes, etc.). Nowadays, many SMT systems employ a phrase-based beam search decoder because of the good performance results achieved (in terms of accuracy and efficiency). We used the decoder provided by `Moses` in our thesis experiments.

2.3.3.6 Overview of the architecture used in SMT systems

Most current state of the art SMT systems use log-linear models with generative submodels in combination with Minimum Error Rate Training (MERT) in order to optimize whatever error function is chosen for evaluation. An overview of the architecture used in these systems is shown in Figure 2.5.

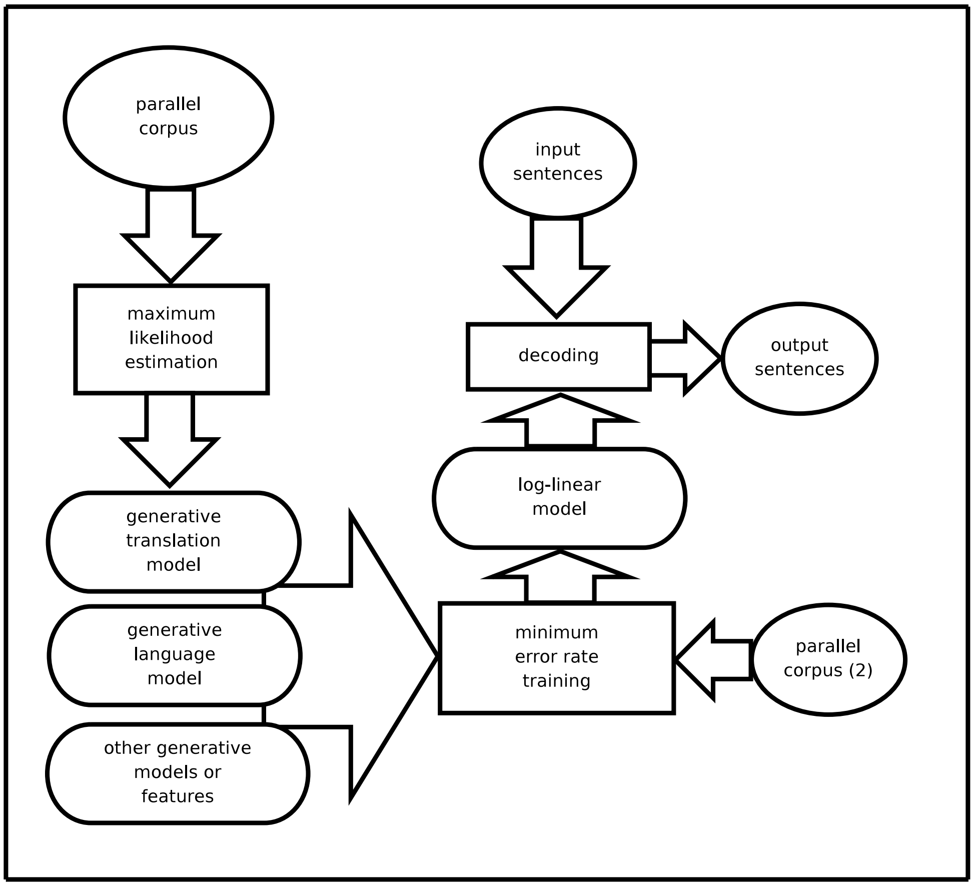

parallel corpus
maximum likelihood estimation
generative translation model
generative language model
other generative models or features
input sentences
decoding
output sentences
log-linear model
minimum error rate training
parallel corpus (2)

Figure 2.5: Overview of the architecture used in SMT systems: the flow of data, models, and process commonly involved

2.3.3.7 Evaluation in SMT

There are many good ways to translate the same sentence, thus it is difficult to define objective criteria for translation evaluation. Many methods have been proposed to evaluate MT output.

Traditionally accepted measures of MT evaluation have required examination of MT system's output by human judges, who rank the adequacy of the translation in conveying the source language meaning and the fluency of expression in the target language. More ideal than this are measures that determine how well some human task can be performed when the human subject is provided with machine-translated text. Unfortunately, human evaluation requires time and money. This usually rules out its use in iterative system development, where there is a need to perform regular evaluation to determine if changes are beneficial to performance. Then, the next thing is to develop automatic metrics that closely correlate with human judgement.

Usually, the automatic evaluation is performed by producing some kind of similarity measure between the translation hypothesis and a set of human reference translations, which represent the expected solution of the system. Therefore, a common element of automatic metrics is their use of a set of test sentences for which human translations, called *reference translations*, are already available. They can come from a parallel corpus, although we must be cautious and use a separate set of sentences from the set we used for training. The intuition behind metrics based on reference sentences is that MT must be good if it closely resembles a human translation of the same sentence [Papineni et al., 2002]. These metrics are based on partial string matching between the output and the reference translations. However, the use of a single reference may bias the evaluation towards a particular translation style. In order to mitigate against this and reflect the diversity of possible good translations, we may use multiple references. This requires the use of human translators to produce the additional references, but it is a one-time cost.

The fact that there are several correct alternative translations for any input sentence adds complexity to this task, and whereas the higher the correlation with the human references the better quality, theoretically we cannot guarantee that incorrelation with the available set of references means bad translation quality, unless we have all possible correct translations available.

Therefore, in general it is accepted that all automatic metrics comparing hypotheses with a limited set of manual reference translations are pessimistic. Yet, instead of an absolute quality score, automatic measures are claimed to capture progress during system development and to statistically correlate well with human intuition.

So far, no automatic translation evaluation measure has been generally accepted, so various measures are typically used instead. Some commonly used measures are:

- **WER (Word Error Rate) or mWER (multi-Reference Word Error Rate)**: the WER is the minimum number of substitution, insertion and deletion operations that have to be performed to convert the generated sentence into the reference target sentence. For the mWER, a whole set of reference translations is used. In this case, for each translation hypothesis, the edit distance to the most similar sentence is calculated.

- **PER (Position-independent word Error Rate) or mPER (multi-reference Position-independent word Error Rate)**: it is similar to WER (and mWER) but does not penalize reorderings, because it regards the output and reference sentences as unordered sets rather than totally ordered strings [Och et al., 1999]

- **BLEU (BiLingual Evaluation Understudy) score**: this score measures the precision of unigrams, bigrams, trigrams, and four-grams with respect to a whole set of reference translations, and with a penalty for too short sentences [Papineni et al., 2001]. BLEU measures accuracy, thus larger BLEU scores are better. As this is the metric used in our thesis we will detail it in the next paragraph.

- **NIST score**: the NIST evaluation metric, introduced in [Doddington, 2002], is based on the BLEU matrix, but with some alterations. Whereas BLEU simply calculates n-gram precision considering each n-gram of equal importance, NIST calculates how informative a particular n-gram is, and the rarer a correct n-gram is, the more weight it will be given. NIST also differs from BLEU in its calculation of the brevity penalty, and small variations in translation length do not impact the overall score as much.

- **METEOR score**: this score includes a word stemming process of the hypothesis and references to extend unigram matches [Banerjee and Lavie, 2005].

For a good contemporary evaluation of metrics across several language pairs, refer to [Callison-Burch, 2007]. A key element of most research in this area is the identification of metrics that correlate with human judgement in controlled studies [Papineni et al., 2002, Callison-Burch et al., 2007]. It is not always clear when a difference in scores between two systems represents a significant difference in their output. [Koehn, 2004b] describes a method to compute statistical confidence intervals for most automatic metrics using bootstrap resampling.

BLEU score Arguably the most extended evaluation measure as of today, BLEU (acronym for BiLingual Evaluation Understudy) was introduced by IBM in [Papineni et al., 2001], and is always referred to a given n-gram order ($BLEU_n$, n usually being 4).

The metric works by measuring the n-gram co-occurrence between a given translation and the set of reference translations and then taking the weighted geometric mean. BLEU is specifically designed to approximate human judgement on a corpus level and can perform badly if used to evaluate the quality of isolated sentences.

$BLEU_n$ is defined as:

$$BLEU_n = exp\left(\frac{\sum\limits_{i=1}^{n} bleu_i}{n} + length - penalty\right) \tag{2.16}$$

where $bleu_i$ and $length - penalty$ are cumulative counts (updated sentence by sentence) referred to the whole evaluation corpus (test and reference sets). Even though these matching counts are computed on a sentence-by-sentence basis, the final score is not computed as a cumulative score, ie. it is not computed by accumulating a given sentence score.

Equations 2.17 and 2.18 show $bleu_n$ and $length - penalty$ definitions, respectively:

$$bleu_n = log\left(\frac{Nmatched_n}{Ntest_n}\right) \tag{2.17}$$

$$length - penalty = min\left\{0, 1 - \frac{shortest - ref - length}{Ntest_1}\right\} \tag{2.18}$$

Finally, $Nmatched_i$, $Ntest_i$ and $shortest - ref - length$ are also cumulative counts (updated sentence by sentence), defined as:

$$Nmatched_i = \sum_{n=1}^{N} \sum_{ngr \in S} min\left\{N\left(test_n, ngr\right), \max_r \left\{N\left(ref_{n,r}, ngr\right)\right\}\right\} \tag{2.19}$$

where S is the set of Ngrams of size i in sentence $test_n$, $N\left(sent, ngr\right)$ is the number of occurrences of the Ngram ngr in sentence sent, N is the number of sentences to eval, $test_i$ is the i^{th} sentence of the test set, R is the number of different references for each test sentence and $ref_{n,r}$ is the r^{th} reference of the n^{th} test sentence.

$$Ntest_i = \sum_{n=1}^{N} length\left(test_n\right) - i + 1 \tag{2.20}$$

$$shortest - ref - length = \sum_{n=1}^{N} \min_r \left\{length\left(ref_{n,r}\right)\right\} \tag{2.21}$$

From BLEU description, we can conclude that:

- BLEU is a **quality metric** and it is defined in a range between 0 and 1, 0 meaning the worst-translation (which does not match the references in any word), and 1 the perfect translation.

- BLEU is mostly a **measure of precision**, as $bleu_n$ is computed by dividing the matching n-grams by the number of n-grams in the test (not in the reference). In this sense, a very high BLEU could be achieved with a short output, so long as all its n-grams are present in a reference.

- The recall or coverage effect is weighted through the length penalty. However, this is a very **rough approach to recall**, as it only takes lengths into account.

- Finally, the **weight of each effect (precision and recall) might not be clear**, being very difficult from a given BLEU score to know whether the provided translation lacks recall, precision or both.

It is important, when interpreting metrics such as BLEU, to note that they can be used to rank systems relative to each other, but the scores are generally uninterpretable as absolute measures of correctness.

BLEU has been highly influential in SMT research. It has been used as the basis for a number of comparative evaluations [Doddington, 2002, Koehn and Monz, 2005, Koehn and Monz, 2006, Callison-Burch et al., 2007] and it is commonly used in the objective function for minimum error-rate training [Och, 2003].

The use of BLEU score has always been controversial. [Turian et al., 2003] provide counterexamples to its claimed correlation with human judgement and other potential problems have been demonstrated by [Callison-Burch et al., 2006]. Despite controversy, automatic evaluation has had a profound impact on progress in SMT research, and it is likely to continue.

With the proliferation of available metrics, it is not always clear which one to use. Practical considerations such as comparison with previous benchmarks encourages continued use of BLEU, despite criticism.

2.4 Related work

Many directions have been explored aiming to improve alignment and translation systems.

Most of the recent work in word alignment is focused on improving the word alignment quality through better modeling [Och and Ney, 2003, Deng and Byrne, 2005, Martin et al., 2005] or alternative approaches to training [Fraser and Marcu, 2006, Moore, 2005, Ittycheriah and Roukos, 2005]. In word alignment systems for languages with scarce resources, some researchers [Aswani and Gaizauskas, 2005, Lopez and Resnik, 2005, Tufiş et al., 2005] have used language-dependent resources such as dictionaries, thesaurus, and dependency parser to improve word alignment results.

For translation between the language pairs with low resources, [Niessen and Ney, 2004] used morpho-syntactic information and [Vandeghinste et al., 2006, Carl et al., 2008] used translation dictionaries and shallow analysis tools .

2.4.1 Translation system combination

The idea of using multiple source knowledge in translation ties in with the recent work on ensemble combination of SMT systems. [Macherey and Och, 2007] presented an empirical study on how different selections of input translation systems affect translation quality in system combination, where they gave an empirical evidence that the systems to be combined should be of similar quality and need to be almost uncorrelated in order to be beneficial for system combination.

Computing (consensus) translations from the outputs of multiple translation engines has become a powerful means to improve translation quality in many machine translation tasks. A composite translation is computed by voting on the translation outputs of multiple machine translation systems. Depending on how the translations are combined and how the voting scheme is implemented the composite translation may differ from any of the original hypotheses. While elementary approaches simply select for each sentence one of the original translations (hypothesis ranking techniques), more sophisticated methods allow to combine translations on a word or a phrase level (*consensus network decoding*).

The latter, **consensus network decoding** ([Mangu et al., 2000]), attempts to improve translation quality by finding a novel, higher quality hypothesis based on the hypotheses produced by multiple translation systems. Recent research ([Frederking and Nirenburg, 1994, Bangalore et al., 2001, Jayaraman and Lavie, 2005, Rosti et al., 2007]) has explored consensus decoding where all systems translate the same language pair. [Matusov et al., 2006] adopted this approach to a multilingual setting, where pairwise word alignments of the original translation hypotheses were estimated for an enhanced statistical alignment model in order to explicitely capture word re-ordering. Their method resulted in substatial gain: 4.8 BLEU higher than the single best system. [Callison-Burch et al., 2008] reported preliminary results that indicate promising results when applying combination techniques on the multisource "News Commentary" corpus.

Alternatively, **hypothesis ranking techniques** attempt to select the single best hypothesis from a list of output hypotheses produced by different translation systems. Several techniques designed for bilingual sentence-level system combination could be applied with no changes to the multisource task. [Kaki et al., 1999, Callison-Burch and Flournoy, 2001] used only the target language model to rank hypotheses. This approach follows the intuition that the hypothesis with the highest language model score will be the most fluent. [Nomoto, 2004] took this step further by using multiple language models which vote on candidate hypotheses. When integrating multilingual data the systems typically create several candidate sentential target translations for source sentences via languages. A single translation is then selected by finding the candidate that yields the best overall score [Och and Ney, 2001] or by co-training [Callison-Burch and Osborne, 2003], where the information is integrated at the training stage to bootstrap more training data from multiple source documents. [Eisele, 2005] have used simple heuristics to combine both multiple translations of the same source sentence provided by different translation engines and the translations of corresponding parts from different source languages.

[Schwartz, 2008] surveyed the state of the art in techniques to exploit multi-parallel corpora and techniques for using multiple source languages in SMT and presents experiments which show the limitation of existing hypothesis ranking methods.

In this thesis we explore a complementary approach to improve a statistical alignment and translation model using multi-lingual, parallel (or multi-parallel) corpora. Our method is based on pivot languages.

2.4.2 Pivot-based methods

2.4.2.1 Definitions

A *pivot language*, sometimes also called a *bridge language* is an artificial or natural language used as an intermediary language for translation. Using a pivot language avoids the combinatorial explosion of having translators across every combination of the supported languages. The disadvantage of a pivot language is that each step of retranslation introduces possible mistakes and ambiguities.

The *triangulation*, is the process of incorporating multilingual knowledge in a single system, which, in our context, utilizes parallel corpora available in more than two languages.

The idea of using multiple source languages for improving translation quality of the target languages is not new. [Kay, 1997, Kay, 2000] suggests that much of the ambiguity of a text that makes it hard to translate into another language may be resolved if a translation into some third language is available and proposes using multiple source documents as a way of informing subsequent machine translations. He calls the use of existing translations to resolve underspecification in a source text "*triangulation in translation*", but does not offer a method to perform this triangulation. The challenge is to find general techniques that will exploit the information in multiple sources to improve the quality of alignment and machine translation.

2.4.2.2 Pivot methods in related fields

Pivot-based methods have also been used in different related areas, such as **translation lexicon** induction [Mann and Yarowsky, 2001, Schafer and Yarowsky, 2002, Sanfilippo and Steinberger, 1997], **word sense disambiguation** [Diab and Resnik, 2002].

The use of an intermediate language as translation aid has also found application in **cross-lingual information retrieval** (CLIR). Thus, pivot languages are employed to translate queries in (CLIR) [Gollins and Sanderson, 2001, Kishida and Kando, 2003]. These methods only used the available dictionaries to perform word by word translation. In addition, NTCIR 4 workshop organized a shared task for CLIR using pivot language. Machine translation systems are used to translate queries into pivot language sentences, and then into target sentences [Sakai et al., 2004].

Pivot languages have been used in **rule-based machine translation** systems. [Boitet, 1988] discusses the pros and cons of the pivot approaches in multilingual machine translation. [Schubert, 1988] argues that a pivot language needs to be a natural language, due to the inherent lack of expressiveness of artificial languages.

2.4.2.3 Pivot language in alignment

Pivot languages have been used to improve sentence alignment [Simard, 1999] or word alignment [Borin, 2000b, Filali and Bilmes, 2005, Wang et al., 2006].

[Simard, 1999, Simard, 2000] describes experiments showing that a system based on trilingual set texts can yield better bilingual **sentence alignments**, while retaining the same computational complexity, as the common bilingual approach.

[Borin, 2000b] used multilingual corpora to increase **word alignment** coverage. He described a non-statistical approach where a pivot alignment is used to combine direct translation and indirect translation via a third language. The alignment system used [Tiedemann, 1999b, Tiedemann, 1999a] utilized several types of information to align the words in the two texts: distributional information, coocurence statistics, iterative size reduction, 'naive' stemming and string similarity to select and rank word alignment candidates. His conclusion is that in a multilingual parallel corpora, pivot alignment is a safe way to increase word alignment recall without lowering the precision. He observes that the degree of relatedness of the languages in a triad play a role on how well pivot alignment will work for the particular triad and that different pivot languages add different alignments, i.e. there seems to be a cumulative positive effect from adding more languages. Even if he did not have all the data needed to calculate the significance of the results, his conclusions remain suggestive and encouraging.

[Filali and Bilmes, 2005] worked on a statistical alignment procedure, in two steps, that exploits information from parallel translations in more than two languages. Their **alignment-tag model** is a multilingual extension of the IBM and HMM models. The preliminary results on a small subset of the Europarl corpus showed a 7% relative improvement (decrease in alignment error rate) over a state of the art alignment model. They consider that an important future direction of research should consist in investigating whether their gains in multilingual alignment quality carry over and improve learning of phrase translation probabilities.

[Wang et al., 2006] suggested an approach to improve **word alignment for languages with scarce resources**, using bilingual corpora of other language pairs. To perform word alignment between source and target languages, for which there are only small amounts of bilingual data available, they introduced a third language (pivot) and large-scale bilingual corpora in source-pivot and pivot-target languages. Using these corpora they trained two word alignment models (source-pivot and pivot-target) and they built an induced alignment model between source and target languages based on these models. They reported a relative error reduction of 10.41% as compared with the direct method, using the small biligual copora between source and target. In addition they interpolated the induced model with the direct one. This interpolated model further improved word alignment results by achieving a relative error rate reduction of 21.30% as compared with the direct method. As a case study, they used English as the pivot language to improve word alignment between Chinese and Japanese. In terms of future work, they suggest to investigate the effect of the size of corpora on the alignment results and different parameters combinations of the induced model and the direct one.

They consider that another evaluation should be done in a real machine translation system, to examine whether lower word alignment error rate will result in higher translation accuracy. This direction has been investigated in [Wu and Wang, 2007], which will be detailed in the next paragraph.

2.4.2.4 Pivot methods in SMT

SMT with bridge languages is concerned with the way to optimally perform translations from source language to target language, by taking advantage of other available language resources.

Dependant (or overlapping) data versus Independant data experiments Translation with pivot language has recently gained attention as a mean to circumvent the data bottleneck of SMT. For this kind of approaches there are two general assumptions:

1. there is a lack of parallel texts between source language and target language;
2. there exists a third language (pivot) for which there are abundant parallel texts between source and pivot and between pivot and target.

Based on these assumptions a realistic working condition is that the parallel corpora for source-pivot and pivot-target are independent, in the sense that they are not derived from the same set of sentences. As they are based on independant data, they report few comparisons between the performance of the pivot-based methods and the directly trained systems, often only on reduced training sets.

In the meantime, recent research has often focused on the use of parallel corpora which provides multiple translations of the same texts. Such data can be regarded as interesting to perform contrastive experiments, namely to compare translations obtained with and without bridge languages. This could be the first step towards the use of pivot methods in situations where training data is extermely scarce [Utiyama and Isahara, 2007, Wu and Wang, 2007]. Aiming at the evaluation of the performance of the pivot strategies against that of direct SMT systems under controlled experiments, these approaches often provide detailed analyses of different factors that could affect the performance of the pivot methods, such as the size of the training data or the choice of the intermediate language(s) [Cohn and Lapata, 2007]. Complementary to this framework and in order to investigate the effectiveness of the pivot methods in “real situations”, some reseachers [Wu and Wang, 2007] performed additional experiments on independently sourced parallel corpora.

We will detail next some approaches directed by low-density resources.

The pivot-based method in [De Gispert and Mariño, 2006] is motivated by the lack of resources between Catalan and English, for which the translation is bridged through Spanish. The authors compare two coupling strategies: cascading of two translation systems versus training of systems from parallel texts, the targets part of which have been automatically translated from pivot to target. Thus, they created an English-Catalan parallel corpus by automatically translating the Spanish part of an English-Spanish parallel corpus into Catalan with a Spanish-Catalan SMT system. They then directly trained a SMT system on the English-Catalan corpus. They showed that this direct training is superior to the "sentence translation strategy" in translating from Catalan into English (in terms of BLEU score). Their experimental results are promising, as the achieved translation quality is nearly equivalent to that of the Spanish-English language pair.

[Eisele, 2006] proposed that existing bilingual translation systems which share one or more common pivot languages should be coupled to build translation systems for language pairs for which no parallel corpus exists; using this approach for example, existing Arabic-English, Arabic-Spanish, Spanish-Chinese and English-Chinese systems could be used together to effect an Arabic-Chinese translation system.

[Wu and Wang, 2007] reported positive results using a similar technique with a single pivot language in conjunction with a small bilingual training corpus. They experimented their methods in the context of both dependent and independent parallel corpora.

Although our aim is to evaluate the performance of the pivot strategies against that of direct systems under controlled experiments (dependent data) and to analyze how much the pivot strategies can be improved by different factors, we performed in addition, complementary experiments on disjoint parallel texts, in order to estimate their robustness on independent data.

Pivot-based Training versus Pivot-based Decoding The pivot knowledge source could be integrated in the translation process at two different moments: during the training or during the decoding process. If this information is integrated during the training, we will refer to this process as *pivot-based training* (or *bridging at training time*), in the other case we can talk about *pivot-based decoding* (or *bridging at translation time* [Bertoldi et al., 2008]). Often, in the literature, the pivot strategies are divided into phrase translation strategy and sentence translation strategy ([Utiyama and Isahara, 2007]). The *phrase translation strategy* directly constructs a phrase translation table from a source-pivot phrase table and a pivot-target phrase table. It then uses this phrase table in a phrase-based SMT system. The *sentence translation strategy* first translates a source language sentence into n pivot sentences and translate these n sentences separately into target language sentences. Then, it selects the highest scoring sentence from the target language sentences.

As a generalisation, we can divide the pivot-based methods into:

- **Pivot methods in training (or at training time)**: this means that the parallel training corpora source-pivot and pivot-target are used to train a translation system from source to target. The pivot information could be integrated into alignment or directly in the phrase table (as described before). This will generate a translation model source-target that can be fed directly into the decoder. In this case the triangulation is part of the translation model.

- **Pivot methods in decoding (at decoding time)**: in this case, the methods should integrate or couple two translation models in the same decoding process. This requires to combine hypotheses from different systems, which will lead to a system combination framework that has already been mentioned in the subsection 2.4.1.

Typically, the pivots methods in training are working with words [Wang et al., 2006] or phrases [Cohn and Lapata, 2007, Wu and Wang, 2007, Chen et al., 2008] at the model level while the pivot methods in decoding cope with sentences at the hypothesis level [Utiyama and Isahara, 2007].

There are different ways to integrate multilingual data in the training process.

[Callison-Burch et al., 2006, Callison-Burch, 2007] used pivot language(s) to paraphrase extraction to handle the unseen phrases for phrased-based SMT. Their method acquires paraphrases by identifying phrases in the source language, translating them into multiple target languages, and then back to the source. Thus, they use paraphrases to deal with unknown source language phrases and to improve coverage and translation quality.

[Cohn and Lapata, 2007] presents another pivot approach based on phrase tables, where the scores of the new phrase-table are computed by combining corresponding translation probabilities in the source-pivot and pivot-target phrase tables.

An approach based on phrase table multiplication is also discussed in [Wu and Wang, 2007], where they compare it with the word-based pivot method proposed in [Wang et al., 2006] (for which the pivot data is integrated at the alignment level). They demonstrate that the phrase method performs better than the word method.

A different strategy is adopted in [Chen et al., 2008], who worked also at the phrase level but focused on the efficiency of the translation process in which they aimed at reducing the model size, by filtering out the less probable entries based on testing correlation using additional training data in a pivot language.

[Kumar et al., 2007] presented a pivot method at training time: they incorporated pivot languages to construct word alignments (that essentially means a word-based pivot method). They showed that this technique can be used to obtain higher quality bilingual word alignments than traditional bilingual word alignment techniques. They performed, in addition, an evaluation by combining the direct method with the pivot-based one(s). The coupling was made at the decoding time, using a consensus decoding technique presented in [Macherey and Och, 2007], that produced a single output hypothesis from multiple systems.

When the triangulation is part of the decoding, pivot-based methods refer to the system combinations based on multilingual data. As in the consensus translation, the systems typically create several candidate sentential target translations for source sentences via languages, from which a composite translation is computed by voting. This composite translation may differ from any of the original hypotheses. The simplest and most straightforward way is to return to one of the original candidate translations, that yields the best overall score [Utiyama and Isahara, 2007, Bertoldi et al., 2008], but there are approaches that combine smaller units, such as words or phrases from different hypotheses.

[Utiyama and Isahara, 2007] compare the two pivot strategies: a phrase-based pivot methods (pivot at training time) and a sentence-based pivot strategy (pivot at decoding time). They report that the phrase translation strategy significantly outperformed the sentence translation strategy, with a relative performance of 0.92 to 0.97 compared to directly trained SMT systems.

Our research explores two pivot-based methods at training time and their variants and compares them with pivot-based method at decoding time.

2.4.3 Relevant approaches

We will here detail works in the literature which are relevant to our approach (pivot-based methods in phrase-based SMT).

"Pivot Language Approach for Phrase-Based Statistical Machine Translation" **- Wu and Wang**

[Wu and Wang, 2007] addressed the translation problem for language pairs with scarce resources by bringing in a pivot language, at training time, via which they can make use of large bilingual corpora.

They calculated a pivot phrase-table and an interpolated phrase table which is a combination of the pivot and the direct one. Their experiments were conducted on Europarl corpus [Koehn, 2005] proposed for the shared task of the NAACL/HLT 2006 Workshop on SMT [Koehn and Monz, 2006], in which four languages were involved: English, French, Spanish and German. They chose English as pivot language, because in general, for most of the languages there exists bilingual corpora between these languages and English. They experimented training data with different sizes and they studied the performance of the interpolated system based on two pivot languages. Additionally experiments on Chinese-Japanese translation using English as pivot language were carried on to investigate the effectiveness of their method on independently sourced parallel corpora.

The results on both the Europarl corpus and Chinese-Japanese translation indicate that the interpolated models achieve the best results. Results also indicate that their pivot language approach is suitable for translation on language pairs with a small bilingual corpus: the less source-target bilingual corpus there is, the bigger the improvement is.

In terms of BLEU score their method achieves an absolute improvement of 0.06 (22.13% relative) as compared with the standard model trained with 5000 source-target sentence pairs for French - Spanish translation (via English). The translation quality is comparable with that of the model trained with a bilingual corpus of 30000 source-target sentence pairs. Moreover, the translation quality is further boosted by using both the small source-target bilingual corpus and the large source-pivot and pivot-target corpora.

"A Comparison of Pivot Methods for Phrase-based Statistical Machine Translation" **- Utiyama and Isahara**

[Utiyama and Isahara, 2007] presented and compared two pivot-based methods, the former integrated at training time (named phrase translation strategy) and the latter applied at decoding time (called sentence translation strategy). The phrase translation strategy builds the source-target pivot table from the source-pivot and pivot-target phrase tables, by multiplication: the scores of the new phrase table are computed by combining corresponding translation probabilities in the source-pivot and pivot-target phrase-tables. The sentence translation strategy is a system cascading technique.

Their experiments were also conducted on the Europarl data for the NAACL/HLT 2006 Workshop on SMT [Koehn and Monz, 2006], that consists in three parallel corpora: French-English, Spanish-English and German-English (which design English as the only possible pivot language).

They showed that the phrase translation strategy consistently outperformed the sentence translation strategies in controlled experiments. They explained this by the fact that the phrase-tables constructed while using the phrase translation strategy can be integrated into the decoder as well as the directly extracted phrase-tables, so the "phrase translation" systems can fully exploit the power of the decoder. This led to better performance even when the induced phrase-tables were noisy. They observed that the relative performance of the pivot systems seems to be related to the BLEU scores for the direct systems.

The relative performance of the phrase translation strategy compared to directly trained systems was 0.92 (Spanish-French via English) to 0.97 (German-Spanish via English).

"Improving Word Alignment with Bridge Languages" **- Kumar, Och and Machery**

[Kumar et al., 2007]) described an approach to improve SMT performance using multi-lingual, parallel, sentence-aligned corpora in several bridge languages. Their approach consists of a simple method for utilizing a bridge language to create a word alignment system, by multiplying posterior probability matrices for source-pivot and pivot-target, and a procedure for combining word alignment systems from multiple bridge languages, by linear interpolation of their posterior probability matrices.

The final translation is obtained by consensus decoding that combines hypothesis obtained using all bridge language word alignments. Thus, their approach combines pivot-methods at the training time with pivot-methods at decoding.

Their alignment combination system is based on word alignement posterior probability matrices, that can be generated by any underlying statistical alignment model.

Therefore, this method can be used to combined word alignments generated by fairly dissimilar word alignment systems, as long as the systems can produce posterior probabilities.

They present experiments showing that multilingual, parallel text in Spanish, French, Russian, and Chinese can be utilized in this framework to improve translation performance on an Arabic-to-English task. The experiments were performed in the open data track of the NIST Arabic-to-English machine translation task [12]. They report the alignment performance in AER: the direct method outperform any of the bridge systems. The alignment obtained by combining the direct system (Arabic-English) with all the bridge systems (via Spanish, French, Russian, Chinese) outperforms all the bridge alignments, but is weaker than the alignment without any bridge language. Their hypothesis is that a good choice of interpolation weights would reduce AER of the combination (issue that is not investigated in the paper).

The translation performance is measured using the NIST implementation of the case-sensitive BLEU-4 (on true-cased translations). They show that the system combinations techniques enable improvements relative to the direct system baseline: alignment combination (by linear interpolation of posterior probability matrices) gives a small gain (0.2 points), while the consensus translation results in a larger improvement (0.8 points).

The performance of the hypothesis consensus combination system steadily increases as bridge systems get added to the direct baseline. Therefore, they conclude that while the bridge language systems are weaker than the direct model, they can provide complementary sources of evidence. Furthermore, experiments on blind test (compared with the test set) show that the bridge systems continue to provide orthogonal evidence at different operating points.

In terms of future work they consider extensions to their framework that lead to more powerful combination strategies using multiple bridge languages.

"Phrase-Based Statistical Machine Translation with Pivot Languages" - **Bertoldi et al.**

[Bertoldi et al., 2008] present a theoretical formulation of SMT, with pivot languages, that embraces several approaches from the literature and an original method based on the random sampling of training data.

Their method consists in generating a parallel corpus source-target (S,T), by random sampling, from a source-pivot corpus (S,P) and using a translation system pivot to target (that was trained on the pivot-target texts). For each sentence pair (s_i, p_i) in the source-pivot corpus they generate a random sample of m translations $t_{ij\ j=1,\ldots,m}$ of p_i, according to the distribution $\tilde{P}(t \mid p)$. The idea is to get a sample that contains the most probable translations with possible duplicates. Given the newly created corpus $(S,T) = \{(s_i, t_{ij}) \mid_{j=1,\ldots,m}\}$ they build a translation system from source to target. This way the most reliable word alignments are reinforced during training as well as phrase-pairs using words of the most probable translation. They compare the performance

[12] http://www.nist.gov/speech/tests/mt/

of this method with a sentence translation strategy and a phrase translation strategy based on pivot language.

They present experimental results on Chinese-Spanish translation via English, on a benchmark provided by the 2008 International Workshop on Spoken Language Translation (IWSLT 2008)[13]. They compare performances of each bridging method when using corpora that are either disjoint, or overlappped on the pivot language side.

Their method for generating training data through random sampling proves to perform as well as the best methods used on the coupling of translation systems (in terms of case sensitive BLEU% score).

All systems trained on the overlapping text achieve significantly larger BLEU scores. In this case the direct system has a score comparable with the method based on the phrase translation strategy, but clearly below the score of the other two pivot-based systems. The authors give a possible explanation for this behaviour. They claim that it is related to the nature of the three languages involved. Translating from Chinese to Spanish requires introducing significant morphology information and word re-ordering. In some sense, pivoting through English results is a nice factorization of the issues: Chinese-English translation copes with most of the word-reordering but little morphology, while English-Spanish translation implies little word re-ordering but more morphology. This factorization probably has a positive impact in terms of less data sparseness in the training data and results in better statistical models. An additional experiment between Chinese and English via Spanish, provides an evidence to their claim.

Their discussion highlights the importance of the nature (relatedness) of the languages in a triad when using a pivot-based method.

"Machine Translation by Triangulation: Making Effective Use of Multi-Parallel Corpora" **- Cohn and Lapata**

[Cohn and Lapata, 2007] present a method that alleviates the coverage problem over source and target phrases, by exploiting multiple translation of the same source phrases. They create a larger table by incorporating one obtained via a pivot language. This way, lexical gaps in the original training data are filled by training data from the third language.

They offer a generative formulation which treats triangulation as part of the translation model itself: the pivot information is integrated at the phase-level (during the training). They show how triangulated phrase-table can be used in conjunction with a standard phrase-table to improve the translation estimates for both seen and unseen phrase-table entries.

They also demonstrate that triangulation can be used on its own, without a source-target distribution, and still yield acceptable translation output. Therefore, it provides a means of translation between the "low-density" language pairs, for which there are none source-target bitexts yet .

[13] `http://www.slc.atr.jp/IWSLT2008/`

They use Europarl corpus [Koehn, 2005] for experimentation. It consists of 700 000 sentences of parliamentary proceedings from the European Union in eleven languages (Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese, Swedish). While employing a large number of intermediate languages, in their experiments they explore the following questions:

1. How do different training requirements affect the performance of the triangulated models?
2. How does the choice of the intermediate language influence the MT output?
3. What is the quality of the triangulated phrase-table?

They show that the triangulation can produce high quality translations, and in conjunction with the standard phrase-table improve over the standard (direct) system in most instances. They claim that the triangulation provides better robustness to noisy alignments and better estimates to low-count events.

They observe large performance gains when translating with triangulated models trained on small datasets. Furthermore, when combined with a standard phrase-table, their models also yield performance improvements on larger datasets.

They show that triangulation benefits from a large set of intermediate languages. Their findings suggest that "intermediate" languages which exhibit a high degree of similarity with the source and target are desirable. They conjecture that this is a consequence of better automatic word alignments and a generally easier translation task, as well as better preservation of information between aligned sentences.

The important future directions suggested for exploration lie in combining triangulation with richer means of conventional smoothing and using triangulation to translate between low density language pairs.

"Improving Statistical Machine Translation Efficiency by Triangulation" **- Chen, Eisele and Kay**

[Chen et al., 2008] present two approaches to phrase tables filtering for more efficient translations. They use multi-parallel data to reduce the computation costs without harming the translation quality of phrase-based SMT.

They describe an attempt to reduce the model size by filtering out the less probable entries using additional training data in an intermediate third language. Considering the efficiency of the process as a whole, their aim is to remove from the table the entries that are not supported by the pivot language based on testing correlation. While previous approaches, aiming to improve the quality of translation, effectively took the union of a pair of phrase tables, they work with the intersection. Essentially, they retain a pair in the original table only if a pair with the same output string appears in the table coming from the third language. They introduce two specific methods for phrase-table filtering that look for phrases in the bridge language that can connect phrase pairs in the phrase table to be filtered. The first method requires strict matches of complete phrases, while in the second the constraints are relaxed by scoring over vocabulary overlap.

To evaluate their approach they conduct experiments using Europarl corpus, performing translation from Spanish to English with French or German as pivot language. The results show that filtering would not reduce the BLEU scores in most of the cases. The performance of models filtered through pivot actually converges when the original phrase table becomes larger. They observe that the performance of the filtered models greatly relates to the choice of the bridge language.

Their approaches reduce the sizes of the models used for SMT and thereby reduce the time and space costs required for translation tasks. The reduction of the model size can be up to 70% while the translation quality is being preserved.

They give some potential directions to continue their work. They suggest that the selection of the intermediate language needs to be studied more systematically. Another potential work is the refinement of the correlation measure for which the current design of the scoring scheme is still ad hoc. As a new future direction, they suggest to scale up their methods to hypotheses level, at which they work with complete sentences rather than phrases. In this situation, resources in the third language could help to eliminate implausible translation candidates.

2.4.4 Conclusions

Multi-parallel texts provide a rich source of information which could be exploited to reduce the noise and to increase the coverage of alignment and translation models. Despite significant research into system combination, relatively little is known about the best way to translate when multiple parallel source languages are available.

The survey of the pivot-based techniques that we previously presented shows that the subject has recently gained attention in SMT, as an additional source of knowledge. Thereby, pivot in translation has been used as a mean to circumvent the data bottleneck, to resolve alignment errors, to reduce the ambiguity, to improve translation choice and the coverage of translation models. Although the existing approaches have just scratched the surface of the possibilities for the framework, their results are encouraging.

To summarize, the main research directions in pivot-based alignment and translation from these previous works, that represents an interest for our study, are the following.

First, different training conditions should be experimented in order to define the effectiveness of a pivot method. This includes the size of the training data and the type of parallel bitext available, i. e. that presents overlapping or not on the pivot language side. The findings suggest that the nature of languages in a triad is a factor that could affect the performance of pivot-based methods. Thus, the degree of relatedness of the languages in a triad seems to play a role on how well pivot alignment or translation will work for the particular triad. Furthermore, it seems that the more languages one add the better the results become, i.e. different additional languages complement each other. More experiments, including using more than one intermediate languages are important before drawing any general conclusions related to the pivot language choice.

Another direction deals with when and how to integrate the pivot method: analyzing the correlation factors, smoothing methods, interpolation weights, combination strategy for pivot-based techniques are suggested as important issues to be further explored.

2.5 Our approach

Although related to [Cohn and Lapata, 2007]'s approach, our method is slightly different in the way we integrate the pivot information, in terms of the implementation and the large coverage of languages. We propose two methods and their variants, one at the alignment level, and the other at the phrase-table level, both focusing on translation improvement. They are compared with a pivot method at decoding time.

Furthermore, our experiments cover a large number of language pairs and intermediate languages and constitute the basis for studying different factors that influence the alignment via a pivot language: the training corpus size, the type of the intermediate language (the relatedness of the pivot language with the source and target language, poor or rich morphology). We have designed a set of experiments that demonstrate the importance of each of these features and show how pivot alignments or phrase-tables can be combined with the standard ones to improve the output of a statistical translation system.

The factors to be studied are:

1. when and how to integrate the pivot information : in the alignment process, in the phrase table, during the decoding

2. pivot language choice (depending on the source and target) and the nature of the triad in general

3. training conditions : training data size (source-target, source-pivot, pivot-target corpora) and the type of data (overlapping versus disjoint data on the pivot side)

We performed experiments that shows the improvement brought by the usage of a pivot language and the influence of different factors on our models.

Part II

JRC-Acquis corpus and its subcorpora

Chapter 3

Corpus description

3.1 Introduction

In many ways, progress in natural language research is driven by the availability of data. This is particularly true to the field of Statistical Machine Translation (SMT), which needs large quantity of parallel text: text paired with its translation in a second language. The harvesting of these resources has allowed the continued improvement of statistical machine translation systems that challenge the state of the art in MT for many language pairs.

JRC-Acquis [Steinberger et al., 2006] is a unique and freely available parallel corpus containing European Union (EU) documents of mostly legal nature. To our knowledge, the JRC Collection of the Acquis Communautaire available currently in 22 official EU languages is the only parallel corpus of its size available in so many languages. The current version of the JRC-Acquis is distributed in TEI-compliant XML format. It is accompanied by paragraph segmentation and information on segment alignment using both Vanilla and HunAlign. It is furthermore accompanied by EUROVOC subject domain information for most texts.

The JRC Acquis corpus has been compiled within the Joint Research Center of the European Commission, while working for the Language Technology group. This work has been carried out in the framework of the Exploratory project "Achieving massive multilinguality", in collaboration with the Romanian Academy of Science and the Slovenian Jozef Stefan Institute. The project started in 2005 and the first version of JRC-Acquis was made publicly available in May 2006. The current version 3.0, that will be described and used in this thesis has been compiled and released in April 2007.

In the next sections, after reminding our motivation for the corpus compilation (see section 3.2), we will explain what the JRC-Acquis is (see section 3.3), its composition, its format and the domain coverage.

We will end the chapter with a short description of the DGT Translation Units, multilingual Translation Memory for the Acquis Communautaire, that has been provided by the European Commission's Directorate-General for Translation (DGT) and was publically released by JRC Language Technology group in November 2007 .

Both sub-corpora of JRC-Acquis corpus and from the DGT Translation Units have been used in our experiments.

3.2 Motivation

Parallel corpora are widely sought after, for instance:

1. to train automatic systems for Statistical Machine Translation [Koehn, 2005] or multilingual categorisation.
2. to produce multilingual lexical or semantic resources such as dictionaries or ontologies [Giguet and Luquet, 2006].
3. to train and test multilingual information extraction software [Ignat et al., 2003].
4. for automatic translation consistency checking.
5. for the training of multilingual subject domain classifiers [Pouliquen et al., 2003, Civera and Juan, 2006].
6. to test and benchmark sentence (and other) alignment softwares because such softwares may perform unevenly well for different language pairs.

Most available parallel corpora exist for a small number of languages and mainly involving at least one widely-spoken language, such as the *French-English Hansards* [Germann, 2001] or the *English-Norwegian Parallel Corpus* (ENPC)[1]. Parallel corpora in more languages are available either for small amounts of text and/or for very specialised texts, such as the *bible* [Resnik et al., 1999] or the novel *"1984"* by George Orwell). To our knowledge, the currently most multilingual corpus with a considerable size and variety is *Europarl* [Koehn, 2005], which exists in eleven European languages. Europarl is offered with bilingual alignments in all language pairs involving English. However, this corpus does not contain any of the languages of the new Member States.

The JRC-Acquis corpus contains bilingual alignment information for all the 231 language pairs, including rare language combinations such as Estonian-Greek and Maltese-Danish. The main interest in exploiting this highly multilingual parallel corpus stems from the fact that it includes the new EU languages. For some of these, only few linguistic resources are available.

An additional feature of the JRC-Acquis is the fact that most texts have been manually classified into subject domains according to the EUROVOC thesaurus ([EUROVOC, 1995]), which is a classification system with over 6000 hierarchically organised classes. Knowing the subject domain(s) of texts can be exploited to produce domain-specific terminology lists, as well as to test and train document classification

[1] `http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc`

softwares [Ignat and Rousselot, 2006a, Ignat and Rousselot, 2006b] and automatic indexing systems. Due to the combination of multi-linguality and subject domain coding of the JRC-Acquis, such systems cannot only be trained multi-monolingually for more than 20 languages, but new approaches, that exploit data from more than one language at a time, can be developed.

A possible exploitation of the corpus could be to extract general and domain-specific terminology lists and to align these terminology lists across languages to produce multilingual term dictionaries. In JRC applications, these resources could be used to link similar texts across languages [Steinberger et al., 2004b, Pouliquen et al., 2004, Steinberger et al., 2004a], to improve further the automatic multilingual and cross-lingual news analysis system NewsExplorer [2] [Steinberger et al., 2005], and to offer cross-lingual glossing applications, i.e. to identify known terms in foreign language texts and to display these terms to the users in their own language [Ignat et al., 2005].

The corpus description follows in the next section.

3.3 Corpus presentation

JRC-Acquis [Steinberger et al., 2006] is, as mentioned earlier, a unique and freely available parallel corpus containing European Union (EU) documents of mostly legal nature. To our knowledge, the corpus with more than 20 European languages is the most multilingual parallel corpus of its size currently in existence. It is available in 22 languages (from 23 official EU languages): Bulgarian (bg), Czech (cz), Danish (da), German (de), Greek (el), English (en), Spanish (es), Estonian (et), Finnish(fi), French (fr), Hungarian (hu), Italian (it), Lithuanian (lt), Latvian (lv), Maltese (mt), Dutch (nl), Polish (pl), Portuguese (pt), Romanian (ro), Slovakian (sk), Slovene (sl), Swedish (sv).

The corpus consists of almost 20 000 documents per language, with an average size of nearly 48 million words per language. It is encoded in XML, according to the Text Encoding Initiative Guidelines TEI P4 [Sperberg-McQueen and Burnard, 2002]. It includes marked-up texts and bilingual alignment information for all the 231 language pair combinations. Pair-wise paragraph alignment information was produced by two different aligners (`Vanilla` and `HunAlign`). Most texts have been manually classified according to the EUROVOC subject domains so that the collection can also be used to train and test multi-label classification algorithms and keyword-assignment software.

The European Commission's Office for Official Publications OPOCE manages the distribution rights of this aligned multilingual parallel corpus. OPOCE agreed that the corpus can be given to research partners for non-commercial use.

The corpus, related alignment information and documentation are freely available for research purposes and can be downloaded from `http://langtech.jrc.it/JRC-Acquis.html`.

[2] Accessible at http://press.jrc.it/NewsExplorer

3.4 Corpus composition

EU/EC *Acquis Communautaire* (AC) is the French and most widely used term to name the body of common rights and obligations which bind all the Member States together within the European Union (EU) (formerly European Community EC). We will refer to this collection as the AC or the Acquis. The Acquis is constantly evolving and comprises: the contents, principles and political objectives of the Treaties; EU legislation; declarations and resolutions; international agreements; acts and common objectives. Countries wanting to join the EU have to accept and adopt the Acquis. By definition, translations of this document collection are therefore available in all twenty-three official EU languages. The current corpus version contains texts in 22 official EU languages. For the 23rd official EU language, Irish, the translations are not yet available.

Most EU documents are uniquely identifiable by their *CELEX* code, which consists of a one-digit document type, four-digits to express the year, one letter, four digits and optionally brackets containing a one or two-digit number. An example for a decision that entered into force in 1999 is 21999D0624(01). The translations of each document have the same unique CELEX identifier.

While a defining list of AC documents should theoretically exist, we have not been able to get hold of this, so we had to infer which documents available on the EU and other web sites are part of the collection. We decided to select all those documents which exist in at least ten of the twenty-two languages and which are available for at least three of the languages of the Member States who joined the EU in 2004 or in 2007 (Bulgarian (bg), Czech (cz), Estonian (et), Hungarian (hu), Lithuanian (lt), Latvian (lv), Maltese (mt), Polish (pl), Romanian (ro), Slovakian (sk), Slovene (sl)). As the corpus we compiled is not exactly identical with the legally binding document collection, we use the term JRC Collection of the Acquis Communautaire (short: JRC-Acquis) to refer to the documents contained in our corpus.

All documents of the version 3.0 were downloaded from the Commission's CELEX web pages[3]. For a given CELEX Code and a given language, the text was downloaded using the following URL: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:*CELEXCODE*:*LG*:HTML, where the two parameters *CELEXCODE* and *LG* should be replaced by their respective values.

The Romanian and Bulgarian documents were available only in Microsoft Word format[4]. These documents have been processed by the team of the Research Institute for Artificial Intelligence of the Romanian Academy, who converted them from their original format to the XML format of the JRC-Acquis corpus. For some of the documents, only preliminary translations were available.

For some reason, not all language versions are available for all AC documents, and some documents have a non-English title but the text body is in English, and vice-versa. An automatic language recognition tool was therefore used to filter out those

[3] `http://europa.eu.int/eur-lex/lex`

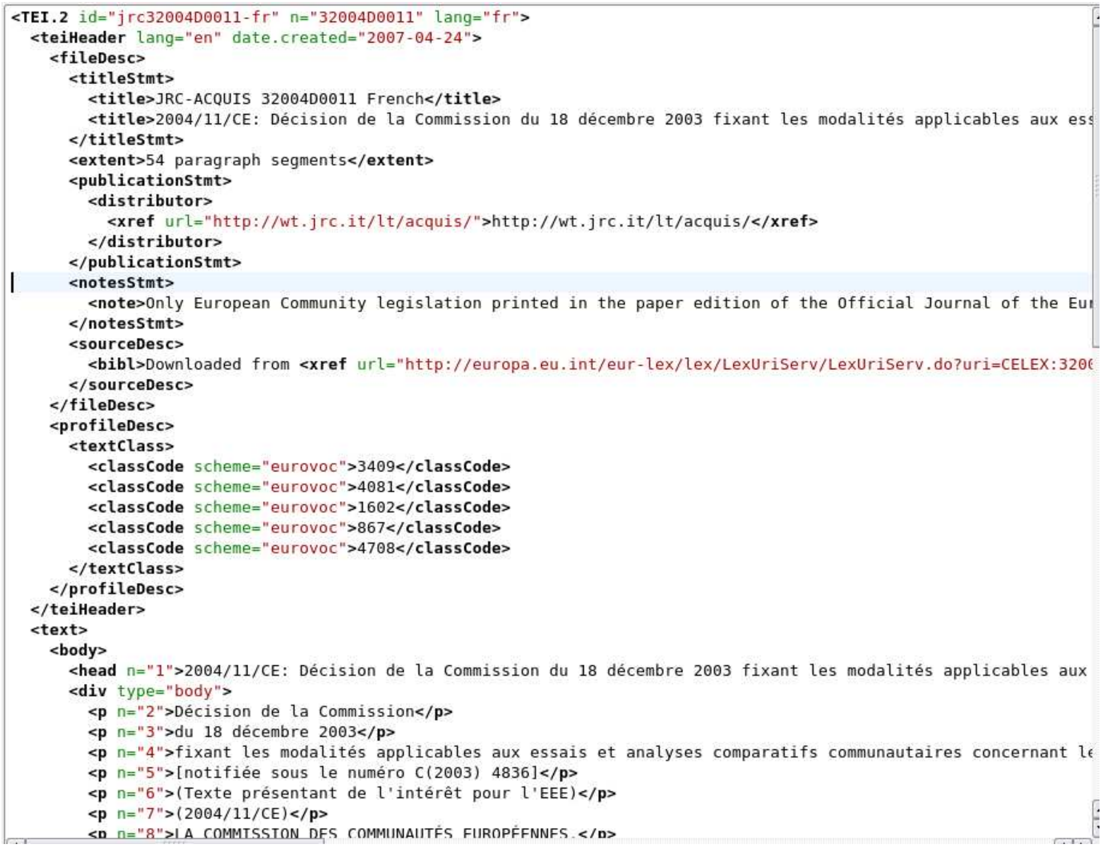
[4] `http://ccvista.taiex.be`

texts that are displayed as being one language, but which are actually English. No manual checking was carried out.

The different steps of corpus compilation and alignment will be detailed in chapter 4. The size of the current version 3.0 of the AC collection for the various languages can be seen in table 3.2.

3.5 Document structure

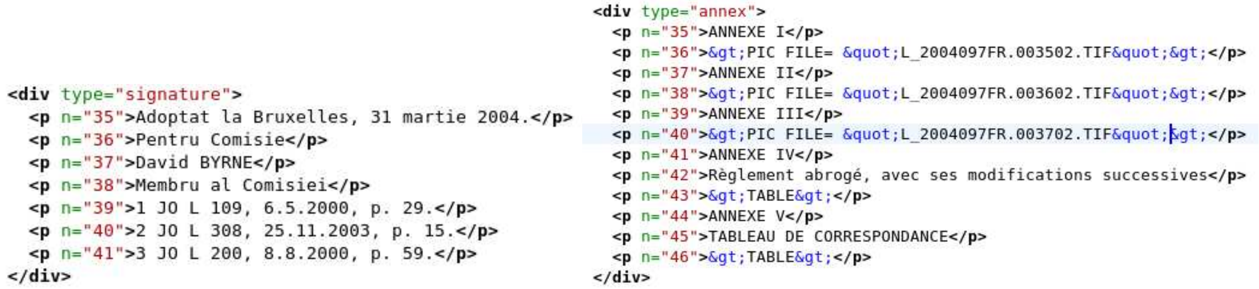
Each document was split into numbered paragraph chunks, based on the original HTML divisions of the documents. As the Acquis texts are consistent and well-structured, these paragraph chunks are mostly the same across languages. Each of these paragraphs can contain a small number of sentences, but they sometimes contain sentence parts (ending with a semicolon or a comma) because legal documents frequently specify their scope with a single sentence spanning over several paragraphs. For an example see Figure 3.1. As a result, each paragraph of the text collection can be uniquely identified using the language, the CELEX identifier and the paragraph number.



```
<TEI.2 id="jrc32004D0011-fr" n="32004D0011" lang="fr">
  <teiHeader lang="en" date.created="2007-04-24">
    <fileDesc>
      <titleStmt>
        <title>JRC-ACQUIS 32004D0011 French</title>
        <title>2004/11/CE: Décision de la Commission du 18 décembre 2003 fixant les modalités applicables aux ess
      </titleStmt>
      <extent>54 paragraph segments</extent>
      <publicationStmt>
        <distributor>
          <xref url="http://wt.jrc.it/lt/acquis/">http://wt.jrc.it/lt/acquis/</xref>
        </distributor>
      </publicationStmt>
      <notesStmt>
        <note>Only European Community legislation printed in the paper edition of the Official Journal of the Eur
      </notesStmt>
      <sourceDesc>
        <bibl>Downloaded from <xref url="http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:320
      </sourceDesc>
    </fileDesc>
    <profileDesc>
      <textClass>
        <classCode scheme="eurovoc">3409</classCode>
        <classCode scheme="eurovoc">4081</classCode>
        <classCode scheme="eurovoc">1602</classCode>
        <classCode scheme="eurovoc">867</classCode>
        <classCode scheme="eurovoc">4708</classCode>
      </textClass>
    </profileDesc>
  </teiHeader>
  <text>
    <body>
      <head n="1">2004/11/CE: Décision de la Commission du 18 décembre 2003 fixant les modalités applicables aux
      <div type="body">
        <p n="2">Décision de la Commission</p>
        <p n="3">du 18 décembre 2003</p>
        <p n="4">fixant les modalités applicables aux essais et analyses comparatifs communautaires concernant le
        <p n="5">[notifiée sous le numéro C(2003) 4836]</p>
        <p n="6">(Texte présentant de l'intérêt pour l'EEE)</p>
        <p n="7">(2004/11/CE)</p>
        <p n="8">LA COMMISSION DES COMMUNAUTÉS EUROPÉENNES,</p>
```

Figure 3.1: Sample of the TEI header and of the first few lines of a French JRC-Acquis document in XML format

The main body of the Acquis texts frequently ends with place and date of signature of the document, lists of person names and references to other documents (Fig. 3.2). Approximately half of the documents furthermore contain an annex, which can consist of plain texts, lists of addresses, lists of goods, etc. In order to allow users to easily make use of the different sections, they have been identified and marked up as body, signature and annex (Fig. 3.1 and 3.2). This division into three document parts allows users to concentrate their effort on the text type that is most useful for them: While the text body, for instance, rather reliably, contains text, the signatures (which are frequently multilingual) contain many named entities (persons, places, dates, references to other documents) so that they could be a good object for named entity recognition tasks. Note that signatures and annexes are usually marked up, but as they were not always clearly identifiable, we have missed the mark-up on some of them.



```
<div type="signature">
  <p n="35">Adoptat la Bruxelles, 31 martie 2004.</p>
  <p n="36">Pentru Comisie</p>
  <p n="37">David BYRNE</p>
  <p n="38">Membru al Comisiei</p>
  <p n="39">1 JO L 109, 6.5.2000, p. 29.</p>
  <p n="40">2 JO L 308, 25.11.2003, p. 15.</p>
  <p n="41">3 JO L 200, 8.8.2000, p. 59.</p>
</div>
```

```
<div type="annex">
  <p n="35">ANNEXE I</p>
  <p n="36">&gt;PIC FILE= &quot;L_2004097FR.003502.TIF&quot;&gt;</p>
  <p n="37">ANNEXE II</p>
  <p n="38">&gt;PIC FILE= &quot;L_2004097FR.003602.TIF&quot;&gt;</p>
  <p n="39">ANNEXE III</p>
  <p n="40">&gt;PIC FILE= &quot;L_2004097FR.003702.TIF&quot;&gt;</p>
  <p n="41">ANNEXE IV</p>
  <p n="42">Règlement abrogé, avec ses modifications successives</p>
  <p n="43">&gt;TABLE&gt;</p>
  <p n="44">ANNEXE V</p>
  <p n="45">TABLEAU DE CORRESPONDANCE</p>
  <p n="46">&gt;TABLE&gt;</p>
</div>
```

Figure 3.2: Typical signature and annex of JRC-Acquis document

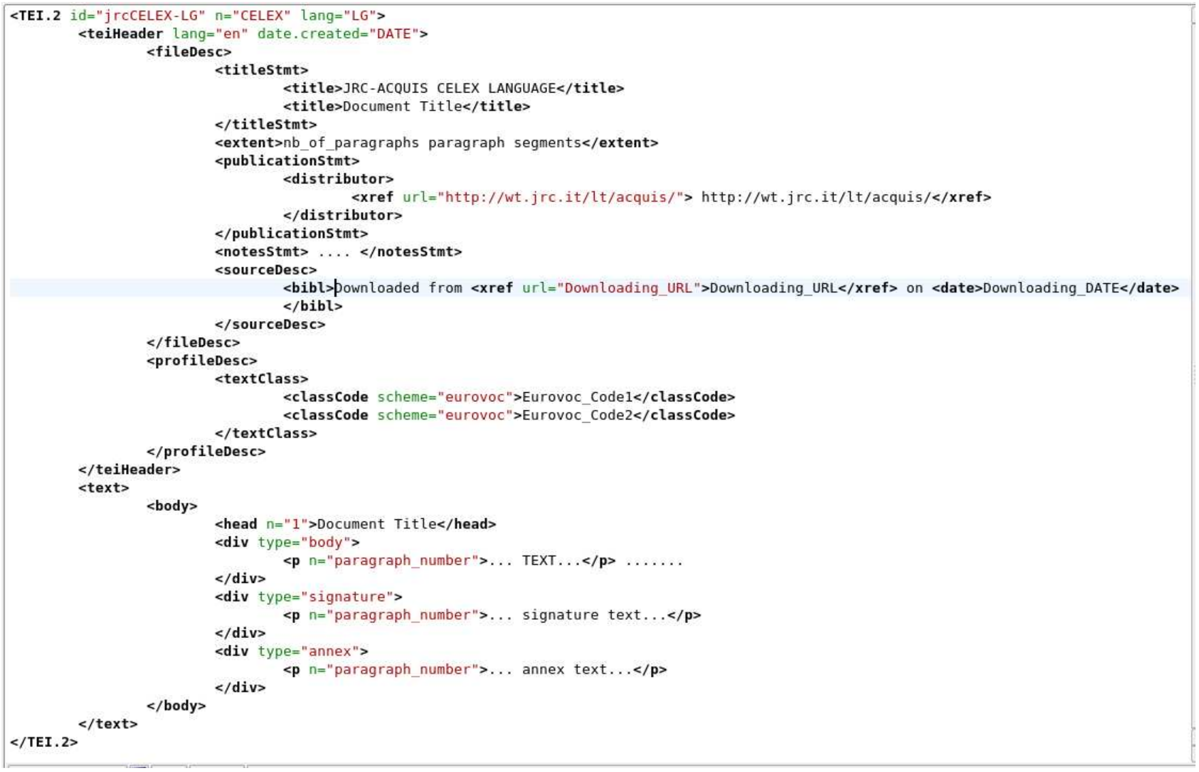
We noticed that for a part of English and French texts the annexes have not been included in the HTML version of the documents, but only referenced by a link to an image file (pdf, pic, tif). For this reason we did not align the annexes. We show in the next section that the average number of words by document with and without annex confirm this choice.

3.6 Alignments

The corpus is distributed with the paragraph alignment information for all 231 language pair combinations using two different aligners, `Vanilla` [Gale and Church, 1991b] and `HunAlign` [Varga et al., 2005]. The alignment results are stored for each language pair in an XML document that does not contain actual texts, but only pointers to the aligned paragraphs. In the corpus distribution we provide a Perl script that can be used to generate a bilingual aligned corpus from any of the 231 language pairs. Figure 4.2 shows an English-Italian sample alignment.

Often the alignments are produced at the sentence level. In JRC-Acquis case we consider the logical structure of the document, the "paragraph" level, as alignment unit. We remind that the Acquis texts are consistent and well-structured and the paragraph chunks are mostly the same across languages. They can contain a small number of sentences or sometimes, sentence parts, because legal documents can specify their scope with a single sentence spanning over several paragraphs. The alignment processing will be described in section 4.2.

3.7 Format / Encoding



```
<TEI.2 id="jrcCELEX-LG" n="CELEX" lang="LG">
        <teiHeader lang="en" date.created="DATE">
                <fileDesc>
                        <titleStmt>
                                <title>JRC-ACQUIS CELEX LANGUAGE</title>
                                <title>Document Title</title>
                        </titleStmt>
                        <extent>nb_of_paragraphs paragraph segments</extent>
                        <publicationStmt>
                                <distributor>
                                        <xref url="http://wt.jrc.it/lt/acquis/"> http://wt.jrc.it/lt/acquis/</xref>
                                </distributor>
                        </publicationStmt>
                        <notesStmt> .... </notesStmt>
                        <sourceDesc>
                                <bibl>Downloaded from <xref url="Downloading_URL">Downloading_URL</xref> on <date>Downloading_DATE</date>
                                </bibl>
                        </sourceDesc>
                </fileDesc>
                <profileDesc>
                        <textClass>
                                <classCode scheme="eurovoc">Eurovoc_Code1</classCode>
                                <classCode scheme="eurovoc">Eurovoc_Code2</classCode>
                        </textClass>
                </profileDesc>
        </teiHeader>
        <text>
                <body>
                        <head n="1">Document Title</head>
                        <div type="body">
                                <p n="paragraph_number">... TEXT...</p> .......
                        </div>
                        <div type="signature">
                                <p n="paragraph_number">... signature text...</p>
                        </div>
                        <div type="annex">
                                <p n="paragraph_number">... annex text...</p>
                        </div>
                </body>
        </text>
</TEI.2>
```

Figure 3.3: The format of JRC-Acquis document

The JRC-Acquis is available in UTF-8-encoded XML format, according to the Text Encoding Initiative Guidelines TEI P4 [Sperberg-McQueen and Burnard, 2002]. The corpus consists of two parts, the documents and the alignments.

The documents are grouped according to language; all the texts from one language constitute one TEI corpus, which consists of the TEI header, giving extensive information about the language corpus, and the actual documents. Each document contains, again, a TEI header, giving for instance the download URL, the EUROVOC codes and the text, which consists of the title and a series of paragraphs.

The two-way alignments are, for each language pair, stored as a TEI-compliant XML document. However, the document does not contain actual texts, but only pointers to the aligned paragraphs. As explained above, these can be converted into in-place alignments with the help of the included program. It should be noted that the headers are also available in HTML, and thus enable the introduction and documentation of the corpus in the distribution.

The documents have the format as illustrated in 3.3. The DTD for this format is also provided with the distribution.

Note that the title, body text, signature and annex further contain *<p>...</p>* tags. Each tag contains as attribute (n) its sequential number in the document, which is used in the paragraph alignment.

3.8 Statistics on JRC-Acquis

The JRC-Acquis corpus (version 3.0) is currently available in 22 languages with the distribution showed in the table 3.2.

The low number of Romanian texts is explained by the fact that the translations were not yet available at the downloading time (as Romania joined the EU only in 2007), and that the overlapping with the selected CELEX codes was quite reduced. The current version (March 2009) includes a new Romanian corpus that contains 19211 documents (182 631 277 characters and 30 832 212 words). Out of the total number of Romanian documents, 11469 are common with the English documents (they have the same CELEX code). As this version was not available when we started the experiments we took into consideration only the Romanian documents from the previous version.

The annexes for some languages are “longer” than for others as illustrated in the Figure 3.4. We notice, for instance, that the average number of words by annex for Romanian, Maltese and Bulgarian are respectively 3351.06, 3089.51 and 2636, while for English and French remain 1960 and 2186.35. This confirm our supposition that the Romanian documents include the translation of the annexes, while the English and French documents often contain only references (Fig. 3.2).

Some alignment statistics will be presented as well in section 4.2 in the next chapter.

3.9 EUROVOC Subject Domain Classification

Like most other official documents of the European Commission and the European Parliament, the Acquis texts have been manually classified according to the multilingual, hierarchically organised EUROVOC thesaurus [EUROVOC, 1995]. The main subject domains assigned to the document collection, listed in Table 3.4, show that the texts cover various subject domains, including economy, health, information technology, law, agriculture, food, politics and more.

Lg ISO	**Nº of**	**Text body**			**Signat.**	**Annexes**	**Total**
code	**texts**	**Nº wrd**	**Nº char**	**Avg. wrd**	**Nº wrd**	**Nº wrd**	**Nº wrd**
bg	11 384	16 140 819	104 522 671	1 417.85	2 170 075	14 114 612	32 425 506
cs	21 438	22 843 279	148 972 981	1 065.55	7 225 300	16 763 733	46 832 312
da	23 624	31 459 627	213 468 135	1 331.68	2 629 786	16 855 213	50 944 626
de	23 541	32 059 892	232 748 675	1 361.87	2 542 149	16 327 611	50 929 652
el	23 184	36 453 749	**239 583 543**	1 572.37	2 973 574	16 459 680	55 887 003
en	23 545	34 588 383	210 692 059	1 469.03	3 198 766	17 750 761	55 537 910
es	23 573	38 926 161	283 016 756	1 651.30	**3 490 204**	19 716 243	**62 132 608**
et	23 541	24 621 625	192 700 704	1 045.90	1 336 051	14 995 748	40 953 424
fi	23 284	24 883 012	212 178 964	1 068.67	2 677 798	12 547 171	40 107 981
fr	**23 627**	**39 100 499**	234 758 290	1 654.91	3 021 013	**19 978 920**	62 100 432
hu	22 801	28 602 380	213 804 614	1 254.44	2 529 488	15 056 496	46 188 364
it	23 472	35 764 670	230 677 013	1 523.72	3 120 797	18 331 535	57 217 002
lt	23 379	26 937 773	199 438 258	1 152.22	2 436 585	15 018 484	44 392 842
lv	22 906	27 592 514	196 452 051	1 204.60	1 673 124	15 437 969	44 703 607
mt	10 545	20 926 909	128 906 748	**1 984.53**	1 336 042	15 620 611	37 883 652
nl	23 564	35 265 161	231 963 539	1 496.57	3 039 580	18 467 115	56 771 856
pl	23 478	29 713 003	214 464 026	1 265.57	2 513 141	17 027 393	49 253 537
pt	23 505	37 221 688	227 499 418	1 583.56	3 034 308	19 350 227	59 606 203
ro	6 573	9 186 947	60 537 301	1 397.68	514 296	11 185 842	20 887 085
sk	21 943	26 792 637	179 920 434	1 221.01	3 227 852	16 190 546	46 211 035
sl	20 642	27 702 305	178 651 767	1 342.04	3 103 193	16 837 717	47 643 215
sv	20 243	29 433 037	199 004 401	1 453.99	2 575 771	14 965 384	46 974 192
Total	**463 792**	**636 216 050**	**4 288 962 348**	**1 387.23**	**60 368 893**	**358 999 011**	**1 055 583 954**

Table 3.2: Size of the JRC-Acquis corpus in each of the 22 official EU languages

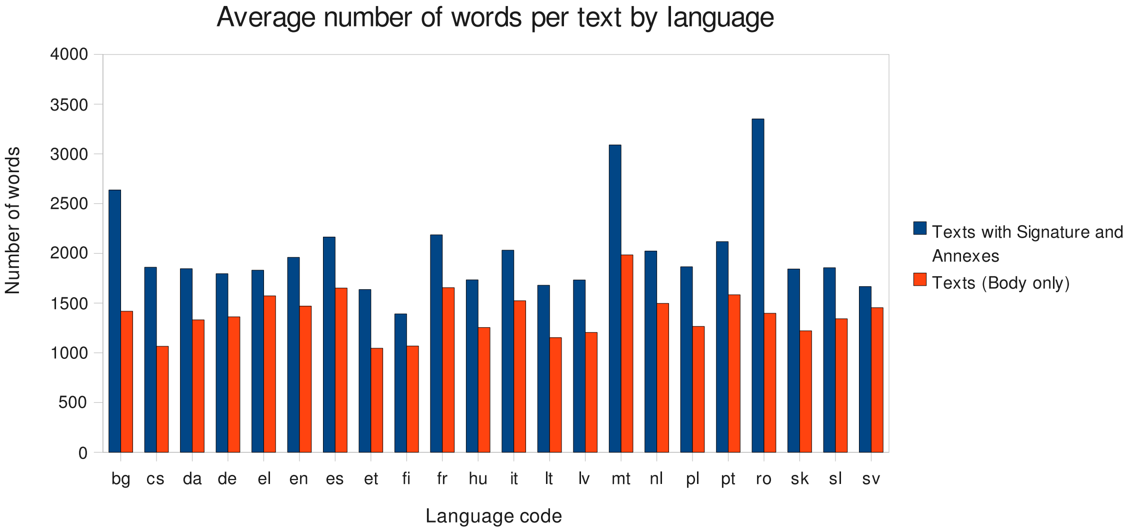


Figure 3.4: JRC-Acquis: the average size of text with and without annexes, by language

The EUROVOC thesaurus [EUROVOC, 1995] exists in one-to-one translations in approximately twenty languages and distinguishes about 6,000 hierarchically organised descriptors (subject domains). Where available, we included the numerical EUROVOC codes into the header of the Acquis documents Fig. 3.1.

The current version of JRC-Acquis contains 20521 classified CELEX codes from 23701 total CELEX codes. The language distribution of documents with EUROVOC descriptors is shown in Figure 3.5.

The EUROVOC subject domain classification in combination with the JRC-Acquis can be used for at least two purposes:

1. the automatic generation of subject domain-specific monolingual or multilingual terminologies [Giguet and Luquet, 2006].

IMPORT	INFORMATION TRANSFER	VETERINARY INSPECTION
PREVENTION OF DISEASE	MARKETING	FOODSTUFF
ORIGINATING PRODUCT	APPROXIMATION OF LAWS	AMNESTY INTERNATIONAL
THIRD COUNTRY	EC COUNTRIES	ANIMAL PRODUCT
HEALTH CERTIFICATE	AGRICULTURAL PRODUCT	PLO
MARKETING STANDARD	TARIFF QUOTA	FISHERY PRODUCT

Table 3.4: Most frequently used EUROVOC descriptors in the JRC-Acquis collection, indicating the most important subject domains of the JRC-Acquis

Language	Language code	Number of documents
Bulgarian	BG	8 259
Czech	CS	18 319
Danish	DA	20 487
German	DE	20 384
Greek	EL	20 153
English	EN	20 382
Spanish	ES	20 479
Estonian	ET	20 389
Finnish	FI	20 426
French	FR	20 462
Hungarian	HU	19 632
Italian	IT	20 312
Lithuanian	LT	20 247
Latvian	LV	19 754
Maltese	MT	7434
Dutch	NL	20 409
Polish	PL	20 311
Portuguese	PT	20 426
Romanian	RO	3 857
Slovakian	SK	18 922
Slovene	SL	17 503
Swedish	SV	17 361

Table 3.5: Number of JRC-Acquis documents with EUROVOC descriptors by language

2. the training of automatic multi-label document classifiers and keyword indexing systems [Civera and Juan, 2006, Pouliquen et al., 2003, Ráez, 2006].

Based on EUROVOC descriptors we selected a Health-related sub-corpora of JRC-Acquis, that was used in our experiments (see section 4.3.1 in the next chapter).

3.10 Acquis Communautaire Translation Memory: DGT Translation Units

3.10.1 Description

As of November 2007, the European Commission's Directorate-General for Translation (DGT) made publicly accessible its multilingual Translation Memory for the Acquis Communautaire, the body of EU law.

A translation memory is a collection of small text segments and their translation (*translation units*). These segments can be sentences or sentence parts. Translation memories are used to support translators by ensuring that pieces of text that have already been translated do not need to be translated again.

The aligned sentences, named “translation units” have been provided by the DGT of the EC by extraction from one of its large shared translation memories in *Euramis* (*European Advanced Multilingual Information System*). This memory contains most, although not all, of the documents of the Acquis Communautaire, as well as some other documents which are not part of the Acquis.

In order to cut down the size, the extraction takes English as the source language. The sequence in the extracted files is not necessarily the same as in the underlying documents, and redundancies of text segments like "Article 1" are inevitable. The documents in the files are identified by the document number (CELEX code) of the original legislative document in the EUR-Lex database, but it should be noted that these documents have been modified. The documents are in TMX format and the texts are encoded in UTF-16 Little Endian. The source language of the documents and sentences is not known, but many of the documents were originally written in English and then translated into the other languages.

3.10.2 Statistics on DGT Translation Units

The DGT Translation Memory is currently available in 22 languages. Table 3.6 shows the coverage, expressed in the total number of translation units available for each language. The number of aligned translation units differs for each language pair.

3.10.3 What is the difference between the DGT Translation Memory and the JRC-Acquis?

The two resources are rather similar in nature as they are both based on the Acquis Communautaire, but they are not identical and can both serve different purposes. The main differences are the following:

- The collection of documents of both resources should mostly be the same, but they are not identical as both resources were collected in different ways. None of the resources is exactly equivalent to the Acquis Communautaire. The criteria for the collection of the JRC-Acquis were rather loose (all the documents which were collected were available in at least ten languages of which at least three “new” EU languages) so that the JRC-Acquis is bigger.

- The DGT Translation Memory is a collection of translation units, from which the full text cannot be reproduced. The JRC-Acquis is mostly a collection of full texts with additional information on which sentences are aligned with each other.

- Most parts of the DGT Translation Memory have been corrected manually using the Euramis alignment editor, while the alignment of the JRC-Acquis docu-

Language	Language code	Number of units
English	EN	2 187 504
Bulgarian	BG	708 658
Czech	CS	890 025
Danish	DA	433 871
German	DE	532 668
Greek	EL	371 039
Spanish	ES	509 054
Estonian	ET	1 047 503
Finnish	FI	514 868
French	FR	1 106 442
Hungarian	HU	1 159 975
Italian	IT	542 873
Lithuanian	LT	1 126 255
Latvian	LV	1 120 835
Maltese	MT	1 021 855
Dutch	NL	502 557
Polish	PL	1 052 136
Portuguese	PT	945 203
Romanian	RO	650 735
Slovakian	SK	1 065 399
Slovene	SL	1 026 668
Swedish	SV	555 362

Table 3.6: Size of DGT's Translation Memory expressed as the total number of translation units per language for each of the 22 official EU languages

ments was done using the two alternative alignment software tools `Vanilla` and `HunAlign`, without manual correction.

- For the cleaning and pre-processing of the texts, different methods and tools were used.

We use sub-corpora from both JRC-Acquis and DGT Translation Units for running our experiments and evaluate our approach.

3.11 Conclusions

Both parallel texts and translation memories are an important linguistic resource that can be used for a variety of purposes, including:

- training automatic systems for statistical machine translation (SMT);

- producing monolingual or multilingual lexical and semantic resources such as dictionaries and ontologies;
- training and testing multilingual information extraction software;
- checking translation consistency automatically;
- testing and benchmarking alignment software (for sentences, words, etc.).

Generally speaking, parallel corpora are useful for all types of cross-lingual research. The value of a parallel corpus grows with its size and the number of languages for which translations exist. While parallel corpora for some languages exist abundantly, there are few or no parallel corpora for most other language pairs. To our knowledge, the Acquis Communautaire is the biggest parallel corpus in existence, if we take into consideration both its size and the large number of languages involved. The most outstanding advantage of the Acquis Communautaire - apart from being freely available - is the number of rare language pairs (e.g. Maltese-Estonian, Slovene-Finnish, etc.).

We will next detail the important steps in JRC-Acquis corpus compilation and we will present the sub-corpora selected from JRC-Acquis and DGT Translation Memory that were used in our experiments.

Chapter 4

Corpus compilation and processing

In the next sections, we will explain how we compiled the JRC-Acquis corpus (section 4.1) and converted it into clean UTF-8 encoded XML texts with paragraph marking (section 4.1.2), enriched with EUROVOC descriptors. We will then summarise the effort to paragraph-align the JRC-Acquis (section 4.2) using two alternative approaches.

The processing presented in these following sections was done to prepare the data for the experiments of this thesis. We created three different subcorpora for that purpose, the first two selected from JRC-Acquis (section 4.3), the third one from DGT Translation Memory (section 4.4), for which we have tokenised the texts (section 4.5).

Finally, the last section will summarise the work on JRC-Acquis and DGT Translation Memory in the context of our thesis.

4.1 Corpus compilation

The work on JRC-Acquis corpus was carried out in the Joint Research Center (JRC) of the European Commission, by the Language Technology team[1], where I worked between 2003 and 2008. The corpus compilation started in 2005 and three different versions were provided up to now, the last one being released in April 2007.

As mentioned in chapter 3, we have attempted to identify the documents which are part of the Acquis Communautaire (AC), have downloaded them and converted them to XML format. The Bulgarian and Romanian documents were processed by the Romanian Academy of Sciences[2]. In further processing steps, the texts were cleaned of their footers and annexes, and enriched with the Eurovoc descriptors.

4.1.1 Gathering the documents

The process consisted in the following steps:

[1] http://langtech.jrc.it

[2] http://www.racai.ro/

1. **Downloading the documents (in HTML format)**

   It is possible to locate the Acquis Communautaire texts via their *CELEX ID* or *CELEX CODE* (unique identifier given for every EU official document). Most documents in the official EU languages could be found in HTML format on the Commissions web site[3], and they can be downloaded with the following URL: `http://europa.eu.int/eur-lex/lex/LexUriServ/LexUriServ.do?uri=CELEX:CELEXCODE:LG:HTML`, where the two parameters *`CELEXCODE`* and *`LG`* should be replaced by their respective values for the CELEX code and the two-digit language code.

   Not all documents (CELEX codes) are translated into each language, so the size of the various language parts can vary considerably.

   Documents in Romanian and Bulgarian languages, for which a translation exists, were only available in Microsoft Word format. Thus, the Romanian and Bulgarian texts of the JRC-Acquis have been downloaded, using the URL: `http://ccvista.taiex.be/Fulcrum/CCVista/$LG/$CELEXCODE-$LG.doc`.

2. **XML conversion (from HTML)**

   After having crawled the mentioned EC web sites and downloaded the selected HTML documents, we converted them to UTF-8-encoded XML format. Each document was then split into numbered paragraph chunks, using the <*BR*> or <*P*> tags from the original HTML documents. As the Acquis texts are consistent, these paragraph chunks are mostly the same across the different languages. They can contain a small number of sentences, but they sometimes contain sentence parts (ending with a semicolon or a comma).

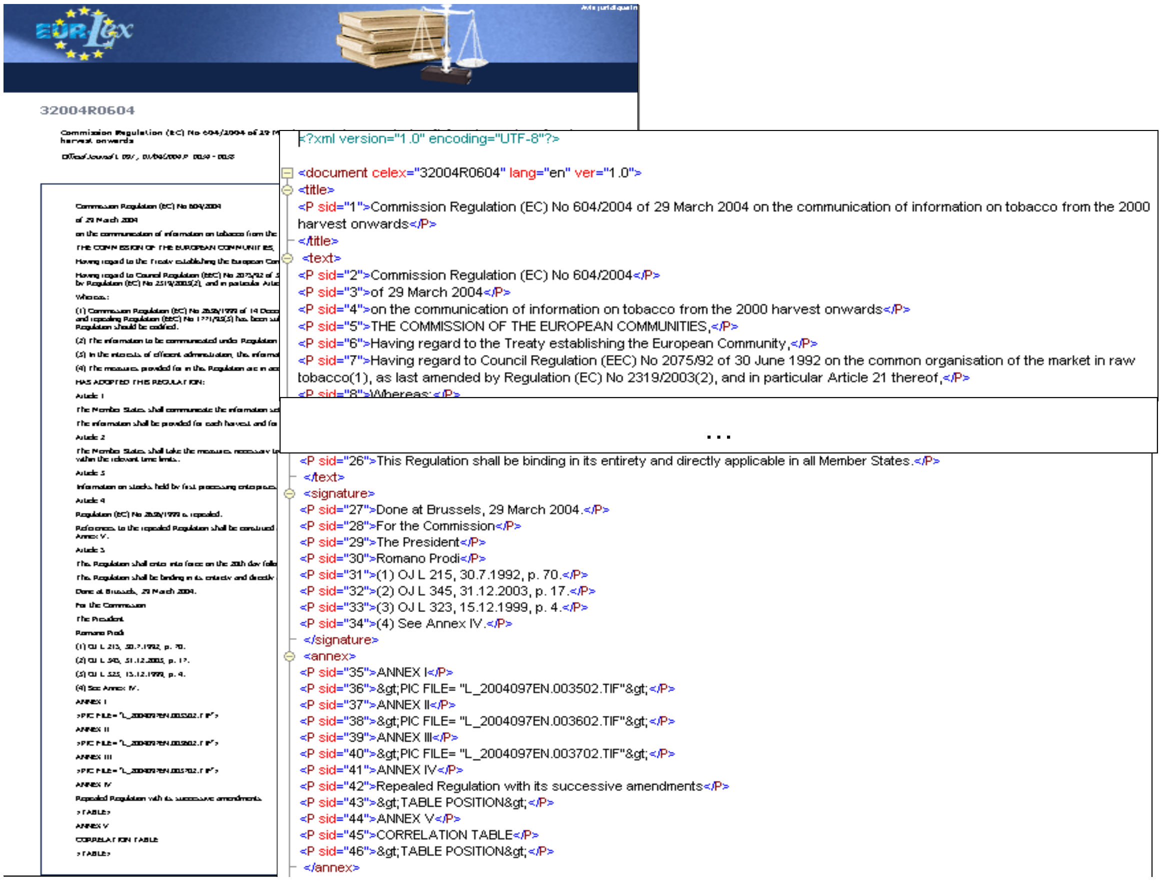
   As a result, each paragraph of the text collection can be uniquely identified using the language, the CELEX identifier and the paragraph number, that will be used in the alignment process.

   The Romanian and Bulgarian documents were converted from their original Microsoft Word format to the xml format of the JRC-Acquis corpus. During the automatic conversion, the translators' annotations and some of the footnotes were discarded. Documents on the ccvista-server do not have an official status yet and the translations may still change.

3. **Language identification on the documents**

   For a small percentage of the documents, the text purportedly in one language is in fact untranslated English text. We verified the language using an n-gram-based in-house language guessing software and we discarded those documents that were not in the expected language.

[3] `http://europa.eu.int/`



```
<?xml version="1.0" encoding="UTF-8"?>

<document celex="32004R0604" lang="en" ver="1.0">
<title>
<P sid="1">Commission Regulation (EC) No 604/2004 of 29 March 2004 on the communication of information on tobacco from the 2000
harvest onwards</P>
</title>
 <text>
<P sid="2">Commission Regulation (EC) No 604/2004</P>
<P sid="3">of 29 March 2004</P>
<P sid="4">on the communication of information on tobacco from the 2000 harvest onwards</P>
<P sid="5">THE COMMISSION OF THE EUROPEAN COMMUNITIES,</P>
<P sid="6">Having regard to the Treaty establishing the European Community,</P>
<P sid="7">Having regard to Council Regulation (EEC) No 2075/92 of 30 June 1992 on the common organisation of the market in raw
tobacco(1), as last amended by Regulation (EC) No 2319/2003(2), and in particular Article 21 thereof,</P>
<P sid="8">Whereas:</P>

                                        ...
 <P sid="26">This Regulation shall be binding in its entirety and directly applicable in all Member States.</P>
  </text>
  <signature>
 <P sid="27">Done at Brussels, 29 March 2004.</P>
 <P sid="28">For the Commission</P>
 <P sid="29">The President</P>
 <P sid="30">Romano Prodi</P>
 <P sid="31">(1) OJ L 215, 30.7.1992, p. 70.</P>
 <P sid="32">(2) OJ L 345, 31.12.2003, p. 17.</P>
 <P sid="33">(3) OJ L 323, 15.12.1999, p. 4.</P>
 <P sid="34">(4) See Annex IV.</P>
  </signature>
  <annex>
 <P sid="35">ANNEX I</P>
 <P sid="36">&gt;PIC FILE= "L_2004097EN.003502.TIF"&gt;</P>
 <P sid="37">ANNEX II</P>
 <P sid="38">&gt;PIC FILE= "L_2004097EN.003602.TIF"&gt;</P>
 <P sid="39">ANNEX III</P>
 <P sid="40">&gt;PIC FILE= "L_2004097EN.003702.TIF"&gt;</P>
 <P sid="41">ANNEX IV</P>
 <P sid="42">Repealed Regulation with its successive amendments</P>
 <P sid="43">&gt;TABLE POSITION&gt;</P>
 <P sid="44">ANNEX V</P>
 <P sid="45">CORRELATION TABLE</P>
 <P sid="46">&gt;TABLE POSITION&gt;</P>
  </annex>
```

Figure 4.1: JRC-Acquis document processing: from HTML to XML, with annex and signature mark up

4.1.2 Reformating with annex and signature detection

The text can be usefully decomposed into the title, body of the text, the signature (e.g. "Done at Brussels, 24 September 2004, for the commission, etc") and annexes (containing tables or lists of codes, usually not translated in all languages). It is the body that will contain most of the "useful" text, yet the backmatter can include a considerable portion of the documents.

These divisions were identified by Perl regular expressions over the texts, using language specific patterns (including Romanian and Bulgarian). We marked them up as body, signature and annex and the resulting corpus was stored as XML (Fig. 4.1). This division into three document parts allows users to concentrate their effort on the text type that is most useful for them.

Note that signatures and annexes are usually marked up, but as they were not always clearly identifiable, we will have missed the mark-up on some of them. We have noticed that with some documents the signature pattern occurs at the beginning. In this case the whole text following the signature pattern was included in the signature division which led to some alignment errors.

4.1.3 Enriching with EUROVOC descriptors

Most CELEX documents have been manually classified into subject domain classes using the EUROVOC thesaurus [EUROVOC, 1995]. Where available, we included the numerical EUROVOC codes into the header of the Acquis documents (Fig. 3.1). A list with all CELEX documents for which we provide the EUROVOC descriptors is also publically available (in tab-separated value format). The latest version (3.0) contains 23701 CELEX documents, from which 20521 CELEX codes present EUROVOC descriptors.

In our experiments we used CELEX documents related to the health domain, by selecting all the CELEX codes that have associated Health-related descriptors.

To prepare our experiments we checked the list of CELEX documents against the EURLEX website[4] to increase the number of EUROVOC descriptors associated. For each CELEX code we downloaded the information available from the URL `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:$celex:en:NOT` (where *$celexcode* should be replaced by the corresponding CELEX code). The EUROVOC descriptors for each Celex code have been identified by Perl scripts. We increased the number of CELEX codes with EUROVOC descriptors up to 23 639 (from 23701), which means only 62 CELEX documents were not classified.

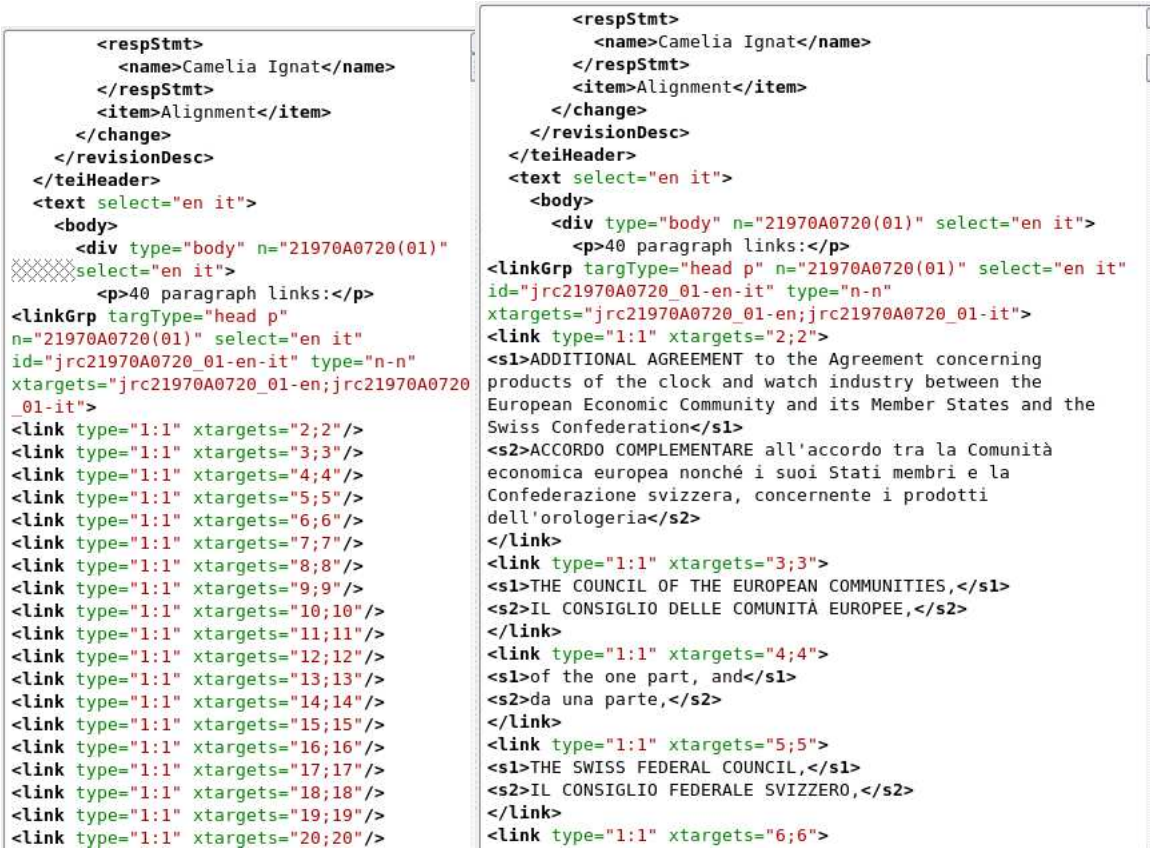
4.2 Paragraph alignment

In further processing steps, the texts were paragraph-aligned. Instead of using a single pivot language, all possible language pair combinations (231) were aligned individually. This is useful due to the n-to-n relationship between aligned sentences, which often differs depending on the language pair involved.

For the paragraph alignment, we used two different tools to align all texts: `Vanilla`, which implements the [Gale and Church, 1991b] alignment algorithm; and `HunAlign` [Varga et al., 2005]. The results for the alignments are available with the distribution of the corpus so that users can use the alignment that suits them best, or for benchmarking exercises. We have not yet been able to carry out a comparative quantitative evaluation of the performance of both tools.

The alignments results were stored for each language pair as TEI-compliant XML file. These documents do not contain actual text, but only pointers to the aligned paragraphs (Fig. 4.2). In the corpus distribution we provide a Perl script that can be used to generate a bilingual aligned corpus for any of the 231 language pairs. The script reads the stand-off alignments and extracts the required paragraphs from the documents in the corpus (or in a selection list) for the language pair of interest, and outputs them as in-place alignments. Figure 4.2 shows an English-Italian sample example.

[4] `http://eur-lex.europa.eu`



```
        <respStmt>
          <name>Camelia Ignat</name>
        </respStmt>
        <item>Alignment</item>
      </change>
    </revisionDesc>
  </teiHeader>
  <text select="en it">
    <body>
      <div type="body" n="21970A0720(01)"
      select="en it">
        <p>40 paragraph links:</p>
<linkGrp targType="head p"
n="21970A0720(01)" select="en it"
id="jrc21970A0720_01-en-it" type="n-n"
xtargets="jrc21970A0720_01-en;jrc21970A0720
_01-it">
<link type="1:1" xtargets="2;2"/>
<link type="1:1" xtargets="3;3"/>
<link type="1:1" xtargets="4;4"/>
<link type="1:1" xtargets="5;5"/>
<link type="1:1" xtargets="6;6"/>
<link type="1:1" xtargets="7;7"/>
<link type="1:1" xtargets="8;8"/>
<link type="1:1" xtargets="9;9"/>
<link type="1:1" xtargets="10;10"/>
<link type="1:1" xtargets="11;11"/>
<link type="1:1" xtargets="12;12"/>
<link type="1:1" xtargets="13;13"/>
<link type="1:1" xtargets="14;14"/>
<link type="1:1" xtargets="15;15"/>
<link type="1:1" xtargets="16;16"/>
<link type="1:1" xtargets="17;17"/>
<link type="1:1" xtargets="18;18"/>
<link type="1:1" xtargets="19;19"/>
<link type="1:1" xtargets="20;20"/>
```

```
        <respStmt>
          <name>Camelia Ignat</name>
        </respStmt>
        <item>Alignment</item>
      </change>
    </revisionDesc>
  </teiHeader>
  <text select="en it">
    <body>
      <div type="body" n="21970A0720(01)" select="en it">
        <p>40 paragraph links:</p>
<linkGrp targType="head p" n="21970A0720(01)" select="en it"
id="jrc21970A0720_01-en-it" type="n-n"
xtargets="jrc21970A0720_01-en;jrc21970A0720_01-it">
<link type="1:1" xtargets="2;2">
<s1>ADDITIONAL AGREEMENT to the Agreement concerning
products of the clock and watch industry between the
European Economic Community and its Member States and the
Swiss Confederation</s1>
<s2>ACCORDO COMPLEMENTARE all'accordo tra la Comunità
economica europea nonché i suoi Stati membri e la
Confederazione svizzera, concernente i prodotti
dell'orologeria</s2>
</link>
<link type="1:1" xtargets="3;3">
<s1>THE COUNCIL OF THE EUROPEAN COMMUNITIES,</s1>
<s2>IL CONSIGLIO DELLE COMUNITÀ EUROPEE,</s2>
</link>
<link type="1:1" xtargets="4;4">
<s1>of the one part, and</s1>
<s2>da una parte,</s2>
</link>
<link type="1:1" xtargets="5;5">
<s1>THE SWISS FEDERAL COUNCIL,</s1>
<s2>IL CONSIGLIO FEDERALE SVIZZERO,</s2>
</link>
<link type="1:1" xtargets="6;6">
```

Figure 4.2: Alignment example (using Vanilla aligner): English-Italian paragraph alignment, with and without the text included

4.2.1 Alignment using Vanilla

Vanilla[5] is a purely statistical aligner which bases its alignment guesses exclusively on sentence length. It implements dynamic time warping by comparing the character counts of possibly aligned sentences [Gale and Church, 1991b]. The Church & Gale's implementation, written in C programming language [Danielsson and Ridings, 1997] was adapted to JRC-Acquis format.

The aligner is provided with the two files split into hard regions, which have to match among the files, and soft regions which are aligned according to the parities 1-1 (one-to-one), 1-2 (splitting), 2-1 (combination), 1-0 (sentence deletion), 0-1 (sentence insertion) and 2-2. In our case each document text corresponds to one hard region. Soft regions are typically sentences, but in our case paragraphs, which, do however tend to be rather short corresponding to one or two sentences or even partial sentences.

As an average for all language pairs, 85.43% of the paragraphs of the JRC-Acquis collection was aligned 1-1, which is roughly in line with the sentence alignment results

[5] http://nl.ijs.si/telri/Vanilla/

of 89% reported by [Gale and Church, 1993]. We report an average of 18 833 aligned documents per language, with an average of 1 052 759 links per language pair.

A brief analysis of the results suggested that:

- the alignment is made more complicated by the fact that some English documents on the Web are previous versions of the ones that served as a source for the translation.
- some alignments errors come from missing mark up of annexes and signatures or other errors in amendments detection. In this case the size of amendments in term of text percentage is not that large but it does raise the error rate of the aligner significantly.
- it would be relatively easy to introduce a pre-processing step that would take into account enumeration tokens (e.g. *1)*, *a)*,...) and declare them as the hard regions for the aligner. This would most likely significantly localise and reduce the alignment errors.

4.2.2 Alignment using HunAlign

The corpus has been processed by the Budapest Technical University, Media Research Centre, using HunAlign, a language-independent sentence aligner [Varga et al., 2005]. Unlike `Vanilla`, `HunAlign` does not emit 2-2 segments, but it can deal with the splitting of a sentence into more than two sentences. For a fixed choice of language pair, the `HunAlign` algorithm runs in three phases.

First, it builds alignments using a simple similarity measure. This measure is based on sentence length and the ratio of identical words. Number tokens are treated specially: similarity of the sets of number tokens in the two sentences is considered. This special treatment is especially useful for legal texts: in the Acquis corpus, 6.5 percent of the tokens are numbers. The one-to-one segments found in this first round of alignment are randomly sampled (10 000 sentence pairs in the case of the Acquis corpus) to feed the second phase of the algorithm: a simple automatic lexicon-building. In the third phase the alignment is re-run, this time also considering similarity information based on the automatically constructed bilingual lexicon. We note that after incremental changes to the corpus, it is not necessary to re-run the first two phases.

4.3 JRC-Acquis sub-corpora

We have created two different subcorpora of JRC-Acquis. For the first one the selection was based on thematic-domain and the second on language availability. The first one, *Health-JRC-Acquis* is a health-related subcorpus, for which the size of the various language parts varies considerably. For the second subcorpus, *Acquis-22*, we selected only the CELEX codes where documents in all the 22 languages were available.

For both subcorpora we proceed with the following processing steps:

- "Cleaning" procedure to remove tables and files references, and all the typographic signs coming from tables.
- Text selection only from the body and the signature of each document (we discard titles and annexes)
- Paragraph alignement using the `Vanilla` aligner, done on the "clean" selected texts resulted from the first two steps, for each language pair combination.
- Text tokenisation using the in-house multilingual tokeniser, `mlToken`, that will be described in the section 4.5.

On *Health-Acquis* subcorpus we ran preliminary domain-specific experiments. The *Acquis-22* subcorpus was used to validate our approach: we have randomly generated different sized sub-corpora of *Acquis-22*, that were used in our experiments.

4.3.1 *Health JRC-Acquis*

This subcorpora includes all the documents of JRC-Acquis corpus, that have been (manually) classified into "health" and "health-related" domain according to the EUROVOC thesaurus.

The steps performed for the sub-corpora compilation are the following:

- Selection of health-related descriptors (from EUROVOC site): we extracted all the descriptors found under the "health" hierarchy (code 2841) and their related terms
- Selection of all CELEX codes that contain at least one of these "health" descriptors
- Generating "health" subcorpus for each language, based on the selected celex codes
- Document "cleaning", text selection, paragraph alignment and tokenisation (as described above)

The resulted corpus includes almost 90 000 documents in all 22 languages. It contains 137 million tokens (114 million words[6]) with an average of 6 million tokens per language, but with a non-uniform language repartition, varying between 2.6 million tokens (in 1788 documents) for Romanian and 7.9 million tokens (in 4400 documents) for each of French and Spanish languages. The distribution per language is shown in the table 4.1.

Thus, the parallel bitexts for different language pairs has different sizes varying from 75 000-85 000 aligned sentences for Greek-Romanian and Bulgarian-Romanian to 254 000 aligned sentences for Estonian-Lithuanian and Estonian-Polish.

[6] The number of words is calculated before tokenisation. We gave the size in words to allow comparison with the JRC-Acquis corpus which has not been tokenised. The size in token allows to compare with other corpora used in SMT.

Language	Nb. of Sentences	Nb. of Words	Nb. of Tokens	Av. Words by Doc.	Av. Token by Doc.
bg	236 453	3 995 646	4 760 569	1 296.45	1 544.64
cs	256 572	5 021 311	6 038 739	1 207.92	1 452.67
da	265 577	5 564 459	6 616 561	1 259.78	1 497.98
de	262 452	5 590 422	6 543 071	1 267.96	1 484.03
el	291 344	6 547 435	7 612 761	1 509.32	1 754.9
en	261 268	6 122 823	7 058 926	1 389.03	1 601.39
es	264 817	7 002 111	7 974 751	1 587.06	1 807.51
et	266 301	4 215 773	5 251 757	955.96	1 190.87
fi	263 950	4 345 908	5 286 048	993.8	1 208.79
fr	261 249	6 573 929	7 949 855	1 486.98	1 798.2
hu	260 465	4 928 771	6 059 758	1 168.79	1 436.98
it	262 273	6 165 606	7 235 537	1 405.75	1 649.69
lt	268 258	4 841 266	5 966 122	1 110.89	1 369
lv	261 811	4 732 429	6 016 057	1 102.1	1 401.04
mt	180 178	3 426 057	5 138 074	1 321.78	1 982.28
nl	261 385	6 242 923	7 162 924	1 418.52	1 627.57
pl	266 285	5 273 619	6 344 440	1 201.28	1 445.2
pt	259 924	6 411 635	7 429 730	1 457.19	1 688.58
ro	138 904	2 312 179	2 688 432	1 293.16	1 503.6
sk	252 881	5 027 410	6 033 791	1 205.32	1 446.61
sl	260 708	5 068 782	6 154 745	1 248.78	1 516.32
sv	254 845	5 318 602	6 125 982	1 327.33	1 528.82
Total	5 557 900	114 729096	137 448 630		
Av. per lg.	252 631.82	5 214958.91	6 247 665		

Table 4.1: Size of the *Health JRC-Acquis* corpus in each of the 22 official EU languages

4.3.2 *Acquis-22*

Acquis22 sub-corpora has been selected on the language availability basis: we have extracted all the documents that have translations in all the 22 languages of the JRC-Acquis corpus.

For its compilation we performed the following processing:

- Selection of the CELEX codes of JRC-Acquis, for which the translation is available in all 22 languages excluding CELEX codes presented in the Acquis development set (section 4.4.2)

- Language subcorpus generation: extracting corresponding documents for each language, based on the CELEX codes

- Document “cleaning”, text selection, paragraph alignment and tokenisation (as described above)

The *Acquis22* corpus includes 114906 documents, that means about 5200 documents for each language. It contains 186 million tokens (156 million words), with an average of 8.4 million tokens (7.1 million words) per language. There are 7.0 million sentences with an average of 360 000 sentences per language. Detailed statistics with the language repartition are presented in Table 4.2.

Language	Nb. of Sentences	Nb. of Words	Nb. of Tokens	Av. Words by Doc.	Av. Token by Doc.
bg	434 864	8 029 841	9 467 753	1 537.4	1 814.44
cs	352 485	6 420 210	7 676 456	1 229.22	1 471.15
da	354 886	7 187 550	8 387 969	1 376.13	1 607.51
de	350 154	7 138 377	8 220 432	1 366.72	1 575.4
el	410 034	8 848 909	10 230 695	1 694.22	1 960.65
en	346 417	8 048 709	9 156 510	1 541.01	1 754.79
es	362 432	9 313 987	10 400 741	1 783.26	1 993.24
et	350 775	5 152 918	6 362 791	986.58	1 219.39
fi	361 476	5 511 001	6 613 098	1 055.14	1 267.36
fr	346 439	8 385 442	9 967 597	1 605.48	1 910.23
hu	354 904	6 255 465	7 637 319	1 197.68	1 463.65
it	353 975	8 059 621	9 186 329	1 543.1	1 760.51
lt	356 470	6 023 252	7 375 204	1 153.22	1 413.42
lv	354 644	6 023 095	7 646 581	1 153.19	1 465.42
mt	351 083	6 508 892	9 700 873	1 246.2	1 859.12
nl	350 227	8 195 670	9 245 117	1 569.15	1 771.77
pl	353 093	6 549 606	7 866 374	1 253.99	1 507.55
pt	345 073	8 292 234	9 515 982	1 587.64	1 823.68
ro	366 167	7 164 729	8 266 982	1 372.29	1 584.93
sk	353 734	6 512 213	7 772 540	1 247.07	1 489.85
sl	350 109	6 384 014	7 725 111	1 222.29	1 480.47
sv	349 345	6 927 240	7 877 667	1 326.3	1 509.71
Total	7 908 786	156 932975	186 300 121		
Av. by lg.	359 490.27	7 133317.05	8 468187.32		

Table 4.2: Size of the *Acquis-22* corpus in each of the 22 official EU languages

4.4 Acquis Translation Units sub-corpora

The JRC-Acquis was compiled and aligned using a completely automatic procedure, with no manual checking of the results. Although, in theory, one should find a consistent correspondance between paragraphs of the same CELEX document in different languages, in practice it is difficult to obtain perfect paragraph alignment. Furthermore, as we are interested in exploiting the multilinguality of the corpus it is even more difficult to get one-to-one paragraph aligment across many languages.

It was for these reasons that we proceeded with compiling a sub-corpora of DGT Translation Memory. We must keep in mind that this corpus is composed by manually checked aligned paragraphs or translation units.

The number of aligned translation units differs for each language pair, that means not all paragraphs have a translation in all 22 languages. For our experiments, we have selected only the translation units available in the all 22 languages and we have built a parallel corpus composed by these one-to-one aligned segments. See Appendix A for an example extracted from this corpus, of a sentence translated in all 22 languages (table A.1 and A.2).

For the triangulation, this is an important resource, as it provides exact sentence-aligned parallel data. In this case, the pivot method can be applied at the alignment level. Thus, the corpus is used to study the different phases when we can use the pivot information (at the alignment level or at the phrase-table generation level).

Last, but not least, to allow the comparison of machine translation systems, it is necessary to define a common test set. Therefore, to be able to compare system performances on different language pairs, we extracted part of this parallel data: a set of sentences (paragraphs) that are aligned with one other across all 22 languages. Thus, we create a development set that includes a test set and a tuning set (necessary to tune the tool, `Moses` decoder, used in our experiments).

In conclusion, we have created two sub-corpora issued from the translation units available in 22 languages: the first one used for training to build translation models (*Translation-Units-22*), and the second one used for tuning and testing the models created (*Acquis-TU DevSet*). The process of sub-corpora compilation consists of the following steps:

- Extracting the CELEX codes and paragraph IDs that are translated in the 22 languages.
- Selecting the CELEX codes and paragraph IDs intented to be part of the developement set, based on some heuristics (paragraph average length in tokens, capital letter at the begging of the paragraph, etc...) and generating the list of paragraphs IDs for each corpus (*Translation-Units-22* and *Acquis-TU DevSet*).
- Generating for each language two corpora in UTF-8 XML format, based on the lists obtained at the second step
- Tokenising the texts

We will describe next the sub-corpora extracted.

4.4.1 Translation-Units-22

The corpus includes around 450 000 sentences, 8.7 million tokens (7.6 million words) for all the languages. It consists of 20729 sentences per language, exactly aligned between the 22 languages, which is a very rare resource. It contains almost 400 000 tokens (350 000 words) per language.

We present the size for each of the 22 languages in the Table 4.3.

Language	Nb. of Words	Nb. of Tokens
bg	370 015	424 524
cs	316 722	360 689
da	341 701	390 810
de	341 957	385 403
el	393 353	436 848
en	389 789	429 340
es	434 935	476 943
et	254 484	296 284
fi	259 601	300 683
fr	408 497	477 180
hu	314 152	365 415
it	382 982	425 759
lt	291 774	343 874
lv	295 276	356 594
mt	329 085	471 897
nl	392 654	434 331
pl	329 734	380 217
pt	398 729	444 134
ro	375 247	419 681
sk	325 252	371 583
sl	321 543	372 170
sv	341 136	375 345
Total	7 608 618	8 739 704
Av. by lg.	345 846.27	397 529.27

Table 4.3: Size of the *Translation-Unit-22* corpus in each of the 22 official EU languages

4.4.2 Acquis Development Set (*Acquis-TU DevSet*)

The development set was built for the quality tuning of tools used in our experiments (`Moses` system) and for the testing. Thus, we split it into two parts: the *test set* and the *tuning set*.

The test set contains almost 2 million tokens. It includes 2000 sentences for each languages with an average of 87667 tokens by language.

The tuning set contains 660 sentences with an average of 26056 tokens by language, and a total size of about 500 000 tokens.

Both corpora are in text UTF-8 format, as required by the testing procedure with `Moses`. The test set has been also reformated in XML (SGML) required by the evaluation tool used in our experiments.

4.5 Tokenisation

We performed text tokenisation in a multilingual setting on the subcorpora described above in order to prepare the data for our experiments. The tokenisation module has been developped with the aim to address language-specific tokenisation issues.

Our multilingual tokenisation module `mlToken` is written in Perl, and in addition to splitting the text input string into tokens has also the following features:

- It assigns to each token its token type. The types distinguish not only between words and punctuation marks but also mark digits, abbreviations, left and right splits (i.e. clitics, e.g. *s*), enumeration tokens (e.g. *a)*), as well as URLs and email addresses.

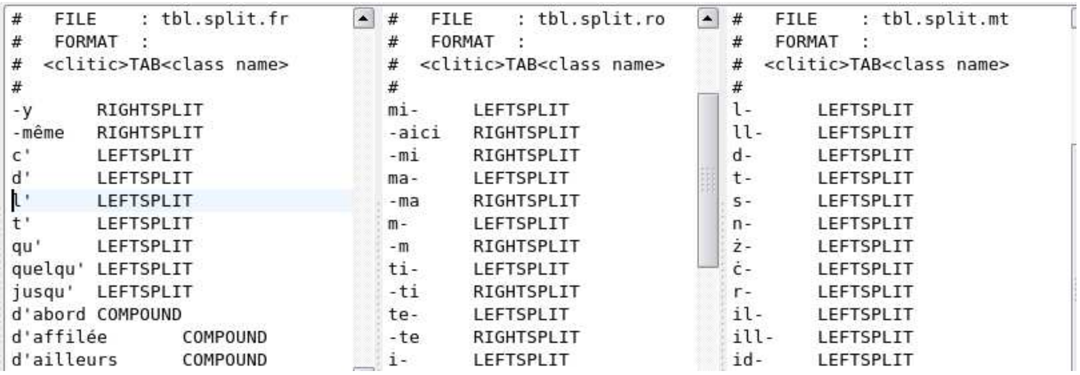
- It marks the end of paragraphs and the end of sentence punctuation, where the sentence internal periods are distinguished from the sentence final ones.

- It preserves (subject to a flag) the inter-word spacing of the original document, so that the input can be reconstituted from the output. This consideration is important when several tokenisers are applied to a text, either for evaluation or production purposes.

The model used for our tokeniser was `mtseg`, the tokeniser (and segmenter) developed in the MULTEXT project [Di Cristo, 1996]; as with `mtseg`, `mlToken` also stores the language dependent features in resource files; in the case of `mlToken` we use abbreviations and split / merge patterns. Figure 4.4 presents the split file for French, Romanian and Maltese.

In the absence of a certain language resource, the tokeniser uses default resource files in order to achieve best results, however, resource files for a language need to be written - this task is helped by having pre-tokenised corpora for the language.

The tokenisation is an important step to prepare data for the translation model training as it alleviates the data sparseness problem. Providing language specific resources might help in this sense for certain kind of languages. Table 4.5 shows the number of different tokens compared with the number of different words for *Translation-Units-22* corpus in French, Romanian, Maltese, English, Finnish, Slovene.



```
#   FILE    : tbl.split.fr
#   FORMAT  :
#  <clitic>TAB<class name>
#
-y      RIGHTSPLIT
-même   RIGHTSPLIT
c'      LEFTSPLIT
d'      LEFTSPLIT
l'      LEFTSPLIT
t'      LEFTSPLIT
qu'     LEFTSPLIT
quelqu' LEFTSPLIT
jusqu'  LEFTSPLIT
d'abord COMPOUND
d'affilée       COMPOUND
d'ailleurs      COMPOUND
```

```
#   FILE    : tbl.split.ro
#   FORMAT  :
#  <clitic>TAB<class name>
#
mi-     LEFTSPLIT
-aici   RIGHTSPLIT
-mi     RIGHTSPLIT
ma-     LEFTSPLIT
-ma     RIGHTSPLIT
m-      LEFTSPLIT
-m      RIGHTSPLIT
ti-     LEFTSPLIT
-ti     RIGHTSPLIT
te-     LEFTSPLIT
-te     RIGHTSPLIT
i-      LEFTSPLIT
```

```
#   FILE    : tbl.split.mt
#   FORMAT  :
#  <clitic>TAB<class name>
#
l-      LEFTSPLIT
ll-     LEFTSPLIT
d-      LEFTSPLIT
t-      LEFTSPLIT
s-      LEFTSPLIT
n-      LEFTSPLIT
ż-      LEFTSPLIT
ċ-      LEFTSPLIT
r-      LEFTSPLIT
il-     LEFTSPLIT
ill-    LEFTSPLIT
id-     LEFTSPLIT
```

Table 4.4: Tokeniser's resources for French, Romanian and Maltese: split and merge patterns

Language	Nb of tokens	Nb of different tokens	Nb of words	Nb of different words
French	477180	13845	408497	27617
Romanian	419681	17988	375247	31631
Maltese	471897	19380	329085	43205
English	429340	12036	389789	23057
Finnish	300683	36406	259601	50212
Slovene	372170	24809	321543	39649

Table 4.5: *Translation-Units-22* corpus: Number of different tokens compared with the number of different words for French, Romanian, Maltese, English, Finnish and Slovene

4.6 Contributions

We presented the compilation of the highly multilingual corpus JRC-Acquis which was accomplished during our stay at the Joint Research Center of the European Commission. In this context, we carried out the compilation of the corpus and the alignment using `Vanilla` aligner. The first publicly released version of JRC-Acquis was described in [Steinberger et al., 2006].

The tokenizer `mlToken` was also developed in JRC, and used in different multilingual contexts in in-house applications. It was also integrated in the corpus annotation tool `totale` described in [Erjavec et al., 2005].

The subcorpora presented in section 4.3 and section 4.4 *(Acquis-22, Health-Acquis, Translation-Units-22, Acquis-TU-Devset)* have been created in the context of our thesis, in order to study and validate the pivot SMT approach. These subcorpora will be (probably) publicly available in the short future.

Part III

Alignment and translation models

Chapter 5

Translation models based on Acquis Communautaire

This chapter presents the application of the Acquis subcorpora (described in 4.3 and 4.4), to the task of statistical machine translation. We used the corpora *Translation-Units-22* and *Acquis22* to build 462 machine translation systems for all the possible language pairs in both directions. To perform phrase-based SMT, we used `Moses` tool. We evaluated the quality of the system with the widely used BLEU metric (as described in 2.3.3.7), which measures overlap with a reference translation. We tested on the *Acquis-TU test set* drawn from the Translation Units corpus (described in 4.4.2).

The resulting systems and their performances demonstrate the different challenges presented to statistical machine translation for different (non-traditional) language pairs.

Our approach relies on the phrase-based statistical machine translation framework described by [Koehn et al., 2003]. We will present it briefly in the next section, followed by the description of the `Moses` toolkit, and the main steps of building a translation system based on it.

The section 5.3 will explain why and how we have created the translation models based on Acquis corpus and will further discuss the challenges raised by these models.

5.1 Building a translation model

A statistical translation model [Brown et al., 1993, Och and Ney, 2003] describes the relationship between a pair of sentences in the source (s) and target (t) languages using a translation probability $p(t|s)$[1].

Statistical machine translation systems are based on probabilistic models automatically induced from corpora. The principle on which they rely to generate grammatical

[1] In the next chapters, we will use the notation s for the source and t for target language segments, although in the state-of-the-art, for historical reasons, we make use of the notation f (Foreign) and e (English) for source and target respectively.

sentences in the target language is a calculation of the cheapest cost for the best combination of hypotheses out of a range of possibilities.

Classic SMT systems implement the noisy channel model: given a sentence in the source language s, we try to choose the translation in language t that maximises $p(t|s)$. According to Bayes rule, this can be rewritten as:

$$arg\,\underset{t}{max}\,p(t|s) = arg\,\underset{t}{max}\,p(s|\,t)p(t)$$

where $p(t)$ is materialised with a language model – typically, a smoothed n-gram language model in the target language – and $p(s|t)$ with a translation model – a model induced from parallel corpora – aligned documents which are the translations of one other.

Several different methods have been used to implement the translation model, and additional models such as fertility and distortion / reordering models have also been employed, as in among the first translation schemes proposed by the IBM Models 1 through 5 in the late 1980's [Brown et al., 1993].

The decoder is the algorithm that calculates the most probable translation out of several possibilities, derived from the models at hand.

The phrase-based statistical machine translation model we present here was defined by [Koehn et al., 2003]. The alternative phrase-based methods differ in the way the phrase table is created.

5.1.1 The formal model

We have described the "Phrase-based model in SMT" in the "Framework chapter", section 2.3.3, in its historical context, as a promising extension of word-based models.

In this section, we will define the phrase-based machine translation model formally, as described by Koehn, Och and Marcu. This translation model is based on the noisy channel model, it uses the Bayes rule to reformulate the translation probability for translating a source sentence s into target t as

$$arg\,\underset{t}{max}\,p(t|s) = arg\,\underset{t}{max}\,p(s|\,t)p(t) \tag{5.1}$$

This allows for a **language model** $p\,(t\,)$ and a separate **translation model** $p(s|t)$.

During decoding, the input sentence s is segmented into a sequence of I phrases $\overline{s}_1^I$. We assume a uniform probability distribution over all possible segmentations. Each source phrase s_i in $\overline{s}_1^I$ is translated into a target phrase t_i. The target phrases may be reordered.

While the equation 5.1 gives the **generative framework** used in the training of the phrase-based model, the decoder is based on a **log-linear formulation** which breaks the probability down into an arbitrary number of weighted feature functions (see equation2.14 in section 2.3.3.3):

$$\hat{t} = arg\,\max_{t} p\,(t|s) = arg\,\max_{t} \sum_{m=1}^{M} \lambda_m h_m\,(t,s) \qquad (5.2)$$

This gives a mechanism, during the decoding, to break down the assignation of cost in a modular way based on different aspects of translation.

The SMT systems use a log-linear model of $p(t|s)$ that incorporates geneartive models as feature functions.

5.1.1.1 Generative framework

First, we detail the language model and the lexical translation model in a generative framework.

Language model and word penalty In order to calibrate the output length, a factor ω (called word cost) was introduced for each generated target laguage word, in addition to the trigram language model p_{LM}. This is a simple means to optimize performance. Usually, this factor is larger than 1, biasing toward longer output.

Translation model The translation model includes the lexical translation model (the phrase table) and the reordering model.

Lexical translation model (phrase-table) Phrase translation is modelled by a probability distribution $\phi(s_i|t_i)$. According to the Bayes rule, the translation direction is inverted from a modelling standpoint. The phrase translation probability distribution is estimated by relative frequency:

$$\phi(\overline{s}\,|\overline{t}) = \frac{count(\overline{s},\overline{t})}{\sum_{\overline{s}} count(\overline{s},\overline{t})} \qquad (5.3)$$

Lexical weights: One way to validate the quality of a phrase translation pair is to check how well its words translate into each other. For this, a lexical translation probability distribution w(f|e) is used . This is estimated by relative frequency from the same word alignments as the phrase model.

$$w(s|t) = \frac{count(s,t)}{\sum_{s'} count(s',t)} \qquad (5.4)$$

A special target NULL token is added to each target sentence and aligned to each unaligned source word.

Given a phrase pair $\overline{s}$, $\overline{t}$ and a word alignment a between the source word positions $i = 1,\ldots,n$ and the target word positions $j = 0,1,\ldots,m$ we compute the lexical weight p_w by

$$p_w\left(\overline{s}|\overline{t},a\right)=\prod_{i=1}^{n}\frac{1}{\left|\{j|(i,j)\in a\}\right|}\sum_{\forall(i,j)\in a}w\left(s_i|t_j\right) \tag{5.5}$$

If there are multiple alignments a for a phrase pair $(\overline{s},\overline{t})$, we use the one with the highest lexical weight.

Reodering model Usually, reordering of the target output phrases is modelled by a relative distortion probability distribution $d(start_i, end_{i-1})$, where $start_i$ denotes the start position of the source phrase that was translated into the i-th target phrase, and end_{i-1} denotes the end position of the source phrase that was translated into the $(i-1)$-th target phrase. A simple distortion model $d(start_i, end_{i-1}) = \alpha^{|start_i - end_{i-1} - 1|}$ is used, with an appropriate value for the parameter α.

We are using a more complex reordering model, that will be detailed at the end of this section: lexicalized reordering model.

To summarise, in a **generative framework**, the best target language output sentence t_{best} given a source input sentence s according to the model is:

$$t_{best}=arg\,\max_t p(t|s)=arg\,\max_t p(s|t)p_{LM}(t)\omega^{length(t)} \tag{5.6}$$

where $p(s|t)$ is decomposed into

$$p(\overline{s}_1^I|\overline{t}_1^I)=\prod_{i=1}^{I}\phi(\overline{s}_i|\overline{t}_i)d(start_i,end_{i-1})$$

When we use the lexical weight p_w during translation as an additional factor, this means that the model $p(s|t)$ is extended to

$$p(\overline{s}_1^I|\overline{t}_1^I)=\prod_{i=1}^{I}\phi(\overline{s}_i|\overline{t}_i)d(start_i,end_{i-1})p_w(\overline{s}_i|\overline{t}_i,a)^{\lambda} \tag{5.7}$$

The parameter λ defines the strength of the lexical weight p_w. Good values for the parameter are around 0.25 (after [Koehn et al., 2003]).

5.1.1.2 Log-linear framework

In a log-linear model the formula 5.6 becomes:

$$t_{best}=arg\,\max_t p(t|s)=arg\,\max_t\sum_{m=1}^{M}\lambda_m h_m(t\,,\,s)) \tag{5.8}$$

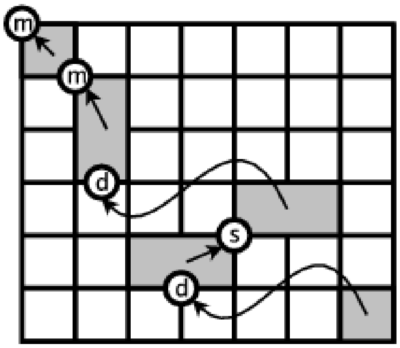


Figure 5.1: Reordering types considered by the lexicalized reordering model: (m) monotone order, (s) switch with previous phrase and (d) discontinous.

where $h_m(t, s)$ is a feature function and λ_m is a weight. The model uses a total of eight feature functions: a trigram language model probability of target language, two phrase translation probabilities (both directions), two lexical translation probabilities (both directions), a word penalty, a phrase penalty, and a linear reordering penalty [Koehn et al., 2003, Koehn, 2004a]. To set the weights λ_m, the minimum error rate training [Och and Ney, 2003] is carried out using BLEU [Papineni et al., 2002] as an objective function.

The phrase-based model, as described above has been implemented by `Moses`, a state of the art SMT system, that we will describe in the following section.

Lexicalized reordering model

The standard reordering model for phrase-based statistical machine translation is only conditioned on movement distance. However, some phrases are reordered more frequently than others. A French adjective like *extérieur* is typically switched with the preceding noun, when translated into English.

Therefore, additional conditional reordering models may be built. These are conditioned on specified factors (in the source and target language), and learn different reordering probabilities for each phrase pair (or just the source phrase).

We are using a lexicalized reordering model that conditions reordering on the actual phrases. One concern, of course, is the problem of sparse data. A particular phrase pair may occur only a few times in the training data, making it hard to estimate reliable probability distributions from these statistics.

Therefore, in the lexicalized reordering model, only three reordering types are considered: (m - mono) monotone order, (s - swap) switch with previous phrase, or (d) discontinuous. See Figure 5.1 for an illustration of these three different types of orientation of a phrase.

The reordering model p_o predicts an orientation type m, s, d given the phrase pair currently used in translation: $p_o(orientation \,|\, s, t)$, where $orientation \in m, s, d$.

The probability distribution can be learnt from the training data. Given the word alignment table, an orientation type can be extracted for each phrase pair, defined as follows:

- monotone: if a word alignment point to the top left exists,
- swap: if a word alignment point to the top right exists,
- discontinuous: if neither a word alignment point to the top left nor to the top right exists, (it is neither monotone order, nor a swap).

We count how often each extracted phrase pair is found with each of the three orientation types. The probability distribution p_o is then estimated based on these counts using the maximum likelihood principle:

$$p_o(orientation\,|\,s,t) = \frac{count\,(orientation, t, s)}{\sum\limits_{o} count\,(o, t, s)} \tag{5.9}$$

Given the sparse statistics of the orientation types, we can smooth the counts of the unconditioned maximum-likelihood probability distribution with a factor σ, as follows:

$$p(orientation) = \frac{\sum\limits_{s}\sum\limits_{t} count\,(orientation, t, s)}{\sum\limits_{o}\sum\limits_{s}\sum\limits_{t} count\,(o, t, s)} \tag{5.10}$$

$$p_o(orientation\,|\,s,t) = \frac{\sigma\, p\,(orientation) + count\,(orientation, t, s)}{\sigma + \sum\limits_{o} count(o, t, s)} \tag{5.11}$$

There is a number of variations of the lexicalized reordering model based on orientation types:

- bidirectional: For each phrase, the ordering of itself with respect to the previous is considered. For bidirectional models, the ordering of the next phrase with respect to the currect phrase is also modelled.
- f and e: Out of sparse data concerns, we may want to condition the probability distribution only on the source (foreign - f) phrase or the target (English - e) phrase. The model may be conditioned on the source phrase (f - Foreign), or on both the source phrase and target phrase
- monotonicity: To further reduce the complexity of the model, we might merge the orientation types swap and discontinuous, leaving a binary decision about the phrase order. Monotonicity models consider only monotone or non-monotone types.

These variations have shown to be occasionally beneficial for certain training corpus sizes and language pairs.

In a lexicalized reordering model "bidirectional", "fe" and non "monotonicity", the log-linear formula 5.8 will take into account new weights: the linear reordering penalty should be replaced by six other scores: the probability for three orientation types (mono, swap, discontinous) for the current phrase with respect to the previous and for the next phrase with respect to the current one.

5.1.2 Moses SMT system

5.1.2.1 Description

The toolkit is a complete out-of-the-box translation system for academic research. It consists of all the components needed to preprocess data, train the language models and the translation models.

It relies upon several models, including the language and translation models described above, and a decoding algorithm. The translation model used by `Moses` is trained from parallel corpora using word alignment methods, and includes a probability distribution over phrase pairs (rather than just single words) of source and target languages. Additional models (a distortion/reordering model and word penalty) are included in the best translation calculation, which is searched for by beam-search decoding.

It also contains tools for tuning these models using minimum error rate training [Och, 2003] and evaluating the resulting translations using the BLEU score [Papineni et al., 2002]. `Moses` uses standard external tools for some of the tasks to avoid duplication, such as `GIZA++` [Och and Ney, 2003] for word alignments and `SRILM` [Stolcke, 2002] for language modelling.

5.1.2.2 GIZA++

`GIZA++` [Och and Ney, 2003] is a software for learning word-by-word alignments between corresponding bisentences and was developed by Franz Joseph Och and Hermann Ney as an enhancement of the `GIZA` tool written in 1999 (at Summer workshop hosted by the Center for Language and Speech Processing (CLSP) at John Hopkins University). `GIZA++` implements partly refined versions of all five IBM models [Brown et al., 1993] and is freely available. It is required to use the training scripts provided by the `Moses` SMT system.

5.1.2.3 SRI Language Modelling Toolkit

The `SRI Language Modelling Toolkit` (`SRILM`) was developed by Andreas Stolcke to build and apply statistical language models. It received some advancements during the CLSP Summer Workshops between 1995 and 2002 at John Hopkins University.

The SRILM package includes a set of C++ libraries, executable programs as well as miscellaneous scripts, all aiming at tasks related to training LMs and their usage. The capabilities and design of the software are described in [Stolcke, 2002]. SRILM is recommended for use with Moses as the latter depends on some of its class libraries for compilation. Moses provides other components for language modelling which we have not used so far.

5.1.2.4 Lexical phrase-based translation with Moses

In Moses, the calculation of the best translation is mainly based on a translation model and a language model. These models are implemented with a phrase translation table, where translation probabilities for phrase pairs are stored, and a smoothed n-gram language model of the target language. In addition, a reordering model and a word penalty model are computed.

$$p(t|s) = p_{\phi}(s|t)^{\lambda_{\phi}} \times p_{LM}(t)^{\lambda_{LM}} \times p_{D}(t,s)^{\lambda_{D}} \times \omega^{length(t)\lambda_{w(t)}}$$

As can be seen above, these models are weighted, and their product enable the system to rank translation hypotheses according to their probability of representing a correct translation in the target language. The algorithm which performs that calculation, the decoder, expands a space of hypotheses based on the probabilities from the models, and performs a search through this space for the best hypotheses. This search is maximised using hypothesis recombination, but also pruning methods such as future cost estimation.

5.2 Moses' processing steps

5.2.0.5 Overview

Figure 5.2 gives an overview of the translation model building process with Moses.

Moses provides the main fonctionalities of a SMT system:

- The training module for building the Translation Model (TM), which consists in a lexicalized translation model (phrase-table) and a lexicalized reordering model.
- A tool for building the Language Model (LM)
- A tuning tool (which is not represented in our schema) that can realise the tuning for quality of the system.
- The decoder that performs the translation based on the translation model and the language model.

Each phase will be detailed in the next subsections.

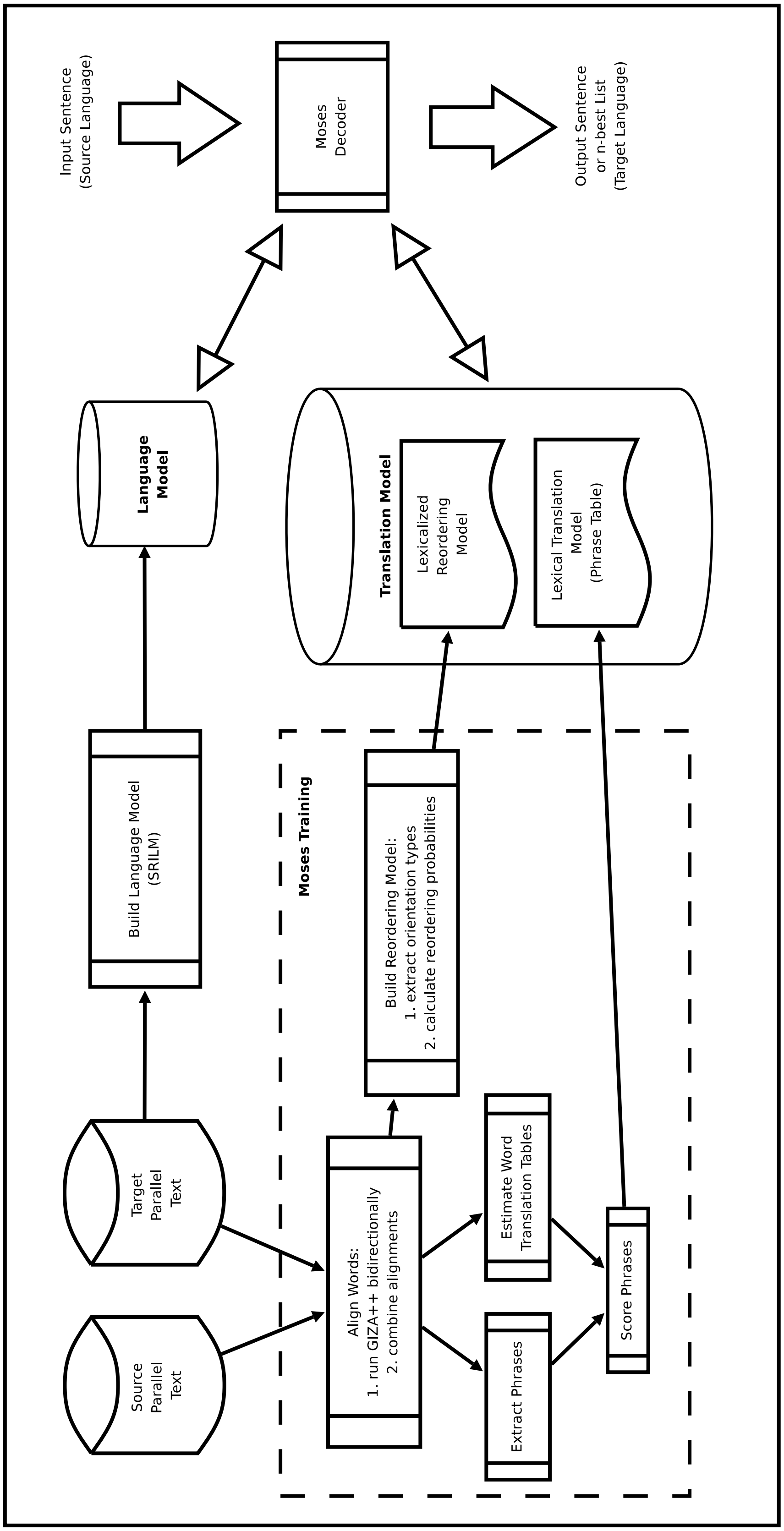

Source Parallel Text
Target Parallel Text
Build Language Model (SRILM)
Language Model
Moses Training
Align Words:
1. run GIZA++ bidirectionally
2. combine alignments
Extract Phrases
Estimate Word Translation Tables
Score Phrases
Build Reordering Model:
1. extract orientation types
2. calculate reordering probabilities
Translation Model
Lexicalized Reordering Model
Lexical Translation Model (Phrase Table)
Input Sentence (Source Language)
Moses Decoder
Output Sentence or n-best List (Target Language)

Figure 5.2: Overview of `Moses` SMT system: building the translation model, building the language model and decoding with `Moses`

5.2.0.6 Training

By the training process, Moses generates the translation model used by the decoder. As mentioned earlier, the translation model in Moses is composed of a translation table and a distortion or reordering model. These are automatically induced from a parallel corpus. Phrase translation tables represent phrases in the source language and their possible translations into the target language, graded with probabilities as automatically learned from the parallel corpus.

Word alignment The algorithm used for word alignment is the EM (Expectation-Maximization) algorithm proposed in GIZA++ (see 2.3.3.1 - Parameter Estimation). This algorithm aligns tokens in sentence pairs extracted from the parallel corpus and finds the most likely word alignment by iterative search. Moses makes use of bidirectional runs of GIZA++: this is because one run of the algorithm can only generate one-to-many translation, from target to source language.

To establish word alignments based on the two GIZA++ alignments, a number of heuristics may be applied. The default heuristic *grow-diag-final* starts with the intersection of the two alignments and then adds additional alignment points from the union of the two runs (see 2.3.3.2 - Symmetrizing word alignments). Other alternative alignment methods can be specified and used depending on the application (*intersect*, *union*, *grow*, *grow-diag*, *srctotgt*, *tgttosrc*).

Lexical translation model (the phrase-table) The phrase pairs that are consistent with the word alignment are collected. The heuristics used to **extract phrases** from the word alignment are described in 2.3.3.2. The translation table, which represents the probability of source (s) language phrases translation into target (t) language phrases (or $\phi(t|s)$) is then built by computing a probability distribution by relative frequency over these phrase pairs:

$$\phi(\overline{s}\,|\overline{t}) = \frac{count(\overline{s}, \overline{t})}{\sum_{\overline{s}'} count(\overline{s}', \overline{t})}$$

It shall be noted that no smoothing is performed on the translation table, relegating the sparse data problem to lexical weighting.

Next to phrase translation probability distributions $\phi(\overline{s}|\overline{t})$ and $\phi(\overline{t}|\overline{s})$, additional phrase translation scoring functions can be computed, e.g. lexical weighting, word penalty, phrase penalty.

In order to calculate the lexical weighting, a maximum likelihood lexical **word translation table** is extracted from the alignment. The lexical translation probability $w(t|s)$, as well as the inverse $w(s|t)$ are estimated, and the lexical weights are calculated based on the alignment and on the lexical probabilities using the formula 5.5.

Currently, five different **phrase translation scores** are computed:

- phrase translation probability $\phi(\overline{s}|\overline{t})$

- lexical weighting $w_p(\overline{s}|\overline{t})$
- phrase translation probability $\phi(\overline{t}|\overline{s})$
- lexical weighting $w_p(\overline{t}|\overline{s})$
- phrase penalty (always $exp(1) = 2.718$)

Lexicalized reordering model Reordering is modelled by a relative distortion probability distribution over the sentence pairs.

By default, only a distance-based reordering model is included in final configuration. This model gives a cost linear to the reordering distance. For instance, skipping over two words costs twice as much as skipping over one word.

However, additional conditional reordering models may be built - different lexicalized reordering models (as described above in 5.1.1.2). We are using a lexicalized reordering model, which is generated from the word alignments, in two steps. The first extracts the ordering type for each phrase and the second calculates the reordering probabilities and generates the reordering model.

The possible configurations are:

- msd vs. monotonicity. MSD models consider three different orientation types: monotone, swap, and discontinous. Monotonicity models consider only monotone or non-monotone, in other words swap, and discontinous are placed together.
- f vs. fe. The model may be conditioned on the source phrase (f - Foreign), or on both the source phrase and target phrase (fe - ForeignEnglish).
- unidirectional vs. bidirectional. For each phrase, the ordering of itself in respect with the previous is considered. For bidirectional models, also the ordering of the next phrase with respect to the current phrase is modelled.

Moses allows the arbitrary combination of these decisions to define the reordering model type (e.g. bidrectional-monotonicity-f).

5.2.0.7 Building the language model

A language model is a statistical model the parameters of which are learned from corpora: word sequences (or n-gram) probabilities are estimated by computing their relative frequency in the corpus. The language model toolkit we used in our experiments is the freely available `SRILM` toolkit [Stolcke, 2002].

5.2.0.8 Tuning for quality: Minimum Error Rate Training

Minimum Error Rate Training, or MERT [Och and Ney, 2002], optimises translation quality by setting the model weight parameters. This is done by taking a held-out section of the parallel corpus, running the decoder with its current translation model on the source language text, and then automatically evaluating the output's translation quality by comparing it to real translation (using automatic methods such as BLEU and word error rate). The weights attributed to the current models are then adjusted accordingly, and the process is iterated until convergence.

5.2.0.9 Decoding

Filtering the phrase table Filtering the phrase table according to the test set we intend to use enables us to tune the decoding process for memory usage. Indeed, by limiting the phrase table to phrases that appear in the test data and their potential translations, we avoid loading the entire phrase table.

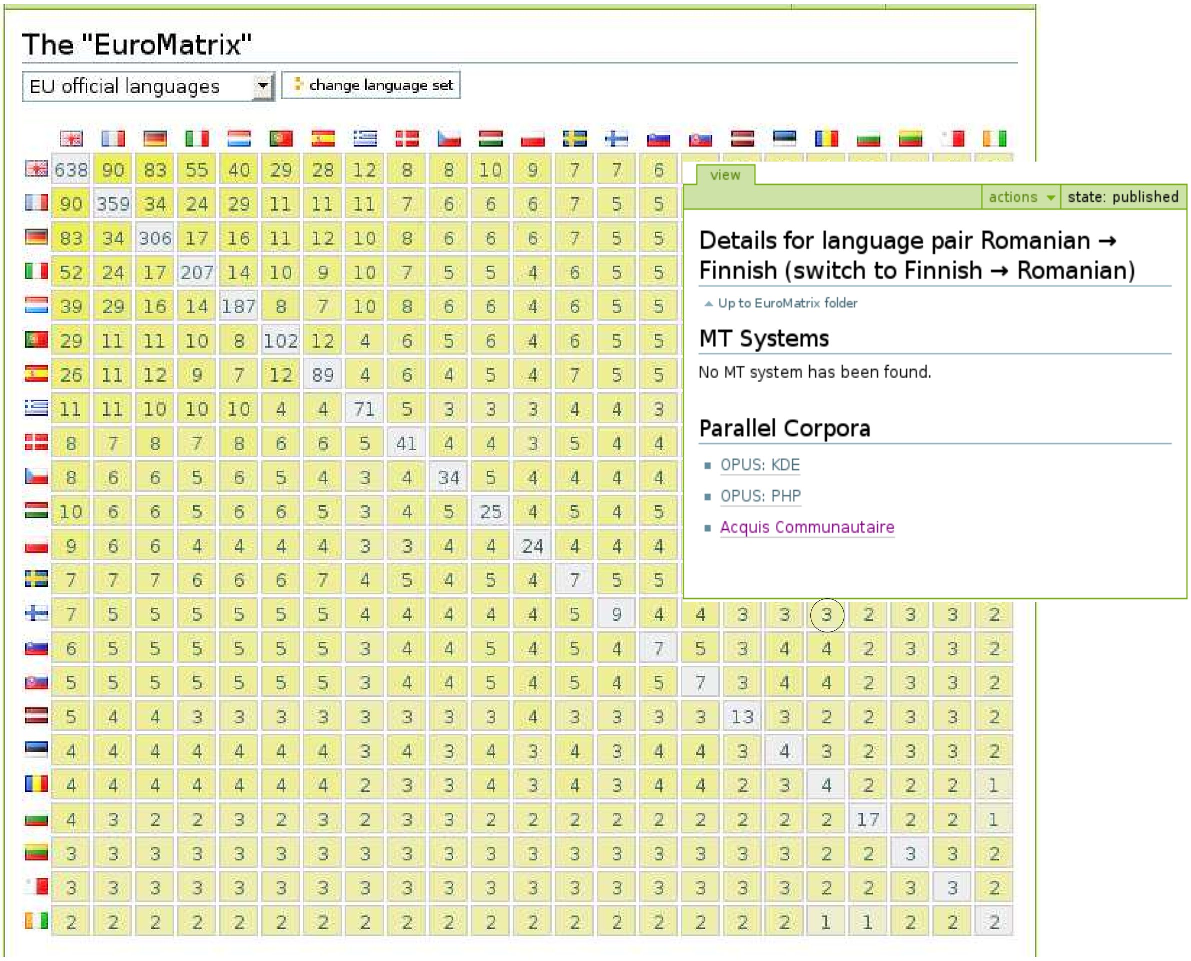
Beam search decoding `Moses`' decoder can translate files one sentence per line in the source language. To translate a sentence, the decoder generates a first hypothesis, or partial translation of a phrase in the input. Then, another hypothesis is generated, based on the previous: the decoder keeps a stack of the best partial translations until now. The notion of "best", or "low cost" is equivalent to "most probable", where probabilities for a hypothesis are the product of probabilities given by the models discussed above.

The decoder uses several methods to limit the search space, including recombination of hypotheses, which is risk-free, and beam search, which risks the pruning of good translation hypotheses. This search algorithm estimates hypothesis cost based on both the future cost (a possibly pre-computed calculation of the part of the sentence which has not yet been decoded, including the language model and translation model factors) and the cost so far, and prunes out more costly hypotheses to only expand those that are likely to succeed. The future cost calculation does not however take into consideration the reordering cost; also, it only gives an estimate of the language model cost. It is thus prone to error. Eventually, the best scoring final translation is outputted. The decoder reads from a configuration file which indicates where the translation models are located, as well as the different weights to these models.

5.2.0.10 Evaluation

We have calculated the BLEU and NIST scores with NIST BLEU scoring tool `mteval-v11b.pl`[2].

[2] `http://www.itl.nist.gov/iad/mig//tools/`



The "EuroMatrix"

EU official languages ▾ | change language set

	🇬🇧	🇫🇷	🇩🇪	🇮🇹	🇳🇱	🇵🇹	🇪🇸	🇬🇷	🇩🇰	🇨🇿	🇭🇺	🇵🇱	🇸🇪	🇫🇮	🇸🇮	🇸🇰	🇱🇻	🇪🇪	🇷🇴	🇧🇬	🇱🇹	🇲🇹	🇮🇪
🇬🇧	638	90	83	55	40	29	28	12	8	8	10	9	7	7	6								
🇫🇷	90	359	34	24	29	11	11	11	7	6	6	6	7	5	5								
🇩🇪	83	34	306	17	16	11	12	10	8	6	6	6	7	5	5								
🇮🇹	52	24	17	207	14	10	9	10	7	5	5	4	6	5	5								
🇳🇱	39	29	16	14	187	8	7	10	8	6	6	4	6	5	5								
🇵🇹	29	11	11	10	8	102	12	4	6	5	6	4	6	5	5								
🇪🇸	26	11	12	9	7	12	89	4	6	4	5	4	7	5	5								
🇬🇷	11	11	10	10	10	4	4	71	5	3	3	3	4	4	3								
🇩🇰	8	7	8	7	8	6	6	5	41	4	4	3	5	4	4								
🇨🇿	8	6	6	5	6	5	4	3	4	34	5	4	4	4	4								
🇭🇺	10	6	6	5	6	6	5	3	4	5	25	4	5	4	5								
🇵🇱	9	6	6	4	4	4	4	3	3	4	4	24	4	4	4								
🇸🇪	7	7	7	6	6	6	7	4	5	4	5	4	7	5	5								
🇫🇮	7	5	5	5	5	5	5	4	4	4	4	4	5	9	4	4	3	3	3	2	3	3	2
🇸🇮	6	5	5	5	5	5	5	3	4	4	5	4	5	4	7	5	3	4	4	2	3	3	2
🇸🇰	5	5	5	5	5	5	5	3	4	4	5	4	5	4	5	7	3	4	4	2	3	3	2
🇱🇻	5	4	4	3	3	3	3	3	3	3	3	4	3	3	3	3	13	3	2	2	3	3	2
🇪🇪	4	4	4	4	4	4	4	3	4	3	4	3	4	3	4	4	3	4	3	2	3	3	2
🇷🇴	4	4	4	4	4	4	4	2	3	3	4	3	4	3	4	4	2	3	4	2	2	2	1
🇧🇬	4	3	2	2	3	2	3	2	3	3	2	2	2	2	2	2	2	2	2	17	2	2	1
🇱🇹	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3	2
🇲🇹	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	2	2	3	3	2
🇮🇪	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	1	2	2	2

Details for language pair Romanian → Finnish (switch to Finnish → Romanian)

▲ Up to EuroMatrix folder

MT Systems

No MT system has been found.

Parallel Corpora

- OPUS: KDE
- OPUS: PHP
- Acquis Communautaire

Figure 5.3: EuroMatrix inventory of available tools, lingware and data for the EU official languages (including MT systems): the number of tools and data for each language pair with the details for Romanian-Finnish

5.3 Translation models based on Acquis subcorpora

We used the Acquis sub-corpora, parallel in 22 languages to create 462 translation systems for all possible language pairs. The resulting systems and their performances reveal the different challenges for the statistical machine translation.

5.3.1 Motivation

Insufficient language coverage in MT Although automatic translation has been one of the core applications of computational linguistics from its very beginning, it may not come as a surprise that only a very few languages are covered by MT systems.

Figure 5.3, taken from EuroMarix project website[3] (March 2009), gives an overview

[3] http://www.euromatrix.net/

of the existing resources, including MT systems for the EU official language pairs. The Compendium of Translation Software directory of commercial machine translation systems and computer-aided translation support tools compiled by John Hutchins[4] (15th edition, January 2009) shows that most existing translation directions evolve around a small number of languages, with English being the most frequently utilised one and that 10 languages are almost completely interconnected while all others are associated with only a few other languages.

Our experiments produced 462 translation systems for all the combinations of EU language pairs (except Irish), which include combinations of non-standard language pairs like Finnish-Maltese or Bulgarian-Hungarian.

Building baseline models At the same time, we wished to investigate `Moses`' current performance, based on direct translation models. We then looked for ways to improve this performance using different pivot models: models combined at the alignment level, or at the phrase table level.

Baseline models were established for the 231 language pairs in both direction (total of 462 translation models). `Moses`' phrase-based translation models were trained on different sizes of the *Acquis22* parallel corpus (on 10 000 sentences, and on the whole corpus, around 300k sentences) to investigate the effect of scarce data on our models. These models were studied through the evaluation of their output by using BLEU metric score.

5.3.2 Experimental design

5.3.2.1 Data

Training corpus The parallel corpora used for these experiments is the *Acquis22* and *Translation-Units-22* corpus, for which the sizes were presented in sub-sections 4.3.2 and 4.4.1.

The corpus *Acquis22* contains a total number of 186 million tokens. It includes around 8.4 million words, and an average of 360 000 sentences, for each language. We perform experiments on subcorpora of different sizes of *Acquis22*: a randomly generated sample of 10 000 sentences (*Acquis22-sample10k*) and the whole corpus.

The corpus *Translation-Units-22* includes 8.7 million tokens with an average of about 400 000 tokens per language. It contains around 20 000 sentences per language, exactly aligned between all language pairs (see tables A.1 and A.2 in Appendix A with an example of sentence translated in 22 languages).

The corpora has been pre-processed for use with Moses system including "sentence" (paragraph) splitting and tokenisation, as well as lower-casing (to avoid training separate models on uppercase and lowercase words). We extracted only sentences that have a length of less than 100 tokens (as this is a limit imposed by `GIZA++` training).

[4] `http://www.hutchinsweb.me.uk/Compendium.htm`

A number of 462 baselines were built for each subcorpora.

Development corpus (Devset) Development data were described in the subsection 4.4.2. They consist of 2600 sentences in the same domain as the training data, but which were not part of this data. They are separated in a tuning set and a test set.

Tuning corpus The tuning set includes 660 sentences for each language.

Test corpus The test set contains 2000 sentences for each language. For computational reasons, we used only the first 1000 sentences, that includes a total number of 1.1 million tokens, with an average of about 50000 tokens per language.

5.3.2.2 Moses' parameters

Training We used the default training parameters:

`GIZA++` was performed in both directions with the default parameters. Then we applied the *grow-diag-final* heuristics to combine unidirectional alignments outputted by `GIZA++`.

Each phrase table contains extracted phrases of maximum 7 tokens, including the phrase probabilties and the lexical weights in both directions (and the word penalty).

We use a lexicalized reordering model "*msd-bidirectional-fe*". Note that this reordering model is conditioned on the pair of phrases source - target (*fe*) for which three orientation types are considered, mono, swap and discontinouous (*msd*), calculated for the current phrase with respect to the previous and for the next phrase with respect to the current one (*bidirectional*).

Language models We created 5-gram language models for our baselines, learnt from the union of *Acquis22* and *Translation-Units-22* corpora in each target language: it is important that the language model is of the same domain as the translation model and the test set. Discounting and smoothing methods (interpolation and Kneser-Ney smoothing) were used to deal with the problem of unforeseen events.

Quality tuning (MERT) For part of the models, the Minimum Error Rate Training was applied to refine them, using a tuning parallel corpus between 300 and 600 sentences. The MERT tuning is very time and resource consuming, taking about 10 hours for a language pair (en-ro) trained on a sample of 10k sentences of *Acquis22* corpus, when the training set includes 500 sentences.

Therefore, the final results for all the language pair combinations were obtained without quality tuning.

Decoding The beam size can be defined with a threshold or by histogram pruning: we used the default threshold, which cuts off probabilities that are less than 0.00001. We did not set a maximum stack size for holding hypotheses. We used a standard distortion limit (maximum distance between two input phrases to two neighbouring output phrases) of 6, as well as a lexicalised reordering model. The word penalty was introduced to the model for each generated target word, in addition to the language model.

Evaluation We use only the first 1000 sentences of the test data to evaluate our translation models. Translation tables were filtered to adjust to the test data. Finally, each baseline model was tested using `Moses` decoders and the BLEU scores (and the NIST scores) was calculated for each system.

The next section presents the evaluation of our translation models, followed by a discussion of the results.

5.3.3 Evaluation of the translation models

We present in this section the performance of the translation systems trained on *Translation-Units-22* (TU22) corpus. Similar results for the other subcorpora used in our experiments*, Acquis22* and *Acquis22-sample10k,* are displayed in the Appendix B (tables B.1 and B.2).

The BLEU scores for the 462 translation systems trained on *Translation-Units-22* (TU22) corpus are shown in Table 5.1 : the higher the score, the better performance.

According to these numbers, the easiest translations directions are Maltese-English (BLEU score of 0.5952) and Portuguese-French (BLEU score of 0.5807) and the hardest are Maltese-Estonian (BLEU score 0.1617) and Maltese-Finnish (BLEU score 0.1713).

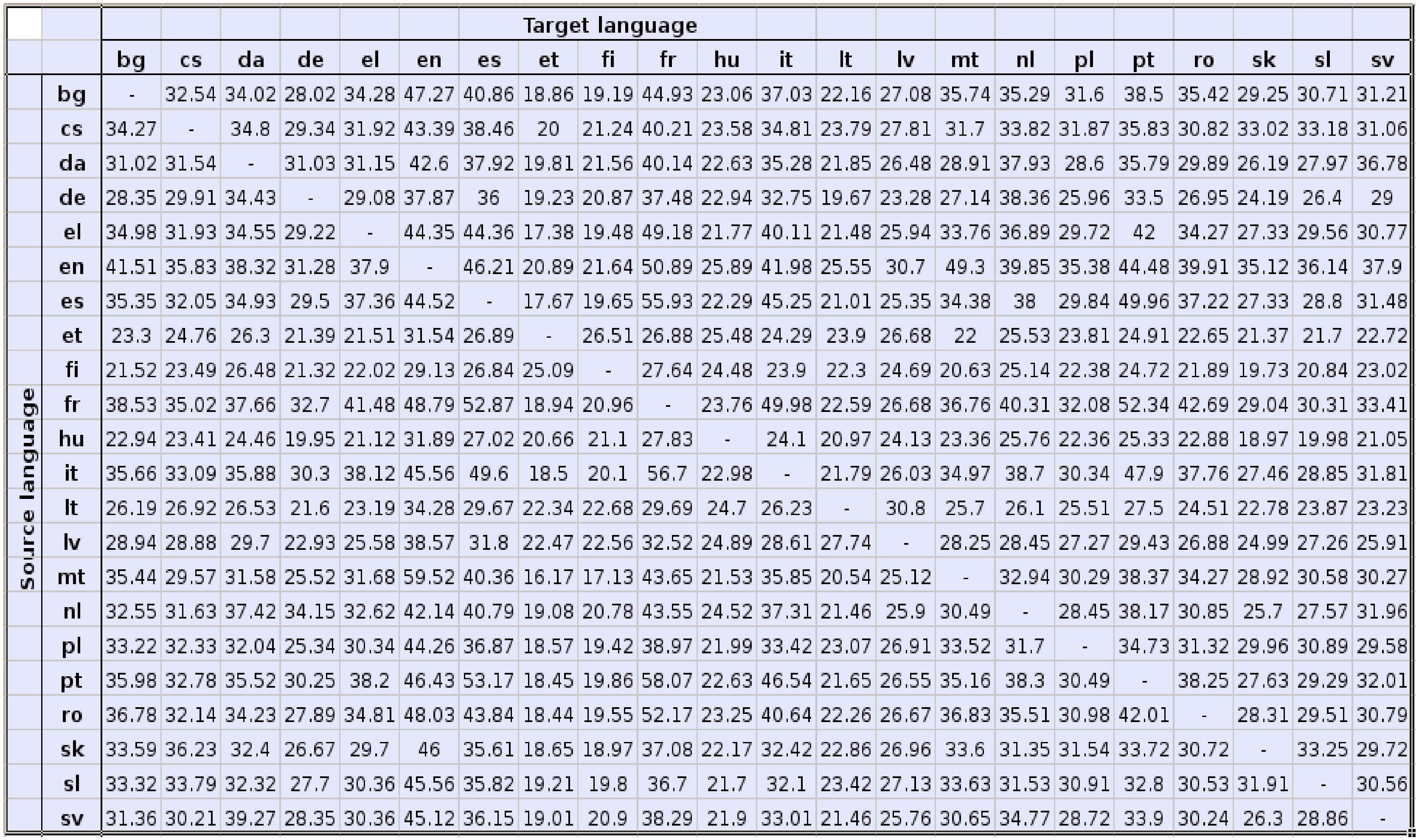
Histograms in Figure 5.4 show the translation scores into and from specific languages (French, English, Romanian, Slovene, Finnish, and German).

The Appendix A presents a sample output of the different translation systems trained on *Translation-Units-22* corpus, translating into French, English and Romanian. Thus, the tables A.3 and A.4 listed one sentence translated into French from all the other 21 languages. In the tables A.5 and A.6 we present the same sentence when translating into English from all the other languages. See tables A.7 and A.8 for the Romanian translations. The reference sentences across all the 22 language are given in the table A.1 and A.2 of the same appendix.

In the next section, we will discuss the results obtained by evaluating our translation systems.

5.3.4 Discussion

The wide range quality for the different SMT systems illustrates the different challenges of statistical translation.



		Target language																					
		bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
Source language	bg	-	32.54	34.02	28.02	34.28	47.27	40.86	18.86	19.19	44.93	23.06	37.03	22.16	27.08	35.74	35.29	31.6	38.5	35.42	29.25	30.71	31.21
	cs	34.27	-	34.8	29.34	31.92	43.39	38.46	20	21.24	40.21	23.58	34.81	23.79	27.81	31.7	33.82	31.87	35.83	30.82	33.02	33.18	31.06
	da	31.02	31.54	-	31.03	31.15	42.6	37.92	19.81	21.56	40.14	22.63	35.28	21.85	26.48	28.91	37.93	28.6	35.79	29.89	26.19	27.97	36.78
	de	28.35	29.91	34.43	-	29.08	37.87	36	19.23	20.87	37.48	22.94	32.75	19.67	23.28	27.14	38.36	25.96	33.5	26.95	24.19	26.4	29
	el	34.98	31.93	34.55	29.22	-	44.35	44.36	17.38	19.48	49.18	21.77	40.11	21.48	25.94	33.76	36.89	29.72	42	34.27	27.33	29.56	30.77
	en	41.51	35.83	38.32	31.28	37.9	-	46.21	20.89	21.64	50.89	25.89	41.98	25.55	30.7	49.3	39.85	35.38	44.48	39.91	35.12	36.14	37.9
	es	35.35	32.05	34.93	29.5	37.36	44.52	-	17.67	19.65	55.93	22.29	45.25	21.01	25.35	34.38	38	29.84	49.96	37.22	27.33	28.8	31.48
	et	23.3	24.76	26.3	21.39	21.51	31.54	26.89	-	26.51	26.88	25.48	24.29	23.9	26.68	22	25.53	23.81	24.91	22.65	21.37	21.7	22.72
	fi	21.52	23.49	26.48	21.32	22.02	29.13	26.84	25.09	-	27.64	24.48	23.9	22.3	24.69	20.63	25.14	22.38	24.72	21.89	19.73	20.84	23.02
	fr	38.53	35.02	37.66	32.7	41.48	48.79	52.87	18.94	20.96	-	23.76	49.98	22.59	26.68	36.76	40.31	32.08	52.34	42.69	29.04	30.31	33.41
	hu	22.94	23.41	24.46	19.95	21.12	31.89	27.02	20.66	21.1	27.83	-	24.1	20.97	24.13	23.36	25.76	22.36	25.33	22.88	18.97	19.98	21.05
	it	35.66	33.09	35.88	30.3	38.12	45.56	49.6	18.5	20.1	56.7	22.98	-	21.79	26.03	34.97	38.7	30.34	47.9	37.76	27.46	28.85	31.81
	lt	26.19	26.92	26.53	21.6	23.19	34.28	29.67	22.34	22.68	29.69	24.7	26.23	-	30.8	25.7	26.1	25.51	27.5	24.51	22.78	23.87	23.23
	lv	28.94	28.88	29.7	22.93	25.58	38.57	31.8	22.47	22.56	32.52	24.89	28.61	27.74	-	28.25	28.45	27.27	29.43	26.88	24.99	27.26	25.91
	mt	35.44	29.57	31.58	25.52	31.68	59.52	40.36	16.17	17.13	43.65	21.53	35.85	20.54	25.12	-	32.94	30.29	38.37	34.27	28.92	30.58	30.27
	nl	32.55	31.63	37.42	34.15	32.62	42.14	40.79	19.08	20.78	43.55	24.52	37.31	21.46	25.9	30.49	-	28.45	38.17	30.85	25.7	27.57	31.96
	pl	33.22	32.33	32.04	25.34	30.34	44.26	36.87	18.57	19.42	38.97	21.99	33.42	23.07	26.91	33.52	31.7	-	34.73	31.32	29.96	30.89	29.58
	pt	35.98	32.78	35.52	30.25	38.2	46.43	53.17	18.45	19.86	58.07	22.63	46.54	21.65	26.55	35.16	38.3	30.49	-	38.25	27.63	29.29	32.01
	ro	36.78	32.14	34.23	27.89	34.81	48.03	43.84	18.44	19.55	52.17	23.25	40.64	22.26	26.67	36.83	35.51	30.98	42.01	-	28.31	29.51	30.79
	sk	33.59	36.23	32.4	26.67	29.7	46	35.61	18.65	18.97	37.08	22.17	32.42	22.86	26.96	33.6	31.35	31.54	33.72	30.72	-	33.25	29.72
	sl	33.32	33.79	32.32	27.7	30.36	45.56	35.82	19.21	19.8	36.7	21.7	32.1	23.42	27.13	33.63	31.53	30.91	32.8	30.53	31.91	-	30.56
	sv	31.36	30.21	39.27	28.35	30.36	45.12	36.15	19.01	20.9	38.29	21.9	33.01	21.46	25.76	30.65	34.77	28.72	33.9	30.24	26.3	28.86	-

Table 5.1: BLEU scores for the 462 translation systems trained on *Translation-Units-22* corpus

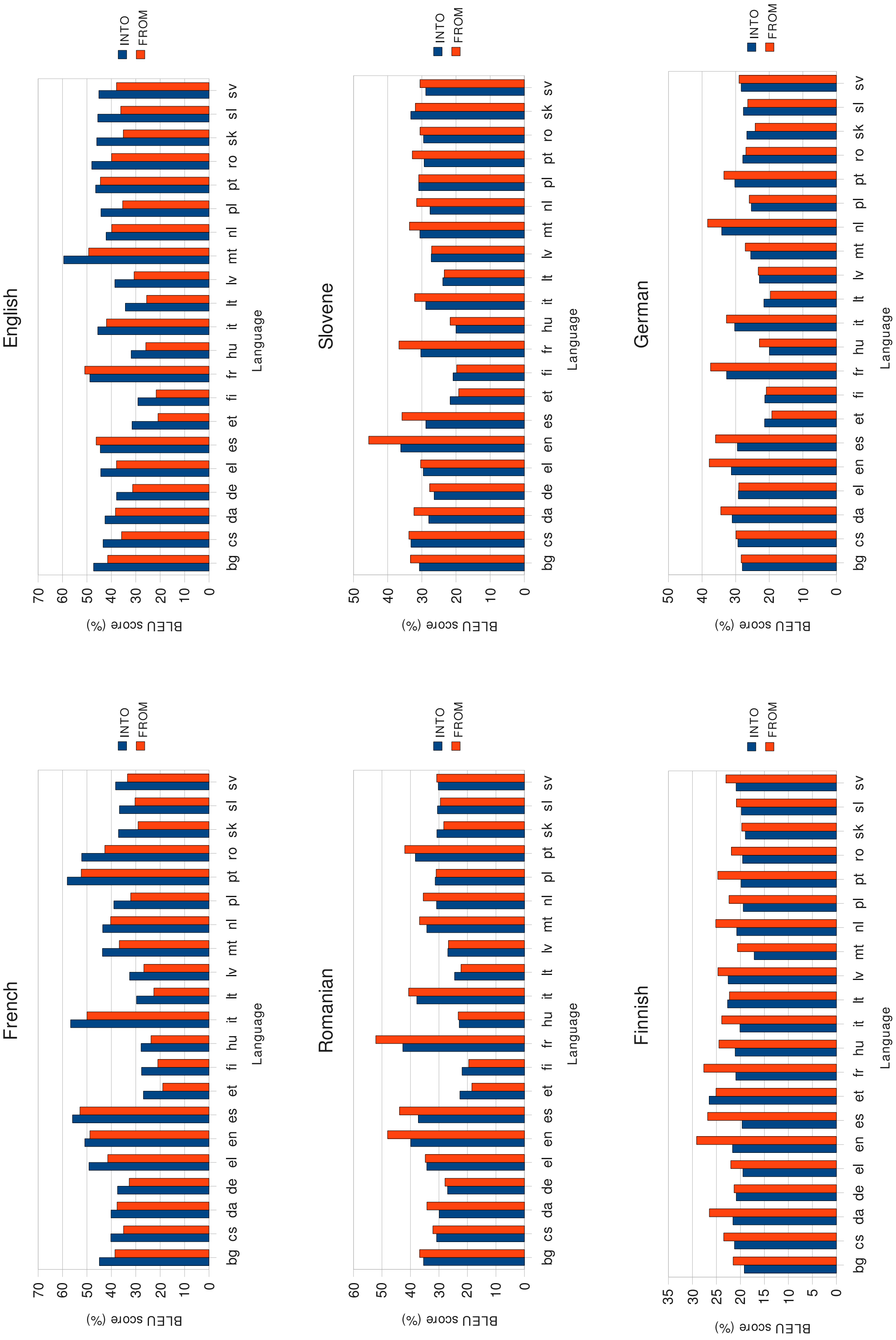


Figure 5.4: Histograms showing the translation scores INTO and FROM the following languages: French, English, Romanian, Slovene, Finnish and German

Language relatedness We note that the performance scores reflect the relatedness of language pairs. Translation from Portuguese to French (58.07) is relatively easy while translating from Romanian to Estonian is relatively hard (BLEU score 18.44).

Intuitively, languages that are related are easier to translate into one other. Calculating the correlation between the vectors of BLEU scores for each language pair (also as source and as target languages), we observe that languages in the same family are strongly correlated, either as target or as source languages. Table B.3 and table B.4 in Appendix B present the correlation values between the BLEU score vectors "INTO" and "FROM" of the twenty-two European languages.

Language	Correlation between BLEU score vectors "INTO"
bg	**pl (96.6%)**, mt (95.0%), ro (94.6%), el (92.6), sl (90.5%), cs (90.2%)
cs	**pl (94.1%)**, sl (93.3%), sk (91.5%), bg (90.2%)
da	**sv (95.7%)**, de (93.7%), nl (91.7%)
de	**nl (96.9%)**, da (93.7%), sv (90.3%)
el	**it (98.2%)**, es (97.8%), ro (97.7%), pt (97.3%), fr (96.6%), nl (91.7%)
en	**mt (97.4%)**
es	**pt (99.8%)**, it (99.5%), fr (99.1%), el (97.8%), ro (95.7%)
et	**fi (90.1%)**
fi	**et (90.1%)**
fr	**it (99.3%)**, pt (99.2%), es (99.1%), el (96.6%), ro (94.9%)
hu	***fi (75,3%)***, *et (71,6%)*
it	**es (99.5%)**, pt (99.5%), fr (99.3%), el (98.2%), ro (95.9%)
lt	***lv (88,6%)***
lv	***lt (88,6%)***
mt	**en (97.4%)**, bg (95.0%), pl (93.5%)
nl	**de (96.9%)**, da (91.7%), el (91.7%), *sv (89%)*
pl	**sl (97.1%)**, bg (96.6%), sk (96.2%), cs (94.1%), mt (93.5%)
pt	**es (99.8%)**, it (99.5%), fr (99.2%), el (97.3%), ro (95.8%)
ro	**el (97.7%)**, it (95.9%), pt (95.8%), es (95.7%), fr (94.9%), bg (94.6%)
sk	**sl (98.8%)**, pl (96.2%), cs (91.5%)
sl	**sk (98.8%)**, pl (97.1%), cs (93.3%), bg (90.2%)
sv	**de (95.7%)**, da (90.3%), *nl (89%)*

Table 5.2: Best correlations given by the BLEU score vectors "INTO" by language.

The correlation as target language, given by the BLEU scores vectors "INTO" are better indicators of the language behaviour in a translation system than the vectors "FROM". From this table (Table B.3) we have extracted the correlation values greater than 90% and thus, we present in the Table 5.2, for each language, the strongly correlated languages via the BLEU score vectors "INTO". A more suggestive graphical representation is given by the figure 5.5, where we found that languages in the same family are correlated to one another. An interesting finding is the strong correlation between Romanian and Bulgarian. Remark also the strong correlation between Maltese and English (for which language pair we obtained the highest BLEU score). The Greek language seems to make a link between Romance, Germanic and Slavic languages. Hungarian has no strong correlation with any of the European languages, but the highest scores are with Finnish (75.6%) and Estonian (71.6%). We remark also that Lithuanian and Latvian languages are correlated at 88%, followed by Hungarian with a quite low correlation (58,8% Hungarian - Lithuanian and 46,8% Hungarian - Latvian).

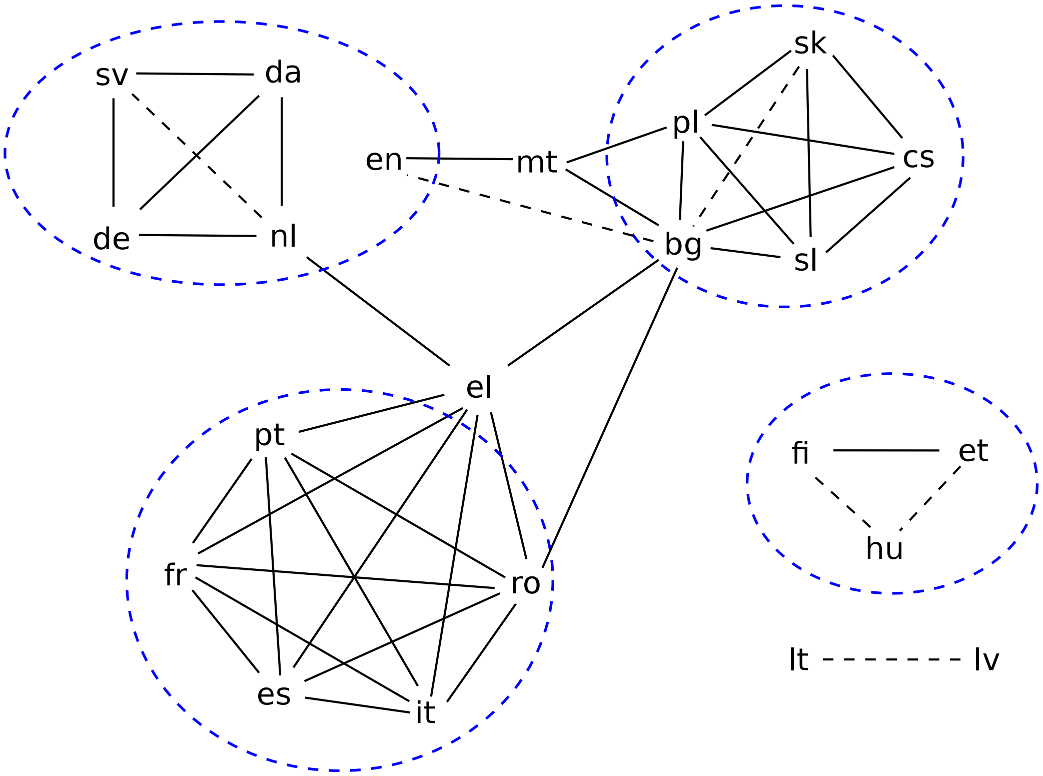


Figure 5.5: Correlations between languages (more than 90%) given by the BLEU score vectors "INTO"

Note that the language relatedness is not the only explanation for translation difficulty (or easiness).

Translation direction Some languages are easier to translate into or easier to translate from. Table 5.3 presents the average scores obtained translating from one language into all the others and into one language from all the others. We calculate the difference (DIFF) between "INTO" and "FROM" scores that gives an idea of the difference of difficulty when we change the translation direction. The last value in the table represents the average between the "FROM" and "INTO" scores, that represents a global indicator of the language performance regarding our translation models. Nevertheless, the scores are dependent on the language set on which they are calculated (because each "FROM" and "INTO" score is relative to the other languages of the set).

LG	FROM	INTO	DIFF	AVER
bg	32.24	32.13	0.11	32.19
cs	31.66	30.86	0.80	31.26
da	30.72	32.99	-2.27	31.86
de	28.73	27.35	1.38	28.04
el	32.33	31.08	1.25	31.70
en	36.51	42.71	-6.20	39.61
es	33.23	38.62	-5.39	35.92
et	24.48	19.50	4.98	21.99
fi	23.68	20.66	3.02	22.17
fr	35.57	41.36	-5.79	38.47
hu	23.30	23.24	0.06	23.27
it	33.91	35.03	-1.12	34.47
lt	25.91	22.45	3.46	24.18
lv	27.79	26.51	1.28	27.15
mt	31.40	31.74	-0.34	31.57
nl	31.29	33.63	-2.34	32.46
pl	30.40	28.96	1.44	29.68
pt	34.15	36.47	-2.32	35.31
ro	33.08	31.42	1.66	32.25
sk	30.63	26.93	3.70	28.78
sl	30.51	28.36	2.15	29.44
sv	30.22	29.73	0.49	29.98

Table 5.3: Average translation scores for systems when translating FROM and INTO a language

Intuitively, translating from an information-rich to an information-poor language is easier than the other way around. Note that translating into and from English is among the easiest. French and other Romance languages also have quite high scores.

English has the best global score (average “FROM” - “INTO”).

Linguistic factors - morphology Some languages are “better” modelled by the statistical translation model than others. The translation model does not take into account different language specific phenomenon. Therefore, the translation systems perform with more difficulty on a language with richer morphology. This is reflected in the results, as we are using no morphological processing. We observe that the SMT models tend to perform much better when translating to morphologically simpler languages.

The poor performance of systems involving Finnish and Estonian can be attributed to its agglutinative morphology. This increases the size of the vocabulary and leads to the problem of sparse data when collecting statistics for word and phrase translation.

We found a high negative correlation between the number of different tokens of the training data and the overall performance of a translation system (correlation value: -0.95).

We suggest that fine-tuning of parameters and dedication processing for each language could improve results.

Noisy training data or scarce training data Not all training data can be expected to be of high quality. The question is whether a machine translation degrades when trained on noisy data. Wang [2002] addressed this question by artificially adding noise to a clean training corpora: a certain percentage of sentence alignments were distorted to simulate misaligned training data. His results suggest that the quality of the translation system only starts to significantly degrade, if half of the training data is distorted this way: in his experiments distortion of up to 25% of training data reduces performance, as measured by the BLEU score, only by about 10%.

The performance of systems trained on *Acquis22* is perfectly coherent with the scores obtained on the reduced, clean corpus "*Translation-Units-22*". However, it is not the case of *Acquis22-sample10k*, which has been randomly generated from *Acquis22*. The results for some languages (Greek, Bulgarian) are less good compared to those obtained on the other two corpora. This suggests that either the "noisy" alignment percentage is quite high with respect to the corpus size, or that is not enough training data.

5.3.5 Conclusions

We used the Acquis sub-corpora, parallel in 22 languages to create 462 translation systems for all possible language pairs. The resulting systems and their performances reveal the different challenges for the statistical machine translation.

We analyse the correlation between the BLEU score vectors "INTO" that reveals how easy or difficult the translation between certain language pairs will be.

We note the importance of the language relatedness in a translation system: the language which are related are easier to translate into one another. On the other hand, the SMT models tend to perform much better when translating to morphologically simpler languages. We found a high correlation between the number of different tokens of the training data (vocabulary size) and the overall performance of a translation system (when translating into English).

Since the research community is primarily occupied with translation into English, interesting problems associated with translation into morphologically rich languages have been neglected. We suggest that fine-tuning of parameters and dedication processing for each language could improve results.

Chapter 6

Alignment and translation models using a pivot language

6.1 Introduction

In this chapter, we will expose first the reasons for our approach, then we will briefly present all the methods that we explored in this thesis.

Each pivot system proposed will be detailed in the following sections, we present the formal model and how each element of the model in a generative framework was built. There are three "main" models and each of them present some variants that will be described in this chapter and then evaluated in our experiments.

The system combination (pivot(s) and direct) will be explained in the section 6.5.

The section 6.7 will briefly expose the factors that we considered to affect the performance of a pivot system. The last section will conclude the chapter.

6.2 Motivation

We argue that the redundancy introduced by a large suite of languages can correct errors in the word alignments and also provide greater generalisation, since the translation distribution is estimated from a richer set of data-points. In general we expect that a wider range of possible translations are found for any source phrase, simply due to the extra layer of indirection.

Thus, the motivation for the pivot approach is two-fold. First, we believe that parallel corpora available in several languages provide better training material for alignment systems relative to bilingual corpora. Word alignment systems trained on different language pairs make errors which are somewhat orthogonal. In such cases, incorrect alignment links between a sentence-pair can be corrected when a translation in a third language is available. Thus it can help to resolve errors in word alignment. We combine word alignments and translation models using several bridge languages with the aim to correct some of these errors.

The second advantage to this approach concerns the problem of data coverage. Current phrase-based statistical machine translation (SMT) systems perform poorly when using small training sets. When there are only small bilingual corpora between low-density language-pairs (like Romanian and Finnish), the triangulation allows the use of a much wider range of parallel corpora for training. Therefore, pivot alignment could be expected to make a positive and safe contribution in a word alignment system, i.e. increasing recall without lowering precision.

[Kay, 2000] suggests that much of the ambiguity of a text that makes it hard to translate into another language may be resolved if a translation into some third language is available and proposes using multiple source documents as a way to inform subsequent machine translations. He calls the use of existing translations to resolve underspecification in a source text "triangulation in translation", but does not offer a method to go about performing this triangulation. The challenge is to find general techniques that will exploit the information in multiple source to improve the quality of alignment and machine translation.

6.3 Building pivot translation systems

We will explore here different heuristics for combining translation models using a pivot language.

We can perform triangulation at different levels of the translation process: in training (at alignment level or at the phrase-table level) and in decoding. We considered three procedures with their variants, one at each of these levels.

As using `Moses`, our lexical scores are estimated on a training corpus which is automatically aligned using `GIZA++` in both directions between source and target and symmetrised using the growing heuristic. Our first heuristic proposes a procedure where this symmetrised alignment table between a language pair is combined with the alignment tables between the source and the pivot language and between the pivot and the target language. Thus, we evaluate the enhancement produced by an intermediate language to alignment.

The second heuristic combines phrase tables. For a triad of languages we create the phrase tables between the source and the pivot language and between the pivot and the target language. For each phrase entry we identify their translations into the intermediate language and then into the target language and we generate the triangulated phrase table.

Each model presents variations that will be described in the section 6.4.

The two methods require different training conditions. While the "phrase-table" pivot method can be performed on training data with different overlap at the pivot level, the "alignment" pivot method requires exact aligned sentences for all the languages in a triad, which is a resource quite difficult to find.

If triangulation is intuitively appealing, it may suffer from a few problems. Firstly, as with any SMT approach, the translation estimates are based on noisy automatic word alignments. This leads to many errors and omission in the phrase-table.

- With a standard source-target phrase-table these errors are only encountered once. However, with triangulation they are encountered twice, and therefore the errors will compound. This leads to larger number of noisy estimates than in source-target phrase-table.

- Secondly, the increased exposure to noise means that triangulation will omit a greater proportion of large or rare phrases than the standard method. An alignment error in either source-pivot or pivot-target bitexts can prevent the extraction of source-target phrase pairs.

These problems can be reduced by using the triangulated phrase-table in conjunction with a standard phrase-table. We merge the phrase-tables by linear interpolation. This interpolated model will be described in section 6.5.

The previously presented methods process the pivot information at the training time, to build a translation model from source to target, that is used like a "direct" translation model by the decoder. We call this "triangulation at training time".

We also compare these pivot methods with a third one, where the pivot information is used directly by the decoder (at the decoding time). In this case, two translation systems are built independently: between source and pivot and between pivot and target. The decoder has to utilise both systems at the decoding time. The input sentence, in the source language, is translated firstly in the pivot language and then in the target language. This is called "triangulation at decoding time", and it will be described in the section 6.6, with its variations. In our experiments we evaluated the performance of both "pivot-at-training" and "pivot-in-decoding" methods. The comparison will be presented in the next chapter.

6.4 Triangulation at training time: pivot translation models

We present two pivot models that integrate the pivot information during the training process. For the first, triangulation is performed at the alignment level, generating a pivot alignment model. For the second, triangulation is done at the phrase-table level.

For each of them we will present firstly the formal model, then the procedure proposed to build the translation model.

6.4.1 Triangulation at alignment level

6.4.1.1 Formal model

We formalise our model in the word alignment framework.

Let us recall that a statistical translation model describes the relationship between a pair of sentences in the source and target languages ($s = s_1^I$, $t = t_1^J$) using a translation probability $p(s \mid t)$. Alignment models introduce a hidden alignment variable $a = a_1^I$ to specify a mapping between source and target words; $a_j = i$ indicates that the j-th source word is linked to the i-th target word.

Alignment models assign a probability $p(s, a \mid t)$ to the source sentence and alignment conditioned on the target sentence. Translation probability is related to the alignment model as: $p(s \mid t) = \sum\limits_a p_\theta(s, a \mid t)$, where θ is a set of parameters. Given a sentence-pair (s, t), the most likely (Viterbi) word alignment is found as:

$$\hat{a} = arg\,\underset{a}{max}\, p(s, a \mid t)$$

We assume that we have triples of sentences that are translations of one another in languages S (source), T (target) and the pivot language Piv: $s = s_1^I$, $t = t_1^J$ $piv = piv_1^K$. Our goal is to obtain the most likely word alignment for the sentence-pair in ST: (s, t), using the alignment estimates for the sentence pairs in SPiv: (s, piv) and PivT: (piv, t). The word alignments between the above sentence-pairs are referred to as a^{ST} , a^{SPiv} , and a^{PivT} respectively; the notation a^{ST} indicates that the alignment maps a position in S to a position in T.

We start by modelling the pivot sentence, piv, and the alignment between pivot and target sentences, a^{PivT}, as hidden variables:

$$\hat{a}^{ST} = arg\,\underset{a^{ST}}{max}\, p(s, a^{ST} \mid t) = arg\,\underset{a^{ST}}{max} \sum_{piv} p(s, a^{ST}, piv \mid t)$$

$$= arg\,\underset{a^{ST}}{max} \sum_{piv} \sum_{a^{PivT}} p(s, a^{ST}, piv, a^{PivT} \mid t)$$

Firstly, we marginalised the pivot variable piv, and then the alignment a^{PivT} :

$$\hat{a}^{ST} = arg\,\underset{a^{ST}}{max} \sum_{piv, a^{PivT}} p(s, a^{ST}, a^{PivT} \mid piv, t)\, p(piv \mid t)$$

$$= arg\,\underset{a^{ST}}{max} \sum_{piv, a^{PivT}} p(s, a^{ST} \mid a^{PivT}, piv, t)\, p(a^{PivT} \mid piv, t)\, p(piv \mid t) \qquad (6.1)$$

We now make some assumptions to simplify the above formula. First, there is exactly one translation piv in pivot language corresponding to the sentence pair (s, t).

Next, we consider the alignment source-pivot, a^{SPiv}, that will produce the alignment source-target, a^{ST} when composed with the alignment pivot-target, a^{PivT}: $a^{SPiv}a^{PivT} = a^{ST}$. Formally, a^{SPiv} is defined as:

$$\left\{ a_i^{SPiv} | a^{PivT}_{a_i^{FPiv}} = a_i^{ST}, \forall i \in 1, \dots, I \right\} \tag{6.2}$$

The first distribution in 6.1 can be expressed using this alignment a^{SPiv}, as follows:

$$p(s, a^{ST} \mid a^{PivT}, piv, t) = p(s, a^{SPiv} \mid piv, t) = p(s, a^{SPiv} \mid piv)$$

knowing that alignments in a^{SPiv} do not depend on t.

Finally, we can express: $p(a^{PivT} \mid piv, t)\, p(piv \mid t) = p(a^{PivT}, piv \mid t)$.

Under these assumptions, we arrive at the final expression:

$$\hat{a}^{ST} = arg \max_{a^{SPiv}a^{PivT}=a^{ST}} \sum_{a^{PivT}} p(s, a^{SPiv} \mid piv)\, p(piv, a^{PivT} \mid t)$$

$$\hat{a}^{ST} = \widehat{a^{SPiv}a^{PivT}} \approx arg \max_{a^{SPiv}} \max_{a^{PivT}} p(s, a^{SPiv} \mid piv)\, p(piv, a^{PivT} \mid t) \tag{6.3}$$

Notice that in the last step we apply the maximum approximation, to reduce the complexity of the search procedure.

The above expression states that in the pivot alignment model, the best alignment should maximise the product probability between source-pivot and pivot-target. The maximisation should be applied at each iteration step when estimating the best alignment.

For simplification, we will use only the combination of the best alignments for each model S-Piv and Piv-T. Thus the best alignment is calculated separately for each of them. In this case, formula 6.3 becomes:

$$\widehat{a}^{ST} = \widehat{a^{SPiv}a^{PivT}} \approx \widehat{a}^{SPiv}\widehat{a}^{PivT} = arg \max_{a^{SPiv}} p(s, a^{SPiv} \mid piv)\, arg \max_{a^{PivT}} p(piv, a^{PivT} \mid t) \tag{6.4}$$

We will describe in the next subsection how we build the pivot alignment model and the pivot translation model based on equation 6.4.

6.4.1.2 Building the pivot translation model

The alignment between source and target is built from the source-pivot and pivot-target alignments, as indicated by formula 6.2:

$$a_i^{ST} = a^{PivT}_{a_i^{FPiv}}, \forall i \in 1, \dots, I \tag{6.5}$$

or

$$a^{ST} = \left\{ (s, t) \mid \exists piv \,:\, (s, piv) \in a^{SPiv} \,\wedge\, (piv, t) \in a^{PivT} \right\} \tag{6.6}$$

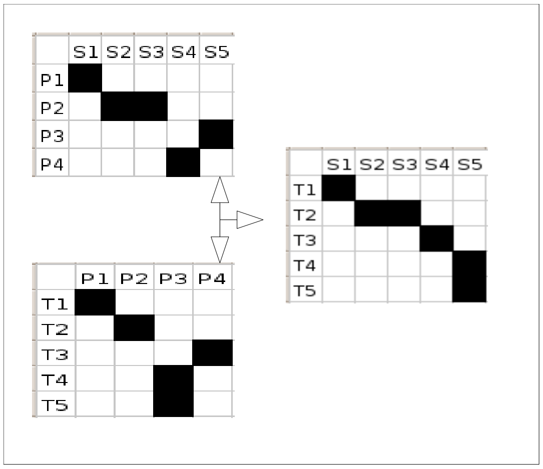


Figure 6.1: Building source-target alignment using source-pivot and pivot-target alignments

An example is shown in figure 6.1

The translation model is induced from the pivot word alignment model built as in 6.6. Thus, the translation model and the reordering model are generated in the same way as the direct model, based on the word alignment.

There are two variants for this method. The first one is combining uni-directional alignments outputted by `GIZA++`, while the second is combining the alignments produced after the symmetrisation heuristic.

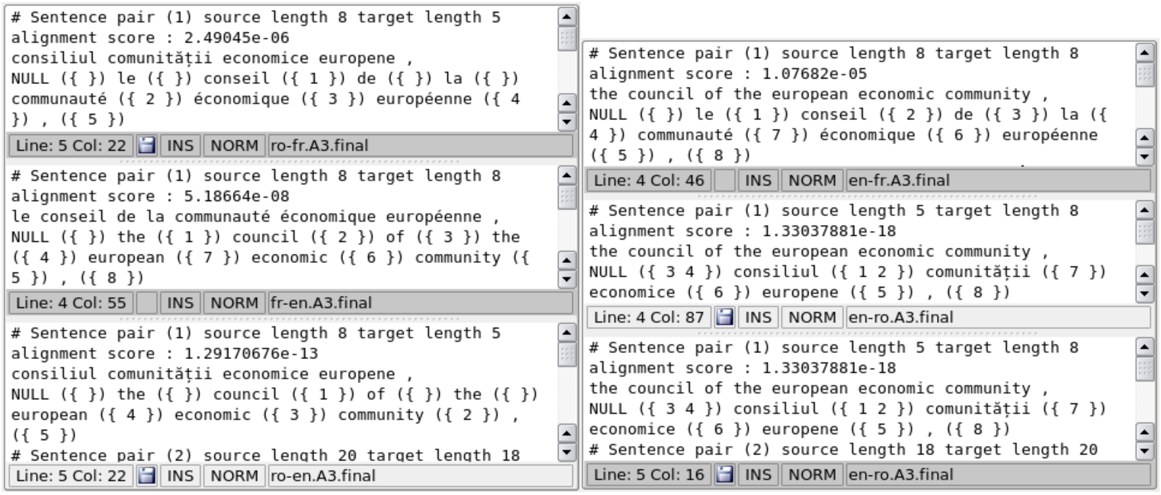
Combining uni-directional alignments In this case, we are processing directly the output of GIZA++ for each direction. Thus, we are combining the uni-directional alignments to obtain source-target alignment via pivot and target-source alignment via pivot.

The alignments source-pivot and pivot-target present the following particularity: they contain only one-to-many word alignments in a direction. This means that, when considering a given direction, a word in the initial language could be translated into zero (NULL), one or more words in the second language, but not the contrary. The growing heuristics of `Moses` tool will combine both directions to symmetrise the alignment.

See Figure 6.2 for an alignment combination example. We have generated English-Romanian alignment using French as a pivot language. The example shows both directions.

Combining symmetrised alignments In this case, the triangulation takes place after the alignment symetrisation. We combined the resulting alignments as described by equation 6.6.

Although the two variants are very close to each other the pivot alignment produced is not the same. Intuitively, we suppose that the first will have a lower recall, which in



```
# Sentence pair (1) source length 8 target length 5
alignment score : 2.49045e-06
consiliul comunității economice europene ,
NULL ({ }) le ({ }) conseil ({ 1 }) de ({ }) la ({ })
communauté ({ 2 }) économique ({ 3 }) européenne ({ 4
}) , ({ 5 })
Line: 5 Col: 22  INS  NORM  ro-fr.A3.final
```

```
# Sentence pair (1) source length 8 target length 8
alignment score : 5.18664e-08
le conseil de la communauté économique européenne ,
NULL ({ }) the ({ 1 }) council ({ 2 }) of ({ 3 }) the
({ 4 }) european ({ 7 }) economic ({ 6 }) community ({
5 }) , ({ 8 })
Line: 4 Col: 55  INS  NORM  fr-en.A3.final
```

```
# Sentence pair (1) source length 8 target length 5
alignment score : 1.29170676e-13
consiliul comunității economice europene ,
NULL ({ }) the ({ }) council ({ 1 }) of ({ }) the ({ })
european ({ 4 }) economic ({ 3 }) community ({ 2 }) ,
({ 5 })
# Sentence pair (2) source length 20 target length 18
Line: 5 Col: 22  INS  NORM  ro-en.A3.final
```

```
# Sentence pair (1) source length 8 target length 8
alignment score : 1.07682e-05
the council of the european economic community ,
NULL ({ }) le ({ 1 }) conseil ({ 2 }) de ({ 3 }) la ({
4 }) communauté ({ 7 }) économique ({ 6 }) européenne
({ 5 }) , ({ 8 })
Line: 4 Col: 46  INS  NORM  en-fr.A3.final
```

```
# Sentence pair (1) source length 5 target length 8
alignment score : 1.33037881e-18
the council of the european economic community ,
NULL ({ 3 4 }) consiliul ({ 1 2 }) comunității ({ 7 })
economice ({ 6 }) europene ({ 5 }) , ({ 8 })
Line: 4 Col: 87  INS  NORM  en-ro.A3.final
```

```
# Sentence pair (1) source length 5 target length 8
alignment score : 1.33037881e-18
the council of the european economic community ,
NULL ({ 3 4 }) consiliul ({ 1 2 }) comunității ({ 7 })
economice ({ 6 }) europene ({ 5 }) , ({ 8 })
# Sentence pair (2) source length 18 target length 20
Line: 5 Col: 16  INS  NORM  en-ro.A3.final
```

Figure 6.2: Combining unidirectional alignments English-Romanian via French and Romanian-English via French

the STM evalution (using BLEU score) seems to be more important than the precision. In the experiments we compare the two variants of our pivot method.

Concerning the difference with the direct method, we present an example (extracted from *Translation-Units-22* corpus training) where the pivot alignment built with our methods makes an improvement compared to the direct method. Figure 6.3 shows the result of our pivot methods (they have the same result in this case) and the alignment obtained by direct training: using the direct method we obtain the wrong link "*community - economice*" which is replaced by the correct link "*community - comunităţii*" in the pivot alignment.

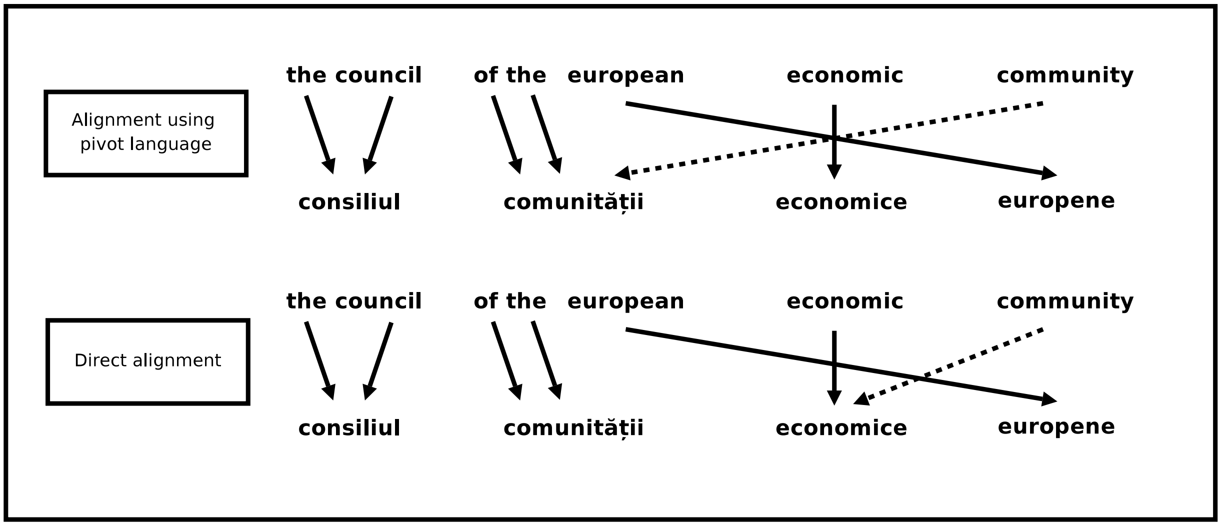


Figure 6.3: Example of two English-Romanian alignments: one obtained by triangulation using French as pivot language, and the other obtained by direct training

6.4.2 Triangulation at phrase-table level

This section introduces the method that performs the triangulation at the phrase-table level, for the language pair S (source) - T (target), using two bilingual corpora of S - Piv (Pivot) and of Piv - T. With these two additional bilingual corpora, we train two translation models for S-Piv and Piv-T, respectively. Based on these two models, we build a pivot translation model for S-T, with Piv as a pivot language. Firstly, we will introduce the formal model, and then we will explain how each element of the translation model is built.

6.4.2.1 Formal model

According to the translation model presented in 5.1.1, given a source sentence s, the best target translation t_{best} can be obtained according to:

$$t_{best} = arg\,\underset{t}{max}\,p(t \mid s) = arg\,\underset{t}{max}\,p(s \mid t)p_{LM}(t)\omega^{length(t)} \tag{6.7}$$

The translation model $p(s \mid t)$ can be decomposed into:

$$p(\overline{s}_1^I \mid \overline{t}_1^I) = \prod_{i=1}^{I} \phi(\overline{s}_i \mid \overline{t}_i)\, p_{reord}(\overline{s}_i \mid \overline{t}_i)\, p_w(\overline{s}_i \mid \overline{t}_i, a)^{\lambda} \tag{6.8}$$

where $\phi(\overline{s}_i \mid \overline{t}_i)$ and $p_{reord}(\overline{s}_i \mid \overline{t}_i)$ denote phrase translation probability and reordering probability (as defined by the lexicalized reordering model), $p_w(\overline{s}_i \mid \overline{t}_i, a)^{\lambda}$ is the lexical weight, and λ is the strength of the lexical weight.

The triangulation is formalised as a generative probabilistic process operating independently on phrase pairs. We start with the conditional distribution over three languages, $p(s,\, piv \mid t)$, where the arguments denote phrases in the source, pivot and target language, respectively. From this distribution, we can find the desired conditional probability over the source-target pair by marginalising out the pivot phrases, as follows:

$$p(s \mid t) = \sum_{piv} p(s,\, piv \mid t) = \sum_{piv} p(s \mid piv, t)\, p(piv \mid t)) \approx \sum_{piv} p(s \mid piv)\, p(piv \mid t) \tag{6.9}$$

where the third formula imposes a simplifying conditional independence assumption: the pivot phrase fully represents the information (semantics, syntax, etc...) in the source phrase, rendering the target phrase redundant in $p(s \mid piv, t)$ ($\approx p(s \mid piv)$).

Equation 6.9 requires that all phrases in the pivot-target bitext be also found in the source-pivot bitext, such that $p(s \mid piv)$ is defined. This supposes that, at decoding time, the translated sentence should share the same segmentation at the pivot level, from the modelling point of view.

A potential problem that may appear is that the independence assumption could be an over simplification and lead to a loss of information.

6.4.2.2 Building the pivot translation model

The translation model includes the translation table (phrase-table) and the lexicalized reordering table. We will explain how we build them in the following paragraphs.

The translation table The phrase table is composed of all the phrase pairs with the alignments information and the translation scores, in the following format:

$$\overline{s} \;|||\; \overline{t} \;|||\; a^{\overline{st}} \;|||\; a^{\overline{ts}} \;|||\; \phi\left(\overline{s} \mid \overline{t}\right) \quad p_w\left(\overline{s} \mid \overline{t}, a^{\overline{st}}\right) \quad \phi\left(\overline{t} \mid \overline{s}\right) \quad p_w\left(\overline{t} \mid \overline{s}, a^{\overline{st}}\right) \quad exp(1)$$

See Table 6.1 for exemplication.

Each phrase pair $\left(\overline{s}, \overline{t}\right)$ (first and second field) is followed by the alignment information in both directions. In the third field each word of the source phrase is associated with the words of the target phrase, or with nothing. Vice versa, in the fourth field. As two word alignments come from one word alignment, the two fields represent the same information. However, they are independent in principle.

The translation scores are the phrase table probabilities ($\phi\left(\overline{s} \mid \overline{t}\right)$ and $\phi\left(\overline{t} \mid \overline{s}\right)$), the lexical weights ($p_w\left(\overline{s} \mid \overline{t}, a^{\overline{st}}\right)$ and $p_w\left(\overline{t} \mid \overline{s}, a^{\overline{st}}\right)$) and the phrase penalty (always $exp(1) = 2.718$).

Phrase pairs selection We select all the phrase pairs $\left(\overline{s}, \overline{t}\right)$ for which

$$\left\{\left(\overline{s}, \overline{t}\right) \mid \exists \overline{piv} : \exists \left(\overline{s}, \overline{piv}\right) \wedge \exists \left(\overline{piv}, \overline{t}\right)\right\}$$

In other words all the source target pairs that have a common pivot phrase in the tables source-pivot and pivot-target, respectively.

Phrase translation probabilities Using the S-Piv and Piv-T bilingual corpora, we train two phrase translation probabilities $\phi\left(\overline{s} \mid \overline{piv}\right)$ and $\phi\left(\overline{piv} \mid \overline{t}\right)$, where $\overline{piv}$ is the phrase in the pivot language Piv. Given the phrase translation probabilities $\phi\left(\overline{s} \mid \overline{piv}\right)$ and $\phi\left(\overline{piv} \mid \overline{t}\right)$, we obtain the phrase translation probability $\phi\left(\overline{s} \mid \overline{t}\right)$ according to the model:

$$\phi\left(\overline{s} \mid \overline{t}\right) = \sum_{piv} \phi\left(\overline{s} \mid \overline{piv}\right) \phi\left(\overline{piv} \mid \overline{t}\right)$$

Table 6.1 shows an example extracted from the model trained on *"Translation-Units-22"* corpus. We show three phrase pairs from the English-Romanian pivot model and the corresponding phrase pairs from English-French and French-Romanian models that were used for their generation.

Alignments The alignment information of the phrase pair $\left(\overline{s}, \overline{t}\right)$ can be induced from the two phrase pairs $\left(\overline{s}, \overline{piv}\right)$ and $\left(\overline{piv}, \overline{t}\right)$ (see Figure 6.1).

Let a^{SPiv} and a^{PivT} represent the word alignment information inside the pairs $\left(\overline{s}, \overline{piv}\right)$ and $\left(\overline{piv}, \overline{t}\right)$ respectively, then the alignment information a^{ST} inside $\left(\overline{s}, \overline{t}\right)$ can be obtained by composing the two alignments a^{SPiv} and a^{PivT}, as follows:

<table>
<tr><th>Phrase-table EN-FR</th></tr>
<tr><td>the council has adopted common rules ||| le conseil a adopté des règles communes ||| (0) (1) (2) (3) (6) (5) ||| (0) (1) (2) (3) () (5) (4) ||| 0.5 0.0967378 0.5 1.74492e-05 2.718</td></tr>
<tr><td>the council has adopted common rules ||| le conseil a adopté un régime commun ||| (0) (1) (2) (3) (6) (5) ||| (0) (1) (2) (3) () (5) (4) ||| 1 0.00382326 0.5 7.14953e-08 2.718</td></tr>
<tr><td>the council adopted common rules ||| le conseil a adopté des règles communes ||| (0) (1) (3) (6) (5) ||| (0) (1) () (2) () (4) (3) ||| 0.5 0.298942 1 1.09667e-07 2.718</td></tr>
</table>

<table>
<tr><th>Phrase-table FR-RO</th></tr>
<tr><td>le conseil a adopté des règles communes ||| consiliul a adoptat norme comune ||| (0) (0) (1) (2) (3) (3) (4) ||| (0,1) (2) (3) (4,5) (6) ||| 1 0.00010888 1 0.00519777 2.718</td></tr>
<tr><td>le conseil a adopté des règles ||| consiliul a adoptat norme ||| (0) (0) (1) (2) (3) (3) ||| (0,1) (2) (3) (4,5) ||| 0.5 0.000333444 1 0.00606406 2.718</td></tr>
</table>

<table>
<tr><th>Pivot phrase-table EN-RO (pivot FR)</th></tr>
<tr><td>the council has adopted common rules ||| consiliul a adoptat norme comune ||| (0) (0) (1) (2) (4) (3) ||| (0,1) (2) (3) (5) (4) ||| 0.5 0.0003801452 0.5 0.0001271746 2.718</td></tr>
<tr><td>the council has adopted common rules ||| consiliul a adoptat un regim comun ||| (0) (0) (1) (2) (5) (4) ||| (0,1) (2) (3) () (5) (4) ||| 1 1.8546e-05 0.5 2.750028e-09 2.718</td></tr>
<tr><td>the council adopted common rules ||| consiliul a adoptat norme comune ||| (0) (0) (2) (4) (3) ||| (0,1) () (2) (4) (3) ||| 0.5 0.01557414 1 1.487071e-05 2.718</td></tr>
</table>

Table 6.1: Building pivot phrase table between English and Romanian using French as pivot language - example extracted from the translation model trained on "*Translation-Units-22*" corpus

$$a^{ST} = \left\{(s,t) \mid \exists piv\ :\ (s,piv) \in a^{SPiv} \wedge (piv,t) \in a^{PivT}\right\} \tag{6.10}$$

Calculating lexical weights Given a phrase pair $\left(\overline{s},\overline{t}\right)$ and a word alignment a between the source word positions $i = 1,\ldots,n$ and the target word positions $j = 1,\ldots,m$, the lexical weight can be estimated according to the following method (presented in section 5.1.1):

$$p_w\left(\overline{s}|\overline{t},a\right) = \prod_{i=1}^{n} \frac{1}{\left|\{j|(i,j)\in a\}\right|} \sum_{\forall(i,j)\in a} w\left(s_i|t_j\right)$$

where the lexical translation probability can be estimated as follows:

$$w(s|t) = \frac{count(s,t)}{\sum_{s'} count(s',t)}$$

Thus, in order to estimate the lexical weight, we need firstly the alignment information a between the two phrases $\overline{s}$ and $\overline{t}$, and then to estimate the lexical translation probability according to the alignment information. The alignments between source and target are generated as above.

Concerning the calculation of the lexical translation probability we propose two methods. The first will estimate the lexical translation probability using the corresponding scores from source-pivot and pivot-target models.

Thus, we can estimate the lexical translation probability with:

$w\left(s \mid t\right) = \sum\limits_{piv} w\left(s \mid piv\right) w\left(piv \mid t\right)$

where $w\left(s \mid piv\right)$ and $w\left(piv \mid t\right)$are two lexical probabilities for the models source-pivot and pivot-target.

The second method we used (proposed by [Wu and Wang, 2007]) will calculate the probability directly from the induced phrase pairs. We estimate the co-occurring frequency of the word pair (s, t) according to the following model.

$count\left(s,t\right) = \sum\limits_{k=1}^{K} \phi_k \left(\overline{s} \mid \overline{t}\right) \sum\limits_{i=1}^{n} \delta\left(s, s_i\right) \delta\left(t, t_{a_i^{ST}}\right)$

where K denotes the number of the induced phrase pairs, and $\phi_k \left(\overline{s} \mid \overline{t}\right)$ is the phrase translation probability for the phrase k. $\delta\left(x,y\right) = 1$ if $x = y$, otherwise $\delta\left(x,y\right) = 0$.

The two methods for the calculation the lexical weight defined two variants of our pivot model.

The first method is based on the lexical translation files generated by the source-pivot and pivot-target model.

The lexical translation file contains the translation probabilities between simple words, out of their semantic context. Thus, the ambiguous words in the pivot language could generate an unreliable association between a source and target word. For example, if we use Romanian as pivot between English and French, the word "*mare*" that has two meanings (*sea* or *big*) could produce high translation scores between "*sea*" and "*grand*" or "*big*" and "*mer*".

The second method was introduced in order to reduce the effects of this problem. The aim is to improve the translation probability estimation, as it generates the lexical translation tables based on the pivot phrase-table, i.e. on the phrase contextual information.

It also alleviates the computational burden of generating the lexical word translation tables which have a more reduced size when generated from phrase table alignments (using the contextual information).

Lexicalized reordering models At first sight, it seems rather difficult to compute the lexicalized reordering model by combining the reordering models of the two training steps source-pivot and pivot-target.

The reordering tables do not contain the necessary data to calculate the distribution for each reordering type (*mono*, *swap*, *discontinuous*). Therefore, we are using intermediate tables generated during the training: the tables that contain all the extracted phrases with the orientation type associated (*extract-orientation* tables). They present for each phrase pair source-target the orientation type of the current phrase with respect to the previous phrase, and the orientation of the next phrase with respect to the current phrase.

As we are using a "*msd-bidirectional-fe*" model, the orientation information extracted from the alignments has the following format:

$source - phrase \mid\mid\mid target - phrase \mid\mid\mid orientation_{current} \quad orientation_{next}$

where $orientation$ can be $mono$, $swap$ or $other$ $(discontinous)$.

We generated a similar table for the pivot model based on the information provided by the source-pivot ($orientation^{SPiv}$) and pivot-target ($orientation^{PivT}$) tables. Firstly, we select all the source-target pairs that "share" the same pivot phrase and then we combine the ordering information of the tables as follows:

$$orientation^{ST}_{current} = \begin{cases} orientation^{SPiv}_{current} & if\ orientation^{PivT}_{current} =' mono' \\ orientation^{SPiv}_{next} & if\ orientation^{PivT}_{current} =' swap' \\ 'indet' & otherwise \end{cases}$$

$$orientation^{ST}_{next} = \begin{cases} orientation^{SPiv}_{current} & if\ orientation^{PivT}_{next} =' mono' \\ orientation^{SPiv}_{next} & if\ orientation^{PivT}_{next} =' swap' \\ 'indet' & otherwise \end{cases}$$

It may be that the orientation information available in the source-pivot and pivot-target tables is not sufficient to establish the orientation type of the source-target phrase pair. In this case, we consider the type as indeterminate and use the value 'indet' for the orientation. This presents a uniform distribution between the three types (*mono*, *swap* or *other*). (It means that this is counted as 1/3 for each orientation type).

Based on this table (pivot extract-orientation table) we calculate the scores for each phrase pair using the following formula (described in section 5.1.1.2)

$$p_o(orientation \mid s, t) = \frac{count\,(orientation, t, s)}{\sum\limits_o count\,(o, t, s)} \tag{6.11}$$

adapted to take into account the indeterminate ('indet') type as follows:

$$p_o(orientation \mid s, t) = \frac{count\,(orientation, t, s) + \frac{1}{3} count\,('indet', t, s)}{\sum\limits_o count\,(o, t, s) + \frac{1}{3} count\,('indet', t, s)} \tag{6.12}$$

and with the smoothing:

$$p_o(orientation \mid s,t) = \frac{\sigma\, p\,(orientation) + count\,(orientation, t, s) + \frac{1}{3} count\,('indet', t, s)}{\sigma + \sum\limits_{o} count(o, t, s) + \frac{1}{3} count\,('indet', t, s)} \tag{6.13}$$

where

$$p(orientation) = \frac{\sum\limits_{s}\sum\limits_{t} (count\,(orientation, t, s) + +\frac{1}{3} count\,('indet', t, s))}{\sum\limits_{o}\sum\limits_{s}\sum\limits_{t} count\,(o, t, s) + \sum\limits_{s}\sum\limits_{t} \frac{1}{3} count\,('indet', t, s)} \tag{6.14}$$

6.5 Interpolated translation models

These pivot methods lead to many errors and omissions in this table, that can be tackled by using the triangulated phrase table in conjunction with a standard table.

Moreover, we can use more than one pivot language to improve translation performance. Different pivot languages may catch different linguistic phenomena, and improve translation quality for the desired language pair S-T in different ways.

We suggest using the linear interpolation to combine two or more phrase tables.

6.5.1 Formal model

Once induced, the triangulated phrase-table can be usefully combined with the standard source-target phrase-table. The simplest approach is to use linear interpolation to combine the two (or more) distributions, as follows:

$$p\,(s,t) = \sum_{i} \lambda_i p_i\,(s,t)$$

where each joint distribution, p_i, has a non-negative weight, λ_i, and the sum of the weights is one. The joint distribution for the triangulated tables is defined by the previously presented pivot methods.

Weighting the contribution of each parallel corpora allows us to place more emphasis on larger parallel corpora, or on more "effective" pivot languages. We suggest that the standard phrase table be allocated a higher weight than triangulated phrase tables as it will be less noisy.

6.5.2 Merging the phrase-tables

If we include n pivot languages, n pivot models can be estimated using a triangulated method at alignment or phrase-table level. The phrase translation probability and the lexical weight are estimated as shown in the following equation:

$\phi\left(\overline{s} \mid \overline{t}\right) = \sum\limits_{i=0}^{n} \alpha_i \phi_i \left(\overline{s} \mid \overline{t}\right)$

$p_w\left(\overline{s} \mid \overline{t}, a\right) = \sum\limits_{i=0}^{n} \beta_i p_{w,i} \left(\overline{s} \mid \overline{t}, a\right)$

where $\phi_0\left(\overline{s} \mid \overline{t}\right)$ and $p_{w,0}\left(\overline{s} \mid \overline{t}, a\right)$ denote the phrase translation probability and lexical weight trained with S-T bitexts; $\phi_i\left(\overline{s} \mid \overline{t}\right)$ and $p_{w,i}\left(\overline{s} \mid \overline{t}, a\right)$ are the phrase translation probability and lexical weight estimated by using pivot languages; α_i and β_i are the interpolation coefficients.

The interpolation coefficient can be tuned using the development set. We will consider the same interpolation coefficients $\alpha_i = \beta_i\ (= \lambda_i)$ for the phrase translation probability and the lexical weight.

6.6 Triangulation at decoding time

This time, the pivot translation system uses two independently trained SMT systems: the S-Piv (source to pivot) translation system and the Piv-T (pivot to target) translation system.

6.6.1 Formal model

Let us recall that we are looking for the best translation given by equation:

$$t_{best} = arg\,\underset{t}{max}\, p\left(t \mid s\right)$$

The corresponding statistical decision can be derived by modelling the pivot sentence as a hidden variable and by assuming the independence between the target, t and the source, s, given the pivot sentence, piv:

$$t_{best} = arg\,\underset{t}{max}\, p\left(t \mid s\right) = arg\,\underset{t}{max} \sum_{piv} p\left(t, piv \mid s\right)$$

$$= arg\,\underset{t}{max} \sum_{piv} p\left(piv \mid s\right) p\left(t \mid piv, s\right) = arg\,\underset{t}{max} \sum_{piv} p\left(piv \mid s\right) p\left(t \mid piv\right)$$

$$\approx arg\,\underset{t}{max}\,\underset{piv}{max}\, p\left(piv \mid s\right) p\left(t \mid piv\right) \tag{6.15}$$

In the last step, we apply the max approximation, to reduce the complexity of the search procedure.

By assuming the standard phrase-based models for each of the probability expressions on the right-hand side of equation 6.15, we extend the search with two other hidden variables: the translation models source-pivot, TM^{SPiv} , and pivot-target, TM^{PivT}.

They model, respectively, phrase segmentation and reordering for each considered translation direction.

$$t_{best} \approx arg \underset{t,TM^{PivT}}{max} \underset{piv,TM^{SPiv}}{max} p\left(piv, TM^{SPiv} \mid s\right) p\left(t, TM^{PivT} \mid piv\right)$$

We can reduce the computational burden of the equation above by limiting the pivot translations piv to a limited subset $BestN_{TM^{SPiv}}$, such as the n-best list produce by source-to-pivot translation system:

$$t_{best} \approx arg \underset{t,TM^{PivT}}{max} p\left(t, TM^{PivT} \mid piv\right) \underset{piv \in BestN_{TM^{SPiv}}, TM^{SPiv}}{max} p\left(piv, TM^{SPiv} \mid s\right)$$

6.6.2 Building the pivot translation model

Our method consists of generating m-best target sentences for the n-best pivot translations generated by the source-pivot system, and re-scoring all the $m \times n$ hypotheses using both source-pivot and pivot-target scores. In this case, the subset $BestN_{TM^{SPiv}} = \{piv_1, \ldots, piv_n\}$

A drawback of this strategy is that translation speed is about $O(n)$ times slower than those of the component SMT systems. This is because we have to run n times for each source sentence. Consequently, we cannot set n very high. Note that when $n = 1$, the above strategy produces the same translation with the simple sequential method that translates a source sentence into pivot language and then translates that sentence into the target language.

The high multilinguality of our resources suggests a "multilingual" version of this method. We propose using the simple sequential method for many pivot languages and combining the results. The simplest way to proceed is to operate at the sentence level and then pick only those sentences from all the generated hypotheses that are the "best" according to some score.

This method could bring improvement to system performance, due to the high number of direct systems available (as described in section 5.3): 20 possible pivot systems for each language pair.

6.7 Factors affecting pivot translation models

We study the different factors that could influence the performance of a pivot translation system.

- Firstly, we consider the factor that constitutes the basis of our pivot methods: the way the pivot information is integrated into the translation system.

- The nature of the languages in a triad is an important factor that affects the translation. The degree of relatedness of the languages in a triad should play a role in how well a pivot alignment will work: a high degree of similarity with the source or target language should make the intermediate language more effective. On the other hand, the complexity of the languages should affect the performance. We assume that the usage of a pivot language more complex or structurally different from the source and / or pivot will not increase the performance. We suggest that the translation from an information-poor language into an information-rich language requires a different pivot than in the opposite direction.

- We also analyse some training requirements, such as the size of the training data: on small data sets, performance should increase with triangulation. The overlapping of the training set at the pivot level will be taken into consideration.

The next chapter will present the analysis of these factors via a set of experiments.

6.8 Conclusion

We presented different pivot-based translation models, that can be distinguished by the way they integrate the pivot information.

Thus, we described two main pivot-at-training methods: one that integrates the pivot information at the alignment level and the other that performs a phrase-table combination. They both present variants. The alignment pivot methods can combine the source-pivot and pivot-target alignments before or after the symmetrisation of the alignments performed during the `Moses` training process. The pivot models that integrates the bridge language at phrase-table level distinguished two heuristics for calculating the lexical scores.

We proposed a simple pivot-at-decoding method with a multi-pivot variant, based on the direct translation systems built for all the European language pairs.

The pivot-based models are evaluated in the next chaper, in a set of experiments designed to study the different factors that could affect their performances.

Chapter 7

Pivot Methods Experiments

The main application for our approach is done in statistical machine translation, the domain in which we performed a set of experiments. We will also describe a preliminary experiment, in which our methods were evaluated in the field of computational lexicography.

All the experiments carried out during our research will be presented in this chapter.

7.1 Preliminary experience

We evaluate the phrase-table based pivot methods in bilingual term extraction domain and more precisely in translation spotting. For this application, we have selected documents in a specific field (*Health*) using Eurovoc descriptors associated to each document (see *Health JRC-Acquis* corpora described in 4.3.1). We trained translation models on these data for different language pairs and we built the pivot models. Given a list of health-related terms, we check the translation produced by our systems and we evaluate the improvement brought by the triangulation.

The reason for this experiment is the initial orientation of our research. The main application that was foreseen at the JRC for the JRC-Acquis corpora was in the field of computational lexicography. The aim was to extract domain-related terms (nuclear-related or health-related) and to generate a bilingual computational dictionary that could be used in cross-language applications [Ignat et al., 2005, Versino et al., 2007].

These experiments open interesting application directions. The drawback is the difficulty in evaluating the translated terms extracted.

7.2 Experimental design

The next sections will present the evaluation of the pivot models in statistical machine translation.

Our experiments and evaluation were motivated by the following questions:

1. What is the best way to integrate the pivot information in the translation system? What is the quality of the pivot systems compared to the direct method?
2. How does the choice of the intermediate language, given a source and a target, influence the translation ?
3. How do different training requirements affect the performance of the pivot systems (size of the training data, the overlap of the training data at the pivot level)

The factors may depend on one other. It is possible that a specific pivot method performs better on a type of triad than on an other, depending on the nature of the languages involved.

7.2.1 Data set

We used the same training data as in the experiments described in 5.3. Our experiments are based on the direct translation models built for 462 language pairs.

7.2.1.1 Training data

The parallel corpora used are the *Acquis22* and *Translation-Units-22* corpus (described in 4.3.2 and 4.4.1).

We remember that the corpus *Acquis22* includes around 8.4 million words, and an average of 360 000 sentences for each language. We perform experiments on sub-corpora of different sizes of *Acquis22*: a randomly generated sample of 10 000 sentences (*Acquis22-sample10k*), of 50 000 sentences (*Acquis22-sample50k*), of 100 000 sentences (*Acquis22-sample100k*).

The corpus *Translation-Units-22* includes 8.7 million tokens with an average of about 400 000 tokens by language. It contains around 20 000 sentences by language, exactly aligned between all language pairs.

It is to be remembered that the corpora has been pre-processed for use with the `Moses` system including "sentence" (paragraph) splitting and tokenisation, as well as lower-casing (to avoid training separate models on uppercase and lowercase words). We extracted only sentences that have a length of less than 100 tokens (as this is a limit imposed by GIZA++ training).

7.2.1.2 Development set

Development data were described in subsection 4.4.2. They consist of 2600 sentences in the same domain as the training data, but which were not part of this data. They are separated in a tuning set and a test set.

Tuning corpus The tuning set includes 660 sentences for each language. For computational reasons we have not used the tuning in our experiments.

Test corpus The test set contains 2000 sentences for each language. For computational reasons (time processing), we used only the first 1000 sentences that includes a total number of 1.1 million tokens, with an average of about 50000 tokens per language.

7.2.2 Description of experiments

We developed specific experiments to study each factor mentioned above. Thus, we can distinguish the following sets of experiments grouped by the envisaged aim.

7.2.2.1 Experiments for comparing pivot methods

We studied the pivot methods described in the previous chapter: two main methods at training time with their variants (that constitutes five methods at training time). They are compared with two methods at decoding time. The methods at the training time integrate the pivot information either at the phrase table level, or at the alignment level. For the first method (at phrase-table level) we compare three variants which differ in the way the lexical weights are calculated (see 6.4.2.2 - Calculating the lexical weights). The two variants of the second method (at alignment level) integrate the pivot information either before, or after the alignment symmetrisation (see 6.4.1.2).

The pivot methods which we have implemented and compared are the following:

1. ***Pivot0* method (at phrase-table level)**

   It is a method at training time, that integrates the pivot data at the phrase table level. The lexical scores are calculated based on the lexical scores provided by the source-pivot and the pivot-target phrase-tables (by multiplication). We have implemented this for comparison reasons, as it is similar with the [Cohn and Lapata, 2007]'s method and for computational reasons as it is the simplest and fastest (computationally) way to calculate the lexical scores.

2. ***Pivot1* method (at phrase-table level)**

   It is a method at training time for which the pivot information is integrated by phrase table combination. The lexical scores are calculated based on the lexical word translation table obtained via the translation tables between source-pivot and pivot-target languages. Computationally, the method has important memory requirements.

3. ***Pivot2* method (at phrase-table level)**

   It is also a method at training time similar with the *Pivot1* method. They differ in the way the lexical scores are calculated. This method is based on the phrase contextual information provided by the phrase alignments, thus the lexical word translation tables are generated based on the pivot phrase table.

4. ***Pivot3* method (at alignment level)**

   It is a method at training time which integrates the pivot information at the alignment level. The symmetrised alignment tables between source-pivot and pivot-target are combined to generate the source-pivot symmetrised table (see 6.4.1.2 - Combining symmetrised alignments).

5. ***Pivot4* method (at alignment level)**

   This is also a method at alignment level for which the pivot information is integrated before the symmetrisation of the alignment tables. Thus, we combine the GIZA++ one-to-many alignments in both directions (see 6.4.1.2 - Combining unidirectional alignments). Then, the tables are symmetrised via the *grow-diag-final* heuristics provided by Moses.

6. ***PaD* method (pivot-at-decoding)**

   This is the direct sequential way to combine two translation systems which translate the source sentence into the pivot language and then the pivot sentence into the target language. This method has been implemented for comparison reasons, as the baseline pivot-at-decoding method.

7. ***mPaD* method (multi-pivot-at-decoding)**

   This is the multilingual version of the pivot-at-decoding method, described in 6.6.2, where we choose the best sentences obtained via the sequential pivot-at-decoding language across multiple pivot languages.

The tools developed for each method are described in the subsection 7.2.3.

The pivot-at-alignment methods require a certain type of training data: sentence-aligned texts across the three languages (source, pivot and target). This type of data is provided by the *Translation-Units-22* corpus, therefore for this experiment the models were trained on the *Translation-Units-22* data set.

7.2.2.2 Comparing interpolated, pivot and direct models

These experiments study the performance of the interpolated translation models, comparing them with the direct and pivot models related. The interpolated models concern only the pivot-at-training methods (*Pivot0*, *Pivot1*, *Pivot2*, *Pivot3*, *Pivot4*) and have been described in the previous chapter, section 6.5. We generated the interpolated models of the pivot systems obtained in the previous experiments and we compare them with the direct and the pivot methods via the BLEU scores. We present two types of interpolated models:

1. **interpolated model**, in which we combine a pivot with a direct model. For each language pair we choose different pivot languages.

2. **multi-pivot interpolated model,** in which we combine the direct model with more pivot models for a given source-target pair.

The simple interpolated model uses the interpolation coefficients equal to 1, meaning that we give the same weight to the direct and the pivot model.

The multi-pivot interpolated model uses different sets of interpolation coefficients.

As these experiments are based on the pivot systems presented in the previous subsection, we used the same data set, *Translation-Units-22* corpus.

7.2.2.3 Experiments for comparing different pivot languages for a source-target language pair

We designed a set of experiments to analyse the performance of different pivot languages for a given source-target language pair. For a given source-target pair we generate the pivot model using different pivot languages. We choose the pivot according to the performance (measured in BLEU score) of the direct systems source-pivot and pivot-target and / or to the relatedness between the pivot and the source or pivot and the target languages. The correlation between languages via the BLEU score vectors “INTO” (see table 5.2 in the section 5.3.4) is also an important criteria in the choice of the triad of our experiments.

The experiments were designed around specific languages (French, Romanian) or language pairs (French - German, Finnish - Maltese).

We studied the improvements brought by the pivot models in the translation systems “from” and “into” French and between Romanian and a Slavic language (in both directions). We tried to find a better translation model between Maltese and Finnish, as it was one of the language pair with the lowest BLEU score for the direct system.

We evaluated English as pivot language for different source-target language pair (where source and target are not English), because this is the language that is the most used in the real life applications as a bridge language in translation.

These models are trained on the *Translation-Units-22* data set.

7.2.2.4 Experiments for comparing different training conditions (corpus size and data overlapping)

We designed a set of experiments with the focus on the size of the pivot training data. We used the *Acquis22* corpus from which we have extracted variously sized portions (*Acquis22-sample10k*, *Acquis22-sample50k, Acquis22-sample100k*), as described in the previous subsection. We trained pivot translation models using each of the data set listed before for the triad: source - German, target - French and pivot - English.

We used the *Pivot0* method to generate the pivot models. We tested how effective the pivot model was at improving the translation quality for translation models trained from all these sets. Because models trained from smaller amounts of training data

Experiments	Source - Target		Pivots
from and into German	fr - de	de - fr	da, en, es, nl, pt, it
	en - de	de - en	nl
French - Romance languages	ro - fr	fr - ro	en, fr, it, pt, (bg, el)
	pt - fr	(fr - pt)	en, fr it pt
Romanian - Slavic Languages	ro - cs	cs - ro	bg, en, fr
	ro - pl	pl - ro	bg, en, fr
	ro - sl	sl - ro	bg, en, fr
	ro - sk	sk - ro	bg, en, fr
Finnish, Estonian (English as pivot)	mt - fi	fi - mt	en
	fr - fi	fi - fr	en, (et, es, pt)
	ro - fi	fi - ro	en
	mt - et	et - mt	en
From English	en -de	-	nl
	en - fi	-	mt
	en - fr	-	ro

Table 7.1: Designed experiments on different source, target and pivot languages

are prone to coverage problems, the expectation was that the translation quality will improve more for smaller training sets and that there was less potential to improve translation quality for larger training sets.

7.2.3 Tools

Our experiments are based on `Moses` tool, but for each pivot method we developed specific modules that have been integrated in the `Moses`' processing workflow. The overview of `Moses`' processing has been presented in the section 5.2 figure 5.2. Here, we will focus on the `Moses` training, as the pivot information has been integrated at this level by most of our pivot methods (*Pivot0* to *Pivot4*). The figure 7.1 details the processing steps of `Moses` training and shows all the intermediate outputs.

Each pivot method integrates the pivot language information in a different way, at a different point of the training.

The first three methods (*Pivot0*, *Pivot1* and *Pivot2*) are based on the phrase-tables source-pivot and pivot-target, from which the source-target phrase table is generated. The *Pivot0* method generates the phrase-table directly from the source-pivot and pivot-target ones in the same time with the alignments, the phrase and lexical scores. The *Pivot1* and *Pivot2* methods use a different approach to calculate the lexical scores. The figures 7.2 and 7.3 detail the steps and the resources of the pivot phrase table generation by *Pivot1* and *Pivot2* methods.

The *Pivot1* method uses the lexical word translation tables source-pivot and pivot-

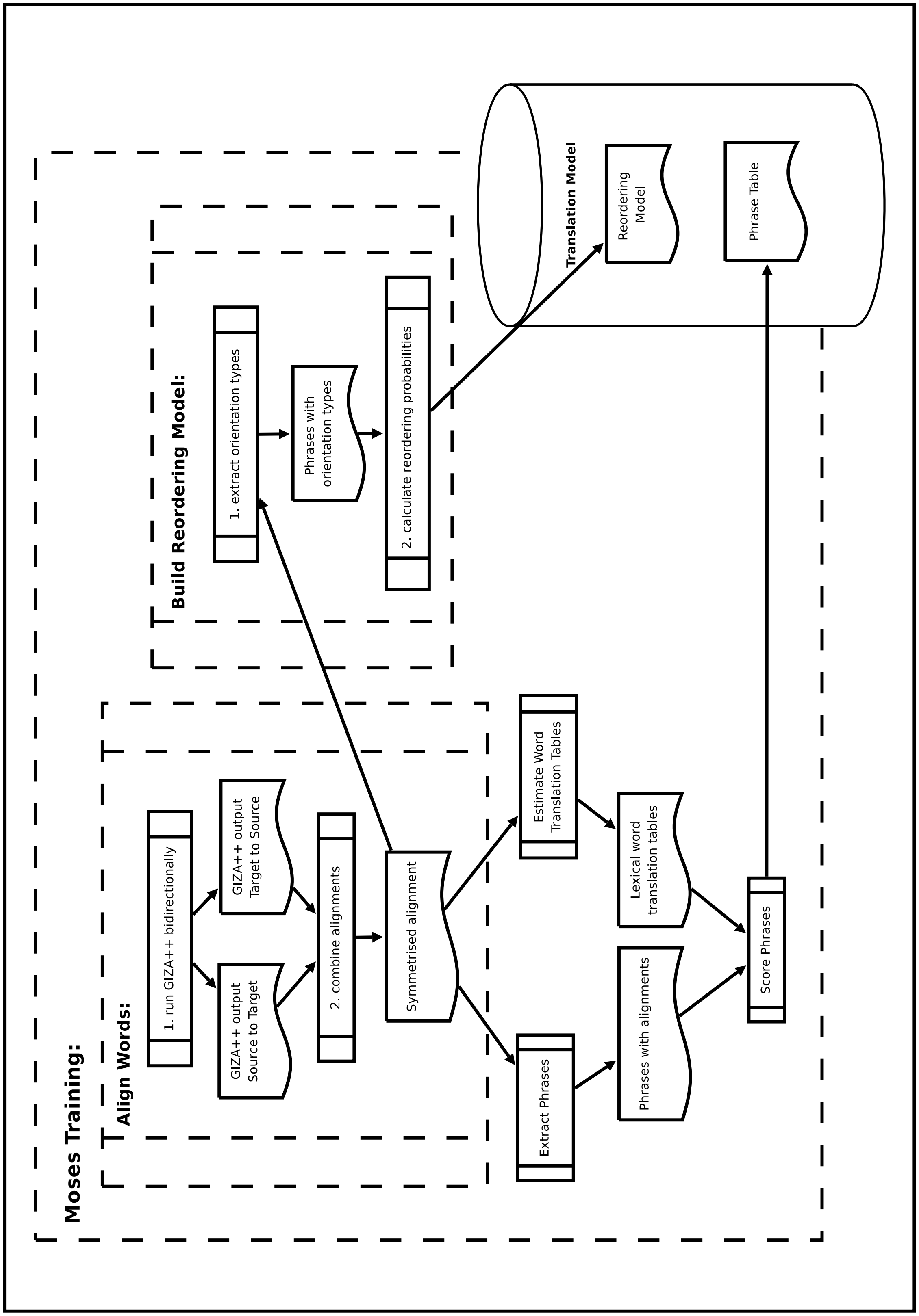

Moses Training:
Align Words:
1. run GIZA++ bidirectionally
GIZA++ output Source to Target
GIZA++ output Target to Source
2. combine alignments
Symmetrised alignment
Extract Phrases
Estimate Word Translation Tables
Phrases with alignments
Lexical word translation tables
Score Phrases
Build Reordering Model:
1. extract orientation types
Phrases with orientation types
2. calculate reordering probabilities
Translation Model
Reordering Model
Phrase Table

Figure 7.1: Zoom on Moses training: the processes and the resources involved

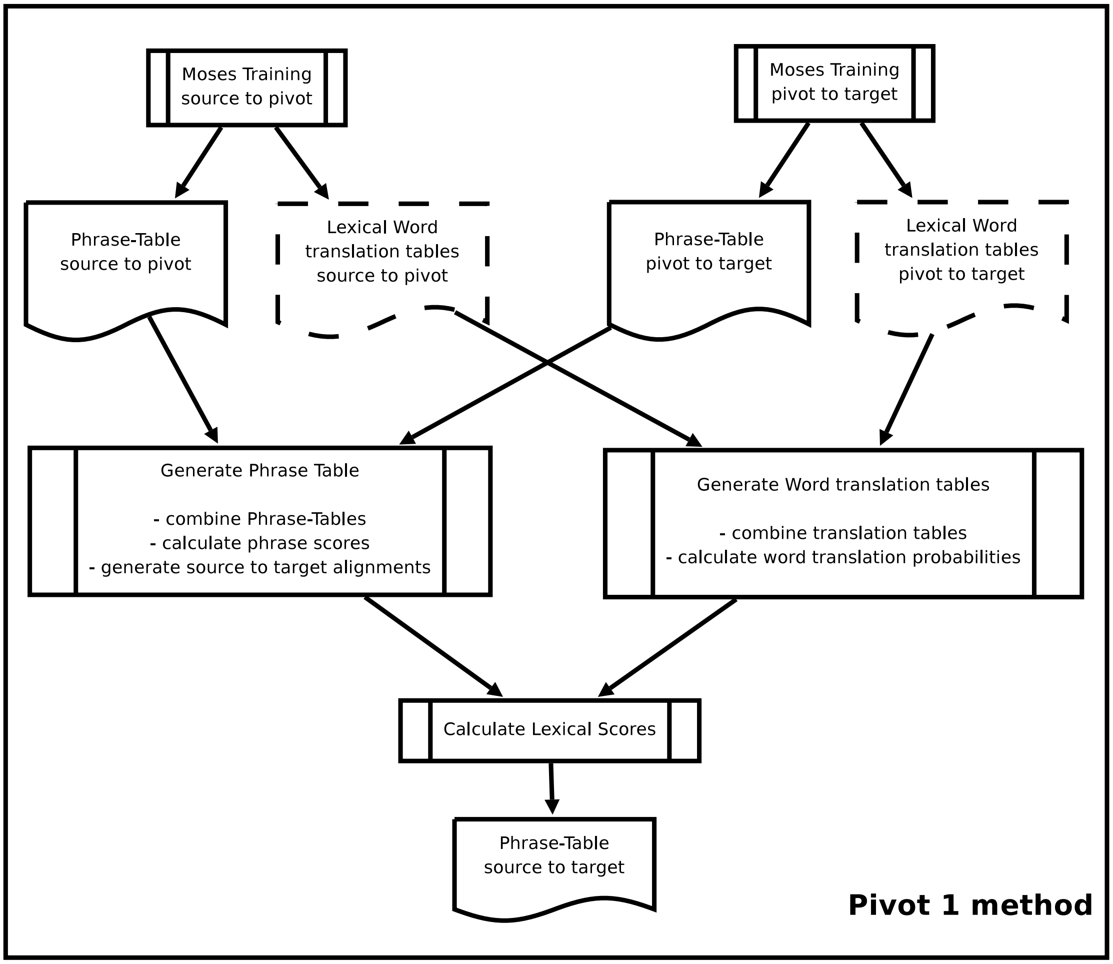

Moses Training source to pivot
Moses Training pivot to target
Phrase-Table source to pivot
Lexical Word translation tables source to pivot
Phrase-Table pivot to target
Lexical Word translation tables pivot to target
Generate Phrase Table
- combine Phrase-Tables
- calculate phrase scores
- generate source to target alignments
Generate Word translation tables
- combine translation tables
- calculate word translation probabilities
Calculate Lexical Scores
Phrase-Table source to target
Pivot 1 method

Figure 7.2: Building the pivot phrase table in *Pivot1* method

target to generate the source-target word translation table. This is used together with the alignment information contained in the source-target phrase-table to calculate the lexical scores. The *Pivot2* method generates the lexical word translation tables based on the alignment information calculated for the source-target phrases. Then, this is used to calculate the lexical scores, as in the *Pivot1* method.

The reordering table is calculated in the same way for all the methods *Pivot0*, *Pivot1* and *Pivot2*. We start with the intermediate outputs of the reordering calculation process: the tables source-pivot and pivot-source that contain the list of extracted phrases with the orientation types (mono, swap, other). We determine the orientation types for the source-target phrases using the procedure described in 6.4.2.2 (where a new orientation type "indeterminate" has been introduced). Then, this generated table is used to calculate the reordering probabilities for the source-target phrases (after equation 6.13). See figure 7.4 for the main steps performed to generate the reordering table of the pivot model.

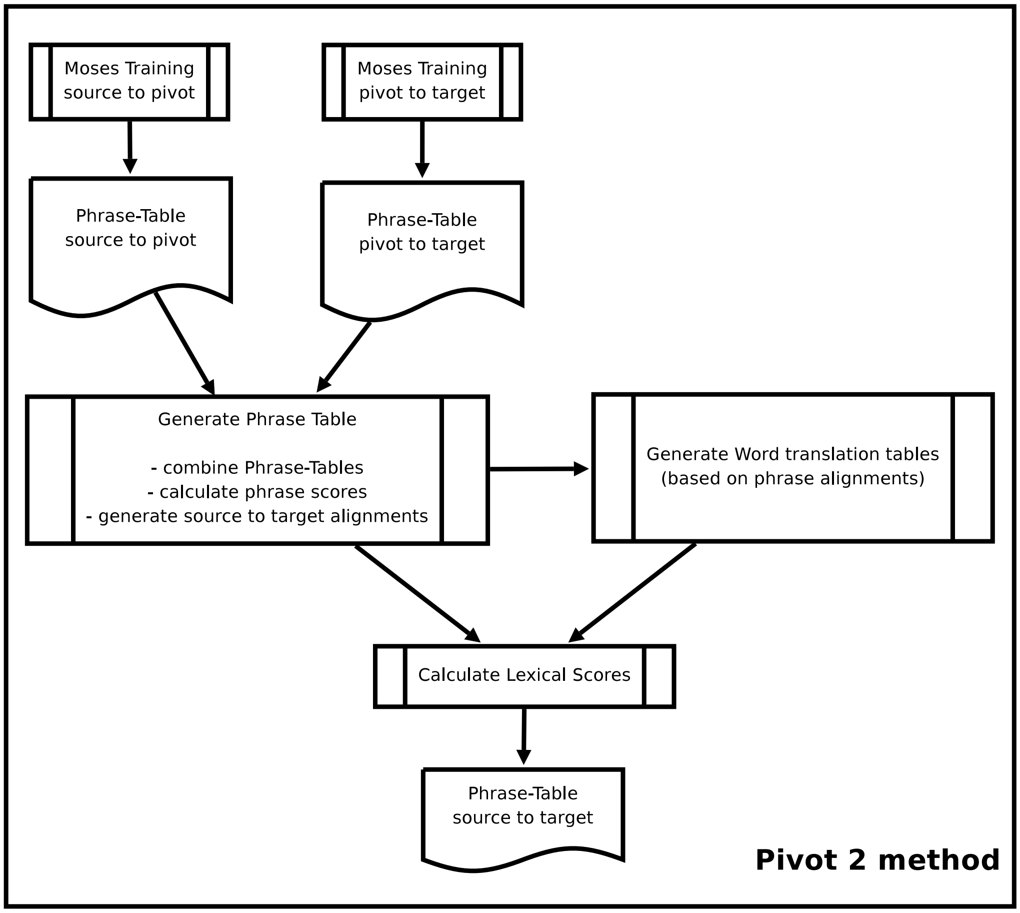


Figure 7.3: Building the pivot phrase table in *Pivot2* method

The *Pivot3* and *Pivot4* methods integrate the pivot information into the `Align words` module. (see the overview of `Moses` training in the figure 7.1). They are both producing the symmetrised alignments source-to-target, that are then used in the `Moses` training workflow.

For the *Pivot3* method, the symmetrised alignments source-to-pivot and pivot-to-target are combined to generate the (symmetrised) source-to-target alignments. The *Pivot4* method uses `GIZA++` outputs and it combines on one hand source-to-pivot with pivot-to-target `GIZA++` outputs to obtain the source-to-pivot uni-directional alignment, and on the other hand, the target-to-pivot with the pivot-to-source `GIZA++` outputs to generate the target-to-source alignments. These one-to-many alignments are then symmetrized using the *grow-diag-and-final* heuristics of `Moses`.

The pivot-at-decoding methods (*PaD* and *mPaD*) directly use the translation systems source-to-pivot and pivot-to-target. Thus, the *PaD* method is the sequential coupling of the two systems that translates firstly the source sentence into pivot language sentence, that is then translated into the target language (see figure 7.5).

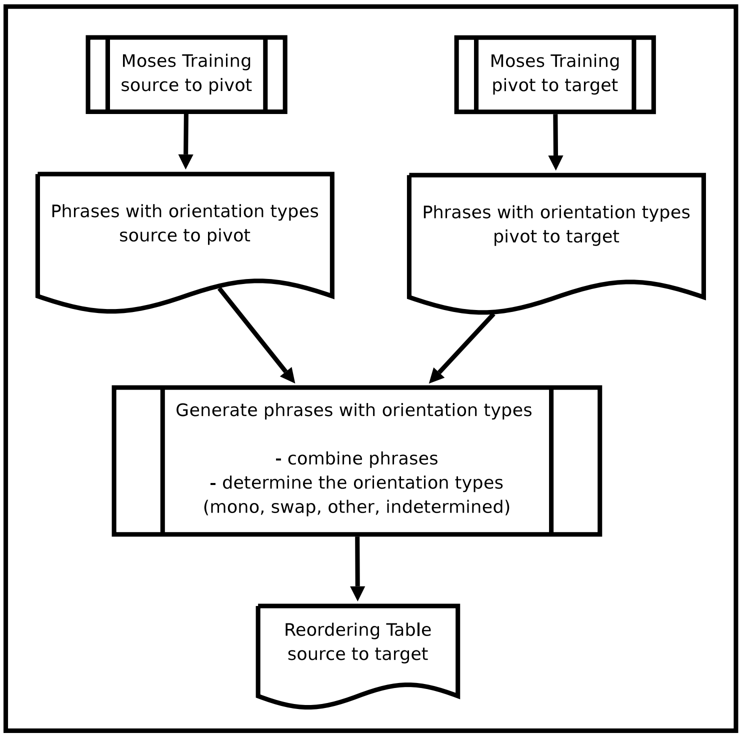

Moses Training source to pivot
Moses Training pivot to target
Phrases with orientation types source to pivot
Phrases with orientation types pivot to target
Generate phrases with orientation types
- combine phrases
- determine the orientation types (mono, swap, other, indetermined)
Reordering Table source to target

Figure 7.4: Building the reordering table in pivot-based models *(Pivot0*, *Pivot1* and *Pivot2* methods)

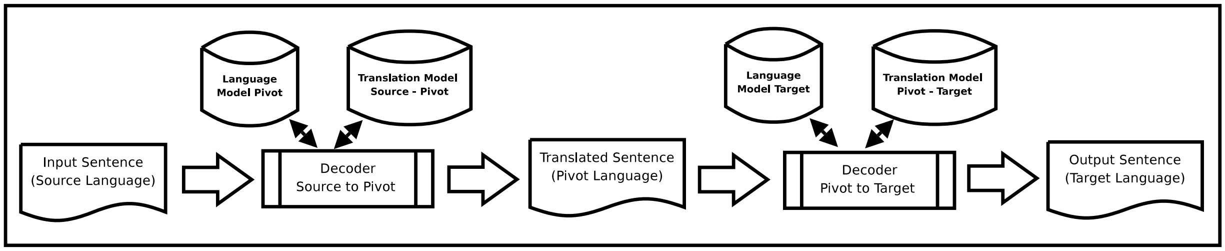

Language Model Pivot
Translation Model Source - Pivot
Language Model Target
Translation Model Pivot - Target
Input Sentence (Source Language)
Decoder Source to Pivot
Translated Sentence (Pivot Language)
Decoder Pivot to Target
Output Sentence (Target Language)

Figure 7.5: The pivot-at-decoding method, *PaD*

The *mPaD* method uses a set of sequential pivot systems, as those generated by the *PaD* method. Each of these parallel systems outputs a translated sentence with its associated score. A `filtering module` chooses the translated sentence with the highest score as an output for the global system *mPaD*. The figure 7.6 gives an overview of this multi-pivot process.

7.3 Results and discussions

In this section, we present and compare the performance of our pivot-based translation systems trained on *Translation-Units-22* corpus.

The Appendix C shows a sample of outputs of pivot-based translation systems, translating into French and Romanian.

Next, we present and discuss the results of the different experiments performed.

7.3.1 Comparing pivot methods

According to our results, it is not possible to choose an overall best method: the performance of a specific model seems to depend on the triad of languages involved. Table 7.2 gives an idea about the performance of the "pivot-at-training" methods and the table 7.3 compares the "pivot-at-decoding" models with *Pivot2* method.

First, we discuss the results for "pivot-at-training" models. Method *Pivot0* has generally the lowest score, but the number of experiments where it was involved is too reduced to provide an evidence on this claim.

Amongst the "phrase-table methods" *Pivot1* and *Pivot2* we cannot distinguish the one that performs better, but from the computational point of view, *Pivot2* method is preferable, as it requires less resources *(Pivot1* requires huge memory resources to calculate the lexical scores).

Amongst "pivot-at-alignment" methods, the *Pivot4* method, where the pivot information is integrated after the symmetrisation seems to obtain higher scores than the *Pivot3* method. We assume that the last one performs the combination of one-to-many uni-directional alignments that leads to a loss of information via the pivot language.

Comparing with the direct translation model between source and target, the performance of the "pivot-at-training" models generally seems to decrease, except for certain triads that get better results for the pivot models. This is the case of the Maltese-to-Finnish and Finnish-to-Maltese systems for which the English pivot language makes significative improvements in the BLEU score comparing it with the direct method. The Romanian-to-Polish system with French or English as bridge languages also brings improvements in the BLEU score, compared to the direct model. The Romanian-to-Polish models with English as pivot are in the same situation.

The pivot-at-decoding methods perform better than the pivot-at-training methods, in the same cases when the pivot-method overscores the direct model. The multi-pivot *mPaD* method has not the best performance among the simple *PaD* models for specific pivot languages. We think this is due to the way of combining and comparing the scores provided by the decoders of these systems. A better way to calculate and filter the scores of sequential system coud improve the results of the multi-pivot model.

7.3.2 Comparing direct, pivot and interpolated methods

We remark above that usually the pivot model performs less well than the direct model, except for certain language combinations of source, pivot and target. However, the interpolated method overscores both of them (on identical training conditions). Table 7.2 shows the BLEU scores of the interpolated models on different language pairs.

7.3.3 Comparing different pivot languages for a source-target language pair

Our experiments have been designed around a language or a language pair source-target.

We will present here the results for each experiment described in 7.2.2.3. (Some results could be displayed in more than one experiment)

Translation into (and from) German In SMT approach, German is quite a difficult language to translate into. We evaluate the impact of a pivot language when translating from French into German (and the opposite direction). We evaluate a set of pivot models based on some Romance and Germanic languages, including English. Using *Pivot2* and *PaD* method, the best bridge language, in this case is Dutch (in both directions), followed by English (see table 7.4).

French - Romance languages The direct translation models between pairs of Romance languages have high BLEU scores among all the combinations of EU official languages. The correlation between these languages is also very close to 1. We study the impact of a pivot language from the same family, on such a system with a good performance. We choose French-Romanian and Portuguese-French models, that are generated by pivoting to other Romance languages. The results are displayed in the table 7.5.

We distinguish Portuguese as the bridge with the highest BLEU score for the model French-Romanian, in both directions. The Portuguese-French model has good performance when pivoting through Spanish or Italian.

We also present the evaluation of the French-Romanian model when using Greek or Bulgarian pivots, as these languages are strongly correlated with Romanian and the French, given the BLUE score vectors "INTO".

Romanian and the Slavic languages We studied the performance of the translation systems Romanian-Czech, Romanian-Polish, Romanian-Slovakian and Romanian-Slovene, in both directions. We have evaluated pivot models using the following pivot languages: Bulgarian (as the Slavic language most correlated with Romanian), English, French (as a Romance language correlated with Romanian).

The results show that the "best" bridge language is English in three out of four systems (see table 7.6).The exception is the Romanian-Polish model, which performs better pivoting via French language. The BLEU scores are higher than those of the direct model, when the translation direction is Romanian to Polish.

Finnish and Estonian pivoting through English Finnish and Estonian are the languages "difficult" to translate into. We try to improve the system involving these languages, by using the pivot model through English. We evaluate the translation from and into Finnish, and the following languages - French, Romanian, and Maltese. We consider also the Maltese-Estonain model, which has the lowest BLEU score among the 462 language pair combinations. The pivot systems are generated using *Pivot2* and/or *pivot-at-decoding* methods.

Table 7.7 presents their performance measured in BLEU score. We note that the pivot systems overscore or have very close scores compared to the direct models. The systems involving Maltese (Estonian-Maltese and Finnish - Maltese) prove significant evidence on this, which we think is due to the strong correlation between English (pivot language) and Maltese (as source or target).

In some sense, pivoting through English results in a nice factorization of the translation model: this probably has a positive impact in terms of less data sparseness in the training data and results in better statistical models. The experiments on Finnish and Estonian pivoting through English, provides an evidence to this claim.

7.4 Conclusions

The evaluation of our pivot-based models has been designed to investigate some main directions. We tried to designate the best way to integrate the pivot information in the translation system and to study the quality of the pivot systems compared to the direct method. On the other hand, we explore how the choice of the intermediate language, given a source and a target, influences the translation.

Given the results of our evaluations, it is not possible to design the overall "best" pivot method, although some general direction exists. Amongst "pivot-at-alignment" methods, the one which integrates the pivot information after the symmetrisation seems to obtain higher scores. We assume that the combination of one-to-many uni-directional alignments may lead to a loss of information via the pivot language. However, the performances of the methods evaluated are dependent on a specific triad.

Generally, the pivot model performs less well than the direct model, but the interpolated method overscores both of them (on identical training conditions). However, for some language pairs the pivot method overscores the direct system, (i.e., Maltese-to-Finnish via English), where the complexity of the translation system Maltese-to-Finnish is better modellised by separating it into two models, Maltese-to-English and English-to-Maltese. In some sense, pivoting through English results in a nice factorization of the translation model: that probably has a positive impact in terms of less data sparseness in the training data and results in better statistical models. The experiments on Finnish and Estonian pivoting through English, also provides an evidence to this claim.

In summary, our experimental results have shown that triangulation is not a mere approximation of the source-target phrase table or the direct model, but that extracts additional useful translation information. We want to highlight the importance of the nature of the languages in a triad when using a pivot language.

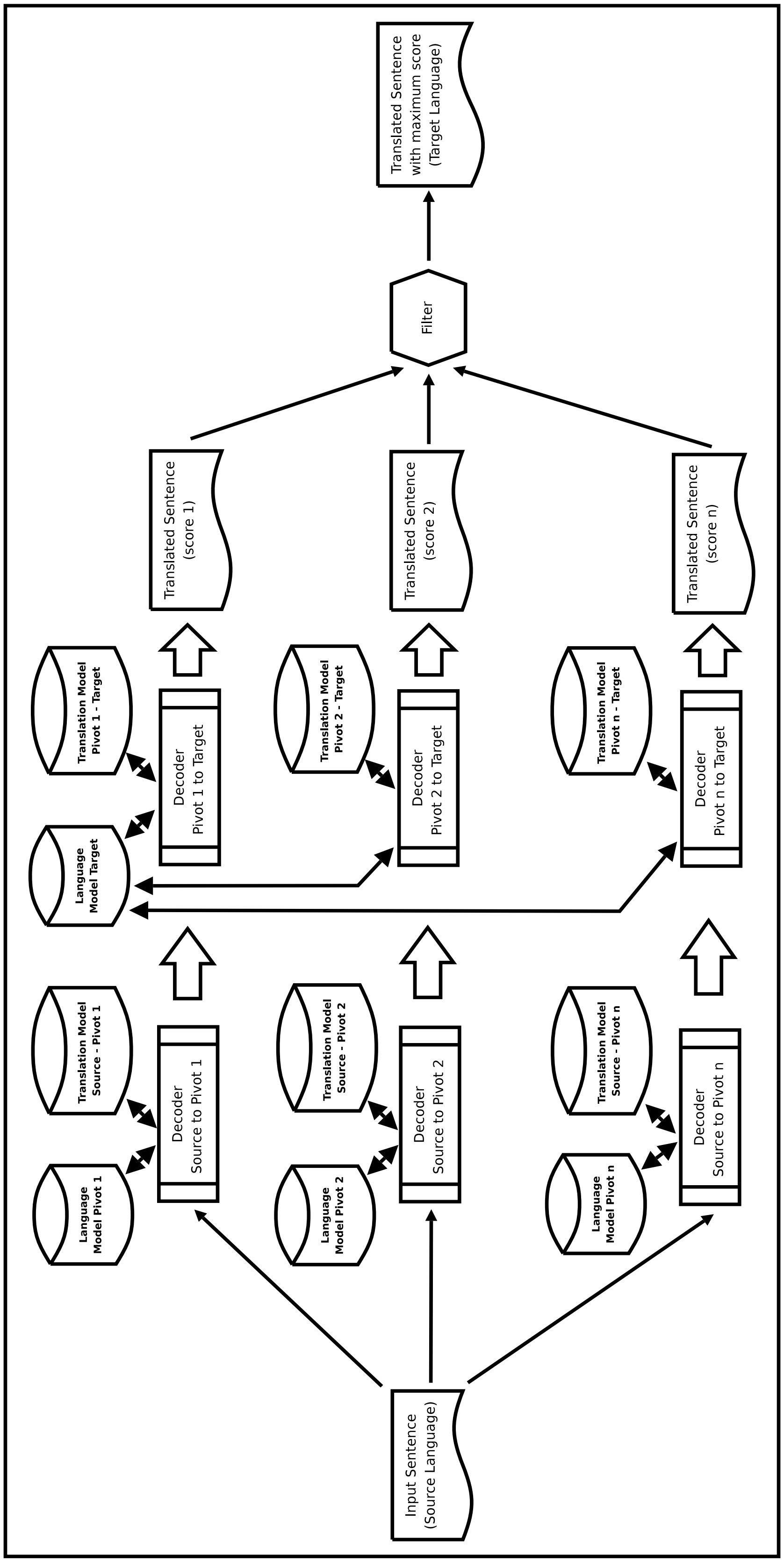


Figure 7.6: Overview of the *mPaD* method, a pivot-at-decoding method with multiple pivot languages

Languages		Direct	Pivot methods						Interpolated methods				
S-T	**Piv**	**method**	***Piv0***	***Piv1***	***Piv2***	***Piv3***	***Piv4***	***Piv5***	***Piv0***	***Piv1***	***Piv2***	***Piv3***	***Piv4***
fr - de	en	*32.7*	31.08	31.41	31.56	32.28	32.49	30.46	33.14	33.49	33.10	32.42	33.39
	nl		31.24	31.65	31.55	32.3	32.15	30.61	32.14	32.45	32.27	32.73	32.81
	da					30.73	30.67					32.60	32.74
mt - fi	en	*17.13*	17.66	18.24	17.7	17.63	17.75	19.54	19.02	19.27	19.03	18.61	18.49
fi - mt	en	*20.63*	21.55	21.21	22.09	22.28	22.02	23.27	22.06	21.89	22.38	22.23	21.98
ro - cs	bg	*32.14*		32.23	32.02	31.86	32			33.33	33.10	33.16	33.10
	en			31.11	32.81	32.33	32.62			33.35	33.23	32.74	32.85
	fr				32.46	32.39	32.72				33.51	33.25	32.82
ro - pl	bg	*30.98*			31	30.72	31.05				31.94	31.67	31.86
	en			31.73	31.43	31.01	31.32			32.26	32.16	31.61	31.72
	fr				31.54	31.23	31.51				31.57	31.20	31.70
ro - sk	bg	*28.31*			28.48	28.18	27.93				29.21	29.16	28.60
	en			29.17	29.2	28.71	29.14			29.81	29.93	29.39	29.21
	fr				28.69	28.11	28.28				29.15	28.67	28.88
ro - sl	bg	*29.51*		29.5	29.16	29.53	29.79			30.66	30.44	30.41	30.44
	en			30.78	31.21	29.35	30.16			31.02	30.61	30.34	30.60
	fr				29.56	29.43	29.83				30.43	30.08	30.22
en - fr	ro	*50.89*		48.41	49.34	49.74				51.01	51.09		
en - de	nl	*31.28*	29.96	30.37	30.83	30.86	30.93		31.97	31.92	32.05	31.88	31.99
en - fi	mt	*21.64*				20.2	21.11					22.60	

Table 7.2: Comparing pivot methods: BLEU scores for different pivot-based models

Languages		**Direct**	**Pivot methods**		
Source-Target	**Pivot**	**method**	***Pivot2***	***PaD***	***mPaD***
fr - de	en	*32.70*	31.56	30.46	31.63
	nl		31.55	30.61	
	es		31.53	30.4	
	pt		31.21	30.8	
fr - ro	en	*42.69*	40.9	37.98	38.7
	nl		41.2	38.85	
	es		41.34	39.22	
	pt		41.37	39.23	
fr - fi	en	*20.96*	20.69	20.84	
fi - fr	en	*27.64*	27.04	26.62	26.57
ro - fi	en	*19.55*	19.04	20.06	
fi - ro	en	*21.89*	21.67	21.67	21.74
mt - fi	en	*17.13*	17.70	19.54	
fi - mt	en	*20.63*	22.09	23.27	
pl - ro	bg	*31.32*	30.33	29.90	31.14
	en		31.11	31.15	
	fr		30.83	31.20	

Table 7.3: Comparing pivot methods: BLEU scores for *Pivot2* and *Pivot-at-Decoding* (*PaD*) methods

		Direct	**Pivot Language**				
Source-Target	**Method**	**Method**	***en***	***es***	***nl***	***pt***	***it***
fr - de	Pivot2	*32.70*	31.56	31.53	31.55	31.21	
	PaD		30.46	30.40	30.61	30.80	30.71
	mPaD		31.63				
de - fr	Pivot2	*37.48*	36.88	36.60	36.97	36.81	

Table 7.4: Comparing pivot languages: BLEU scores for French - German pivot-based models

		Direct	**Pivot Language**						
Source-Target	**Method**	**Method**	***en***	***es***	***it***	***pt***	***ro***	***bg***	***el***
ro - fr	Pivot2	*52.17*	50.86	50.69	50.81	51.03	-		
fr - ro	Pivot2	*42.69*	40.90	41.20	41.34	41.37			
	PaD		37.98	38.85	39.22	39.23	-	36.46	36.74
	mPaD		38.70						
pt - fr	Pivot2	*58.07*	55.25	57.17	57.16	-	55.83		
fr - pt	Pivot2	*52.34*	50.56			-			

Table 7.5: Comparing pivot languages: BLEU scores for Romanian-French and Portuguse-French pivot-based models

	Direct	**Pivot Language**		
Source-Target	**Method**	***bg***	***en***	***fr***
ro - cs	*32.14*	32.02	**32.81**	32.46
cs - ro	*30.82*	30.23	**30.95**	30.56
ro - pl	*30.98*	31.00	31.43	**31.54**
pl - ro	*31.32*	30.33	**31.11**	30.83
ro - sk	*28.31*	28.48	**29.20**	28.69
sk - ro	*30.72*	30.27	**31.16**	30.15
ro - sl	*29.51*	29.16	**30.21**	29.56
sl - ro	*30.53*	29.61	**31.00**	30.07

Table 7.6: Comparing pivot languages: BLEU scores for different pivot-based models (Romanian-Czech, Romanian-Polish, Romanian-Slovakian, Romanian-Slovene)

Source-Target	**Method**	**Direct Method**	**Pivot Language**			
			en	***et***	***es***	***pt***
fr - fi	Pivot2	*20.96*	20.69			
	PaD		20.84			
fi - fr	Pivot2	*27.64*	27.04			
	PaD		26.62	25.49	25.36	25.81
	mPaD		26.57			
ro - fi	Pivot2	*19.55*	19.04			
	PaD		20.06			
fi - ro	Pivot2	*21.89*	21.67			
	PaD		21.74			
mt - fi	Pivot2	*17.13*	17.70			
	PaD		19.54			
fi - mt	Pivot2	*20.63*	22.09			
	PaD		23.27			
mt - et	PaD	*16.17*	18.41			
et - mt	PaD	*22.00*	24.42			

Table 7.7: BLEU scores for different pivot-based models (Finnish or Estonian pivot English)

Part IV

Conclusion

Chapter 8

Conclusions

This final chapter contains our conclusions, a summary of contributions and prospects for future work.

Parallel corpora available in several languages provide better training material for alignment systems relative to bilingual corpora. We combine word alignments using several bridge languages with the aim of correcting some alignment errors and improving the coverage. We provide recipes to use a bridge language to construct a word alignment and a translation model and to combine translation models. We show that parallel corpora available in multiple languages can be exploited to improve the translation performance of a phrase-based translation system.

8.1 Contributions

Compilation of parallel corpora JRC-Acquis and its specific sub-corpora Parallel corpora are useful for all types of cross-lingual research. The value of a parallel corpus grows with its size and the number of languages for which translations exist. While parallel corpora for some languages exist in abundance, there are few or no parallel corpora for most other language pairs. To our knowledge, the JRC-Acquis is the biggest parallel corpus in existence, if we take into consideration both its size and the large number of languages involved. The most outstanding advantage of the JRC-Acquis - apart from being freely available - is the number of rare language pairs (e.g. Maltese-Estonian, Slovene-Finnish, etc.).

We presented the compilation of the highly multilingual corpus JRC-Acquis which was accomplished during my stay at the Joint Research Centre of the European Commission.

The subcorpora presented in section 4.3 and section 4.4 *(Acquis-22, Health-Acquis, Translation-Units-22, Acquis-TU-Devset)* have been created in the context of our thesis, in order to study and validate the pivot SMT approach. These subcorpora will probably be publicly available in the near future.

Translation models for 231 language pairs (in both directions) We used the Acquis sub-corpora parallel in 22 languages to create 462 translation systems for all possible language pairs. The resulting systems and their performances revealed the different challenges for statistical machine translation.

We analysed the correlation between the BLEU score vectors "INTO" that reveals how easy or difficult the translation " between certain language pairs will be.

We note the importance of the language relatedness in a translation system: it is easier to translate languages that are related to one another. On the other hand, the SMT models tend to perform much better when translating to morphologically simpler languages. We found a high correlation between the number of different tokens of the training data (vocabulary size) and the overall performance of a translation system (when translating into English).

Translation models by triangulation We presented different pivot-based translation models, that can be distinguished by the way they integrate the pivot information.

Thus, we described two main pivot-at-training methods: one that integrates the pivot information at the alignment level and the other that performs a phrase-table combination. They both present variants. The alignment pivot methods can combine the source-pivot and pivot-target alignments before or after the symmetrisation of the alignments performed during the `Moses` training process. The pivot models that integrate the bridge language at phrase-table level distinguished two heuristics for calculating the lexical scores.

We proposed a simple pivot-at-decoding method with a multi-pivot variant, based on the direct translation systems built for all the European language pairs.

The pivot-based models have been evaluated in a set of experiments designed to study the different factors that could affect their performances.

Experiments using pivot languages The evaluation of our pivot-based models has been designed to investigate some main directions. We tried to designate the best way to integrate the pivot information in the translation system and to study the quality of the pivot systems compared to the direct method. On the other hand, we explore how the choice of the intermediate language, given a source and a target, influence the translation.

Given the results of our evaluations, it is not possible to design the overall "best" pivot method, although some general direction exists. Amongst "pivot-at-alignment" methods, the one which integrates the pivot information after the symmetrisation seems to obtain higher scores. We assume that the combination of one-to-many uni-directional alignments may lead to a loss of information via the pivot language. However, the performances of the methods evaluated are dependent on a specific triad.

Generally, the pivot model performs less well than the direct model, but the interpolated method overscores both of them (on identical training conditions). However, for some language pairs the pivot method overscores the direct system, (i.e., Maltese-to-Finnish via English), where the complexity of the translation system Maltese-to-Finnish is better modellised by separating it into two models, Maltese-to-English and English-to-Maltese. In some sense, pivoting through English results in a nice factorization of the translation model: that probably has a positive impact in terms of less data sparseness in the training data and results in better statistical models. The experiments on Finnish and Estonian pivoting through English, also provides an evidence to this claim.

In summary, our experimental results have shown that triangulation is not a mere approximation of the source-target phrase table or the direct model, but that extracts additional useful translation information. We want to highlight the importance of the nature of the languages in a triad when using a pivot language.

8.2 Further directions

In our research, the advantages of the pivot-based models and their limits were investigated to define future lines of research.

We have emphasized the importance of the nature of the language in a triad when using a pivot method, therefore more experiments should be performed on other “low density” language pairs.

Since the research community is primarily occupied with translation into English, interesting problems associated with translation into morphologically rich languages have been neglected. We suggest fine-tuning of parameters and dedication processing for each language could improve results. That is a reason for using factored models, that allow for the introduction of linguistic pre-processing (such as lemmatisation) in a translation model.

Using factored translation models Instead of representing phrases only as sequences of words, it should be possible to introduce a more sophisticated representation for phrases. This is the idea of factored translation models, that include multiple levels of information. The advantages of factored representation are that models can employ more sophisticated linguistic information. As a result, they can draw generalisations from the training data and can generate better translations. This has the potential to lead to improved coverage, more grammatical output and better use of existing training data. The factored translation models are supported and implemented by `Moses.`

Tuning for quality A fine-tuning of parameters using MERT should enhance the performance of the baseline and pivot-based systems for certain language pairs. The tuning could emphasize some common features between two languages to optimize the translation output, and thus could change the “preferences” for a pivot language, given a source-target language pair.

Multi-pivot methods In terms of future work we consider extensions to our framework that lead to more powerful combination strategies using multiple bridge languages. We propose to study different weighting methods to combine or interpolate pivot-based models.

Application in terminology extraction A possible exploitation of the corpora that we have compiled could be to extract general and domain-specific terminology lists and to align these terminology lists across languages to produce multilingual term dictionaries. These resources could be used to link similar texts across languages and to offer cross-lingual glossing applications, i.e. to identify known terms in foreign language texts and to display these terms to the users in their own language. The pivot-based methods could be adapted, with the focus on precision, to these kind of applications.

Part V

Appendixes

Appendix A

Sample outputs of direct translation systems

The Appendix A presents results related to the chapter 5, on Translation models based on Acquis.

In the first two tables, we present an example extracted from the *Acquis Translation Units* sub-corpora, more precisely from the *Acquis Development Set* (Test Set), which represents the same sentence across the twenty-two languages (tables A.1 and A.2).

The next tables present a sample output of the different translation systems trained on *Translation-Units-22* corpus. The same reference sentence is translated into French, English and Romanian by our systems.

Thus, the tables A.3 and A.4 list the sentence translated into French from all the other 21 languages.

In the tables A.5 and A.6, we present the same sentence when translating into English from all the other languages.

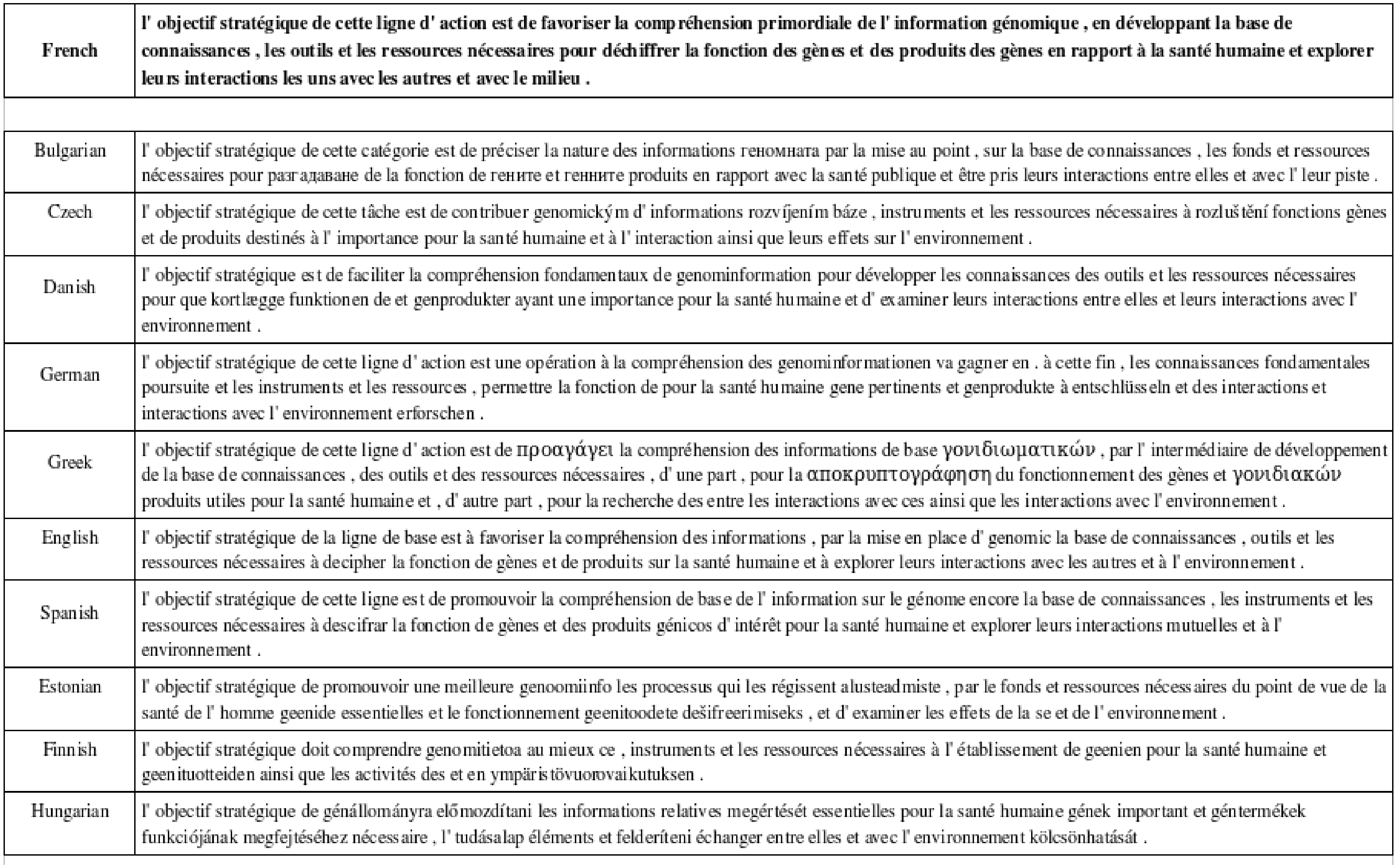
See tables A.7 and A.8 for the Romanian translations.

Bulgarian	стратегическата цел на тази насока е да се изясни същността на геномната информация чрез разработване на база научни познания , средства и ресурси , необходими за разгадаване на функцията на гените и генните продукти във връзка с човешкото здраве и да се проучат взаимодействията помежду им и с околната им среда .
Czech	strategickým cílem tohoto úkolu je podpořit základní porozumění genomickým informacím rozvíjením báze poznatků , nástrojů a zdrojů potřebných k rozluštění funkce genů a genových produktů , které mají význam pro lidské zdraví , a využít jejich vzájemné působení a rovněž účinky na prostředí .
Danish	det strategiske mål er her at fremme den grundlæggende forståelse af genominformation ved at udvikle den videnbase , de redskaber og de ressourcer , som er nødvendige for at kortlægge funktionen af gener og genprodukter , der har betydning for menneskers sundhed og at undersøge deres indbyrdes samspil og deres samspil med miljøet .
German	strategisches ziel dieser forschungsschwerpunkts ist es , ein grundlegendes verständnis der genominformationen zu gewinnen . zu diesem zweck sollen die grundlagenkenntnisse ausgebaut und instrumente und ressourcen geschaffen werden , die es ermöglichen , die funktion der für die menschliche gesundheit relevanten gene und genprodukte zu entschlüsseln und gegenseitige wechselwirkungen sowie wechselwirkungen mit der umwelt zu erforschen .
Greek	στρατηγικός στόχος αυτής της προτεραιότητας είναι να προαγάγει τη βασική κατανόηση των γονιδιωματικών πληροφοριών , μέσω της ανάπτυξης της γνωστικής βάσης , των εργαλείων και πόρων που είναι απαραίτητα , αφενός , για την αποκρυπτογράφηση της λειτουργίας των γονιδίων και γονιδιακών προϊόντων που είναι χρήσιμα για την ανθρώπινη υγεία και , αφετέρου , για τη διερεύνηση των μεταξύ τους αλληλεπιδράσεων καθώς και των αλληλεπιδράσεων αυτών με το περιβάλλον τους .
English	the strategic objective of this line is to foster the basic understanding of genomic information , by developing the knowledge base , tools and resources needed to decipher the function of genes and gene products relevant to human health and to explore their interactions with each other and with their environment .
Spanish	el objetivo estratégico de esta línea es fomentar la comprensión básica de la información sobre el genoma desarrollando la base de conocimientos , los instrumentos y los recursos necesarios para descifrar la función de los genes y de los productos génicos de interés para la salud humana y explorar sus interacciones mutuas y con su medio .
Estonian	selle tegevusliini strateegiline eesmärk on edendada genoomiinfo paremat mõistmist alusteadmiste , vahendite ja ressursside arendamise kaudu , mis on vajalikud inimese tervise seisukohast oluliste geenide ja geenitoodete toimimise dešifreerimiseks , ning uurida nende mõju üksteisele ja keskkonnale .
French	l' objectif stratégique de cette ligne d' action est de favoriser la compréhension primordiale de l' information génomique , en développant la base de connaissances , les outils et les ressources nécessaires pour déchiffrer la fonction des gènes et des produits des gènes en rapport à la santé humaine et explorer leurs interactions les uns avec les autres et avec le milieu .
Finnish	strategisena tavoitteena on ymmärtää genomitietoa paremmin kehittämällä tietämysperustaa , välineitä ja resursseja , jotka ovat tarpeen ihmisten terveyden kannalta merkityksellisten geenien ja geenituotteiden toiminnan selvittämiseksi sekä niiden keskinäisen ja ympäristövuorovaikutuksen tutkimiseksi .
Hungarian	ezen irányzat stratégiai célkitűzése előmozdítani a génállományra vonatkozó információk alapvető megértését az emberi egészség szempontjából fontos gének és géntermékek funkciójának megfejtéséhez szükséges tudásalap , eszközök és források fejlesztésével , és felderíteni azok egymás közötti és a környezettel való kölcsönhatását .

Table A.1: One paragraph aligned accross the twenty-two languages (part 1)

Italian	l'obiettivo strategico di questa linea d'azione è favorire la comprensione delle informazioni genomiche , sviluppando la base delle conoscenze , gli strumenti e le risorse necessarie per decifrare la funzione dei geni e dei prodotti genici in relazione alla salute umana ed esplorare le interazioni tra loro e con il loro ambiente .
Lithuanian	strateginis tikslas yra puoselėti fundamentalų genominės informacijos supratimą , plėtojant žinių bazę , priemones ir išteklius , kurių reikia genų ir genų produktų , svarbių žmogaus sveikatai , funkcijai iššifruoti bei ištirti jų tarpusavio sąveiką ir sąveiką su aplinka .
Latvian	šīs sadaļas stratēģiskais mērķis ir sekmēt pamatizpratni par genoma informāciju , izstrādājot pamatzināšanas , mehānismus un resursus , kas nepieciešami , lai atšifrētu gēnu funkcijas un noteiktu , kādi gēnu produkti saistīti ar cilvēka veselību , izpētītu to savstarpējo mijiedarbību un vidi , kas ir ap tiem .
Maltese	l- gan strateġiku ta ' din il- linja huwa t- tkattir tal- garfien bażiku ta ' l- informazzjoni ġenomika , permezz ta ' l- iżvilupp tal- bażi ta ' konoxxenza , l- istrumenti u r- riżorsi meħtieġa biex jiġi deċifrat il- funzjoni tal- ġeni u tal- prodotti tal- ġeni ta ' importanza għas- saħħa tal- bniedem u sabiex tiġu esplorata l- interazzjoni ta ' bejniethom u ma ' l- ambjent tagħhom .
Dutch	het strategische doel van dit type onderzoek is de vergroting van de fundamentele kennis van de informatie die in het genoom besloten ligt . dit doel wordt nagestreefd door de ontwikkeling van het kennisbestand , het instrumentarium en de middelen die nodig zijn om de functies te bepalen van genen en genproducten die voor de menselijke gezondheid relevant kunnen zijn en om de onderlinge relaties en de interacties met het milieu daarvan te onderzoeken .
Polish	strategicznym celem w tym obszarze jest osiągnięcie podstawowej zdolności zrozumienia informacji genomicznej poprzez rozwój bazy , narzędzi i źródeł wiedzy potrzebnych do rozszyfrowania funkcji genów i produktów genów mających znaczenie dla ludzkiego zdrowia i do zbadania ich wzajemnego oddziaływania i oddziaływania między nimi a środowiskiem , w którym występują .
Portuguese	o objectivo estratégico desta linha é promover o conhecimento básico da informação genómica , desenvolvendo a base de conhecimentos , as ferramentas e os recursos necessários para decifrar as funções dos genes e produtos de genes relevantes para a saúde humana e para explorar as interacções entre estes e com o respectivo ambiente .
Romanian	obiectivul strategic al acestei componente este de a facilita înțelegerea informațiilor genomice , prin dezvoltarea unei baze de cunoștințe , a instrumentelor și resurselor necesare pentru a descifra funcția genelor și a produselor genelor care au relevanță pentru sănătatea umană și pentru a investiga interacțiunile acestora între ele și cu mediul .
Slovakian	strategickým cieľom tohoto smeru je podporovať základné porozumenie genomických informácii rozvíjaním vedomostnej základne , nástrojov a zdrojov potrebných na dešifrovanie funkcie génov a génových produktov , ktoré majú význam pre ľudské zdravie a skúmanie ich vzájomného pôsobenia medzi sebou a so svojím prostredím .
Slovene	strateški cilj tega področja je krepitev osnovnega razumevanja genomskih informacij s pomočjo razvijanja baze znanja , orodij in sredstev , potrebnih za določanje funkcije genov in genskih produktov , ki se nanašajo na zdravje ljudi , ter raziskovanje njihovih medsebojnih interakcij ter interakcij z okoljem .
Swedish	det strategiska målet är att bidra till den grundläggande kunskapen om genom-informationen genom att utveckla den kunskapsbas , de verktyg och de resurser som krävs för att utforska funktionen hos gener och genprodukter som är av betydelse för människors och undersöka deras inbördes samverkan och deras samspel med omgivningen .

Table A.2: One paragraph aligned accross the twenty-two languages (part 2)



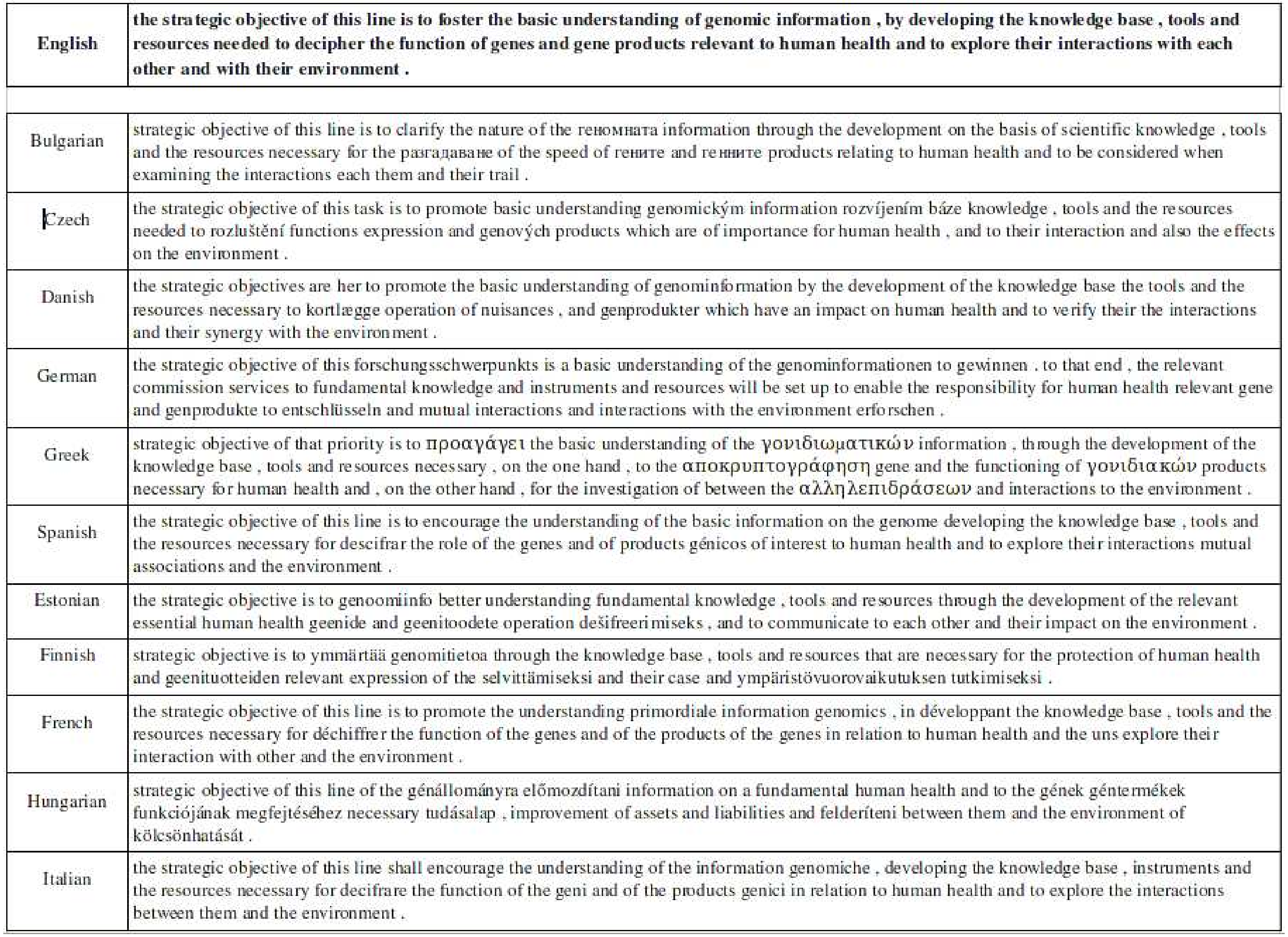
French	**l' objectif stratégique de cette ligne d' action est de favoriser la compréhension primordiale de l' information génomique , en développant la base de connaissances , les outils et les ressources nécessaires pour déchiffrer la fonction des gènes et des produits des gènes en rapport à la santé humaine et explorer leurs interactions les uns avec les autres et avec le milieu .**
Bulgarian	l' objectif stratégique de cette catégorie est de préciser la nature des informations геномната par la mise au point , sur la base de connaissances , les fonds et ressources nécessaires pour разгадаване de la fonction de гените et генните produits en rapport avec la santé publique et être pris leurs interactions entre elles et avec l' leur piste .
Czech	l' objectif stratégique de cette tâche est de contribuer genomickým d' informations rozvíjením báze , instruments et les ressources nécessaires à rozluštění fonctions gènes et de produits destinés à l' importance pour la santé humaine et à l' interaction ainsi que leurs effets sur l' environnement .
Danish	l' objectif stratégique est de faciliter la compréhension fondamentaux de genominformation pour développer les connaissances des outils et les ressources nécessaires pour que kortlægge funktionen de et genprodukter ayant une importance pour la santé humaine et d' examiner leurs interactions entre elles et leurs interactions avec l' environnement .
German	l' objectif stratégique de cette ligne d' action est une opération à la compréhension des genominformationen va gagner en . à cette fin , les connaissances fondamentales poursuite et les instruments et les ressources , permettre la fonction de pour la santé humaine gene pertinents et genprodukte à entschlüsseln et des interactions et interactions avec l' environnement erforschen .
Greek	l' objectif stratégique de cette ligne d' action est de προαγάγει la compréhension des informations de base γονιδιωματικών , par l' intermédiaire de développement de la base de connaissances , des outils et des ressources nécessaires , d' une part , pour la αποκρυπτογράφηση du fonctionnement des gènes et γονιδιακών produits utiles pour la santé humaine et , d' autre part , pour la recherche des entre les interactions avec ces ainsi que les interactions avec l' environnement .
English	l' objectif stratégique de la ligne de base est à favoriser la compréhension des informations , par la mise en place d' genomic la base de connaissances , outils et les ressources nécessaires à decipher la fonction de gènes et de produits sur la santé humaine et à explorer leurs interactions avec les autres et à l' environnement .
Spanish	l' objectif stratégique de cette ligne est de promouvoir la compréhension de base de l' information sur le génome encore la base de connaissances , les instruments et les ressources nécessaires à descifrar la fonction de gènes et des produits génicos d' intérêt pour la santé humaine et explorer leurs interactions mutuelles et à l' environnement .
Estonian	l' objectif stratégique de promouvoir une meilleure genoomiinfo les processus qui les régissent alusteadmiste , par le fonds et ressources nécessaires du point de vue de la santé de l' homme geenide essentielles et le fonctionnement geenitoodete dešifreerimiseks , et d' examiner les effets de la se et de l' environnement .
Finnish	l' objectif stratégique doit comprendre genomitietoa au mieux ce , instruments et les ressources nécessaires à l' établissement de geenien pour la santé humaine et geenituotteiden ainsi que les activités des et en ympäristövuorovaikutuksen .
Hungarian	l' objectif stratégique de génállományra előmozdítani les informations relatives megértését essentielles pour la santé humaine gének important et géntermékek funkciójának megfejtéséhez nécessaire , l' tudásalap éléments et felderíteni échanger entre elles et avec l' environnement kölcsönhatását .

Table A.3: Sample output of the translation systems “LG-fr” trained on *Translation-Units-22* corpus (part 1)

French	**l' objectif stratégique de cette ligne d' action est de favoriser la compréhension primordiale de l' information génomique , en développant la base de connaissances , les outils et les ressources nécessaires pour déchiffrer la fonction des gènes et des produits des gènes en rapport à la santé humaine et explorer leurs interactions les uns avec les autres et avec le milieu .**

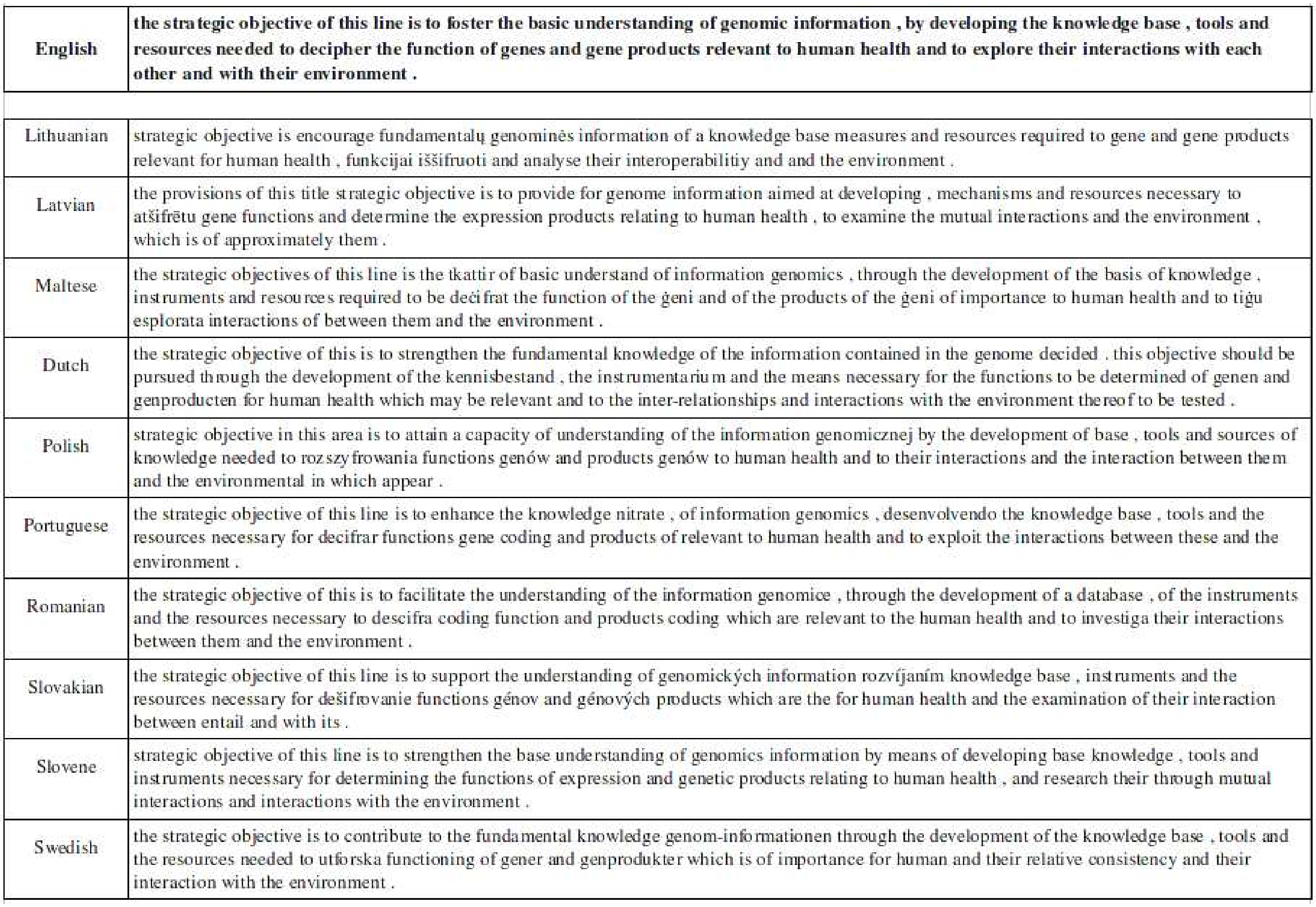
Italian	l' objectif stratégique de cette ligne d' action est de promouvoir la compréhension des informations génomique , mettant la base de connaissances , les instruments et les ressources nécessaires pour decifrare la fonction des gènes et des produits genici par rapport à la santé humaine et explorer les interactions entre elles et avec l' environnement .
Lithuanian	objectif global stratégique est puoselėti fundamentalų genominės d' information , la base de connaissances , et des mesures pour genų et genų des produits et la santé humaine , destiné iššifruoti et examine leur et l' interopérabilité avec l' environnement .
Latvian	l' objectif stratégique de cette est de contribuer à fournir des informations sur les connaissances fondamentales génomique , mettre en place des mécanismes et les ressources nécessaires pour atšifrētu sur des fonctions et qui , sur des produits liées à la santé de l' europe , leur mutuelle et de l' environnement , qui est d' environ les .
Maltese	les objectifs stratégique de cette ligne est de renforcer le domaine de sensibles de base de données génomique , par l' intermédiaire de l' évolution de la base de connaissances , les instruments et les ressources nécessaires à l' decifrat la fonction de gènes et des produits des gènes d' importance pour la santé humaine et pour tigu esplorata interactions entre elles et avec l' environnement .
Dutch	l' objectif stratégique de ce type est de l' augmentation des connaissances fondamentales des informations qui , dans le genoom décidé . cet objectif est poursuivis par le développement de la kennisbestand , le instrumentarium et les moyens nécessaires pour les fonctions à déterminer de gènes et genproducten pour la santé humaine peuvent et pour les interrelations et interactions avec l' environnement à des enquêtes .
Polish	l' objectif stratégique dans ce domaine sont de la réalisation de la compréhension des informations genomicznej par le développement de base , outils et de sources de connaissances nécessaires à rozszyfrowania fonctions gènes et des gènes ayant une importance pour la santé humaine et à examiner leurs interactions et l' interaction entre elles et , dans lequel figurent .
Portuguese	l' objectif stratégique de cette ligne est d' améliorer la connaissance basique de l' information génomique , débouchant sur la base de connaissances , les outils et les ressources nécessaires pour decifrar les fonctions des gènes et produits de gènes pertinents pour la santé humaine et à exploiter les interactions entre ceux-ci et à leur environnement .
Romanian	l' objectif stratégique de cette ligne est de faciliter la compréhension des informations genomice , par la mise en place d' une base de connaissances , des instruments et des ressources nécessaires pour descifra la fonction codant et des produits codant intéressant pour la santé humaine et pour investiga leurs interactions entre elles et avec l' environnement .
Slovakian	l' objectif stratégique de cette action est de soutenir les informations porozumenie genomických rozvíjaním vedomostnej gagneraient , instruments et les ressources nécessaires à dešifrovanie fonctions génov et génových de produits , pour la santé humaine et l' examen de leur les interactions entre les entraîné une et de l' environnement .
Slovene	l' objectif stratégique de ce domaine est la compréhension des informations genomskih conformément au point de bases de données des connaissances , des outils et des ressources nécessaires à l' audit des gènes et génétiques dégradation sur la santé humaine et de la recherche de leurs medsebojnih étude et interactions avec l' environnement .
Swedish	l' objectif stratégique est de contribuer à la formation de genom-informationen par la mise au point d' une , les outils et les ressources nécessaires à la utforska fonction des gener et genprodukter à la cohérence et à leur entre elles et la synergie avec son environnement ;

Table A.4: Sample output of the translation systems "LG-fr" trained on *Translation-Units-22* corpus (part 2)



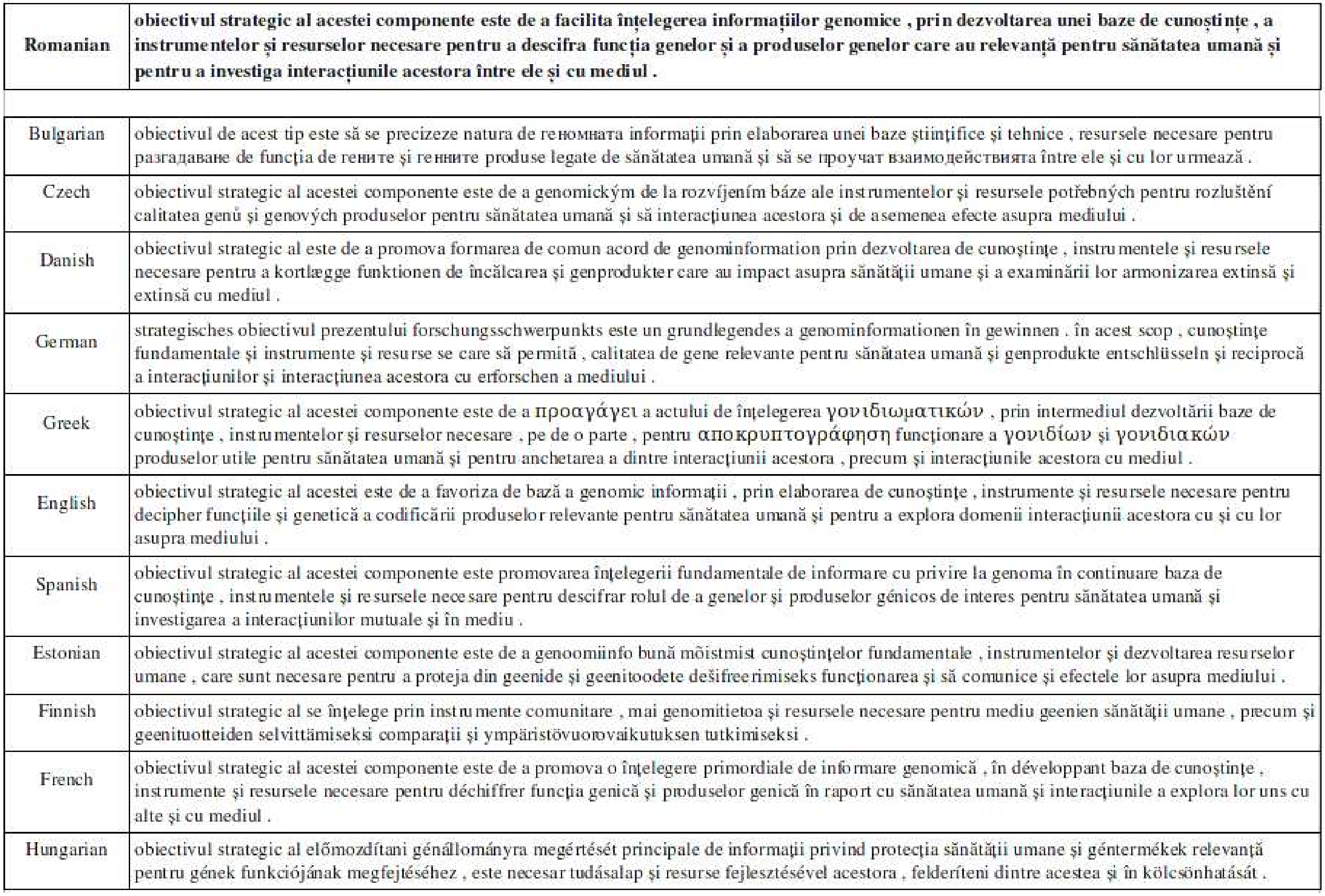
English	**the strategic objective of this line is to foster the basic understanding of genomic information , by developing the knowledge base , tools and resources needed to decipher the function of genes and gene products relevant to human health and to explore their interactions with each other and with their environment .**
Bulgarian	strategic objective of this line is to clarify the nature of the реномната information through the development on the basis of scientific knowledge , tools and the resources necessary for the разгадаване of the speed of рените and ренните products relating to human health and to be considered when examining the interactions each them and their trail .
Czech	the strategic objective of this task is to promote basic understanding genomickým information rozvíjením báze knowledge , tools and the resources needed to rozluštění functions expression and genových products which are of importance for human health , and to their interaction and also the effects on the environment .
Danish	the strategic objectives are her to promote the basic understanding of genominformation by the development of the knowledge base the tools and the resources necessary to kortlægge operation of nuisances , and genprodukter which have an impact on human health and to verify their the interactions and their synergy with the environment .
German	the strategic objective of this forschungsschwerpunkts is a basic understanding of the genominformationen to gewinnen . to that end , the relevant commission services to fundamental knowledge and instruments and resources will be set up to enable the responsibility for human health relevant gene and genprodukte to entschlüsseln and mutual interactions and interactions with the environment erforschen .
Greek	strategic objective of that priority is to προαγάγει the basic understanding of the γονιδιωματικών information , through the development of the knowledge base , tools and resources necessary , on the one hand , to the αποκρυπτογράφηση gene and the functioning of γονιδιακών products necessary for human health and , on the other hand , for the investigation of between the αλληλεπιδράσεων and interactions to the environment .
Spanish	the strategic objective of this line is to encourage the understanding of the basic information on the genome developing the knowledge base , tools and the resources necessary for descifrar the role of the genes and of products génicos of interest to human health and to explore their interactions mutual associations and the environment .
Estonian	the strategic objective is to genoomiinfo better understanding fundamental knowledge , tools and resources through the development of the relevant essential human health geenide and geenitoodete operation dešifreerimiseks , and to communicate to each other and their impact on the environment .
Finnish	strategic objective is to ymmärtää genomitietoa through the knowledge base , tools and resources that are necessary for the protection of human health and geenituotteiden relevant expression of the selvittämiseksi and their case and ympäristövuorovaikutuksen tutkimiseksi .
French	the strategic objective of this line is to promote the understanding primordiale information genomics , in développant the knowledge base , tools and the resources necessary for déchiffrer the function of the genes and of the products of the genes in relation to human health and the uns explore their interaction with other and the environment .
Hungarian	strategic objective of this line of the génállományra előmozdítani information on a fundamental human health and to the gének géntermékek funkciójának megfejtéséhez necessary tudásalap , improvement of assets and liabilities and felderíteni between them and the environment of kölcsönhatását .
Italian	the strategic objective of this line shall encourage the understanding of the information genomiche , developing the knowledge base , instruments and the resources necessary for decifrare the function of the geni and of the products genici in relation to human health and to explore the interactions between them and the environment .

Table A.5: Sample output of the translation systems "LG-en" trained on *Translation-Units-22* corpus (part 1)



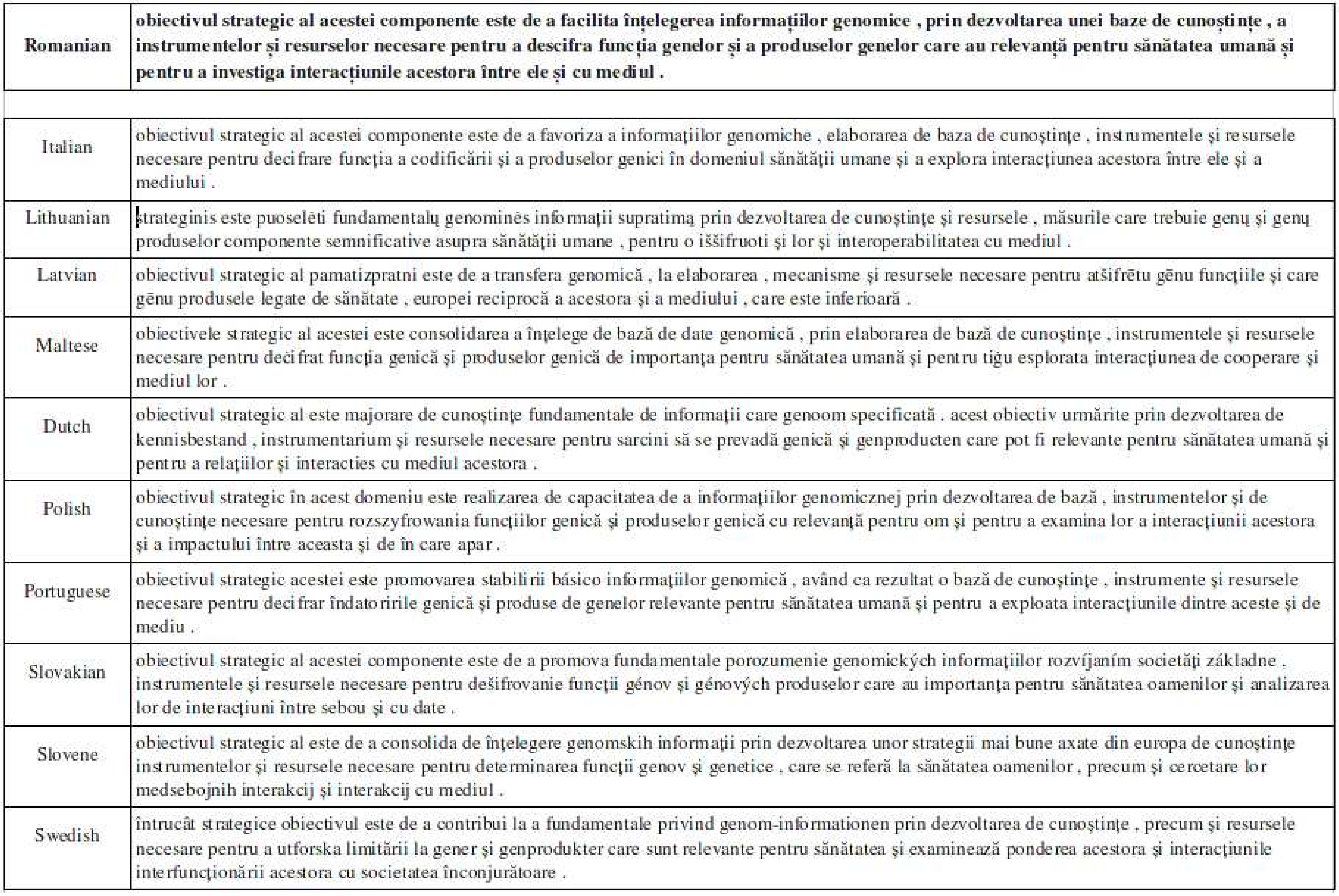
English	**the strategic objective of this line is to foster the basic understanding of genomic information , by developing the knowledge base , tools and resources needed to decipher the function of genes and gene products relevant to human health and to explore their interactions with each other and with their environment .**
Lithuanian	strategic objective is encourage fundamentalų genominės information of a knowledge base measures and resources required to gene and gene products relevant for human health , funkcijai iššifruoti and analyse their interoperabilitiy and and the environment .
Latvian	the provisions of this title strategic objective is to provide for genome information aimed at developing , mechanisms and resources necessary to atšifrētu gene functions and determine the expression products relating to human health , to examine the mutual interactions and the environment , which is of approximately them .
Maltese	the strategic objectives of this line is the tkattir of basic understand of information genomics , through the development of the basis of knowledge , instruments and resources required to be deċifrat the function of the ġeni and of the products of the ġeni of importance to human health and to tiġu esplorata interactions of between them and the environment .
Dutch	the strategic objective of this is to strengthen the fundamental knowledge of the information contained in the genome decided . this objective should be pursued through the development of the kennisbestand , the instrumentarium and the means necessary for the functions to be determined of genen and genproducten for human health which may be relevant and to the inter-relationships and interactions with the environment thereof to be tested .
Polish	strategic objective in this area is to attain a capacity of understanding of the information genomicznej by the development of base , tools and sources of knowledge needed to rozszyfrowania functions genów and products genów to human health and to their interactions and the interaction between them and the environmental in which appear .
Portuguese	the strategic objective of this line is to enhance the knowledge nitrate , of information genomics , desenvolvendo the knowledge base , tools and the resources necessary for decifrar functions gene coding and products of relevant to human health and to exploit the interactions between these and the environment .
Romanian	the strategic objective of this is to facilitate the understanding of the information genomice , through the development of a database , of the instruments and the resources necessary to descifra coding function and products coding which are relevant to the human health and to investiga their interactions between them and the environment .
Slovakian	the strategic objective of this line is to support the understanding of genomických information rozvíjaním knowledge base , instruments and the resources necessary for dešifrovanie functions génov and génových products which are the for human health and the examination of their interaction between entail and with its .
Slovene	strategic objective of this line is to strengthen the base understanding of genomics information by means of developing base knowledge , tools and instruments necessary for determining the functions of expression and genetic products relating to human health , and research their through mutual interactions and interactions with the environment .
Swedish	the strategic objective is to contribute to the fundamental knowledge genom-informationen through the development of the knowledge base , tools and the resources needed to utforska functioning of gener and genprodukter which is of importance for human and their relative consistency and their interaction with the environment .

Table A.6: Sample output of the translation systems “LG-en” trained on *Translation-Units-22* corpus (part 2)



Romanian	**obiectivul strategic al acestei componente este de a facilita înțelegerea informațiilor genomice , prin dezvoltarea unei baze de cunoștințe , a instrumentelor și resurselor necesare pentru a descifra funcția genelor și a produselor genelor care au relevanță pentru sănătatea umană și pentru a investiga interacțiunile acestora între ele și cu mediul .**
Bulgarian	obiectivul de acest tip este să se precizeze natura de геномната informaţii prin elaborarea unei baze ştiinţifice şi tehnice , resursele necesare pentru разгадаване de funcţia de гените şi генните produse legate de sănătatea umană şi să se проучат взаимодействията între ele şi cu lor urmează .
Czech	obiectivul strategic al acestei componente este de a genomickým de la rozvíjením báze ale instrumentelor şi resursele potřebných pentru rozluštění calitatea genů şi genových produselor pentru sănătatea umană şi să interacţiunea acestora şi de asemenea efecte asupra mediului .
Danish	obiectivul strategic al este de a promova formarea de comun acord de genominformation prin dezvoltarea de cunoştinţe , instrumentele şi resursele necesare pentru a kortlægge funktionen de încălcarea şi genprodukter care au impact asupra sănătăţii umane şi a examinării lor armonizarea extinsă şi extinsă cu mediul .
German	strategisches obiectivul prezentului forschungsschwerpunkts este un grundlegendes a genominformationen în gewinnen . în acest scop , cunoştinţe fundamentale şi instrumente şi resurse se care să permită , calitatea de gene relevante pentru sănătatea umană şi genprodukte entschlüsseln şi reciprocă a interacţiunilor şi interacţiunea acestora cu erforschen a mediului .
Greek	obiectivul strategic al acestei componente este de a προαγάγει a actului de înţelegerea γονιδιωματικών , prin intermediul dezvoltării baze de cunoştinţe , instrumentelor şi resurselor necesare , pe de o parte , pentru αποκρυπτογράφηση funcţionare a γονιδίων şi γονιδιακών produselor utile pentru sănătatea umană şi pentru anchetarea a dintre interacţiunii acestora , precum şi interacţiunile acestora cu mediul .
English	obiectivul strategic al acestei este de a favoriza de bază a genomic informaţii , prin elaborarea de cunoştinţe , instrumente şi resursele necesare pentru decipher funcţiile şi genetică a codificării produselor relevante pentru sănătatea umană şi pentru a explora domenii interacţiunii acestora cu şi cu lor asupra mediului .
Spanish	obiectivul strategic al acestei componente este promovarea înţelegerii fundamentale de informare cu privire la genoma în continuare baza de cunoştinţe , instrumentele şi resursele necesare pentru descifrar rolul de a genelor şi produselor génicos de interes pentru sănătatea umană şi investigarea a interacţiunilor mutuale şi în mediu .
Estonian	obiectivul strategic al acestei componente este de a genoomiinfo bună mõistmist cunoştinţelor fundamentale , instrumentelor şi dezvoltarea resurselor umane , care sunt necesare pentru a proteja din geenide şi geenitoodete dešifreerimiseks funcţionarea şi să comunice şi efectele lor asupra mediului .
Finnish	obiectivul strategic al se înţelege prin instrumente comunitare , mai genomitietoa şi resursele necesare pentru mediu geenien sănătăţii umane , precum şi geenituotteiden selvittämiseksi comparaţii şi ympäristövuorovaikutuksen tutkimiseksi .
French	obiectivul strategic al acestei componente este de a promova o înţelegere primordiale de informare genomică , în développant baza de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer funcţia genică şi produselor genică în raport cu sănătatea umană şi interacţiunile a explora lor uns cu alte şi cu mediul .
Hungarian	obiectivul strategic al előmozdítani génállományra megértését principale de informaţii privind protecţia sănătăţii umane şi géntermékek relevanţă pentru gének funkciójának megfejtéséhez , este necesar tudásalap şi resurse fejlesztésével acestora , felderíteni dintre acestea şi în kölcsönhatását .

Table A.7: Sample output of the translation systems “LG-ro” trained on *Translation-Units-22* corpus (part 1)



Romanian	**obiectivul strategic al acestei componente este de a facilita înțelegerea informațiilor genomice , prin dezvoltarea unei baze de cunoștințe , a instrumentelor și resurselor necesare pentru a descifra funcția genelor și a produselor genelor care au relevanță pentru sănătatea umană și pentru a investiga interacțiunile acestora între ele și cu mediul .**
Italian	obiectivul strategic al acestei componente este de a favoriza a informaţiilor genomiche , elaborarea de baza de cunoştinţe , instrumentele şi resursele necesare pentru decifrare funcţia a codificării şi a produselor genici în domeniul sănătăţii umane şi a explora interacţiunea acestora între ele şi a mediului .
Lithuanian	ştrateginis este puoselēti fundamentalų genominės informaţii supratimą prin dezvoltarea de cunoştinţe şi resursele , măsurile care trebuie genų şi genų produselor componente semnificative asupra sănătăţii umane , pentru o iššifruoti şi lor şi interoperabilitatea cu mediul .
Latvian	obiectivul strategic al pamatizpratni este de a transfera genomică , la elaborarea , mecanisme şi resursele necesare pentru atšifrētu gēnu funcţiile şi care gēnu produsele legate de sănătate , europei reciprocă a acestora şi a mediului , care este inferioară .
Maltese	obiectivele strategic al acestei este consolidarea a înţelege de bază de date genomică , prin elaborarea de bază de cunoştinţe , instrumentele şi resursele necesare pentru decifrat funcţia genică şi produselor genică de importanţa pentru sănătatea umană şi pentru tiġu esplorata interacţiunea de cooperare şi mediul lor .
Dutch	obiectivul strategic al este majorare de cunoştinţe fundamentale de informaţii care genoom specificată . acest obiectiv urmărite prin dezvoltarea de kennisbestand , instrumentarium şi resursele necesare pentru sarcini să se prevadă genică şi genproducten care pot fi relevante pentru sănătatea umană şi pentru a relaţiilor şi interacties cu mediul acestora .
Polish	obiectivul strategic în acest domeniu este realizarea de capacitatea de a informaţiilor genomicznej prin dezvoltarea de bază , instrumentelor şi de cunoştinţe necesare pentru rozszyfrowania funcţiilor genică şi produselor genică cu relevanţă pentru om şi pentru a examina lor a interacţiunii acestora şi a impactului între aceasta şi de în care apar .
Portuguese	obiectivul strategic acestei este promovarea stabilirii básico informaţiilor genomică , având ca rezultat o bază de cunoştinţe , instrumente şi resursele necesare pentru decifrar îndatoririle genică şi produse de genelor relevante pentru sănătatea umană şi pentru a exploata interacţiunile dintre aceste şi de mediu .
Slovakian	obiectivul strategic al acestei componente este de a promova fundamentale porozumenie genomických informaţiilor rozvíjaním societăţi základne , instrumentele şi resursele necesare pentru dešifrovanie funcţii génov şi génových produselor care au importanţa pentru sănătatea oamenilor şi analizarea lor de interacţiuni între sebou şi cu date .
Slovene	obiectivul strategic al este de a consolida de înţelegere genomskih informaţii prin dezvoltarea unor strategii mai bune axate din europa de cunoştinţe instrumentelor şi resursele necesare pentru determinarea funcţii genov şi genetice , care se referă la sănătatea oamenilor , precum şi cercetare lor medsebojnih interakcij şi interakcij cu mediul .
Swedish	întrucât strategice obiectivul este de a contribui la a fundamentale privind genom-informationen prin dezvoltarea de cunoştinţe , precum şi resursele necesare pentru a utforska limitării la gener şi genprodukter care sunt relevante pentru sănătatea şi examinează ponderea acestora şi interacţiunile interfuncţionării acestora cu societatea înconjurătoare .

Table A.8: Sample output of the translation systems “LG-ro” trained on *Translation-Units-22* corpus (part 2)

Appendix B

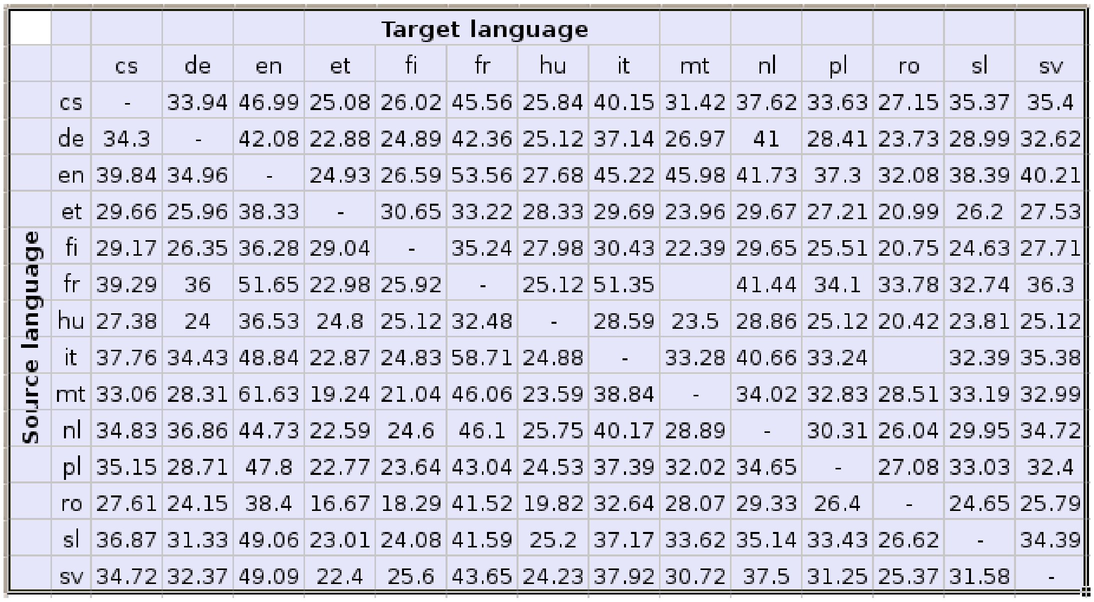
Evaluation tables of direct translation systems trained on Acquis

This appendix is related to chapter 5 and presents the evaluation of our transaltion models, trained on different data sets.

The table B.1 shows the performance of the systems trained on the *Acquis-22* corpus (around 360k sentences per language), measured in BLEU score %.

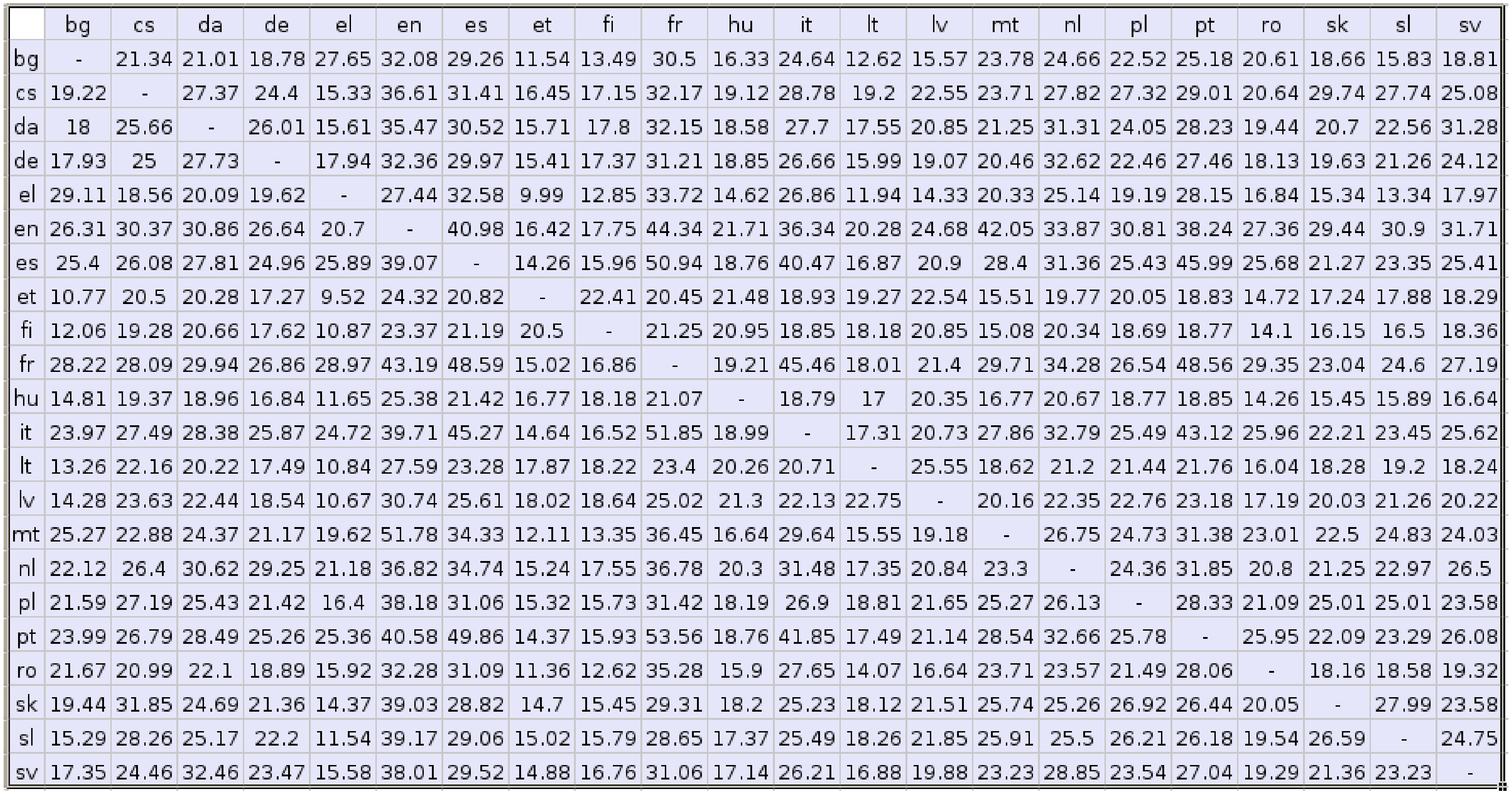
The systems presented in B.2 have been trained on a sample of *Acquis-22*, sized of 10 000 sentences (*Acquis-22-sample10k*), randomly generated for each language.

Table B.3 and table B.4 present the correlation values between the BLEU score vectors “INTO” and “FROM” of the twenty-two European languages. The systems have been trained on *Translation-Units-22* corpus and their performance in BLEU score % is shown in chapter 5, table 5.1.



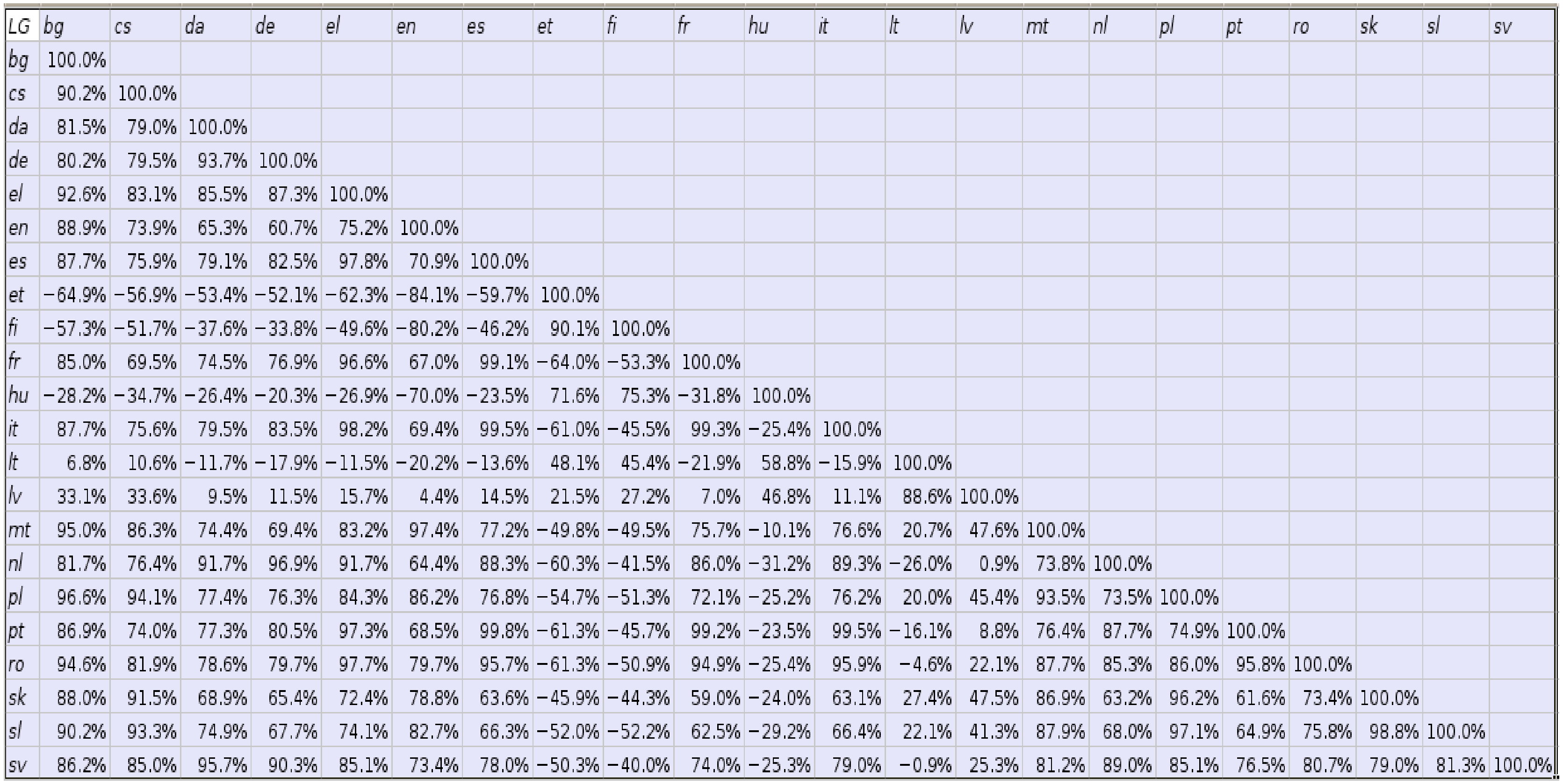
		Target language													
		cs	de	en	et	fi	fr	hu	it	mt	nl	pl	ro	sl	sv
	cs	-	33.94	46.99	25.08	26.02	45.56	25.84	40.15	31.42	37.62	33.63	27.15	35.37	35.4
	de	34.3	-	42.08	22.88	24.89	42.36	25.12	37.14	26.97	41	28.41	23.73	28.99	32.62
	en	39.84	34.96	-	24.93	26.59	53.56	27.68	45.22	45.98	41.73	37.3	32.08	38.39	40.21
	et	29.66	25.96	38.33	-	30.65	33.22	28.33	29.69	23.96	29.67	27.21	20.99	26.2	27.53
Source language	fi	29.17	26.35	36.28	29.04	-	35.24	27.98	30.43	22.39	29.65	25.51	20.75	24.63	27.71
	fr	39.29	36	51.65	22.98	25.92	-	25.12	51.35		41.44	34.1	33.78	32.74	36.3
	hu	27.38	24	36.53	24.8	25.12	32.48	-	28.59	23.5	28.86	25.12	20.42	23.81	25.12
	it	37.76	34.43	48.84	22.87	24.83	58.71	24.88	-	33.28	40.66	33.24		32.39	35.38
	mt	33.06	28.31	61.63	19.24	21.04	46.06	23.59	38.84	-	34.02	32.83	28.51	33.19	32.99
	nl	34.83	36.86	44.73	22.59	24.6	46.1	25.75	40.17	28.89	-	30.31	26.04	29.95	34.72
	pl	35.15	28.71	47.8	22.77	23.64	43.04	24.53	37.39	32.02	34.65	-	27.08	33.03	32.4
	ro	27.61	24.15	38.4	16.67	18.29	41.52	19.82	32.64	28.07	29.33	26.4	-	24.65	25.79
	sl	36.87	31.33	49.06	23.01	24.08	41.59	25.2	37.17	33.62	35.14	33.43	26.62	-	34.39
	sv	34.72	32.37	49.09	22.4	25.6	43.65	24.23	37.92	30.72	37.5	31.25	25.37	31.58	-

Table B.1: BLEU scores for the translation systems trained on *Acquis-22* corpus



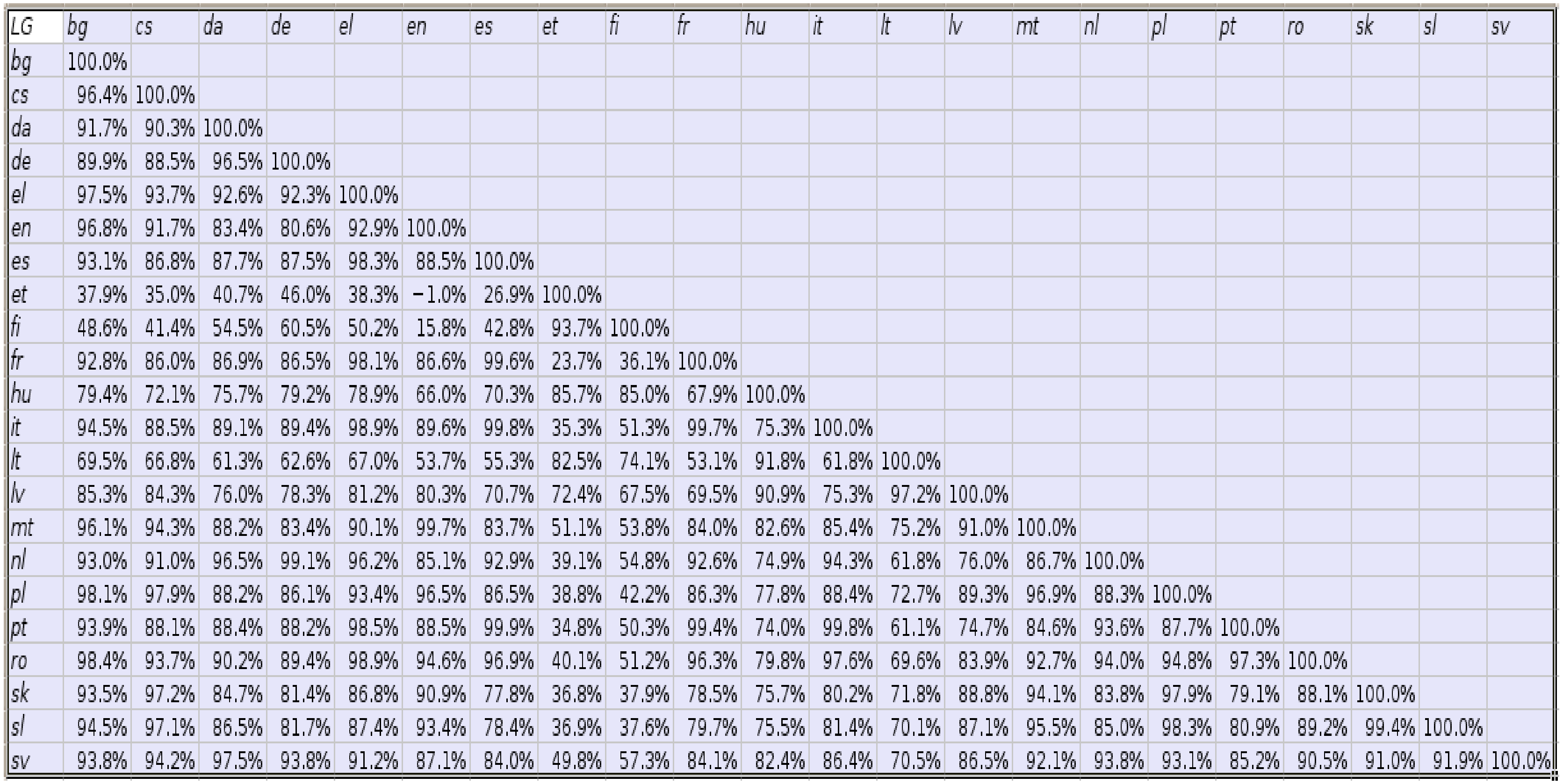
	bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	-	21.34	21.01	18.78	27.65	32.08	29.26	11.54	13.49	30.5	16.33	24.64	12.62	15.57	23.78	24.66	22.52	25.18	20.61	18.66	15.83	18.81
cs	19.22	-	27.37	24.4	15.33	36.61	31.41	16.45	17.15	32.17	19.12	28.78	19.2	22.55	23.71	27.82	27.32	29.01	20.64	29.74	27.74	25.08
da	18	25.66	-	26.01	15.61	35.47	30.52	15.71	17.8	32.15	18.58	27.7	17.55	20.85	21.25	31.31	24.05	28.23	19.44	20.7	22.56	31.28
de	17.93	25	27.73	-	17.94	32.36	29.97	15.41	17.37	31.21	18.85	26.66	15.99	19.07	20.46	32.62	22.46	27.46	18.13	19.63	21.26	24.12
el	29.11	18.56	20.09	19.62	-	27.44	32.58	9.99	12.85	33.72	14.62	26.86	11.94	14.33	20.33	25.14	19.19	28.15	16.84	15.34	13.34	17.97
en	26.31	30.37	30.86	26.64	20.7	-	40.98	16.42	17.75	44.34	21.71	36.34	20.28	24.68	42.05	33.87	30.81	38.24	27.36	29.44	30.9	31.71
es	25.4	26.08	27.81	24.96	25.89	39.07	-	14.26	15.96	50.94	18.76	40.47	16.87	20.9	28.4	31.36	25.43	45.99	25.68	21.27	23.35	25.41
et	10.77	20.5	20.28	17.27	9.52	24.32	20.82	-	22.41	20.45	21.48	18.93	19.27	22.54	15.51	19.77	20.05	18.83	14.72	17.24	17.88	18.29
fi	12.06	19.28	20.66	17.62	10.87	23.37	21.19	20.5	-	21.25	20.95	18.85	18.18	20.85	15.08	20.34	18.69	18.77	14.1	16.15	16.5	18.36
fr	28.22	28.09	29.94	26.86	28.97	43.19	48.59	15.02	16.86	-	19.21	45.46	18.01	21.4	29.71	34.28	26.54	48.56	29.35	23.04	24.6	27.19
hu	14.81	19.37	18.96	16.84	11.65	25.38	21.42	16.77	18.18	21.07	-	18.79	17	20.35	16.77	20.67	18.77	18.85	14.26	15.45	15.89	16.64
it	23.97	27.49	28.38	25.87	24.72	39.71	45.27	14.64	16.52	51.85	18.99	-	17.31	20.73	27.86	32.79	25.49	43.12	25.96	22.21	23.45	25.62
lt	13.26	22.16	20.22	17.49	10.84	27.59	23.28	17.87	18.22	23.4	20.26	20.71	-	25.55	18.62	21.2	21.44	21.76	16.04	18.28	19.2	18.24
lv	14.28	23.63	22.44	18.54	10.67	30.74	25.61	18.02	18.64	25.02	21.3	22.13	22.75	-	20.16	22.35	22.76	23.18	17.19	20.03	21.26	20.22
mt	25.27	22.88	24.37	21.17	19.62	51.78	34.33	12.11	13.35	36.45	16.64	29.64	15.55	19.18	-	26.75	24.73	31.38	23.01	22.5	24.83	24.03
nl	22.12	26.4	30.62	29.25	21.18	36.82	34.74	15.24	17.55	36.78	20.3	31.48	17.35	20.84	23.3	-	24.36	31.85	20.8	21.25	22.97	26.5
pl	21.59	27.19	25.43	21.42	16.4	38.18	31.06	15.32	15.73	31.42	18.19	26.9	18.81	21.65	25.27	26.13	-	28.33	21.09	25.01	25.01	23.58
pt	23.99	26.79	28.49	25.26	25.36	40.58	49.86	14.37	15.93	53.56	18.76	41.85	17.49	21.14	28.54	32.66	25.78	-	25.95	22.09	23.29	26.08
ro	21.67	20.99	22.1	18.89	15.92	32.28	31.09	11.36	12.62	35.28	15.9	27.65	14.07	16.64	23.71	23.57	21.49	28.06	-	18.16	18.58	19.32
sk	19.44	31.85	24.69	21.36	14.37	39.03	28.82	14.7	15.45	29.31	18.2	25.23	18.12	21.51	25.74	25.26	26.92	26.44	20.05	-	27.99	23.58
sl	15.29	28.26	25.17	22.2	11.54	39.17	29.06	15.02	15.79	28.65	17.37	25.49	18.26	21.85	25.91	25.5	26.21	26.18	19.54	26.59	-	24.75
sv	17.35	24.46	32.46	23.47	15.58	38.01	29.52	14.88	16.76	31.06	17.14	26.21	16.88	19.88	23.23	28.85	23.54	27.04	19.29	21.36	23.23	-

Table B.2: BLEU scores for the translation models trained on *Acquis-22-sample10k*



LG	bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	100.0%																					
cs	90.2%	100.0%																				
da	81.5%	79.0%	100.0%																			
de	80.2%	79.5%	93.7%	100.0%																		
el	92.6%	83.1%	85.5%	87.3%	100.0%																	
en	88.9%	73.9%	65.3%	60.7%	75.2%	100.0%																
es	87.7%	75.9%	79.1%	82.5%	97.8%	70.9%	100.0%															
et	−64.9%	−56.9%	−53.4%	−52.1%	−62.3%	−84.1%	−59.7%	100.0%														
fi	−57.3%	−51.7%	−37.6%	−33.8%	−49.6%	−80.2%	−46.2%	90.1%	100.0%													
fr	85.0%	69.5%	74.5%	76.9%	96.6%	67.0%	99.1%	−64.0%	−53.3%	100.0%												
hu	−28.2%	−34.7%	−26.4%	−20.3%	−26.9%	−70.0%	−23.5%	71.6%	75.3%	−31.8%	100.0%											
it	87.7%	75.6%	79.5%	83.5%	98.2%	69.4%	99.5%	−61.0%	−45.5%	99.3%	−25.4%	100.0%										
lt	6.8%	10.6%	−11.7%	−17.9%	−11.5%	−20.2%	−13.6%	48.1%	45.4%	−21.9%	58.8%	−15.9%	100.0%									
lv	33.1%	33.6%	9.5%	11.5%	15.7%	4.4%	14.5%	21.5%	27.2%	7.0%	46.8%	11.1%	88.6%	100.0%								
mt	95.0%	86.3%	74.4%	69.4%	83.2%	97.4%	77.2%	−49.8%	−49.5%	75.7%	−10.1%	76.6%	20.7%	47.6%	100.0%							
nl	81.7%	76.4%	91.7%	96.9%	91.7%	64.4%	88.3%	−60.3%	−41.5%	86.0%	−31.2%	89.3%	−26.0%	0.9%	73.8%	100.0%						
pl	96.6%	94.1%	77.4%	76.3%	84.3%	86.2%	76.8%	−54.7%	−51.3%	72.1%	−25.2%	76.2%	20.0%	45.4%	93.5%	73.5%	100.0%					
pt	86.9%	74.0%	77.3%	80.5%	97.3%	68.5%	99.8%	−61.3%	−45.7%	99.2%	−23.5%	99.5%	−16.1%	8.8%	76.4%	87.7%	74.9%	100.0%				
ro	94.6%	81.9%	78.6%	79.7%	97.7%	79.7%	95.7%	−61.3%	−50.9%	94.9%	−25.4%	95.9%	−4.6%	22.1%	87.7%	85.3%	86.0%	95.8%	100.0%			
sk	88.0%	91.5%	68.9%	65.4%	72.4%	78.8%	63.6%	−45.9%	−44.3%	59.0%	−24.0%	63.1%	27.4%	47.5%	86.9%	63.2%	96.2%	61.6%	73.4%	100.0%		
sl	90.2%	93.3%	74.9%	67.7%	74.1%	82.7%	66.3%	−52.0%	−52.2%	62.5%	−29.2%	66.4%	22.1%	41.3%	87.9%	68.0%	97.1%	64.9%	75.8%	98.8%	100.0%	
sv	86.2%	85.0%	95.7%	90.3%	85.1%	73.4%	78.0%	−50.3%	−40.0%	74.0%	−25.3%	79.0%	−0.9%	25.3%	81.2%	89.0%	85.1%	76.5%	80.7%	79.0%	81.3%	100.0%

Table B.3: Correlation between the BLEU score vectors "INTO" of the 22 official EU languages



LG	bg	cs	da	de	el	en	es	et	fi	fr	hu	it	lt	lv	mt	nl	pl	pt	ro	sk	sl	sv
bg	100.0%																					
cs	96.4%	100.0%																				
da	91.7%	90.3%	100.0%																			
de	89.9%	88.5%	96.5%	100.0%																		
el	97.5%	93.7%	92.6%	92.3%	100.0%																	
en	96.8%	91.7%	83.4%	80.6%	92.9%	100.0%																
es	93.1%	86.8%	87.7%	87.5%	98.3%	88.5%	100.0%															
et	37.9%	35.0%	40.7%	46.0%	38.3%	−1.0%	26.9%	100.0%														
fi	48.6%	41.4%	54.5%	60.5%	50.2%	15.8%	42.8%	93.7%	100.0%													
fr	92.8%	86.0%	86.9%	86.5%	98.1%	86.6%	99.6%	23.7%	36.1%	100.0%												
hu	79.4%	72.1%	75.7%	79.2%	78.9%	66.0%	70.3%	85.7%	85.0%	67.9%	100.0%											
it	94.5%	88.5%	89.1%	89.4%	98.9%	89.6%	99.8%	35.3%	51.3%	99.7%	75.3%	100.0%										
lt	69.5%	66.8%	61.3%	62.6%	67.0%	53.7%	55.3%	82.5%	74.1%	53.1%	91.8%	61.8%	100.0%									
lv	85.3%	84.3%	76.0%	78.3%	81.2%	80.3%	70.7%	72.4%	67.5%	69.5%	90.9%	75.3%	97.2%	100.0%								
mt	96.1%	94.3%	88.2%	83.4%	90.1%	99.7%	83.7%	51.1%	53.8%	84.0%	82.6%	85.4%	75.2%	91.0%	100.0%							
nl	93.0%	91.0%	96.5%	99.1%	96.2%	85.1%	92.9%	39.1%	54.8%	92.6%	74.9%	94.3%	61.8%	76.0%	86.7%	100.0%						
pl	98.1%	97.9%	88.2%	86.1%	93.4%	96.5%	86.5%	38.8%	42.2%	86.3%	77.8%	88.4%	72.7%	89.3%	96.9%	88.3%	100.0%					
pt	93.9%	88.1%	88.4%	88.2%	98.5%	88.5%	99.9%	34.8%	50.3%	99.4%	74.0%	99.8%	61.1%	74.7%	84.6%	93.6%	87.7%	100.0%				
ro	98.4%	93.7%	90.2%	89.4%	98.9%	94.6%	96.9%	40.1%	51.2%	96.3%	79.8%	97.6%	69.6%	83.9%	92.7%	94.0%	94.8%	97.3%	100.0%			
sk	93.5%	97.2%	84.7%	81.4%	86.8%	90.9%	77.8%	36.8%	37.9%	78.5%	75.7%	80.2%	71.8%	88.8%	94.1%	83.8%	97.9%	79.1%	88.1%	100.0%		
sl	94.5%	97.1%	86.5%	81.7%	87.4%	93.4%	78.4%	36.9%	37.6%	79.7%	75.5%	81.4%	70.1%	87.1%	95.5%	85.0%	98.3%	80.9%	89.2%	99.4%	100.0%	
sv	93.8%	94.2%	97.5%	93.8%	91.2%	87.1%	84.0%	49.8%	57.3%	84.1%	82.4%	86.4%	70.5%	86.5%	92.1%	93.8%	93.1%	85.2%	90.5%	91.0%	91.9%	100.0%

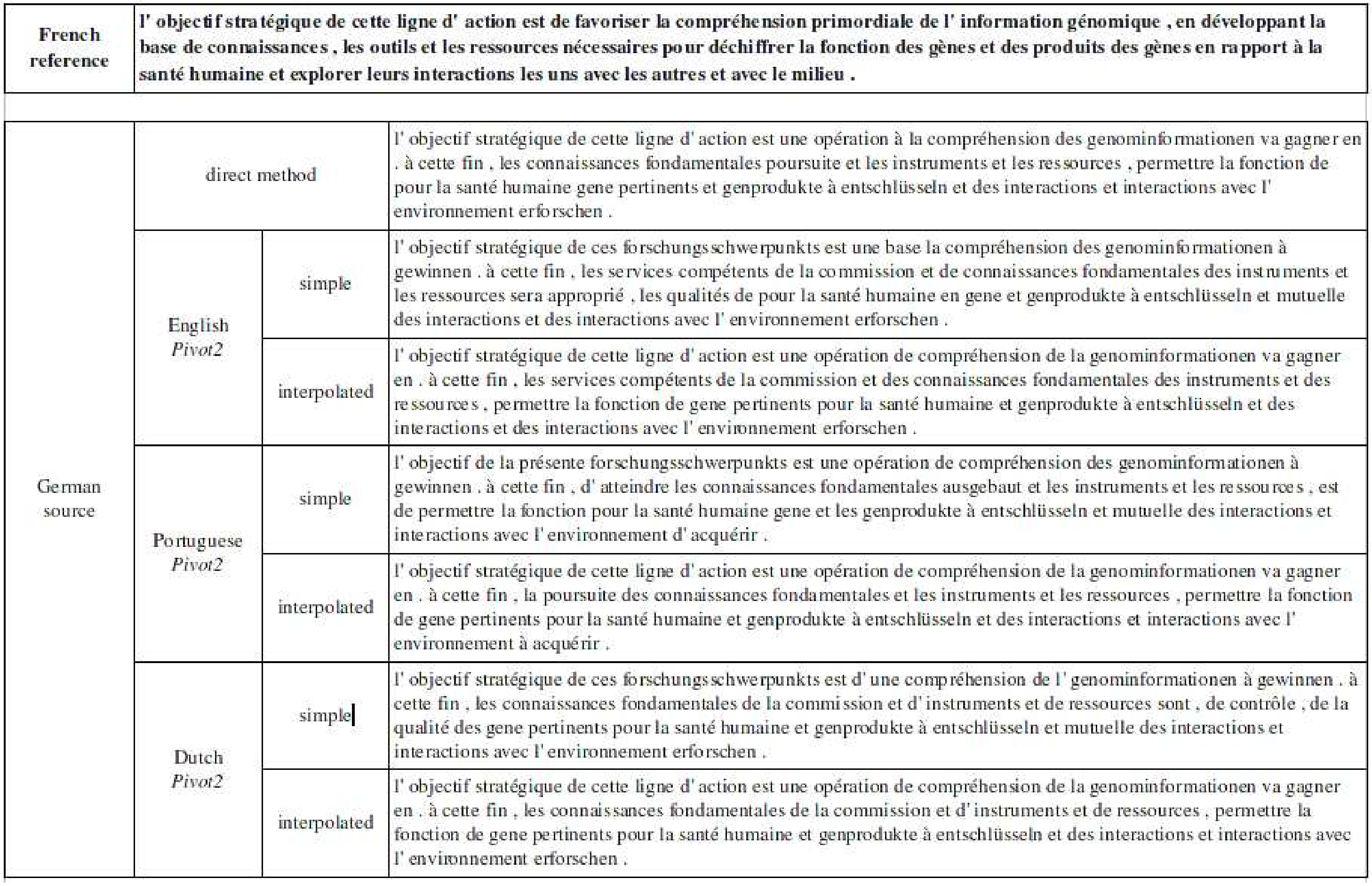
Table B.4: Correlation between the BLEU score vectors “FROM” of the 22 official EU languages

Appendix C

Sample outputs of pivot-based translation systems

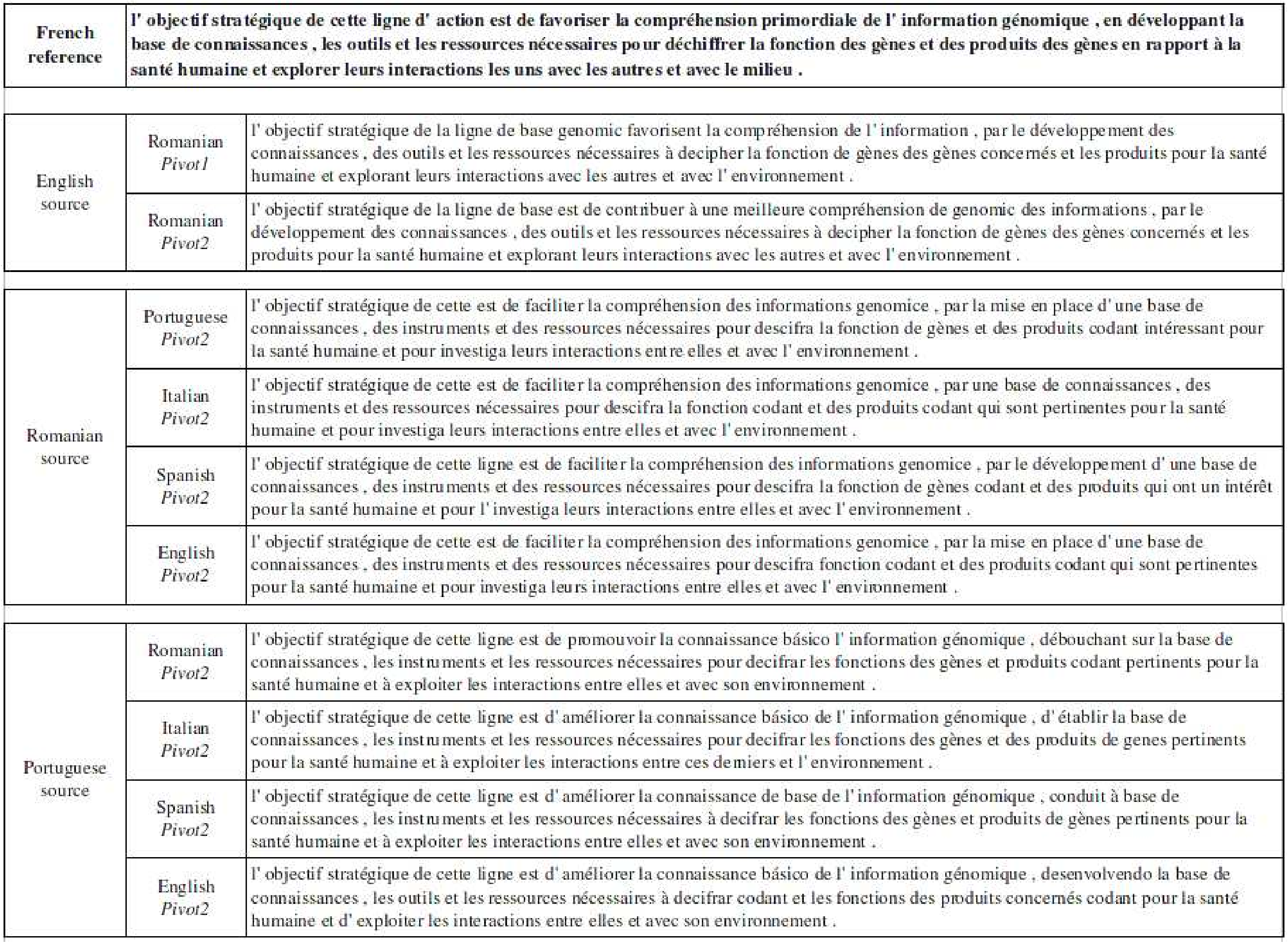
This appendix is related to chapter 7, which describes the pivot-based experiments. It presents some sample translations of our pivot-based models. We consider the sentence listed in Appendix A (tables A.1 and A.2), translated this time by different pivot-based systems.

The tables C.1 and C.2 present the French translations and the table C.3 shows the Romanian ones.



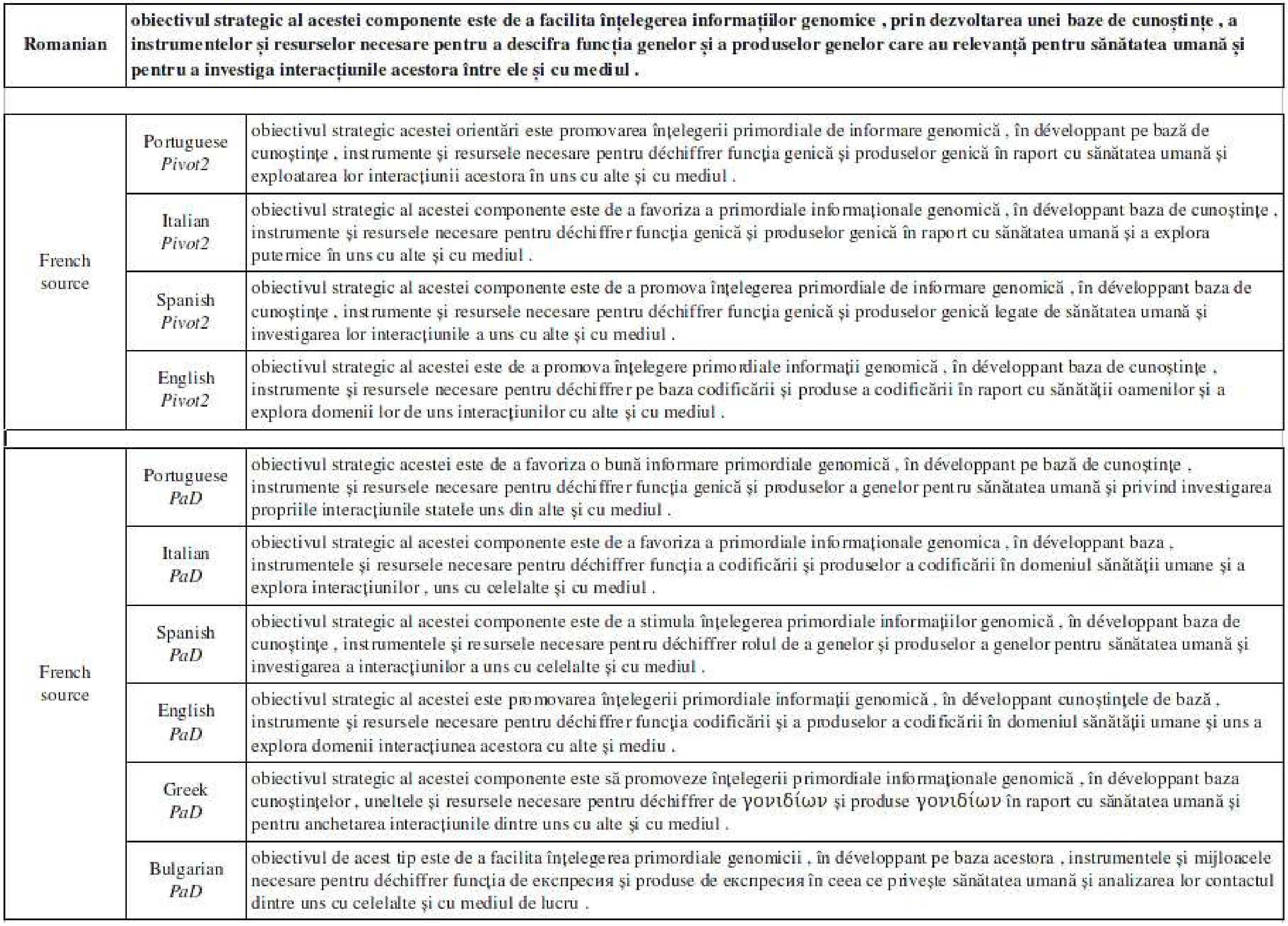
French reference	**l' objectif stratégique de cette ligne d' action est de favoriser la compréhension primordiale de l' information génomique , en développant la base de connaissances , les outils et les ressources nécessaires pour déchiffrer la fonction des gènes et des produits des gènes en rapport à la santé humaine et explorer leurs interactions les uns avec les autres et avec le milieu .**		
German source	direct method		l' objectif stratégique de cette ligne d' action est une opération à la compréhension des genominformationen va gagner en . à cette fin , les connaissances fondamentales poursuite et les instruments et les ressources , permettre la fonction de pour la santé humaine gene pertinents et genprodukte à entschlüsseln et des interactions et interactions avec l' environnement erforschen .
	English *Pivot2*	simple	l' objectif stratégique de ces forschungsschwerpunkts est une base la compréhension des genominformationen à gewinnen . à cette fin , les services compétents de la commission et de connaissances fondamentales des instruments et les ressources sera approprié , les qualités de pour la santé humaine en gene et genprodukte à entschlüsseln et mutuelle des interactions et des interactions avec l' environnement erforschen .
		interpolated	l' objectif stratégique de cette ligne d' action est une opération de compréhension de la genominformationen va gagner en . à cette fin , les services compétents de la commission et des connaissances fondamentales des instruments et des ressources , permettre la fonction de gene pertinents pour la santé humaine et genprodukte à entschlüsseln et des interactions et des interactions avec l' environnement erforschen .
	Portuguese *Pivot2*	simple	l' objectif de la présente forschungsschwerpunkts est une opération de compréhension des genominformationen à gewinnen . à cette fin , d' atteindre les connaissances fondamentales ausgebaut et les instruments et les ressources , est de permettre la fonction pour la santé humaine gene et les genprodukte à entschlüsseln et mutuelle des interactions et interactions avec l' environnement d' acquérir .
		interpolated	l' objectif stratégique de cette ligne d' action est une opération de compréhension de la genominformationen va gagner en . à cette fin , la poursuite des connaissances fondamentales et les instruments et les ressources , permettre la fonction de gene pertinents pour la santé humaine et genprodukte à entschlüsseln et des interactions et interactions avec l' environnement à acquérir .
	Dutch *Pivot2*	simple	l' objectif stratégique de ces forschungsschwerpunkts est d' une compréhension de l' genominformationen à gewinnen . à cette fin , les connaissances fondamentales de la commission et d' instruments et de ressources sont , de contrôle , de la qualité des gene pertinents pour la santé humaine et genprodukte à entschlüsseln et mutuelle des interactions et interactions avec l' environnement erforschen .
		interpolated	l' objectif stratégique de cette ligne d' action est une opération de compréhension de la genominformationen va gagner en . à cette fin , les connaissances fondamentales de la commission et d' instruments et de ressources , permettre la fonction de gene pertinents pour la santé humaine et genprodukte à entschlüsseln et des interactions et interactions avec l' environnement erforschen .

Table C.1: Sample output of pivot-based translation systems, when translating into French, trained on *Translation-Units-22* corpus (part 1)



French reference		**l' objectif stratégique de cette ligne d' action est de favoriser la compréhension primordiale de l' information génomique , en développant la base de connaissances , les outils et les ressources nécessaires pour déchiffrer la fonction des gènes et des produits des gènes en rapport à la santé humaine et explorer leurs interactions les uns avec les autres et avec le milieu .**
English source	Romanian *Pivot1*	l' objectif stratégique de la ligne de base genomic favorisent la compréhension de l' information , par le développement des connaissances , des outils et les ressources nécessaires à decipher la fonction de gènes des gènes concernés et les produits pour la santé humaine et explorant leurs interactions avec les autres et avec l' environnement .
	Romanian *Pivot2*	l' objectif stratégique de la ligne de base est de contribuer à une meilleure compréhension de genomic des informations , par le développement des connaissances , des outils et les ressources nécessaires à decipher la fonction de gènes des gènes concernés et les produits pour la santé humaine et explorant leurs interactions avec les autres et avec l' environnement .
Romanian source	Portuguese *Pivot2*	l' objectif stratégique de cette est de faciliter la compréhension des informations genomice , par la mise en place d' une base de connaissances , des instruments et des ressources nécessaires pour descifra la fonction de gènes et des produits codant intéressant pour la santé humaine et pour investiga leurs interactions entre elles et avec l' environnement .
	Italian *Pivot2*	l' objectif stratégique de cette est de faciliter la compréhension des informations genomice , par une base de connaissances , des instruments et des ressources nécessaires pour descifra la fonction codant et des produits codant qui sont pertinentes pour la santé humaine et pour investiga leurs interactions entre elles et avec l' environnement .
	Spanish *Pivot2*	l' objectif stratégique de cette ligne est de faciliter la compréhension des informations genomice , par le développement d' une base de connaissances , des instruments et des ressources nécessaires pour descifra la fonction de gènes codant et des produits qui ont un intérêt pour la santé humaine et pour l' investiga leurs interactions entre elles et avec l' environnement .
	English *Pivot2*	l' objectif stratégique de cette est de faciliter la compréhension des informations genomice , par la mise en place d' une base de connaissances , des instruments et des ressources nécessaires pour descifra fonction codant et des produits codant qui sont pertinentes pour la santé humaine et pour investiga leurs interactions entre elles et avec l' environnement .
Portuguese source	Romanian *Pivot2*	l' objectif stratégique de cette ligne est de promouvoir la connaissance básico l' information génomique , débouchant sur la base de connaissances , les instruments et les ressources nécessaires pour decifrar les fonctions des gènes et produits codant pertinents pour la santé humaine et à exploiter les interactions entre elles et avec son environnement .
	Italian *Pivot2*	l' objectif stratégique de cette ligne est d' améliorer la connaissance básico de l' information génomique , d' établir la base de connaissances , les instruments et les ressources nécessaires pour decifrar les fonctions des gènes et des produits de genes pertinents pour la santé humaine et à exploiter les interactions entre ces derniers et l' environnement .
	Spanish *Pivot2*	l' objectif stratégique de cette ligne est d' améliorer la connaissance de base de l' information génomique , conduit à base de connaissances , les instruments et les ressources nécessaires à decifrar les fonctions des gènes et produits de gènes pertinents pour la santé humaine et à exploiter les interactions entre elles et avec son environnement .
	English *Pivot2*	l' objectif stratégique de cette ligne est d' améliorer la connaissance básico de l' information génomique , desenvolvendo la base de connaissances , les outils et les ressources nécessaires à decifrar codant et les fonctions des produits concernés codant pour la santé humaine et d' exploiter les interactions entre elles et avec son environnement .

Table C.2: Sample output of pivot-based translation systems, when translating into French, trained on *Translation-Units-22* corpus (part 2)



Romanian	**obiectivul strategic al acestei componente este de a facilita înțelegerea informațiilor genomice , prin dezvoltarea unei baze de cunoștințe , a instrumentelor și resurselor necesare pentru a descifra funcția genelor și a produselor genelor care au relevanță pentru sănătatea umană și pentru a investiga interacțiunile acestora între ele și cu mediul .**	
French source	Portuguese *Pivot2*	obiectivul strategic acestei orientări este promovarea înţelegerii primordiale de informare genomică , în développant pe bază de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer funcţia genică şi produselor genică în raport cu sănătatea umană şi exploatarea lor interacţiunii acestora în uns cu alte şi cu mediul .
	Italian *Pivot2*	obiectivul strategic al acestei componente este de a favoriza a primordiale informaţionale genomică , în développant baza de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer funcţia genică şi produselor genică în raport cu sănătatea umană şi a explora puternice în uns cu alte şi cu mediul .
	Spanish *Pivot2*	obiectivul strategic al acestei componente este de a promova înţelegerea primordiale de informare genomică , în développant baza de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer funcţia genică şi produselor genică legate de sănătatea umană şi investigarea lor interacţiunile a uns cu alte şi cu mediul .
	English *Pivot2*	obiectivul strategic al acestei este de a promova înţelegere primordiale informaţii genomică , în développant baza de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer pe baza codificării şi produse a codificării în raport cu sănătăţii oamenilor şi a explora domenii lor de uns interacţiunilor cu alte şi cu mediul .
French source	Portuguese *PaD*	obiectivul strategic acestei este de a favoriza o bună informare primordiale genomică , în développant pe bază de cunoştinţe , instrumente şi resursele necesare pentru déchiffrer funcţia genică şi produselor a genelor pentru sănătatea umană şi privind investigarea propriile interacţiunile statele uns din alte şi cu mediul .
	Italian *PaD*	obiectivul strategic al acestei componente este de a favoriza a primordiale informaţionale genomica , în développant baza , instrumentele şi resursele necesare pentru déchiffrer funcţia a codificării şi produselor a codificării în domeniul sănătăţii umane şi a explora interacţiunilor , uns cu celelalte şi cu mediul .
	Spanish *PaD*	obiectivul strategic al acestei componente este de a stimula înţelegerea primordiale informaţiilor genomică , în développant baza de cunoştinţe , instrumentele şi resursele necesare pentru déchiffrer rolul de a genelor şi produselor a genelor pentru sănătatea umană şi investigarea a interacţiunilor a uns cu celelalte şi cu mediul .
	English *PaD*	obiectivul strategic al acestei este promovarea înţelegerii primordiale informaţii genomică , în développant cunoştinţele de bază , instrumente şi resursele necesare pentru déchiffrer funcţia codificării şi a produselor a codificării în domeniul sănătăţii umane şi uns a explora domenii interacţiunea acestora cu alte şi mediu .
	Greek *PaD*	obiectivul strategic al acestei componente este să promoveze înţelegerii primordiale informaţionale genomică , în développant baza cunoştinţelor , uneltele şi resursele necesare pentru déchiffrer de γονιδίων şi produse γονιδίων în raport cu sănătatea umană şi pentru anchetarea interacţiunile dintre uns cu alte şi cu mediul .
	Bulgarian *PaD*	obiectivul de acest tip este de a facilita înţelegerea primordiale genomicii , în développant pe baza acestora , instrumentele şi mijloacele necesare pentru déchiffrer funcţia de експресия şi produse de експресия în ceea ce priveşte sănătatea umană şi analizarea lor contactul dintre uns cu celelalte şi cu mediul de lucru .

Table C.3: Sample output of pivot-based translation systems, when translating into Romanian, trained on *Translation-Units-22* corpus

List of Tables

List of Figures

Bibliography

[Ahrenberg et al., 2000] Ahrenberg, L., Andersson, M., and Merkel, M. (2000). A knowledge-lite approach to word alignment. In *J. Véronis (ed.) Parallel Text Processing: Alignment and Use of Parallel Corpora*, pages 97–116.

[Ahrenberg et al., 2002] Ahrenberg, L., Andersson, M., and Merkel, M. (2002). A system for incremental and interactive word linking. In *In Proceedings of the 3rd International Conference on Language Resources and Evaluation*, pages 485–490, Las Palmas, Spain.

[Al-Onaizan et al., 1999] Al-Onaizan, Y., Curin, J., Jahr, M., Knight, K., Lafferty, J., Melamed, D., Och, F., Purdy, D., Smith, N. A., and Yarowsky, D. (1999). Statistical machine translation. Technical report, Final Report, JHU Summer Workshop.

[Arad, 1991] Arad, I. (1991). *A quasi-statistical approach to the automatic generation of linguistic knowledge*. PhD thesis, University of Manchester.

[Aswani and Gaizauskas, 2005] Aswani, N. and Gaizauskas, R. (2005). Aligning words in English-Hindi parallel corpora. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 115–118, Ann Arbor, Michigan, USA.

[Ayan and Dorr, 2006] Ayan, N. and Dorr, B. (2006). Going beyond AER: An extensive analysis of word alignments and their impact on MT. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 9–16, Sydney, Australia.

[Banerjee and Lavie, 2005] Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, USA.

[Bangalore et al., 2001] Bangalore, S., Bordel, G., and Riccardi, G. (2001). Computing consensus translation from multiple machine translation systems. In *Proceedings of Automated Speech Recognition and Understanding Workshop (ASRU)*, pages 351–354, Madonna di Campiglio, Italy.

[Berger et al., 1994] Berger, A. L., Brown, P. F., Della Pietra, S. A., Della Pietra, V. J., Gillett, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., and Ures, L. (1994). The candide system for machine translation. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 157–162, Plainsboro, New Jersey, USA.

[Bertoldi et al., 2008] Bertoldi, N., Barbaiani, M., Federico, M., and Cattoni, R. (2008). Phrase-based statistical machine translation with pivot languages. In *Proceedings of the International Workshop on Spoken Language, Evaluation Campaign on Spoken Language Translation (IWSLT)*, pages 143–149.

[Birch et al., 2006] Birch, A., Callison-Burch, C., and Osborne, M. (2006). Constraining the phrase-based, joint probability statistical translation model. In *Proceedings of the 7th Conference of the Association for Machine Translation of the Americas (AMTA)*, pages 10–18, Cambridge, Massachusetts, USA.

[Blunsom and Cohn, 2006] Blunsom, P. and Cohn, T. (2006). Discriminative word alignment with conditional random fields. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (ACL-COLING 2006)*, pages 65–72, Sydney, Australia.

[Boitet, 1988] Boitet, C. (1988). Pros and cons of the pivot and transfer approaches in multilingual machine translation. In Dan Maxwell, Klaus Schubert, T. W., editor, *New directions in Machine Translation, BSO Conference*, pages 93–106.

[Borin, 2002] Borin, L. . (2002). Alignment and tagging. In *Parallel Corpora, Parallel Worlds: Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999*, pages 207–218.

[Borin, 2000a] Borin, L. (2000a). Enhancing tagging performance by combining knowledge sources. In *Papers from the ASLA symposium Corpora in research and teaching*, pages 19–31, Vaxjo, Sweden.

[Borin, 2000b] Borin, L. (2000b). You'll take the high road and i'll take the low road: Using a third language to improve bilingual word alignment. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 97–103, Saarbrücken, Germany.

[Brown et al., 1990] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Lafferty, J., Mercer, R., and Roossin, P. (1990). A statistical approach to machine translation. *Computational Linguistics*, Vol 16(2):79–85.

[Brown et al., 1988] Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., and Roossin, P. (1988). A statistical approach to language translation. In *Proceedings of the 12th Conference on Computational Linguistics (COLING 1988)*, pages 71–76, Budapest, Hungary.

[Brown et al., 1992] Brown, P., Della Pietra, S., Della Pietra, V., Lafferty, J., and Mercer, R. (1992). Analysis, statistical transfer, and synthesis in machine translation. In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 83–100.

[Brown et al., 1993] Brown, P., Della Pietra, S., Della Pietra, V., and Mercer, R. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

[Brown et al., 1991] Brown, P. F., Lai, J. C., and Mercer, R. L. (1991). Aligning sentences in parallel corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 169–176, Berkeley, California, USA.

[Brown, 1996] Brown, R. D. (1996). Example-based machine translation in the pangloss system. In *Proceedings of the 16th conference on Computational linguistics (COLING 1996)*, pages 169–174, Copenhagen, Denmark.

[Callison-Burch, 2007] Callison-Burch, C. (2007). *Paraphrasing and Translation*. PhD thesis, University of Edinburgh.

[Callison-Burch and Flournoy, 2001] Callison-Burch, C. and Flournoy, R. (2001). A program for automatically selecting the best output from multiple machine translation engines. In *Proceedings of the Machine Translation Summit VIII*, pages 9–13, Santiago de Compostela, Spain.

[Callison-Burch et al., 2007] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (meta-) evaluation of machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

[Callison-Burch et al., 2008] Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2008). Further meta-evaluation of machine translation. In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, USA.

[Callison-Burch et al., 2006] Callison-Burch, C., Koehn, P., and Osborne, M. (2006). Improved statistical machine translation using paraphrases. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, pages 17–24, New York City, USA.

[Callison-Burch and Osborne, 2003] Callison-Burch, C. and Osborne, M. (2003). Co-training for statistical machine translation. In *In Proceedings of the 6th Annual Computational Linguistics UK Research Colloquium (CLUK)*.

[Callison-Burch et al., 2004] Callison-Burch, C., Talbot, D., and Osborne, M. (2004). Statistical machine translation with word- and sentence-aligned parallel corpora. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, pages 175–182, Barcelona, Spain.

[Carl et al., 2008] Carl, M., Melero, M., Badia, T., Vandeghinste, V., Dirix, P., Schuurman, I., Markantonatou, S., Sofianopoulos, S., Vassiliou, M., and Yannoutsou, O. (2008). METIS-II: low resource machine translation. *Machine Translation*, 22(1-2):67–99.

[Chen et al., 2008] Chen, Y., Eisele, A., and Kay, M. (2008). Improving statistical machine translation efficiency by triangulation. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

[Cherry and Lin, 2003] Cherry, C. and Lin, D. (2003). A probability model to improve word alignment. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 88–95, Sapporo, Japan.

[Civera and Juan, 2006] Civera, J. and Juan, A. (2006). Bilingual Machine-Aided Indexing. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 1302–1305, Genoa, Italy.

[Cohn and Lapata, 2007] Cohn, T. and Lapata, M. (2007). Machine translation by triangulation: Making effective use of multi-parallel corpora. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 728–735, Prague, Czech Republic.

[Dagan and Church, 1994] Dagan, I. and Church, K. (1994). Termight: identifying and translating technical terminology. In *Proceedings of the 4th conference on Applied natural language processing (ANLP)*, pages 34–40, Stuttgart, Germany.

[Dagan et al., 1993] Dagan, I., Church, K. W., and Gale, W. A. (1993). Robust bilingual word alignment for machine aided translation. In *In Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 1–8, Columbus, Ohio, USA.

[Danielsson and Ridings, 1997] Danielsson, P. and Ridings, D. (1997). Practical presentation of a "Vanilla" aligner. In *Trans-European Language Resources Infrastructure Newsletter (TELRI)*, volume No 5.

[De Gispert and Mariño, 2006] De Gispert, A. and Mariño, J. B. (2006). Catalan-English statistical machine translation without parallel corpus: Bridging through Spanish. In *Proceeding of the LREC 5th Workshop on Strategies for developing Machine Translation for Minority Languages (SALTMIL)*, pages 65–68, Genova, Italy.

[Deléger et al., 2009] Deléger, L., Merkel, M., and Zweigenbaum, P. (2009). Translating terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, Vol 39(1):1–38.

[Deng and Byrne, 2005] Deng, Y. and Byrne, W. (2005). HMM word and phrase alignment for statistical machine translation. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 169–176, Vancouver, British Columbia, Canada.

[Di Cristo, 1996] Di Cristo, P. (1996). Mtseg: The MULTEXT multilingual segmenter tools. Deliverable MSG 1, Version 1.3.1. CNRS, Aix-en-Provence.

[Diab and Resnik, 2002] Diab, M. and Resnik, P. (2002). An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 255–262, Philadelphia, Pennsylvania, USA.

[Dimitrova et al., 1998] Dimitrova, L., Erjavec, T., Ide, N., Kaalep, H.-J., Petkevič, V., and Tufiş, D. (1998). Multext-east: Parallel and comparable corpora and lexicons for six central and eastern european languages. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 315–319, Montréal, Québec, Canada.

[Doddington, 2002] Doddington, G. (2002). Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second International Conference on Human Language Technology Research (HLT)*, pages 138–145, San Francisco, California, USA.

[Eisele, 2005] Eisele, A. (2005). First steps towards multi-engine machine translation. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 155–158, Ann Arbor, Michigan, USA.

[Eisele, 2006] Eisele, A. (2006). Parallel corpora and phrase-based statistical machine translation for new language pairs via multiple intermediaries. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 845–848, Genoa, Italy.

[Erjavec, 2004] Erjavec, T. (2004). Multext-east version 3: Multilingual morphosyntactic specifications, lexicons and corpora. European Language Resources Association (ELRA).

[Erjavec et al., 1996] Erjavec, T., Ide, N., Petkevič, V., and Véronis, J. (1996). Multext-east: Multilingual text tools and corpora for central and eastern european languages. In *Proceedings of the First TELRI European Seminar: Language Resources for Language Technology*, pages 87–98, Tihany, Hungary.

[Erjavec et al., 2005] Erjavec, T., Ignat, C., Pouliquen, B., and Steinberger, R. (2005). Massive multi lingual corpus compilation: Acquis communautaire and totale. *Archives of Control Sciences*, Vol 15(4):529–540.

[EUROVOC, 1995] EUROVOC (1995). Thesaurus eurovoc - Volume 2:Subject-Oriented Version. Ed 3/English language. Annex to the index of the Official Journal of the EC, Office for Official Publications of the European Communities, Luxembourg.

[Filali and Bilmes, 2005] Filali, K. and Bilmes, J. (2005). Leveraging multiple languages to improve statistical MT word alignments. In *IEEE Automatic Speech Recognition and Understanding (ASRU)*, Cancun, Mexico,.

[Fraser and Marcu, 2006] Fraser, A. and Marcu, D. (2006). Semi-supervised training for statistical word alignment. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL 2006)*, pages 769–776, Sydney, Australia.

[Fraser and Marcu, 2007] Fraser, A. and Marcu, D. (2007). Measuring word alignment quality for statistical machine translation. *Computational Linguistics*, Vol 33(3):293–303.

[Frederking and Nirenburg, 1994] Frederking, R. and Nirenburg, S. (1994). Three heads are better than one. In *In Proceedings of the fourth ACL Conference on Applied Natural Language Processing (ANLP)*, pages 95–100, Stuttgart, Germany.

[Fung and McKeown, 1997] Fung, P. and McKeown, K. (1997). A technical word- and term-translation aid using noisy parallel corpora across language groups. *Machine Translation*, Vol 12(1-2):53–87.

[Gale and Church, 1991a] Gale, W. A. and Church, K. W. (1991a). Identifying word correspondence in parallel texts. In *Proceedings of the workshop on Speech and Natural Language (HLT)*, pages 152–157, Pacific Grove, California, USA.

[Gale and Church, 1991b] Gale, W. A. and Church, K. W. (1991b). A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics (ACL 1991)*, pages 177–184, Berkeley, California, USA.

[Gale and Church, 1993] Gale, W. A. and Church, K. W. (1993). A program for aligning sentences in bilingual corpora. *Computational Linguistics*, Vol 19:177–184.

[Gale et al., 1992] Gale, W. A., Church, K. W., and Yarowsky, D. (1992). Using bilingual materials to develop word sense disambiguation methods. In *Proceedings of the 7th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 101–112, Santa Fe, New Mexico, USA.

[Gaussier, 1998] Gaussier, E. (1998). Flow network models for word alignment and terminology extraction from bilingual corpora. In *Proceedings of the 17th international conference on Computational linguistics (COLING 1998)*, pages 444–450, Montreal, Quebec, Canada.

[Germann, 2001] Germann, U. (2001). Aligned Hansards of the 36th Parliament of Canada - Released 2001-1a. http://www.isi.edu/natural-language/download/hansard/index.html.

[Germann et al., 2001] Germann, U., Jahr, M., Knight, K., Marcu, D., and Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. In *Proceedings of the Association for Computational Linguistic, 39th Annual Meeting and 10th Conference of the European Chapter (ACL 2001)*, pages 228–235, Toulouse, France.

[Giguet and Luquet, 2006] Giguet, E. and Luquet, P.-S. (2006). Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In *Poster Proceedings of the ACL-COLING International Conference*, Sydney, Australia.

[Gollins and Sanderson, 2001] Gollins, T. and Sanderson, M. (2001). Improving cross language retrieval with triangulated translation. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and Development in Information Retrieval*, pages 90–95, New York City, USA.

[Hiemstra, 1998] Hiemstra, D. (1998). Multilingual domain modeling in twenty-one: automatic creation of a bi-directional translation lexicon from a parallel corpus. In *Proceedings of the 8th Computational Linguistics in the Netherlands meeting (CLIN)*, pages 41–58, Leuven, Belgium.

[Hoang and Koehn, 2008] Hoang, H. and Koehn, P. (2008). Design of the "Moses" decoder for statistical machine translation. In *ACL Workshop on Software engineering, testing, and quality assurance for Natural Language Processing*, pages 58–65, Columbus, Ohio, USA.

[Hutchins and Somers, 1992] Hutchins, J. W. and Somers, H. L. (1992). *An introduction to machine translation.* London: Academic Press.

[Ignat et al., 2003] Ignat, C., Pouliquen, B., Ribeiro, A., and Steinberger, R. (2003). Extending an information extraction tool set to central and eastern european languages. In *Proceedings of the International Workshop Information Extraction for Slavonic and other Central and Eastern European Languages held at RANLP*, pages 33–39, Borovets, Bulgaria.

[Ignat et al., 2005] Ignat, C., Pouliquen, B., Steinberger, R., and Erjavec, T. (2005). A tool set for the quick and efficient exploration of large document collections. In *Proceedings of the Symposium on Safeguards and Nuclear Material Management (ESARDA)*, London, United Kingdom.

[Ignat and Rousselot, 2006a] Ignat, C. and Rousselot, F. (2006a). Représentation de textes a l'aide d'étiquettes sémantiques dans le cadre de la classification automatique. *Romanian Review of Linguistics*, Vol 51(3-4):217–240.

[Ignat and Rousselot, 2006b] Ignat, C. and Rousselot, F. (2006b). Un algorithme de génération de profil de document et son évaluation dans le contexte de la classification thématique. In *Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT)*, pages 19–21, Besançon, France.

[Ittycheriah and Roukos, 2005] Ittycheriah, A. and Roukos, S. (2005). A maximum entropy word aligner for Arabic-English machine translation. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 89–96, Vancouver, British Columbia, Canada.

[Jayaraman and Lavie, 2005] Jayaraman, S. and Lavie, A. (2005). Multi-engine machine translation guided by explicit word matching. In *Proceedings of the European Association for Machine Translation Annual Conference (EAMT)*, pages 143–152, Budapest, Hungary.

[Kaki et al., 1999] Kaki, S., Yamada, S., and Sumita, E. (1999). Scoring multiple translations using character n-gram. In *Proceedings of the 5th Natural Language Processing Pacific Rim Symposium (NLPRS)*, pages 298–302, Beijing, China.

[Karlgren et al., 1994] Karlgren, H., Karlgren, J., Nordström, M., Pettersson, P., and Wahrolén, B. (1994). Dilemma: an instant lexicographer. In *Proceedings of the 15th Conference on Computational Linguistics (COLING 1994)*, pages 82–84, Kyoto, Japan.

[Kay, 1997] Kay, M. (1997). The proper place of men and machines inlanguage translation. *Machine Translation*, Vol 12(1-2):3–23.

[Kay, 2000] Kay, M. (2000). Triangulation in translation. Invited talk at the MT 2000 conference, University of Exeter.

[Kay and Röscheisen, 1993] Kay, M. and Röscheisen, M. (1993). Text-translation alignment. *Computational Linguistics*, Vol 19(1):121–142.

[Kishida and Kando, 2003] Kishida, K. and Kando, N. (2003). Two-stage refinement of query translation in a pivot language approach to cross-lingual information retrieval: An experiment at CLEF 2003. In *Proceedings of Cross-Language Evaluation Forum*, pages 253–262, Trondheim, Norway.

[Klavans and Tzoukermann, 1990] Klavans, J. and Tzoukermann, E. (1990). The bicord system: combining lexical information from bilingual corpora and machine readable dictionaries. In *Proceedings of the 13th conference on Computational linguistics (COLING 1990)*, pages 174–179, Helsinki, Finland.

[Knight and Al-Onaizan, 1998] Knight, K. and Al-Onaizan, Y. (1998). Translation with finite-state devices. In *Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 421–437, Langhorne, Pennsylvania, USA.

[Koehn, 2004a] Koehn, P. (2004a). "Pharaoh": A beam search decoder for phrase-based statistical machine translation models. In *Proceedings of The 6th Conference of the Association for Machine Translation in the Americas (AMTA)*, page 115–124, Washington DC, USA.

[Koehn, 2004b] Koehn, P. (2004b). Statistical significance tests for machine translation evaluation. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, Barcelona, Spain.

[Koehn, 2005] Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit 2005*, pages 79–86.

[Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). "Moses": Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007), Demonstration Session*, pages 177–180, Prague, Czech Republic.

[Koehn and Monz, 2005] Koehn, P. and Monz, C. (2005). Shared task: Statistical machine translation between European languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124, Ann Arbor, Michigan.

[Koehn and Monz, 2006] Koehn, P. and Monz, C. (2006). Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the of NAACL Workshop on Statistical Machine Translation*, pages 102–121, New York City, USA.

[Koehn et al., 2003] Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 127–133, Edmonton, Canada.

[Kumar et al., 2007] Kumar, S., Och, F. J., and Macherey, W. (2007). Improving word alignment with bridge languages. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 42–50, Prague, Czech Republic.

[Kupiec, 1993] Kupiec, J. (1993). An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics (ACL 1993)*, pages 17–22, Columbus, Ohio, USA.

[Lambert and Castell, 2004] Lambert, P. and Castell, N. (2004). Alignment of parallel corpora exploiting asymmetrically aligned phrases. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004)*, page 26–29, Lisbon, Portugal.

[Lambert et al., 2005] Lambert, P., de Gispert, A., Banchs, R., and Mariño, J. B. (2005). Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, Vol 39(4):267–285.

[Langlais and El-Beze, 1997] Langlais, P. and El-Beze, M. (1997). Alignement de corpus bilingues : algorithmes et évaluation. In *Actes des 1ères JST FRANCIL (Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-UREF)*, pages 191–197, Avignon, France.

[Langlais et al., 2008] Langlais, P., Yvon, F., and Zweigenbaum., P. (2008). Analogical translation of medical words in different languages. In Ranta, A. and Nordström, B., editors, *6th International Conference on Natural Language Processing, GoTAL 2008*, volume 5221, pages 284–295, Berlin / Heidelberg.

[Lopez, 2007] Lopez, A. (2007). A survey of statistical machine translation. Technical report, Maryland University College Park, Department of Computer Science.

[Lopez, 2008] Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys*, 40(3):1—-49.

[Lopez and Resnik, 2005] Lopez, A. and Resnik, P. (2005). Improved HMM alignment models for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 83–86, Ann Arbor, Michigan, USA.

[Lopez and Resnik, 2006] Lopez, A. and Resnik, P. (2006). Word-based alignment, phrase-based translation: What's the link? In *Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas (AMTA)*, pages 90–99, Cambridge, Massachusetts, USA.

[Macherey and Och, 2007] Macherey, W. and Och, F. J. (2007). An empirical study on computing consensus translations from multiple machine translation systems. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 986–995, Prague, Czech Republic.

[Macklovitch and Hannan, 1996] Macklovitch, E. and Hannan, M. L. (1996). Line 'em up: Advances in alignment technology and their impact on translation support tools. In *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 41–57, Montreal, Quebec, Canada.

[Mangu et al., 2000] Mangu, L., Brill, E., and Stolcke, A. (2000). Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Computer Speech and Language*, Vol 14(4):373–400.

[Mann and Yarowsky, 2001] Mann, G. S. and Yarowsky, D. (2001). Multipath translation lexicon induction via bridge languages. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 151–158, Pittsburgh, Pennsylvania, USA.

[Manning and Schütze, 1999] Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press.

[Marcu and Wong, 2002] Marcu, D. and Wong, W. (2002). A phrase-based, joint probability model for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 133–139, Philadelphia, Pennsylvania, USA.

[Martin et al., 2005] Martin, J., Mihalcea, R., and Pedersen, T. (2005). Word alignment for languages with scarce resources. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 65–74, Ann Arbor, Michigan, USA.

[Matusov et al., 2006] Matusov, E., Ueffing, N., and Ney, H. (2006). Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of the European Chapter of the Association for Computational Linguistics (EACL 2006)*, pages 33–40, Trento, Italy.

[Matusov et al., 2004] Matusov, E., Zens, R., and Ney, H. (2004). Symmetric word alignments for statistical machine translation. In *Proceedings of the International Conference on Computational Linguistics (COLING 2004)*, pages 219–225, Geneva, Switzerland.

[Melamed, 1998a] Melamed, D. (1998a). Annotation style guide for the blinker project, version 1.0.4. Technical Report 98-06, Institute for Research in Cognitive Science (IRCS), University of Pennsylvania, Philadelphia, Pennsylvania, USA.

[Melamed, 2001] Melamed, D. (2001). *Empirical Methods for Exploiting Parallel Texts*. The MIT Press.

[Melamed, 1997a] Melamed, I. D. (1997a). Automatic discovery of non-compositional compounds in parallel data. In *In Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 97–108, Providence, Rhode Island, USA.

[Melamed, 1997b] Melamed, I. D. (1997b). A scalable architecture for bilingual lexicography. Technical Report MS-CIS-91-01, University of Pennsylvania, Department of Computer and Information Science.

[Melamed, 1998b] Melamed, I. D. (1998b). Manual annotation of translational equivalence: The Blinker project. Technical Report 98-07, Institute for Research in Cognitive Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

[Merkel, 1999] Merkel, M. (1999). *Understanding and enhancing translation by parallel text processing*. Linköping studies in science and technology, dissertation no. 607, Department of Computer and Information Science, Linköping University, Linköping, Sweden.

[Merkel et al., 2002] Merkel, M., Andersson, M., and Ahrenberg, L. (2002). The plug link annotator - interactive construction of data from parallel corpora. In *Parallel Corpora, Parallel Worlds: Proceedings of the Symposium on Parallel Corpora, Department of Linguistics, Uppsala University, Sweden, 1999*, pages 151–168.

[Mihalcea and Pedersen, 2003] Mihalcea, R. and Pedersen, T. (2003). An evaluation exercise for word alignment. In *In Proceedings of the HLT-NAACL Workshop on Building and Using Parallel Texts*, pages 1—-10, Edmonton, Alberta, Canada.

[Moore, 2005] Moore, R. C. (2005). A discriminative framework for bilingual word alignment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 81–88, Vancouver, British Columbia, Canada.

[Nagao, 1984] Nagao, M. (1984). A framework of a mechanical translation between japanese and english by analogy principle. In *Proceedings of the international NATO symposium on Artificial and human intelligence*, pages 173–180, Lyon, France.

[Niessen and Ney, 2004] Niessen, S. and Ney, H. (2004). Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, Vol 30(2):181–204.

[Nomoto, 2004] Nomoto, T. (2004). Multi-engine machine translation with voted language model. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics (ACL 2004)*, pages 494–501, Barcelona, Spain.

[Oard et al., 2003] Oard, D., Doermann, D., Dorr, B., He, D., Resnik, P., Weinberg, A., Byrne, W., Khudanpur, S., Yarowsky, D., Leuski, A., Koehn, P., and Knight, K. (2003). Desperately seeking cebuano. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 76–78, Edmonton, Alberta, Canada.

[Oard and Och, 2003] Oard, D. W. and Och, F. J. (2003). Rapid-response machine translation for unexpected languages. In *Proceedings of the MT Summit IX*, pages 277–283, New Orleans, Louisiana, USA.

[Och, 2003] Och, F. J. (2003). Minimum error rate training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 160–167, Sapporo, Japan.

[Och and Ney, 2000] Och, F. J. and Ney, H. (2000). A comparison of alignment models for statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, volume 2, pages 1086–1090, Saarbrüken, Germany.

[Och and Ney, 2001] Och, F. J. and Ney, H. (2001). Statistical multi-source translation. In *Proceedings of the MT Summit VIII*, pages 253–258, Santiago de Compostela, Spain.

[Och and Ney, 2002] Och, F. J. and Ney, H. (2002). Discriminative training and maximum entropy models for statistical machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 295–302, Philadelphia, Pennsylvania, USA.

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

[Och and Ney, 2004] Och, F. J. and Ney, H. (2004). The alignment template approach to statistical machine translation. *Computational Linguistics*, Vol 30(4):417–449.

[Och et al., 1999] Och, F. J., Tillmann, C., and Ney, H. (1999). Improved alignment models for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, pages 20–28, College Park, Maryland, USA.

[Och et al., 2001] Och, F. J., Ueffing, N., and Ney, H. (2001). An efficient a* search algorithm for statistical machine translation. In *Proceedings of the Data-Driven Machine Translation Workshop, 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001)*, pages 55–62, Toulouse, France.

[Papineni et al., 2001] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). Bleu: a method for automatic evaluation of machine translation. Research Report RC22176, IBM.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, Pennsylvania, USA.

[Pouliquen et al., 2003] Pouliquen, B., Steinberger, R., and Ignat, C. (2003). Automatic annotation of multilingual text collections with a conceptual thesaurus. In *Proceedings of the Workshop Ontologies and Information Extraction at the Summer School, The Semantic Web and Language Technology - Its Potential and Practicalities*, pages 33–39, Bucharest, Romania.

[Pouliquen et al., 2004] Pouliquen, B., Steinberger, R., and Ignat, C. (2004). Automatic linking of similar texts across languages. *Current Issues in Linguistic Theory (CILT)*, 260:307–316.

[Resnik and Melamed, 1997] Resnik, P. and Melamed, I. D. (1997). Semi-automatic acquisition of domain-specific translation lexicons. In *Proceedings of the 5th conference on Applied natural language processing (ANLP)*, pages 340–347, Washington DC, USA.

[Resnik et al., 1999] Resnik, P., Olsen, M. B., and Diab, M. (1999). The bible as a parallel corpus: Annotating the book of 2000 tongues. *Computers and the Humanities*, Vol 33(1-2):129–153.

[Ribeiro et al., 2001] Ribeiro, A., Lopes, G. P., and Mexia, J. T. (2001). Extracting translation equivalents from portuguese-chinese parallel texts. *Studies in Lexicography*, Vol 11(1):181–194.

[Rosti et al., 2007] Rosti, A.-V., Ayan, N. F., Xiang, B., Matsoukas, S., Schwartz, R. M., and Dorr, B. J. (2007). Combining outputs from multiple machine translation systems. In *Proceedings of the Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 228–235, Rochester, New York, USA.

[Ráez, 2006] Ráez, A. M. (2006). *Automatic Text Categorization of Documents in the High Energy Physics Domain.* PhD thesis, Granada University, Granada, Spain.

[Sadler, 1989a] Sadler, V. (1989a). The bilingual knowledge bank: A new conceptual basis fot MT. Technical report, Utrecht: BSO-Research.

[Sadler, 1989b] Sadler, V. (1989b). Translating with a simulated bilingual knowledge bank. Technical report, Utrecht: BSO-Research.

[Sakai et al., 2004] Sakai, T., Koyama, M., Kumano, A., and Manabe, T. (2004). Toshiba BRIDJE at NTCIR-4 CLIR: Monolingual/bilingual IR and Flexible Feedback. In *Working Notes of NTCIR-4*, Tokyo, Japan.

[Sanfilippo and Steinberger, 1997] Sanfilippo, A. and Steinberger, R. (1997). Automatic selection and ranking of translation candidates. In *Proceedings of the 7th Conference on Theoretical and Methodological Issues in Machine Translation: "MT Yesterday, Today, and Tomorrow" (TMI)*, pages 200–207, Santa Fe, New Mexico, USA.

[Sato and Nagao, 1990] Sato, S. and Nagao, M. (1990). Toward memory-based translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING 1990)*, pages 247–252, Helsinki, Finland.

[Schafer and Yarowsky, 2002] Schafer, C. and Yarowsky, D. (2002). Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th conference on Natural language learning (COLING 2002)*, pages 1–7, Taipei, Taiwan.

[Schubert, 1988] Schubert, K. (1988). Implicitness as a guiding principle in machine translation. In *Proceedings of the 12th conference on Computational Linguistics*, pages 599–601, Budapest, Hungary.

[Schwartz, 2008] Schwartz, L. (2008). Multi-source translation methods. In *Proceedings of the Association for Machine Translation in the Americas (AMTA)*, Waikiki, Hawaii.

[Simard, 1999] Simard, M. (1999). Text-translation alignment: Three languages are better than two. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP-VLC)*, pages 2–11, College Park, Maryland, USA.

[Simard, 2000] Simard, M. (2000). *Multilingual text alignment. Aligning three or more versions of a text. In Parallel Text Processing, Alignment and Use of Translation Corpora*, chapter 3, pages 49–67. Kluwer Academic Publishers.

[Simard et al., 1993] Simard, M., Foster, G. F., and Isabelle, P. (1993). Using cognates to align sentences in bilingual corpora. In *Proceedings of the conference of the Centre for Advanced Studies on Collaborative research (CASCON)*, pages 1071–1082, Toronto, Ontario, Canada.

[Smadja et al., 1996] Smadja, F., McKeown, K. R., and Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics*, Vol 22(1):1–38.

[Sperberg-McQueen and Burnard, 2002] Sperberg-McQueen, C. and Burnard, L. (2002). (eds.) Guidelines for Electronic Text Encoding and Interchange, The XML version of the TEI Guidelines. The TEI Consortium.

[Steinberger et al., 2004a] Steinberger, R., Bruno, P., and Ignat, C. (2004a). Providing cross-lingual information access with knowledge-poor methods. *Informatica. An international Journal of Computing and Informatics*, 28:415–423.

[Steinberger et al., 2004b] Steinberger, R., Pouliquen, B., and Ignat, C. (2004b). Exploiting multilingual nomenclatures and language-independent text features as an interlingua for cross-lingual text analysis applications. In *Information Society 2004, Proceedings B of the 7th International Multiconference - Language Technologies*, pages 2–12, Ljubljana, Slovenia.

[Steinberger et al., 2005] Steinberger, R., Pouliquen, B., and Ignat, C. (2005). Navigating multilingual news collections using automatically extracted information. In *Proceedings of the 27th International Conference 'Information Technology Interfaces' (ITI)*, pages 27–34, Cavtat / Dubrovnik, Croatia.

[Steinberger et al., 2006] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 2142–2147, Genoa, Italy.

[Stolcke, 2002] Stolcke, A. (2002). SRILM - an Extensible Language Modeling Toolkit. In *Proceedings of the International Conference for Speech and Language Processing (ICSLP)*, pages 901–904, Denver, Colorado, USA.

[Sumita et al., 1990] Sumita, E., H., I., and H., K. (1990). Translating with examples: A new approach to machine translation. In *Proceedings of the Third International Conference on Theoretical and Methodological Issues in Machine Translation,*, pages 203–212, Austin, Texas, USA.

[Sumita and Tsutsumi, 1988] Sumita, E. and Tsutsumi, Y. (1988). A translation aid system using flexible text retrieval based on syntax-matching,. TRL Reseach Report TR-87-1019, Tokyo Research Laboratory, IBM.

[Taskar et al., 2005] Taskar, B., Lacoste-Julien, S., and Klein, D. (2005). A discriminative matching approach to word alignment. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT-EMNLP)*, pages 73–80, Vancouver, British Columbia, Canada.

[Tiedemann, 1998] Tiedemann, J. (1998). Extraction of translation equivalents from parallel corpora. In *Proceedings of the 11th Nordic Conference on Computational Linguistics (NODALIDA)*, Copenhagen, Denmark.

[Tiedemann, 1999a] Tiedemann, J. (1999a). "Uplug" - a modular corpus tool for parallel corpora. In *Parallel Corpora, Parallel Worlds. Proceedings of Parallel Corpus Symposium*, pages 181–197, Uppsala, Sweden.

[Tiedemann, 1999b] Tiedemann, J. (1999b). Word alignment - step by step. In *Proceedings of the 12th Nordic Conference on Computational Linguistics (NODALIDA)*, pages 216–227, Trondheim, Norway.

[Tiedemann, 2003] Tiedemann, J. (2003). *Recycling Translations – Extraction of Lexical Data from Parallel Corpora and their Application in Natural Language Processing.* PhD thesis, Uppsala University, Uppsala.

[Tillmann, 2003] Tillmann, C. (2003). A projection extension algorithm for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).*

[Tillmann and Ney, 2000] Tillmann, C. and Ney, H. (2000). Word re-ordering and dp-based search in statistical machine translation. In *Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, pages 850–856, Saarbrücken, Germany.

[Toma, 1976] Toma, P. (1976). *An Operational Machine Translation System*, pages 247–259. Gardner Press, New York.

[Tufiş, 2007] Tufiş, D. (2007). Exploiting Aligned Parallel Corpora in Multilingual Studies and Applications. In Ishida, T., Fussell, S. R., and Vossen, P. T., editors, *Intercultural Collaboration. First International Workshop (IWIC 2007)*, volume 4568 of *Lecture Notes in Computer Science*, pages 103–117. Springer-Verlag.

[Tufiş et al., 2004a] Tufiş, D., Barbu, A. M., and Ion, R. (2004a). Extracting Multilingual Lexicons from Parallel Corpora. *Computers and the Humanities*, 38(2):163–189. Springer Netherlands.

[Tufiş et al., 2006] Tufiş, D., Ion, R., Ceauşu, A., and Ştefănescu, D. (2006). Improved Lexical Alignment by Combining Multiple Reified Alignments. In Ishida, T., Fussell, S. R., and Vossen, P. T., editors, *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pages 153–160, Trento, Italy.

[Tufiş et al., 2005] Tufiş, D., Ion, R., Ceausu, A., and Stefanescu, D. (2005). Combined word alignments. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 107–110, Ann Arbor, Michigan, USA.

[Tufiş et al., 2004b] Tufiş, D., Ion, R., and Ide, N. (2004b). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 1312–1318, Geneva, Switzerland. Association for Computational Linguistics.

[Tufiş et al., 2008] Tufiş, D., Koeva, S., Erjavec, T., Gavrilidou, M., and Krstev, C. (2008). Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Tadic, M., Dimitrova-Vulchanova, M., and Koeva, S., editors, *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, Dubrovnik, Croatia.

[Turian et al., 2003] Turian, J. P., Shen, L., and Melamed, D. I. (2003). Evaluation of machine translation and its evaluation. In *Proceedings of MT Summit IX*, pages 386–393, New Orleans, Louisiana, USA.

[Utiyama and Isahara, 2007] Utiyama, M. and Isahara, H. (2007). A comparison of pivot methods for phrase-based statistical machine translation. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 484–491, Rochester, New York, USA.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. (1979). *Information Retrieval, 2nd edition.* Butterworths, London.

[Vandeghinste et al., 2006] Vandeghinste, V., Schuurman, I., Carl, M., Markantonatou, S., and Badia, T. (2006). METIS-II: Machine Translation for Low Resource Languages. In *Proceedings of the the 5th international conference on Language Resources and Evaluation (LREC 2006)*, pages 1284–1289, Genoa, Italy.

[Varga et al., 2005] Varga, D., Halacsy, P., Kornai, A., Nagy, V., Nemeth, L., and Tron, V. (2005). Parallel corpora for medium density languages. In *Proceedings of the Iternational Conference on Recent Advances in Natural Language Processing (RANLP)*, pages 590–596, Borovets, Bulgaria.

[Véronis, 1998] Véronis, J. (1998). Tagging guidelines for word alignment. http://aune.lpl.univ-aix.fr/projects/arcade/2nd/word/guide/index.html.

[Véronis, 2000] Véronis, J. (2000). *From the Rosetta stone to the information society. A survey of parallel text processing. In Parallel Text Processing, Alignment and Use of Translation Corpora*, chapter 1, pages 1–24. Kluwer Academic Publishers.

[Véronis and Langlais, 2000] Véronis, J. and Langlais, P. (2000). *Evaluation of parallel text alignement systems. The ARCADE project. In Parallel Text Processing, Text, Speech and Language Technology series*, chapter 19. Kluwer Academic Publishers, Dordrecht, The Netherlands.

[Versino et al., 2007] Versino, C., Ignat, C., and Bril, L.-V. (2007). Open source information for export control. In *Proceedings of the 29th ESARDA Symposium on Safeguards and Nuclear Material Management.*, pages 1–8, Aix en Provence, France.

[Vogel et al., 1996] Vogel, S., Ney, H., and Tillmann, C. (1996). HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, pages 836–841, Copenhagen, Denmark.

[Vogel et al., 2003] Vogel, S., Zhang, Y., Huang, F., Tribble, A., Venogupal, A., Zhao, B., and Waibel, A. (2003). The CMU statistical translation system. In *Proceedings of MT Summit IX*, pages 402–409, New Orleans, Louisiana, USA.

[Wang et al., 2006] Wang, H., Wu, H., and Liu, Z. (2006). Word alignment for languages with scarce resources using bilingual corpora of other language pairs. In *Proceedings of the COLING/ACL Main Conference, Poster Sessions*, pages 874–881, Sydney, Australia.

[Wang and Waibel, 1998] Wang, Y.-Y. and Waibel, A. (1998). Modeling with structures in statistical machine translation. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL 1998)*, pages 1357–1363, Montreal, Quebec, Canada.

[Weaver, 1949] Weaver, W. (1949). Translation. In *Mimeographed*, pages 15–23. MIT Press.

[Wu and Xia, 1994] Wu, D. and Xia, X. (1994). Learning an english-chinese lexicon from a parallel corpus. In *In Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 206–213, Columbia, Maryland, USA.

[Wu and Wang, 2007] Wu, H. and Wang, H. (2007). Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, pages 856–863, Prague, Czech Republic.

[Yarowsky et al., 2001] Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the 1st international conference on Human language technology research*, pages 1–8, San Diego, California, USA.

[Zens et al., 2002a] Zens, R., Och, F., and Ney, H. (2002a). Phrase-based statistical machine translation. In *German Conference on Artificial Intelligence (ZI)*, pages 18–32, Aachen, Germany.

[Zens et al., 2002b] Zens, R., Och, F. J., and Ney, H. (2002b). Statistical machine translation. In *Proceedings of the 6th EAMT Workshop*, pages 18–32, Manchester, England.

Author's Publications

2008

Ralf Steinberger, Pouliquen Bruno & **Camelia Ignat** (2008). *Using language-independent rules to achieve high multilinguality in Text Mining.* In: Fogelman-Soulié Françoise, Domenico Perrotta, Jakub Piskorski & Ralf Steinberger (eds.): Mining Massive Data Sets for Security. pp. 217-240. IOS Press, Amsterdam, The Netherlands. (Overview article explaining the design principles to achieve highly multilingual applications such as NewsExplorer)

2007

Camelia Ignat & François Rousselot. Représentation de textes a l'aide d'étiquettes sémantiques dans le cadre de la classification automatique. Romanian Review of Linguistics, VOL. LI, 2006, Issues 3-4, Ed. Romanian Academy, July-December.

Cristina Versino, **Camelia Ignat**, Louis-Victor Bril (2007). Open Source Information for Export Control. Proceedings of the 29th ESARDA Symposium on Safeguards and Nuclear Material Management. page 1-8, OPOCE (publ.), Luxembourg, 2007.

2006

Camelia Ignat & François Rousselot (2006). Un algorithme de génération de profil de document et son évaluation dans le contexte de la classification thématique. Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT'2006). Besançon, 19-21 April 2006.

Steinberger Ralf, Bruno Pouliquen, Anna Widiger, **Camelia Ignat**, Tomaž Erjavec, Dan Tufiş, Dániel Varga (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Pouliquen Bruno, Marco Kimler, Ralf Steinberger, **Camelia Ignat**, Tamara Oellinger, Ken Blackler, Flavio Fuart, Wajdi Zaghouani, Anna Widiger, Ann-Charlotte Forslund, Clive Best (2006). Geocoding multilingual texts: Recognition, Disambiguation and Visualisation. Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'2006). Genoa, Italy, 24-26 May 2006.

Žižka Jan, Jiří Hroza, Bruno Pouliquen, **Camelia Ignat** & Ralf Steinberger (2006). The selection of electronic text documents supported by only positive examples. Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT'2006). Besançon, 19-21 April 2006.

Pouliquen Bruno, Ralf Steinberger, **Camelia Ignat** & Tamara Oellinger (2006). Building and displaying name relations using automatic unsupervised analysis of newspaper articles. Proceedings of the 8th International Conference on the Statistical Analysis of Textual Data (JADT'2006). Besançon, 19-21 April 2006.

Best Clive, Bruno Pouliquen, Ralf Steinberger, Eric van der Goot, Ken Blackler, Flavio Fuart, Tamara Oellinger & **Camelia Ignat** (2006). Towards automatic event tracking. In: Proceedings of IEEE International Conference on Intelligence and Security Informatics (ISI'2006). San Diego, California, USA, 23-24.05.2006.

2005

Steinberger Ralf, Bruno Pouliquen, **Camelia Ignat** (2005). Navigating multilingual news collections using automatically extracted information. Journal of Computing and Information Technology - CIT 13, 2005, 4, 257-264. ISSN: 1330-1136.

Pouliquen Bruno, Ralf Steinberger, **Camelia Ignat**, Irina Temnikova, Anna Widiger, Wajdi Zaghouani & Jan Žižka (2005). Multilingual person name recognition and transliteration. Revue CORELA - Cognition, Représentation, Langage.

Erjavec Tomaž, **Camelia Ignat**, Bruno Pouliquen & Ralf Steinberger (2005). Massive multilingual corpus compilation: Acquis Communautaire and totale. Journal Archives of Control Sciences, Volume 15(LI), 2005, No. 3, pages 253-264.

Steinberger Ralf, Bruno Pouliquen, **Camelia Ignat** (2005). Navigating multilingual news collections using automatically extracted information. Proceedings of the 27th International Conference 'Information Technology Interfaces' (ITI'2005). Cavtat / Dubrovnik, Croatia, June 20-23, 2005.

Ignat Camelia, Bruno Pouliquen, Ralf Steinberger & Tomaž Erjavec (2005). A tool set for the quick and efficient exploration of large document collections. Proceedings of the Symposium on Safeguards and Nuclear Material Management. 27th Annual Meeting of the European Safeguards Research and Development Association (ESARDA-2005). London, UK, 10-12 May 2005.

Tomaz Erjavec, **Camelia Ignat**, Bruno Pouliquen & Ralf Steinberger (2005). Massive multilingual corpus compilation; Acquis Communautaire and totale. In: 2nd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (L&T'05). Poznań, Poland, 21-23 April 2005.

Pouliquen Bruno, Ralf Steinberger, **Camelia Ignat**, Irina Temnikova, Wajdi Zaghouani & Jan Žižka (2005). Detection of person names and their translations in multilingual news. Colloque Traîtement lexicographique des noms propres, Tours, 24 March 2005.

2004

Pouliquen Bruno, Ralf Steinberger & **Camelia Ignat** (2004). Automatic Linking of Similar Texts Across Languages. In: N. Nicolov, K. Bontcheva, G. Angelova & R. Mitkov (eds.): Current Issues in Linguistic Theory 260 - Recent Advances in Natural Language Processing III. Selected Papers from RANLP'2003. John Benjamins Publishers, Amsterdam.

Steinberger Ralf, Pouliquen Bruno & **Camelia Ignat** (2004). Providing cross-lingual information access with knowledge-poor methods. In: Informatica. An international Journal of Computing and Informatics. Volume 28. Special Issue.

Steinberger Ralf, Pouliquen Bruno & **Camelia Ignat** (2004). Exploiting Multilingual Nomenclatures and Language-Independent Text Features as an Interlingua for Cross-lingual Text Analysis Applications. In: Information Society 2004 (IS'2004) - Proceedings of the 4th Slovenian Language Technologies Conference, pages 2-12. Ljubljana, Slovenia, 13-14 October 2004.

Pouliquen Bruno, Ralf Steinberger, **Camelia Ignat**, Emilia Käsper & Irina Temnikova (2004). Multilingual and Cross-lingual News Topic Tracking. In: Proceedings of the 20th International Conference on Computational Linguistics (CoLing'2004), Vol. II, pages 959-965. Geneva, Switzerland, 23-27 August 2004.

Pouliquen Bruno, Ralf Steinberger, **Camelia Ignat** & Tom de Groeve (2004). Geographical Information Recognition and Visualisation in Texts Written in Various Languages. In: Proceedings of the 19th Annual ACM Symposium on Applied Computing (SAC'2004), Special Track on Information Access and Retrieval (SAC-IAR), vol. 2, pp. 1051-1058. Nicosia, Cyprus, 14 - 17 March 2004.

2003

Pouliquen Bruno, Ralf Steinberger & **Camelia Ignat** (2003). Automatic Identification of Document Translations in Large Multilingual Document Collections. In: Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'2003), pp. 401-408. Borovets, Bulgaria, 10 - 12 September 2003.

Ignat Camelia, Bruno Pouliquen, António Ribeiro & Ralf Steinberger (2003). Extending an Information Extraction Tool Set to Central and Eastern European Languages. In: Proceedings of the International Workshop Information Extraction for Slavonic and other Central and Eastern European Languages (IESL'2003), held at RANLP'2003, pp. 33-39. Borovets, Bulgaria, 8 - 9 September 2003.

Pouliquen Bruno, Ralf Steinberger & **Camelia Ignat** (2003). Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus. In: Proceedings of the Workshop Ontologies and Information Extraction at the Summer School The Semantic Web and Language Technology - Its Potential and Practicalities (EUROLAN'2003). Bucharest, Romania, 28 July - 8 August 2003.